Effective Theory of Deep Neural Networks

Sho Yaida

Meta AI

Effective Theory of Deep Neural Networks

Dan Roberts
Sho Yaida
Boris Hanin

[arXiv:2106.10165]

Meta AI

Effective Theory of Deep Neural Networks

Dan Roberts
Sho Yaida
Boris Hanin

[arXiv:2106.10165~470 pages]

Meta AI

Agenda

1. Overview
2. Neural Networks at Infinite Width
3. Neural Networks at Finite Width
4. The Principles

1. Overview

Deep learning is powerful

Deep learning is powerful

[put an impressive chart]

[put an impressive picture]

[put an impressive text]

Deep learning is powerful & interesting

[put an
impressive chart]

[put an
impressive picture]

[put an
impressive text]

Machine Learning in a Nutshell

Machine Learning in a Nutshell

- Instantiate a model

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

Machine Learning in a Nutshell

- 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

- 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

- 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

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

Machine Learning in a Nutshell

- 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

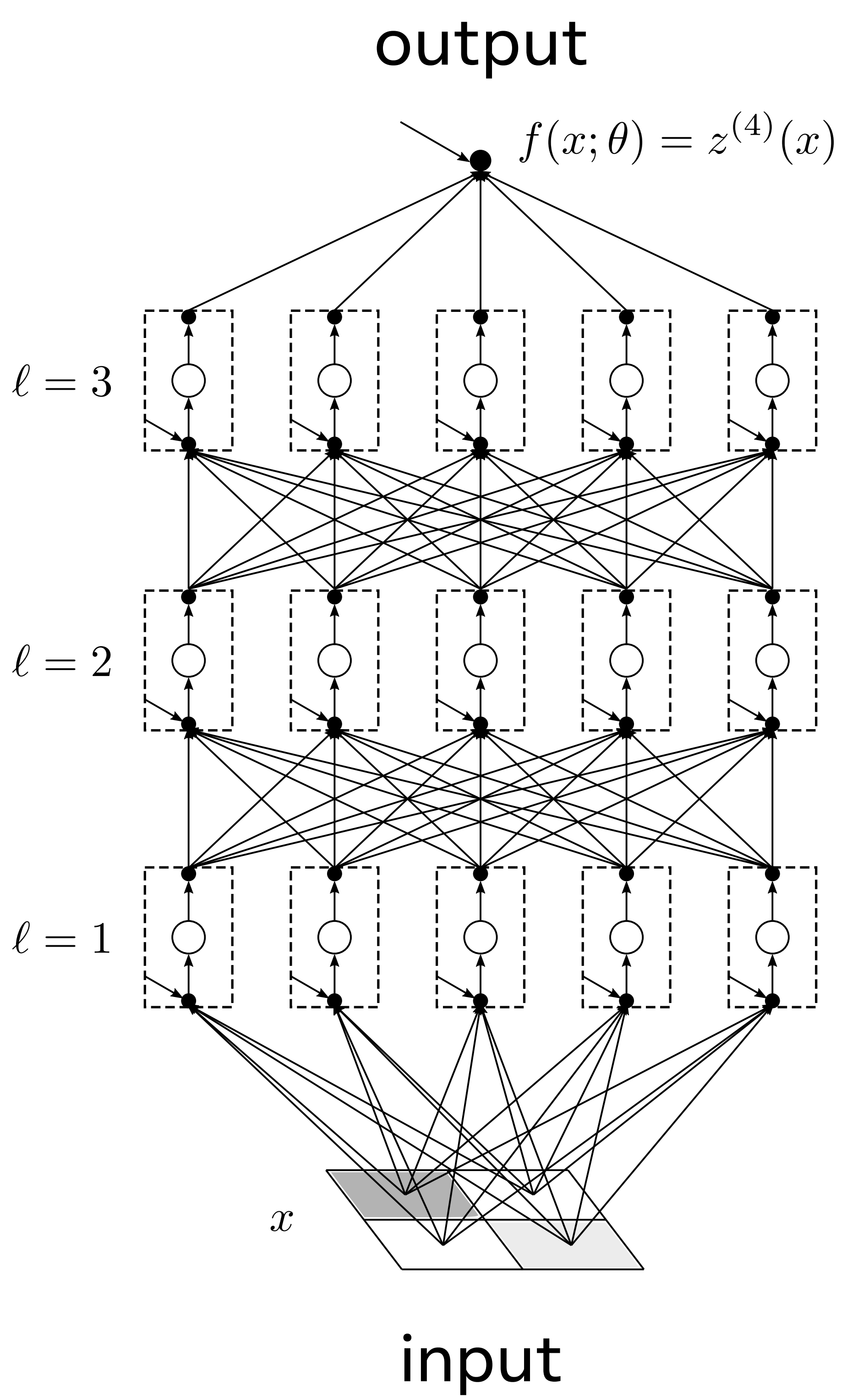
$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \frac{\partial \mathcal{L}}{\partial \theta_\mu}\bigg|_{\theta=\theta(t)}$$

- Use the trained model to make predictions

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

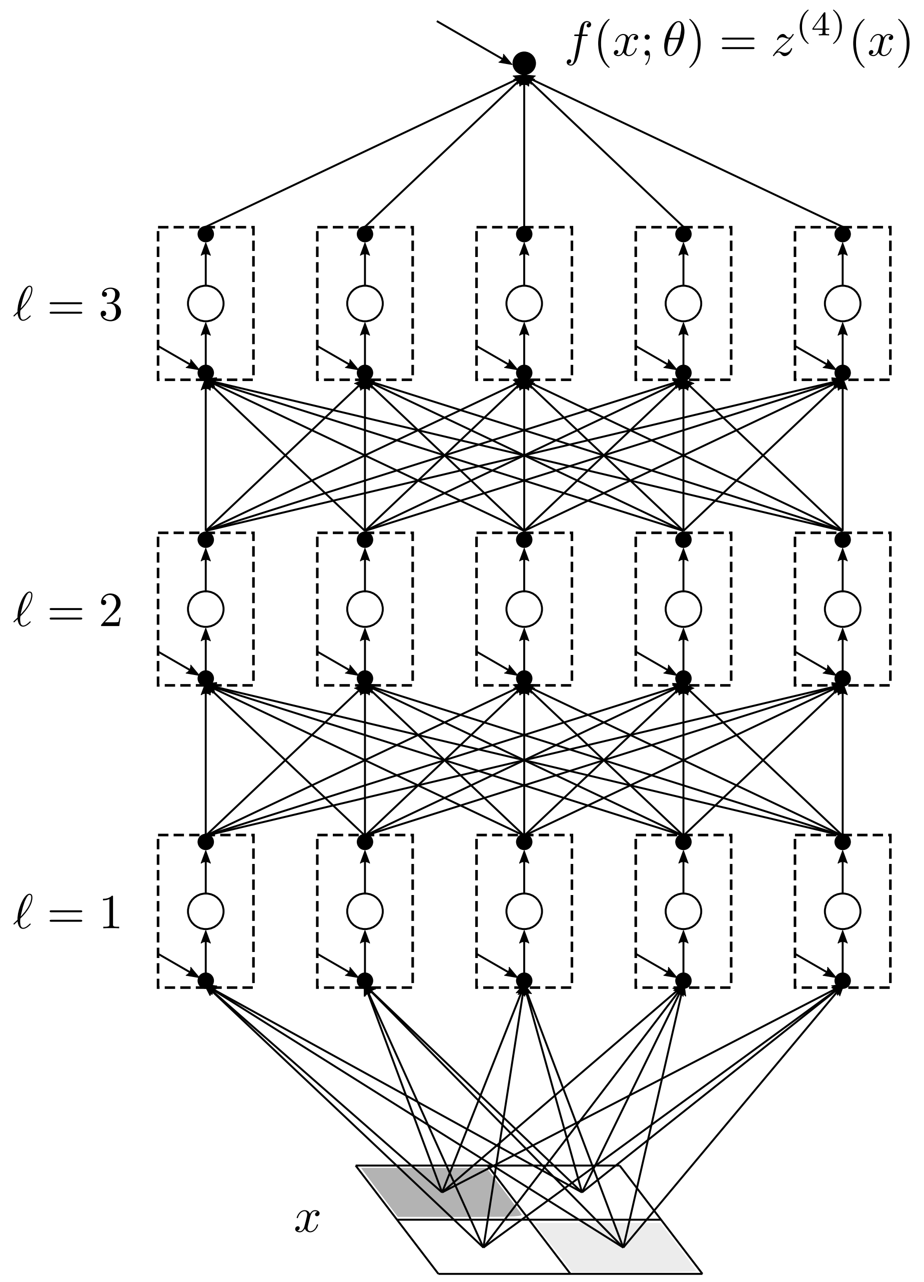
mean, variance, etc.

Neural Networks



Neural Networks

output



input

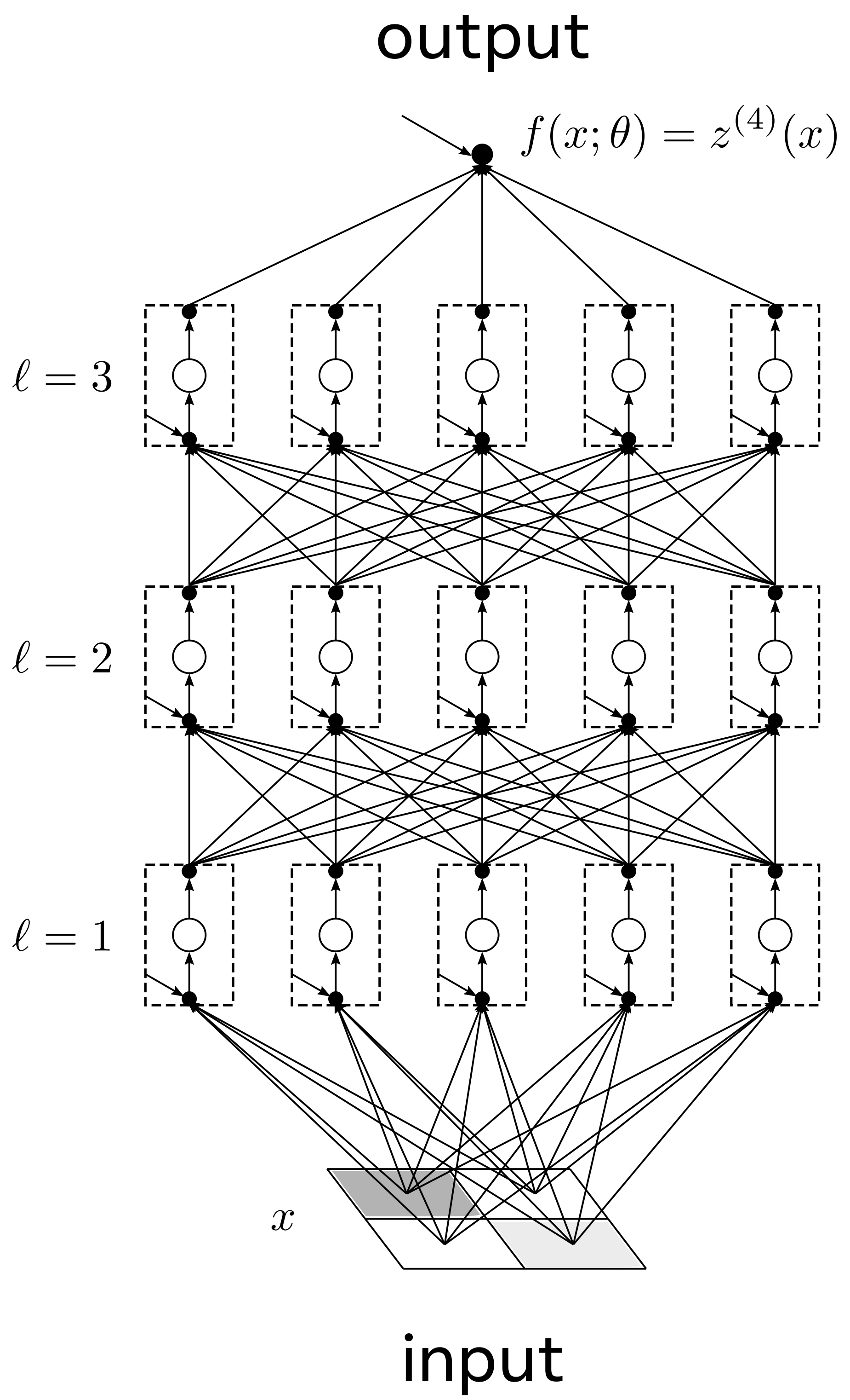
- Function:

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

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

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

Neural Networks



- Function:

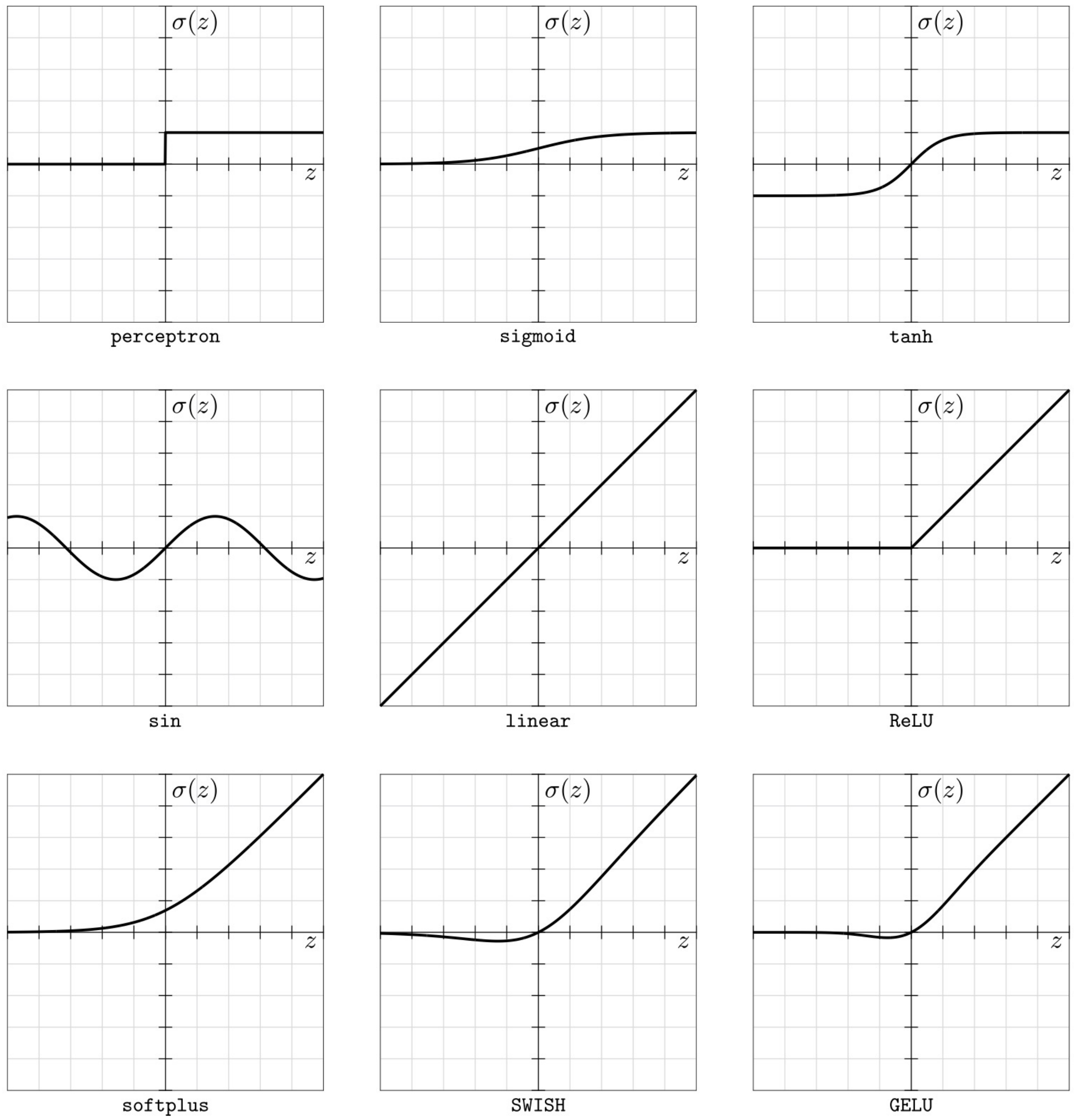
$$z_i^{(1)}(x) \equiv b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_j \quad \text{for} \quad i = 1, \dots, n_1 \,,$$

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

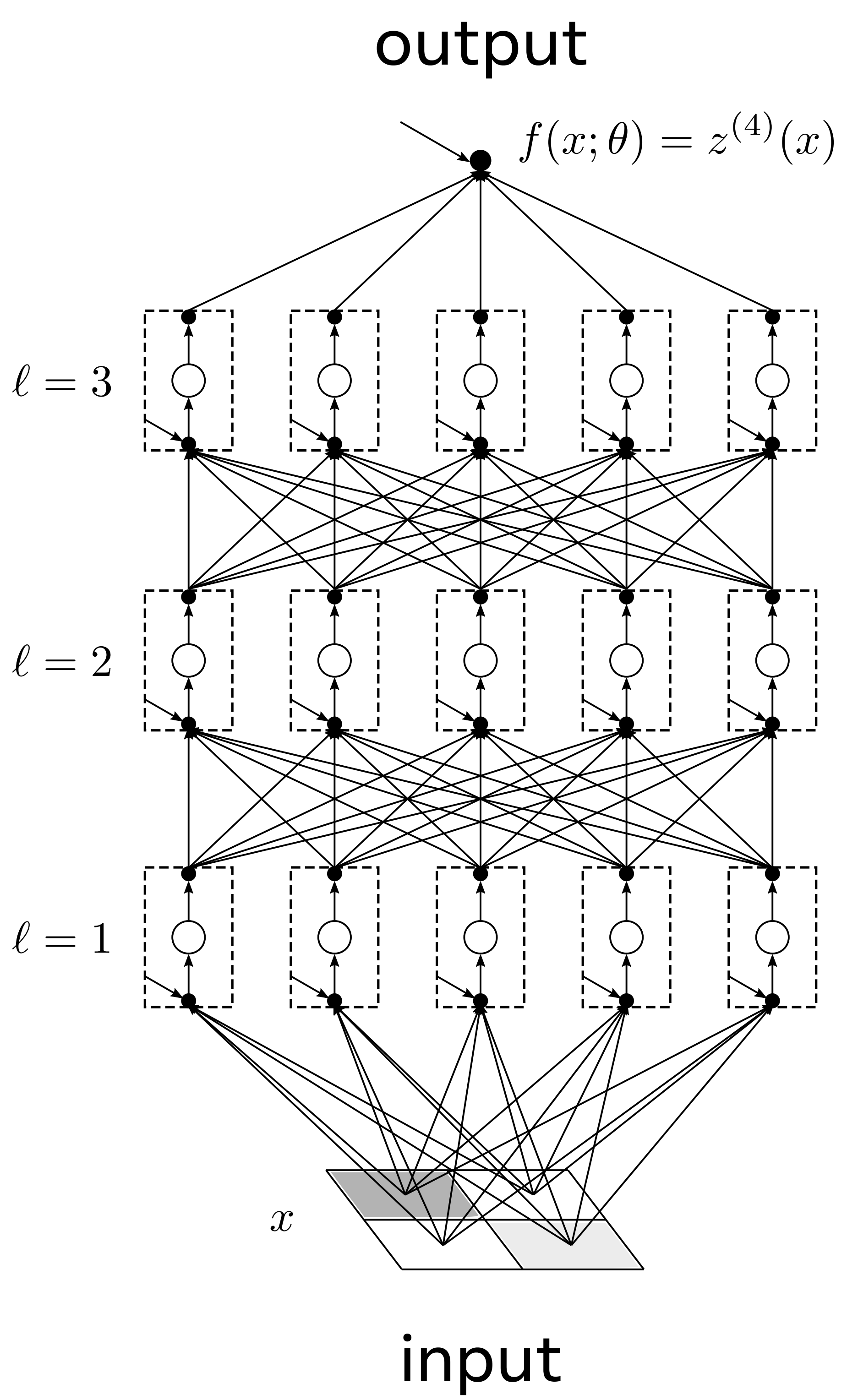
$$f(x;\theta) = z^{(L)}(x)$$

activation function

$$\sigma(z)$$



Neural Networks



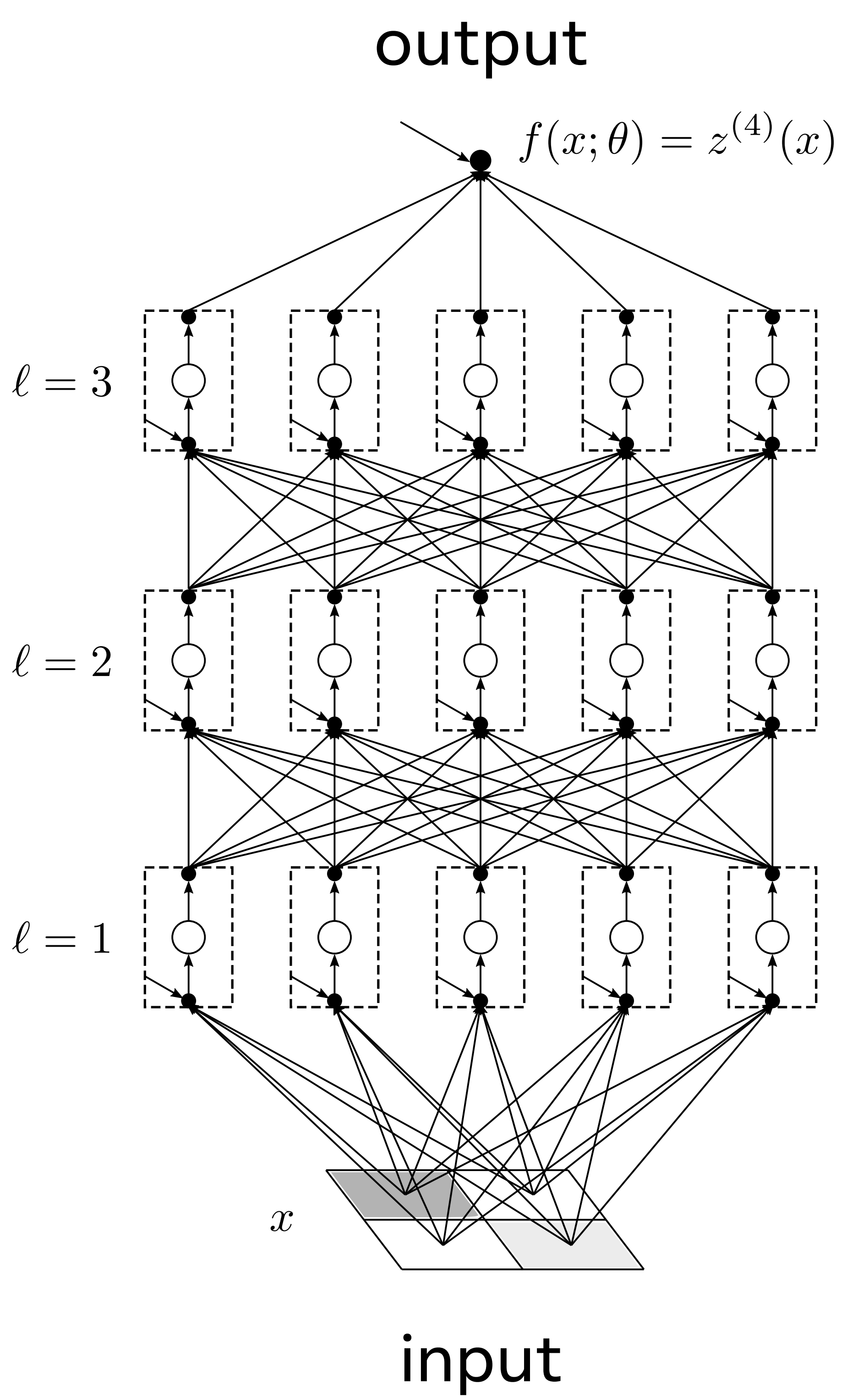
- Function:

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

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

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

Neural Networks



- Function:

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

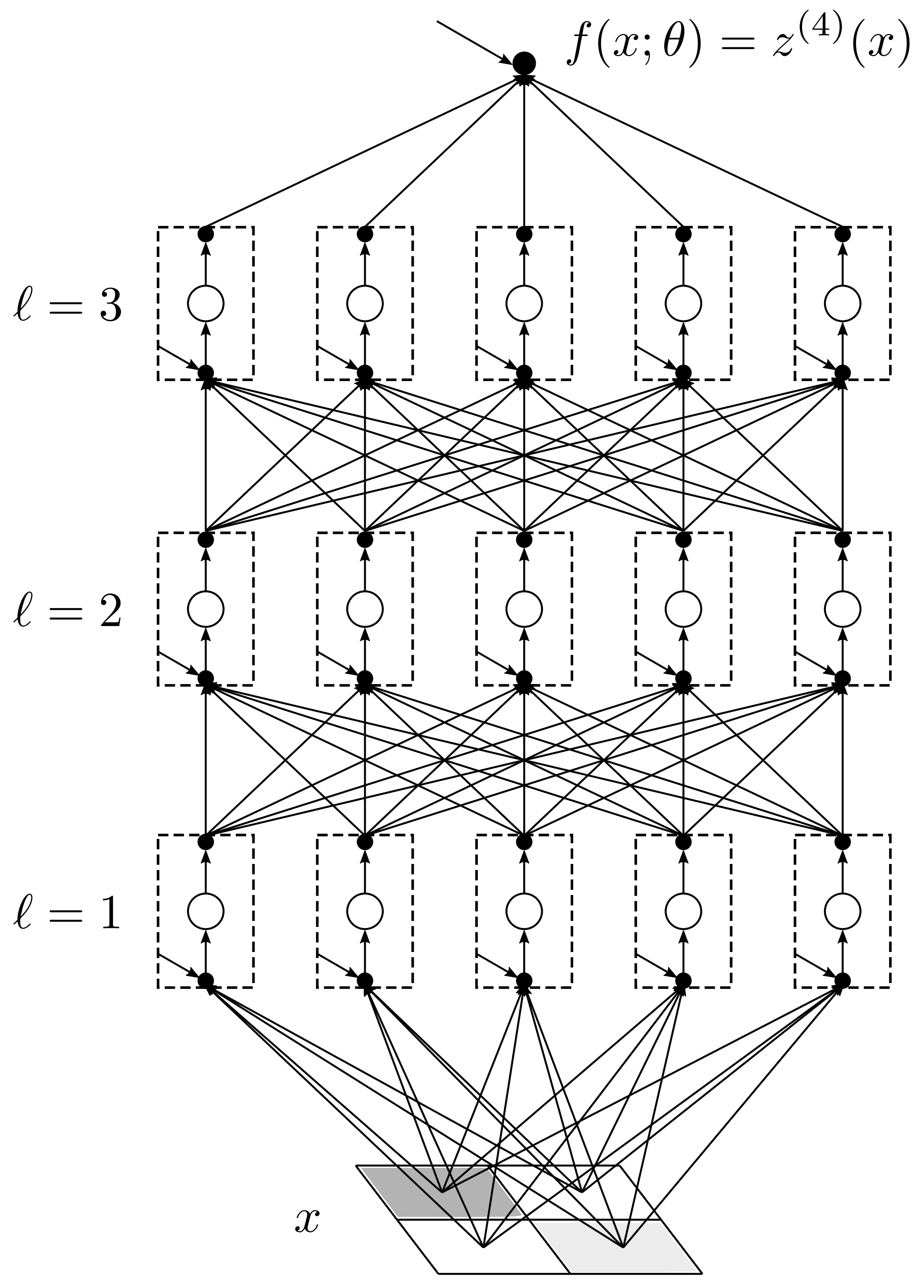
$$z_i^{(\ell+1)}(x) \equiv b_i^{(\ell+1)} + \sum_{j=1}^{n_\ell} W_{ij}^{(\ell+1)} \sigma\Big(z_j^{(\ell)}(x)\Big) \quad \text{for} \quad i = 1, \dots, n_{\ell+1}\,; \ \ell = 1, \dots, L-1$$

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

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

Neural Networks

output



input

- Function:

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

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

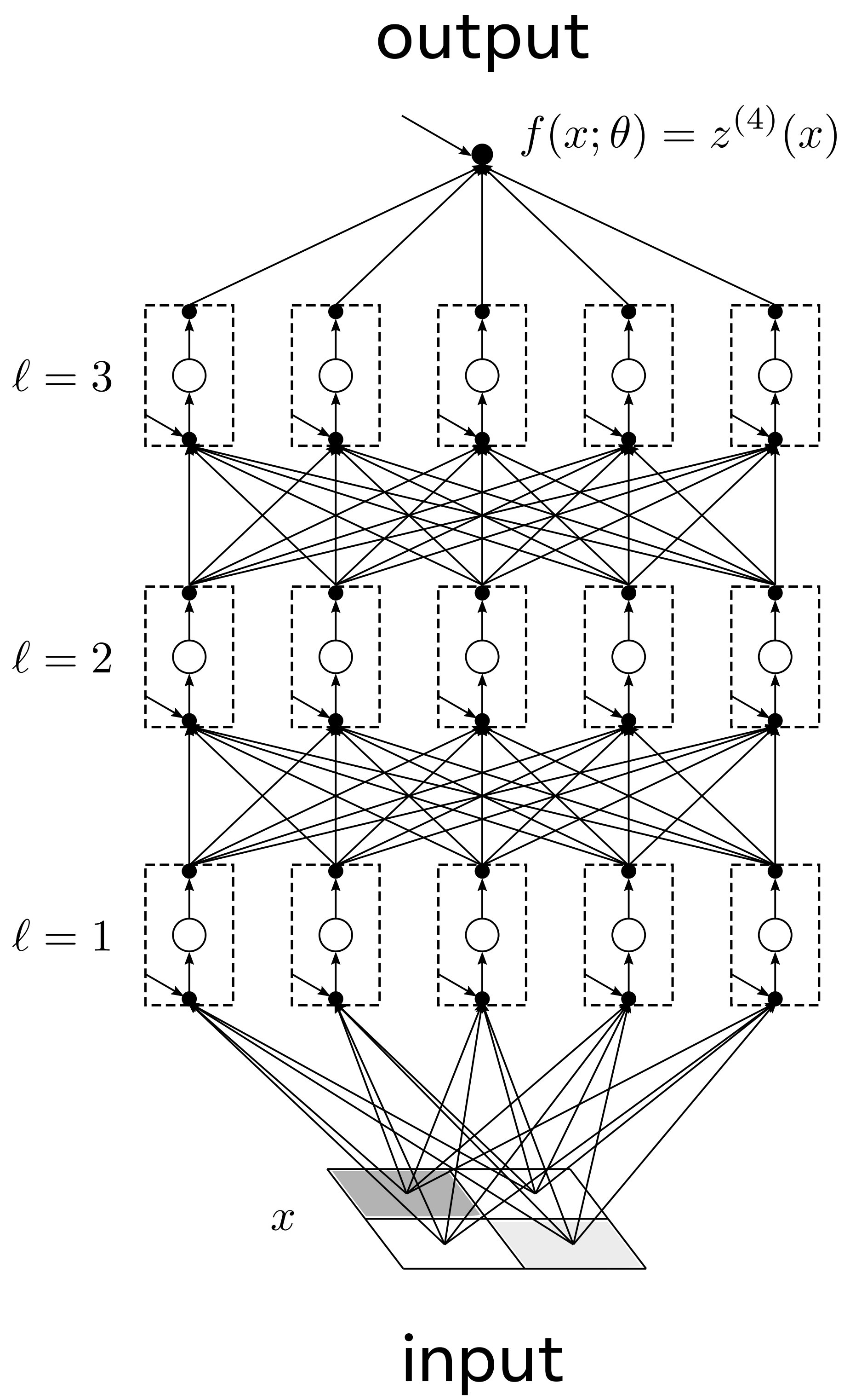
$$f(x;\theta) = z^{(L)}(x)$$

- 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

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

Neural Networks



- Function:

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

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

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

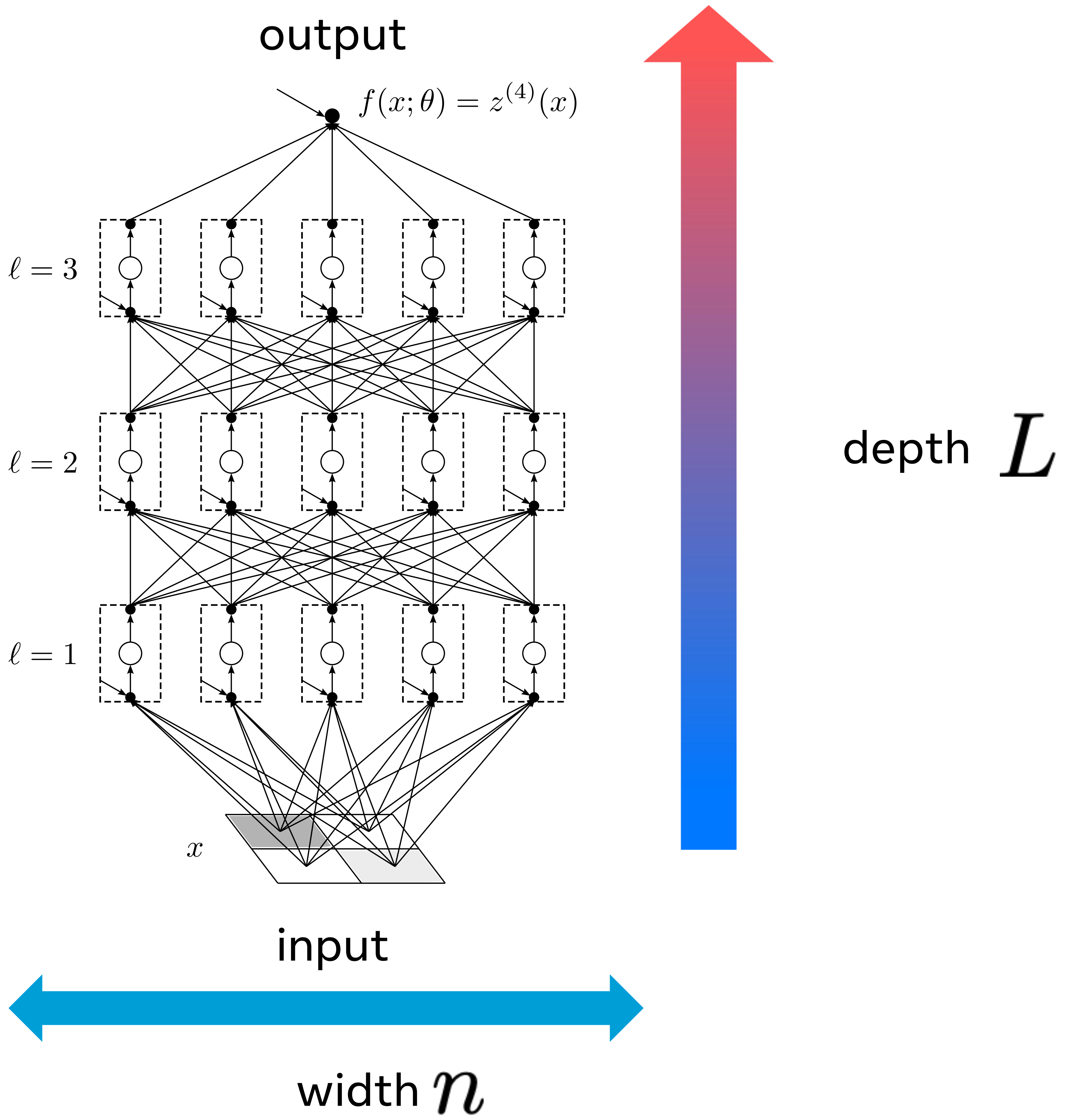
- Model parameters: $\theta_{\mu=1,\dots,P} = \left\{ b_i^{(1)}, W_{ij}^{(1)}, b_i^{(2)}, W_{ij}^{(2)}, \dots, b_i^{(L)}, W_{ij}^{(L)} \right\}$

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

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

good wide limit

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

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

- 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:

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

Problems 1, 2, & 3

Trained function, Taylor-expanded around initialization:

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

- 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:

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

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

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

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

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**)
AND
systematically going beyond that idealized limit
(a.k.a. **perturbation theory**)

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**)
AND
systematically going beyond that idealized limit
(a.k.a. **perturbation theory**)

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

Statistics become *sparse* & dynamics can be truncated

Trailer for part 2: $n = \infty$

Trailer for part 2: $n = \infty$

- Gaussian Prior (& Posterior) [R. Neal (1996), J. Lee+Y. Bahri et al. (ICLR 2018), A. Matthews et al. (ICLR2018)]
- (Neural Tangent) Kernel Learning [A. Jacot, F. Gabriel, & C. Hongler (NeurIPS 2018)]

As simple as we can imagine, but…

Trailer for part 2: $n = \infty$

- Gaussian Prior (& Posterior) [R. Neal (1996), J. Lee+Y. Bahri et al. (ICLR 2018), A. Matthews et al. (ICLR2018)]
- (Neural Tangent) Kernel Learning [A. Jacot, F. Gabriel, & C. Hongler (NeurIPS 2018)]

As simple as we can imagine, but…

- No Representation Learning(*)
- No Algorithm Dependence

too simple to describe real <u>deep</u> neural networks

(*Representation Learning:
the ability of a model to learn useful representations from data)

Trailer for part 2: $n = \infty$

- Gaussian Prior (& Posterior) [R. Neal (1996), J. Lee+Y. Bahri et al. (ICLR 2018), A. Matthews et al. (ICLR2018)]
- (Neural Tangent) Kernel Learning [A. Jacot, F. Gabriel, & C. Hongler (NeurIPS 2018)]

As simple as we can imagine, but…

- No Representation Learning
- No Algorithm Dependence

too simple to describe real <u>deep</u> neural networks

$\boxed{\text{A useful starting point but not the end of the story}}$

Trailer for part 3: $n \gg L$

Trailer for part 3: $n \gg L$

- *Nearly*-Gaussian Prior (& Posterior) [§4 (& §6) of arXiv:2106.10165]
- *Weakly-Nonlinear* Learning Dynamics [§11 & §∞]

A bit more complex but tractable, and…

Trailer for part 3: $n \gg L$

- *Nearly*-Gaussian Prior (& Posterior) [§4 (& §6) of arXiv:2106.10165]
- *Weakly-Nonlinear* Learning Dynamics [§11 & §∞]

A bit more complex but tractable, and…

- Yes Representation Learning $\propto \dfrac{L}{n}$
- Yes Algorithm Dependence

complex enough to capture rich phenomenology of real deep neural networks

Trailer for part 3: $n \gg L$

- *Nearly*-Gaussian Prior (& Posterior) [§4 (& §6) of arXiv:2106.10165]
- *Weakly-Nonlinear* Learning Dynamics [§11 & §∞]

A bit more complex but tractable, and…

- Yes Representation Learning $\propto \frac{L}{n}$
- Yes Algorithm Dependence

complex enough to capture rich phenomenology of real <u>deep</u> neural networks

Qualitatively very different from infinite-width limit

2. Neural Networks at Infinite Width

Some Notations

- Initial outputs: $\widehat{z}_{i;\delta} = z_i^{(L)}(x_\delta; \theta_{\text{init}})$

- Trained outputs: $z^\star_{i;\delta} = z_i^{(L)}(x_\delta; \theta_{\text{trained}})$

  - introduced sample index δ
  - dropped (L)
  - hatted initial
  - starred trained

Training Dynamics

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

Training Dynamics

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\mu} \right)$$

Training Dynamics

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

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

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{d z_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

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

good wide limit

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{d z_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

$$O(1/n) \quad \cancel{+ \cdots}$$

Training Dynamics

Gradient descent: $$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

Neural Tangent Kernel (NTK)

Training Dynamics

Gradient descent:
$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

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

Neural Tangent Kernel (NTK)

Similarly:
$$H_{i_1 i_2;\delta_1\delta_2}(t+1) = H_{i_1 i_2;\delta_1\delta_2}(t) + O\left(\frac{1}{n}\right)$$

Training Dynamics

Gradient descent:
$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\mathrm{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:
$$\begin{aligned} z_{i;\delta}(t+1) =& z_{i;\delta}(t) \\ & - \eta \sum_{j,\tilde{\alpha}} \left(\sum_{\mu,\nu} \lambda_{\mu\nu} \frac{dz_{i;\delta}}{d\theta_\mu} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right) \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \\ & \equiv H_{ij;\delta\tilde{\alpha}}(t) = \widehat{H}_{ij;\delta\tilde{\alpha}} \end{aligned}$$

"frozen" NTK

Similarly:
$$H_{i_1 i_2;\delta_1\delta_2}(t+1) = H_{i_1 i_2;\delta_1\delta_2}(t) + \cancel{O\left(\frac{1}{n}\right)}$$

Training Dynamics

$$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \widehat{H}_{ij;\delta\tilde{\alpha}} \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}}$$

Solving "Problem 3" (Dynamics)

$$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \widehat{H}_{ij;\delta\tilde{\alpha}} \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}}$$

Solving "Problem 3" (Dynamics)

$$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \widehat{H}_{ij;\delta\tilde{\alpha}} \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}}$$

"E.g.," (*) for $\mathcal{L} = \frac{1}{2} \sum_{i,\tilde{\alpha}} \left(z_{i;\tilde{\alpha}} - y_{i;\tilde{\alpha}}\right)^2$

(*It turns out that the detailed forms of the loss/scheduling won't matter: *algorithm independence*, §10.2.2 of [arXiv:2106.10165](https://arxiv.org/abs/2106.10165))

Solving "Problem 3" (Dynamics)

$$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \widehat{H}_{ij;\delta\tilde{\alpha}}[z_{j;\tilde{\alpha}}(t) - y_{j;\tilde{\alpha}}]$$

"E.g.," (*) for $\mathcal{L} = \frac{1}{2}\sum_{i,\tilde{\alpha}}\left(z_{i;\tilde{\alpha}} - y_{i;\tilde{\alpha}}\right)^2$

(*It turns out that the detailed forms of the loss/scheduling won't matter:
algorithm independence, §10.2.2 of arXiv:2106.10165)

Solving "Problem 3" (Dynamics)

$$z_{i;\delta}(t+1) = z_{i;\delta}(t) - \eta \sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \widehat{H}_{ij;\delta\tilde{\alpha}} [z_{j;\tilde{\alpha}}(t) - y_{j;\tilde{\alpha}}]$$

"E.g.," (*) for $\mathcal{L} = \frac{1}{2} \sum_{i,\tilde{\alpha}} \left(z_{i;\tilde{\alpha}} - y_{i;\tilde{\alpha}}\right)^2$

(exponentially) $\boxed{z^{\star}_{i;\delta} = \widehat{z}_{i;\delta} - \sum_{j,k,\tilde{\alpha}_1,\tilde{\alpha}_1} \widehat{H}_{ij;\delta\tilde{\alpha}_1} \left(\widehat{\widetilde{H}}^{-1}\right)^{\tilde{\alpha}_1 \tilde{\alpha}_2}_{jk} \left(\widehat{z}_{k;\tilde{\alpha}_2} - y_{k;\tilde{\alpha}_2}\right)}$

(*It turns out that the detailed forms of the loss/scheduling won't matter: *algorithm independence*, §10.2.2 of arXiv:2106.10165)

Solving "Problem 3" (Dynamics)

$$z^{\star}_{i;\delta} = \widehat{z}_{i;\delta} - \sum_{j,k,\tilde{\alpha}_1,\tilde{\alpha}_1} \widehat{H}_{ij;\delta\tilde{\alpha}_1} \left(\widehat{\tilde{H}}^{-1}\right)_{jk}^{\tilde{\alpha}_1\tilde{\alpha}_2} \left(\widehat{z}_{k;\tilde{\alpha}_2} - y_{k;\tilde{\alpha}_2}\right)$$

$$p(\theta_{\text{init}}) \to p(\widehat{z}, \widehat{H}) \to p(z^{\star})$$

Solutions to “Problems 1 & 2”

$$p(\theta_{\text{init}}) \rightarrow p(\widehat{z}, \widehat{H}) \rightarrow p(z^{\star})$$

Solutions to “Problems 1 & 2”

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

$$\boxed{p(\hat{z}_{i;\delta}) \propto \exp\left[-\frac{1}{2}\sum_{i,\delta_1,\delta_2}(K^{-1})^{\delta_1\delta_2}\,\hat{z}_{i;\delta_1}\,\hat{z}_{i;\delta_2}\right]}$$

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

$$\boxed{\widehat{H}_{i_1 i_2;\delta_1\delta_2} = \delta_{i_1 i_2}\Theta_{\delta_1\delta_2}}$$

$$p(\theta_{\text{init}}) \rightarrow p(\widehat{z}, \widehat{H}) \rightarrow p(z^\star)$$

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ Gaussian distribution

with mean $$m_{i;\delta} = \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2 \in \text{train}} \Theta_{\delta \tilde{\alpha}_1} (\widetilde{\Theta}^{-1})^{\tilde{\alpha}_1 \tilde{\alpha}_2} y_{i;\tilde{\alpha}_2}$$

and variance involving both $K_{\delta_1 \delta_2}$, $\Theta_{\delta_1 \delta_2}$

Solved EVERYTHING but…

$p(z^{\star}_{i;\delta})$ Gaussian distribution

with mean $$m_{i;\delta} = \sum_{\tilde{\alpha}_1, \tilde{\alpha}_2 \in \text{train}} \Theta_{\delta \tilde{\alpha}_1} (\widetilde{\Theta}^{-1})^{\tilde{\alpha}_1 \tilde{\alpha}_2} y_{i;\tilde{\alpha}_2}$$

and variance involving both $K_{\delta_1 \delta_2}$, $\Theta_{\delta_1 \delta_2}$

- No Representation Learning (linear model with random features)
- No Algorithm Dependence (GD, Newton, SGD with decreasing learning rate, …)

too simple to describe real <u>deep</u> neural networks

3. Neural Networks at Finite Width

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_{j} \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

NTK $H(t)$

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

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_{j} \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

NTK $H(t)$

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

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

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

$$+ \ldots$$

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha}\in\mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

NTK $H(t)$

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

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

$O(1/n)$

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

$$- \frac{\eta^3}{6} \sum \left(\sum \lambda_{\mu_1\nu_1}\lambda_{\mu_2\nu_2}\lambda_{\mu_3\nu_3} \frac{d^3 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2} d\theta_{\mu_3}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \frac{dz_{j_3;\tilde{\alpha}_3}}{d\theta_{\nu_3}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \frac{\partial \mathcal{L}}{\partial z_{j_3;\tilde{\alpha}_3}}$$

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

$O(1/n^2)$

$$+ \ldots$$

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\text{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

NTK $H(t)$

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

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

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

$$\cancel{+ \dots}$$

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

Training Dynamics

Gradient descent:

$$\theta_\mu(t+1) = \theta_\mu(t) - \eta \sum_{\nu=1}^{P} \lambda_{\mu\nu} \left(\sum_{\tilde{\alpha} \in \mathcal{B}_{\mathrm{train}}} \sum_j \frac{\partial \mathcal{L}}{\partial z_{j;\tilde{\alpha}}} \frac{dz_{j;\tilde{\alpha}}}{d\theta_\nu} \right)$$

Taylor expansion:

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

NTK $H(t)$

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

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

$O(1/n)$

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

$$- \frac{\eta^3}{6} \sum \left(\sum \lambda_{\mu_1\nu_1} \lambda_{\mu_2\nu_2} \lambda_{\mu_3\nu_3} \frac{d^3 z_{i;\delta}}{d\theta_{\mu_1} d\theta_{\mu_2} d\theta_{\mu_3}} \frac{dz_{j_1;\tilde{\alpha}_1}}{d\theta_{\nu_1}} \frac{dz_{j_2;\tilde{\alpha}_2}}{d\theta_{\nu_2}} \frac{dz_{j_3;\tilde{\alpha}_3}}{d\theta_{\nu_3}} \right) \frac{\partial \mathcal{L}}{\partial z_{j_1;\tilde{\alpha}_1}} \frac{\partial \mathcal{L}}{\partial z_{j_2;\tilde{\alpha}_2}} \frac{\partial \mathcal{L}}{\partial z_{j_3;\tilde{\alpha}_3}}$$

$$\cancel{+ \dots}$$

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

Similarly some dynamical equations for NTK and dNTK (while ddNTK is frozen at this order)

Solving “Problem 3” (Dynamics)

[…long song & dance with dynamical perturbation theory to get $z^{\star}\left(\hat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}\right)$ …]

Solving "Problem 3" (Dynamics)

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

$$+ \,\text{despicable}(y, \widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}; \text{algorithm})$$

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

Solving "Problem 3" (Dynamics)

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

$$+ \,\mathrm{despicable}(y, \widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}; \mathrm{algorithm})$$

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

- NTK now "defrosted": indicative of Representation Learning!
- Solution depends on the algorithm: Algorithm Dependence!

Solving “Problem 3” (Dynamics)

$$
\begin{aligned}
& z_{i;\delta}(t=\infty) \qquad\qquad (\infty.141)\\
\equiv & z^{\mathrm{F}}_{i;\delta}(t=\infty)+z^{\mathrm{I}}_{i;\delta}(t=\infty)\\
= & z_{i;\delta}-\sum_{j,k,\tilde\alpha_1,\tilde\alpha_2}\widehat{H}_{ij;\delta\tilde\alpha_1}\left(\widehat{H}^{-1}\right)^{\tilde\alpha_1\tilde\alpha_2}_{jk}\left(z_{k;\tilde\alpha_2}-y_{k;\tilde\alpha_2}\right)\\
& +\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\delta\tilde\alpha_2}-\sum_{\tilde\alpha_5,\tilde\alpha_6}H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right]\\
& \qquad\times Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& +\sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\delta\tilde\alpha_1\tilde\alpha_2}-\sum_{\tilde\alpha_5,\tilde\alpha_6}H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right]\\
& \qquad\times Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& +\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& +\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& +\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& +\sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}-\sum_{\tilde\alpha_7,\tilde\alpha_8}H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& +O\!\left(\frac{1}{n^2}\right).
\end{aligned}
$$

Solving "Problem 3" (Dynamics)

$$
\begin{aligned}
& z_{i;\delta}(t=\infty) && (\infty.141)\\
\equiv & z^{\mathrm{F}}_{i;\delta}(t=\infty) + z^{\mathrm{I}}_{i;\delta}(t=\infty)\\
= & z_{i;\delta} - \sum_{j,k,\tilde\alpha_1,\tilde\alpha_2} \widehat{H}_{ij;\delta\tilde\alpha_1}\left(\widehat{H}^{-1}\right)^{\tilde\alpha_1\tilde\alpha_2}_{jk}\left(z_{k;\tilde\alpha_2}-y_{k;\tilde\alpha_2}\right)\\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\delta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1 i j_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right]\\
& \qquad\times Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{i j_1 j_2;\delta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{i j_1 j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right]\\
& \qquad\times Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{j_1 i j_2 j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{j_1 j_2 i j_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{i j_1 j_2 j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad\times Z_{\mathrm{IIB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + O\left(\frac{1}{n^2}\right).
\end{aligned}
$$

algorithm projectors

Solving “Problem 3” (Dynamics)

$$
\begin{aligned}
& z_{i;\delta}(t=\infty) \qquad (\infty.141)\\
\equiv\, & z^{\mathrm{F}}_{i;\delta}(t=\infty) + z^{\mathrm{I}}_{i;\delta}(t=\infty)\\
=\, & z_{i;\delta} - \sum_{j,k,\tilde\alpha_1,\tilde\alpha_2} \widehat{H}_{ij;\delta\tilde\alpha_1}\left(\widehat{H}^{-1}\right)^{\tilde\alpha_1\tilde\alpha_2}_{jk}\left(z_{k;\tilde\alpha_2}-y_{k;\tilde\alpha_2}\right)\\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\delta\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{j_1ij_2;\tilde\alpha_1\tilde\alpha_6\tilde\alpha_2}\right]\\
& \qquad \times Z_{\mathrm{A}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& + \sum_{j_1,j_2,\tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4}\left[\widehat{\mathrm{d}H}_{ij_1j_2;\delta\tilde\alpha_1\tilde\alpha_2} - \sum_{\tilde\alpha_5,\tilde\alpha_6} H_{\delta\tilde\alpha_5}\widetilde{H}^{\tilde\alpha_5\tilde\alpha_6}\widehat{\mathrm{d}H}_{ij_1j_2;\tilde\alpha_6\tilde\alpha_1\tilde\alpha_2}\right]\\
& \qquad \times Z_{\mathrm{B}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4}\left(z_{j_1;\tilde\alpha_3}-y_{j_1;\tilde\alpha_3}\right)\left(z_{j_2;\tilde\alpha_4}-y_{j_2;\tilde\alpha_4}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\delta\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{j_1ij_2j_3;\tilde\alpha_1\tilde\alpha_8\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad \times Z_{\mathrm{IA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_I}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad \times Z_{\mathrm{IB}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\left(z_{j_2;\tilde\alpha_5}-y_{j_2;\tilde\alpha_5}\right)\left(z_{j_3;\tilde\alpha_6}-y_{j_3;\tilde\alpha_6}\right)\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\delta\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{j_1j_2ij_3;\tilde\alpha_1\tilde\alpha_2\tilde\alpha_8\tilde\alpha_3}\right]\\
& \qquad \times Z_{\mathrm{IIA}}^{\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3\tilde\alpha_4\tilde\alpha_5\tilde\alpha_6}\left(z_{j_1;\tilde\alpha_4}-y_{j_1;\tilde\alpha_4}\right)\cdots\\
& + \sum_{\substack{j_1,j_2,j_3,\\ \tilde\alpha_1,\tilde\alpha_2,\tilde\alpha_3,\tilde\alpha_4,\tilde\alpha_5,\tilde\alpha_6}}\left[\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\delta\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3} - \sum_{\tilde\alpha_7,\tilde\alpha_8} H_{\delta\tilde\alpha_7}\widetilde{H}^{\tilde\alpha_7\tilde\alpha_8}\widehat{\mathrm{dd_{II}}H}_{ij_1j_2j_3;\tilde\alpha_8\tilde\alpha_1\tilde\alpha_2\tilde\alpha_3}\right]\\
& \qquad \times \cdots
\end{aligned}
$$

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

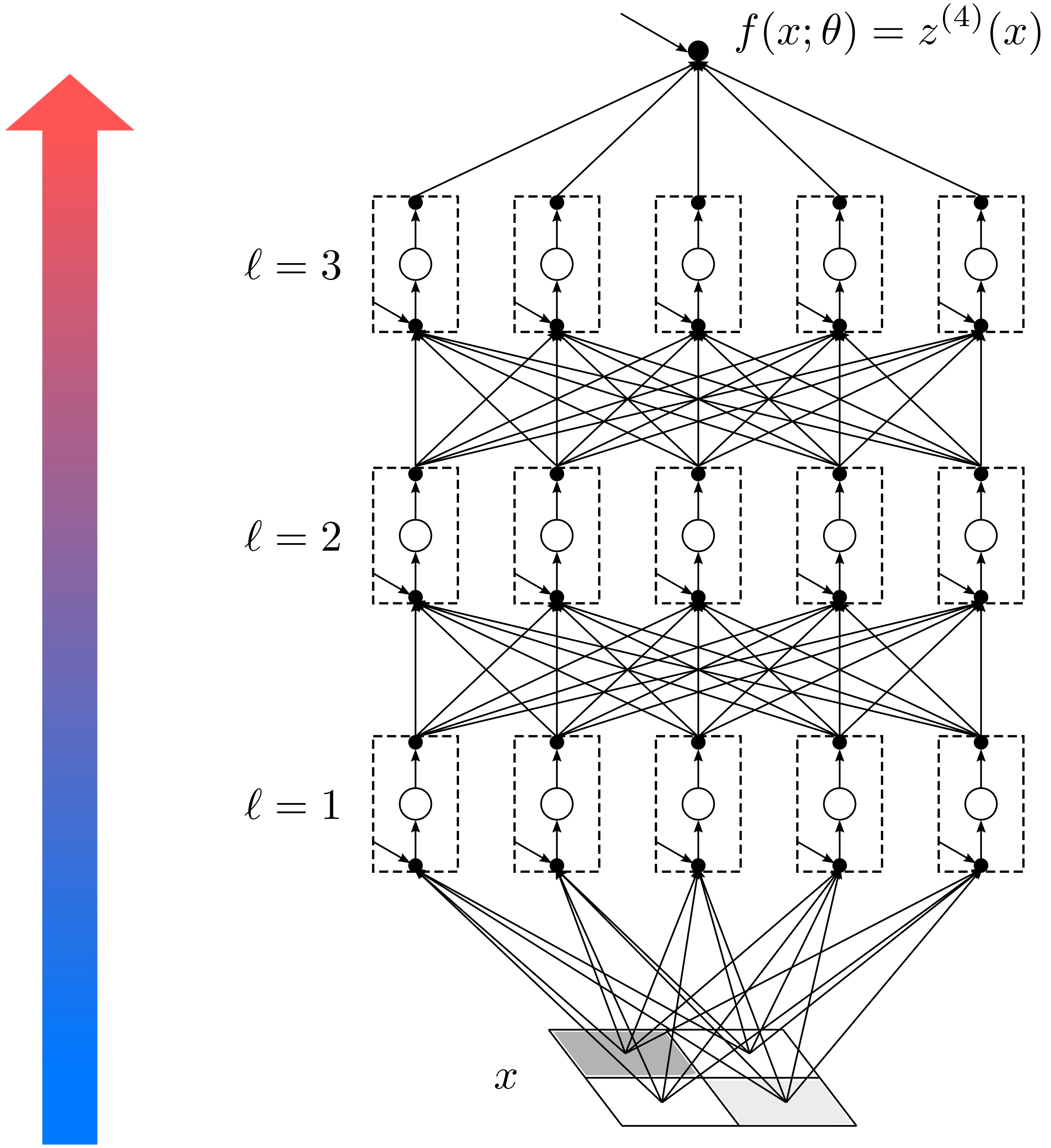
$$p(\theta_{\text{init}}) \to p(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}) \to p(z^\star)$$

Solutions to “Problems 1 & 2”

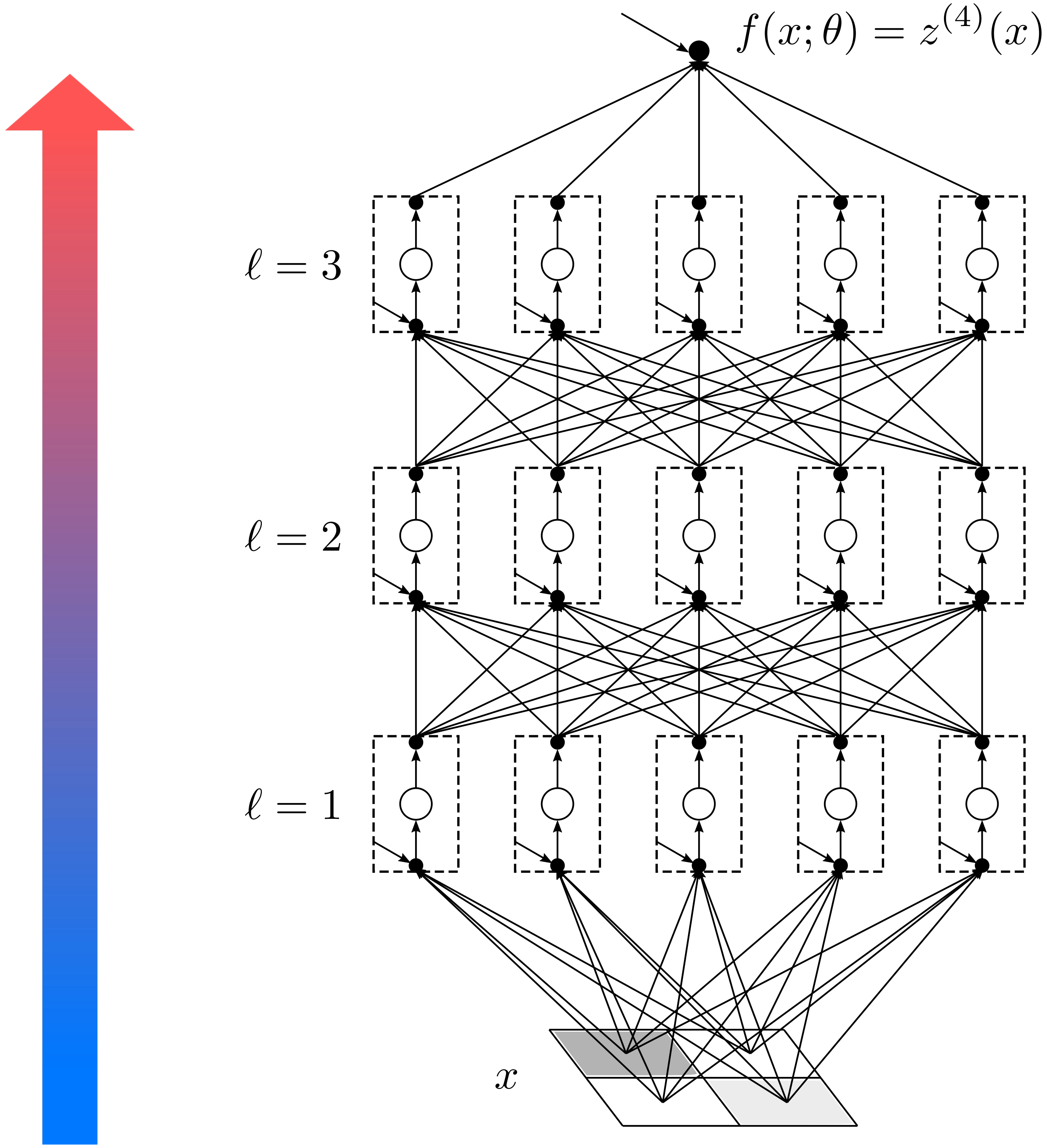
[See: §4, §8, §11.2, & §∞.3 of arXiv:2106.10165]

$$p(\theta_{\text{init}}) \rightarrow p(\widehat{z}, \widehat{H}, \widehat{\mathrm{d}H}, \widehat{\mathrm{dd}H}) \rightarrow p(z^{\star})$$

Strategy: Always Forward

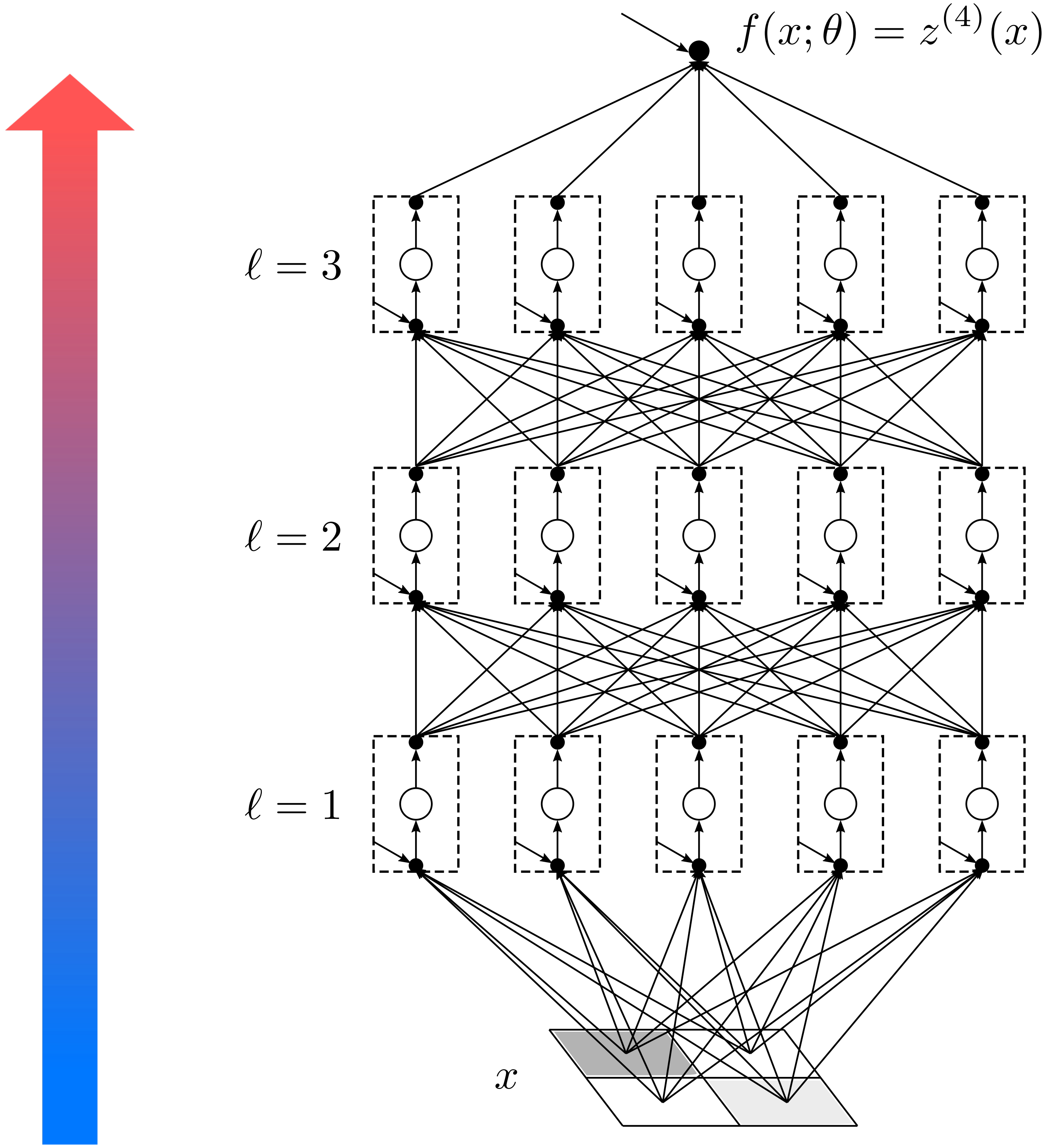


Strategy: Always Forward



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

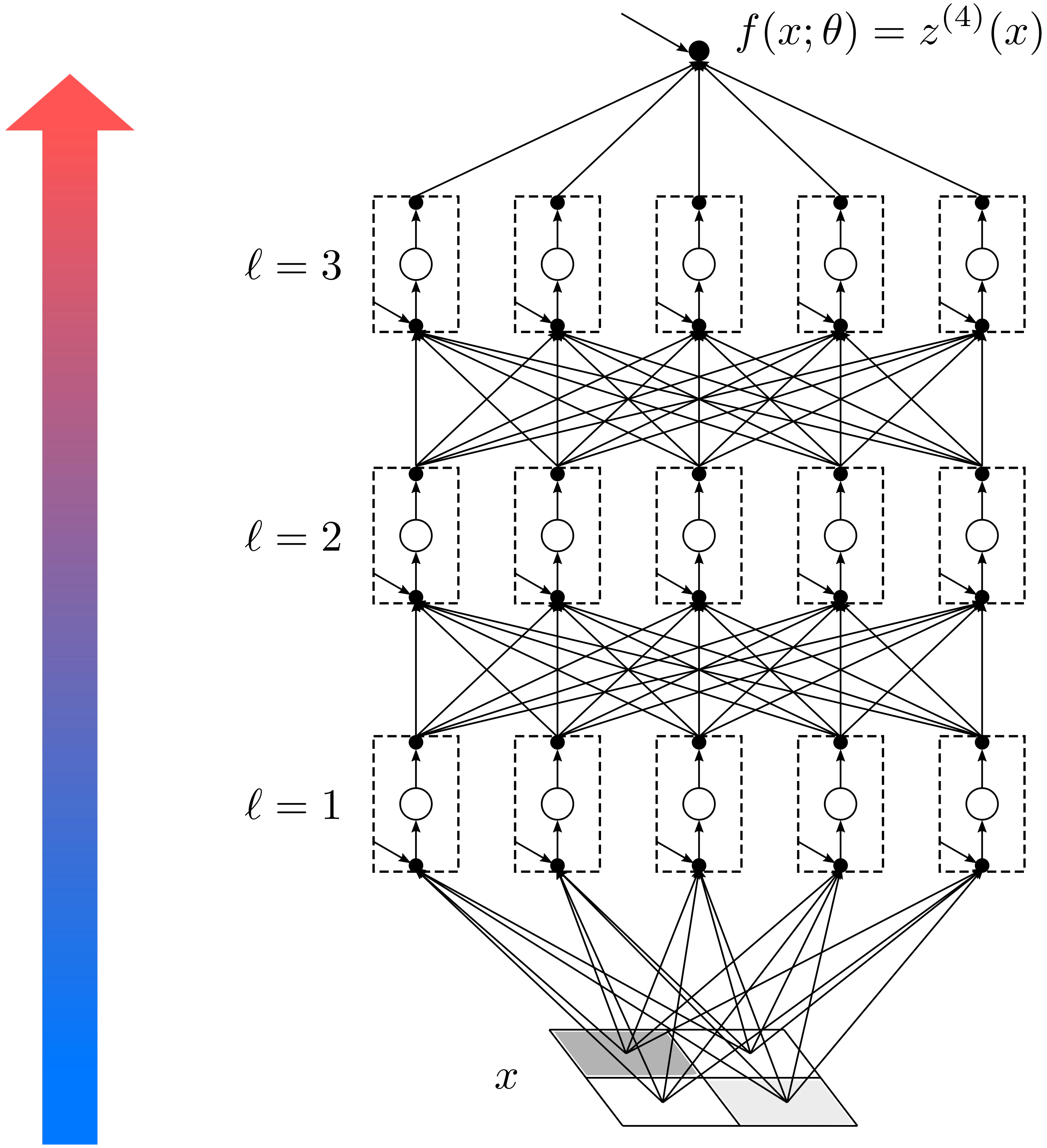
Strategy: Always Forward



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

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

Strategy: Always Forward



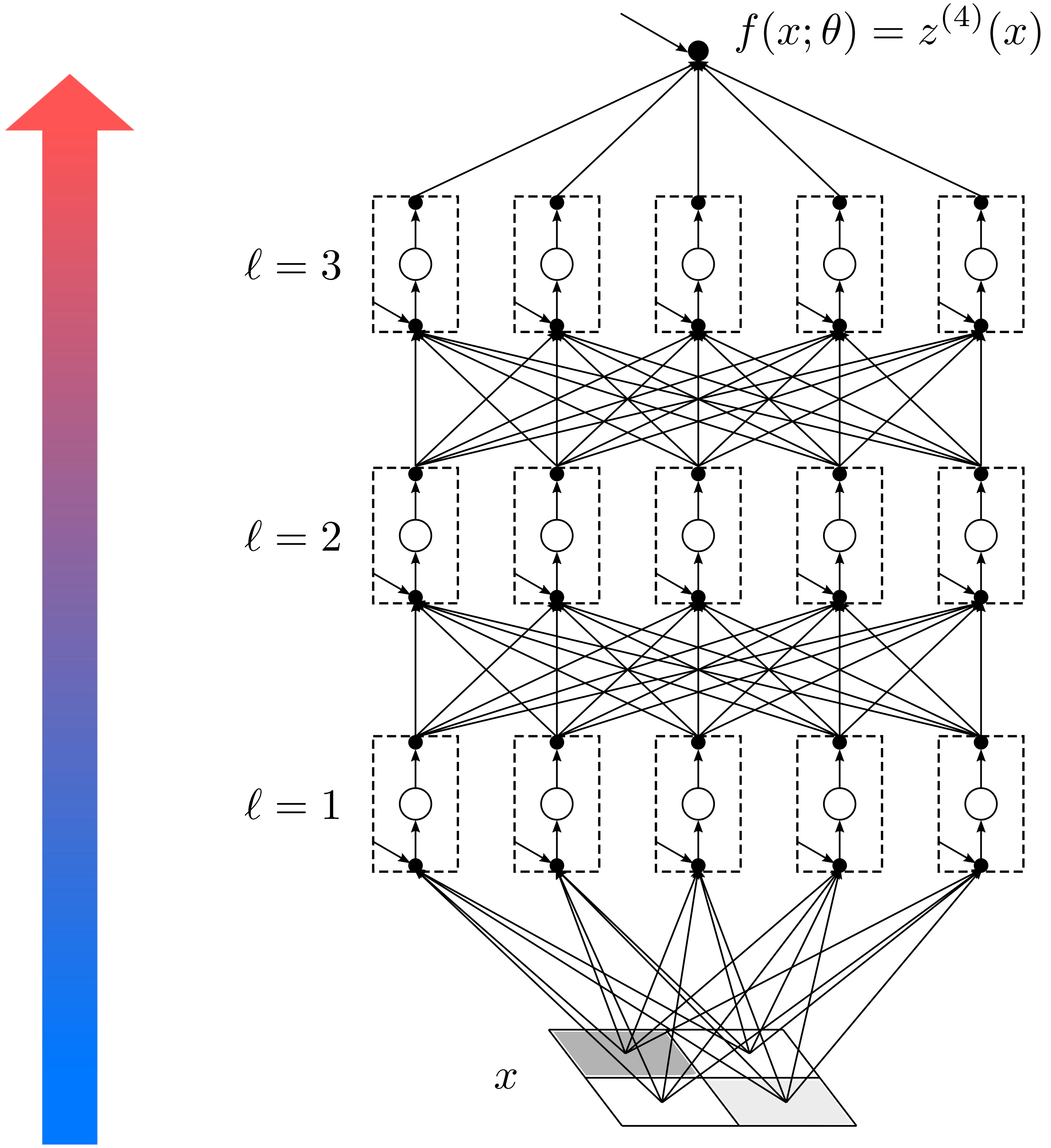
$$p\left(\widehat{z}^{(4)}, \widehat{H}^{(4)}, \widehat{\mathrm{d}H}^{(4)}, \ldots\right)$$

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

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

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

Strategy: Always Forward



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

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

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

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

(*RG-flow interpretation*: §4.6 of [arXiv:2106.10165](https://arxiv.org/abs/2106.10165))

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}, P_{....}, Q_{....}, R_{....}, S_{....}, T_{....}, U_{....}$$

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}, P_{....}, Q_{....}, R_{....}, S_{....}, T_{....}, U_{....}$$

non-Gaussianity

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}, P_{....}, Q_{....}, R_{....}, S_{....}, T_{....}, U_{....}$$

non-Gaussianity

NTK mean

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

NTK fluctuations (agitated NTK)

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}, P_{....}, Q_{....}, R_{....}, S_{....}, T_{....}, U_{....}$$

non-Gaussianity

NTK mean

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, \overbrace{A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}}^{\text{NTK fluctuations (agitated NTK)}}, \underbrace{P_{....}, Q_{....}}_{\text{dNTK (defrosted NTK)}}, \overbrace{R_{....}, S_{....}, T_{....}, U_{....}}^{\text{ddNTK (defrosted NTK)}}$$

non-Gaussianity

NTK mean

Solved EVERYTHING

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2},\, V_{(\delta_1\delta_2)(\delta_3\delta_4)},\, H_{\delta_1\delta_2},\, \overbrace{A_{\delta_1\delta_2\delta_3\delta_4},\, B_{....},\, D_{....},\, F_{....}}^{\text{NTK fluctuations (agitated NTK)}},\, \underbrace{P_{....},\, Q_{....}}_{\text{dNTK (defrosted NTK)}},\, \overbrace{R_{....},\, S_{....},\, T_{....},\, U_{....}}^{\text{ddNTK (defrosted NTK)}}$$

non-Gaussianity

NTK mean

* all recursively computable

Solved EVERYTHING and...

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....}, \underbrace{P_{....}, Q_{....}}_{\text{dNTK}}, \overbrace{R_{....}, S_{....}, T_{....}, U_{....}}^{\text{ddNTK}}$$

- Yes Representation Learning (cubic model with evolving features)
- Yes Algorithm Dependence (encapsulated by algorithm projectors)

$$\propto \frac{L}{n}$$

complex enough to capture rich phenomenology of real <u>deep</u> neural networks

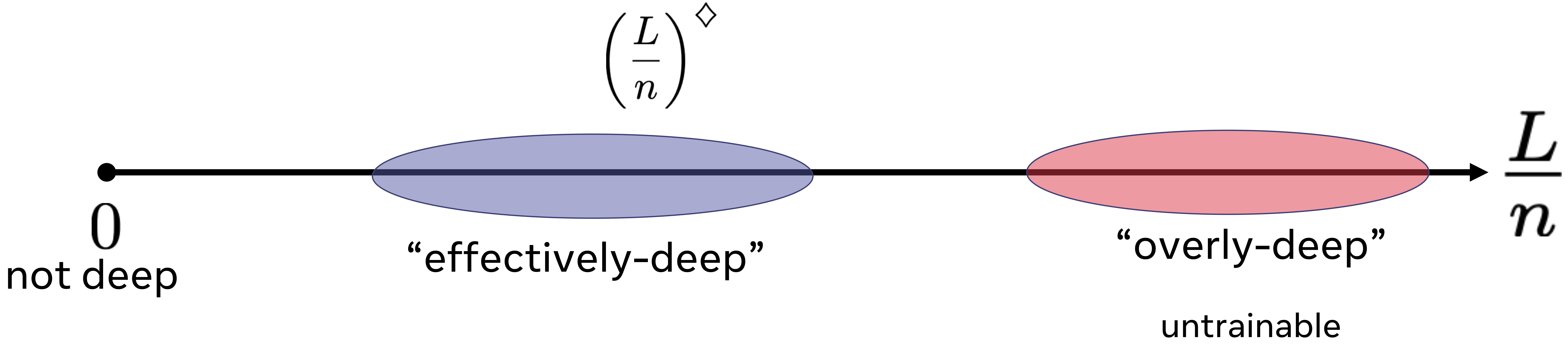
A Word about Overly-Deep Neural Networks

$p(z^{\star}_{i;\delta})$ nearly-Gaussian

$$G_{\delta_1\delta_2}, V_{(\delta_1\delta_2)(\delta_3\delta_4)}, H_{\delta_1\delta_2}, \overbrace{A_{\delta_1\delta_2\delta_3\delta_4}, B_{....}, D_{....}, F_{....},}^{\text{NTK fluctuations}} P_{....}, Q_{....}, R_{....}, S_{....}, T_{....}, U_{....}$$

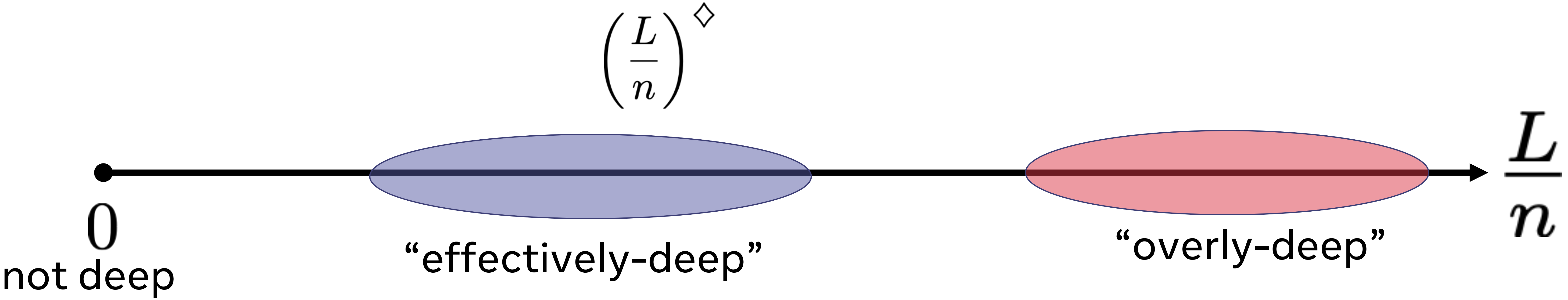
- Instantiation-to-instantiation fluctuations $\propto \frac{L}{n}$

A Word about Overly-Deep Neural Networks



Summary

- $n = \infty$
  simple, no representation learning, no algorithm dependence
- $n \gg L$
  a little more complex but tractable, yes representation learning & algorithm dependence
- $L \gg n$
  too complex, chaotic, untrainable



4. The Principles

The Principle of Sparsity for WIDE Neural Networks

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

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

statistics at *initialization*

statistics *after training*

The Principle of Sparsity for WIDE Neural Networks

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

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

statistics at *initialization* statistics *after training*

- Infinite width:

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

The Principle of Sparsity for WIDE Neural Networks

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

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

statistics at *initialization* statistics *after training*

- Infinite width:

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

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

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

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

The Principle of Criticality for DEEP Neural Networks

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

optimal initialization hyperparameters for <u>deep</u> neural networks

The Principle of Criticality for DEEP Neural Networks

- Taming *exponential* exploding/vanishing kernel problem:
  Poole et al. (NeurIPS2016); Raghu et al. (ICML2016); Schoenholz et al. (ICLR2017); §3 (DLN)+§5 (general) of arXiv:2106.10165
- Taming *exponential* exploding/vanishing gradient problem: §9.4
- Bayesian evidence: §6.3.1
- Generalization error: §10.3
- Mutual information: §A.2

$$\longrightarrow \boxed{(C_b, C_W)^{\text{critical}}}$$

optimal initialization hyperparameters for <u>deep</u> neural networks

The Principle of (Layer) Equivalence

- Taming *polynomial* exploding/vanishing gradient problem: §9.4

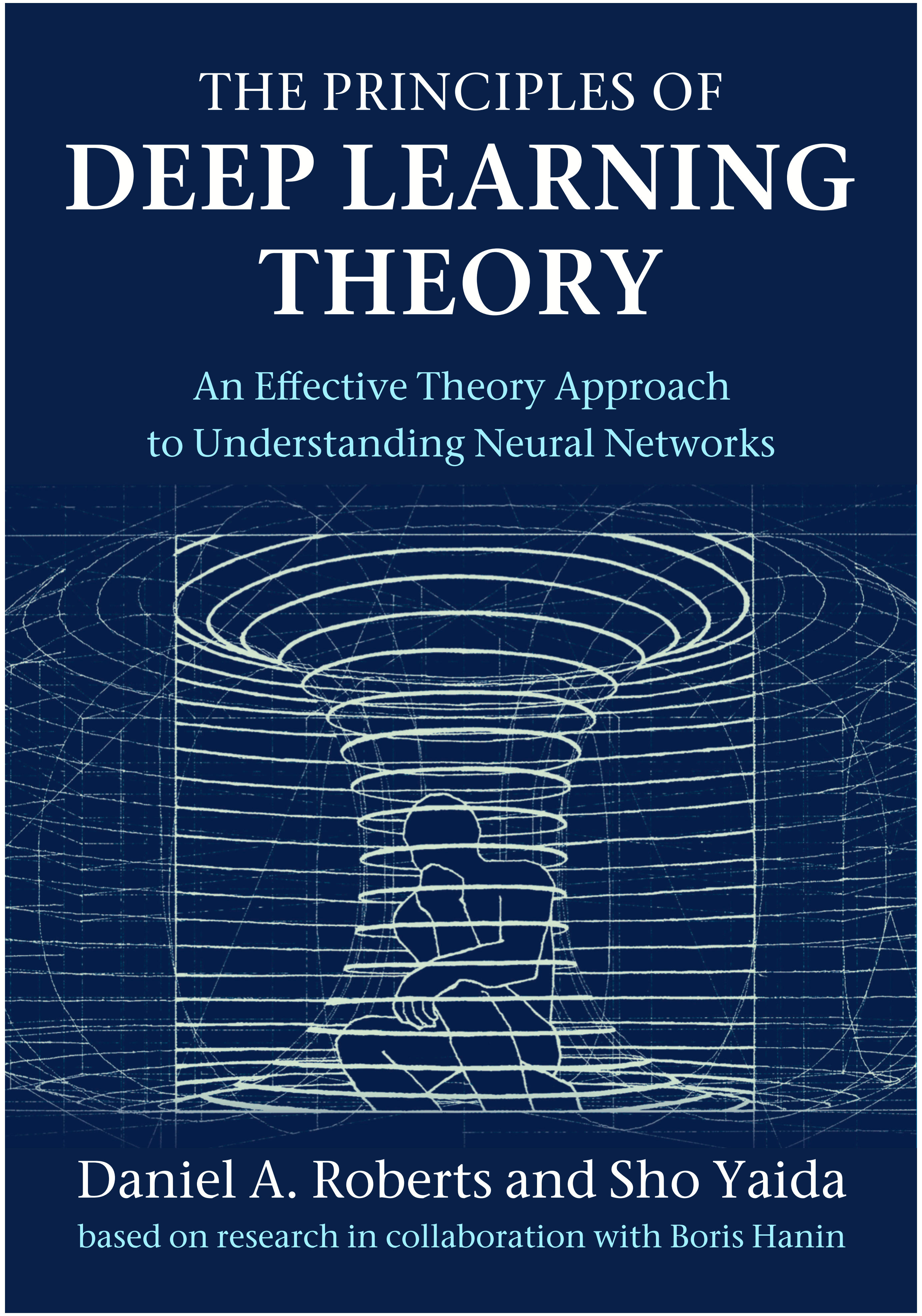
→ How to scale learning rates with depth
such that all groups of model parameters contribute equally

The Principle of Typicality

The way we study things:
analyze statistics and ask what happens *typically*

The Principle of Universality

The study of criticality can be organized into
various universality classes of activation functions



arXiv:2106.10165