Disentangling Trainability and Generalization in Deep Neural Networks

Lechao Xiao, Jeffrey Pennington, Samuel Schoenholz

Notation

Notation

- Fully-connected L-layer network:

$$\boldsymbol{z}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{d_l}} \boldsymbol{W}^{(l+1)} \phi\left(\boldsymbol{z}^{(l)}(\boldsymbol{x})\right) + \sigma_b \boldsymbol{b}^{(l+1)}$$

Notation

- Fully-connected L-layer network:

$$\boldsymbol{z}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{d_l}} \boldsymbol{W}^{(l+1)} \phi\left(\boldsymbol{z}^{(l)}(\boldsymbol{x})\right) + \sigma_b \boldsymbol{b}^{(l+1)}$$

- **Output:** $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{z}^{(L)}(\boldsymbol{x})$

Notation

- Fully-connected L-layer network:

$$\boldsymbol{z}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{d_l}} \boldsymbol{W}^{(l+1)} \phi\left(\boldsymbol{z}^{(l)}(\boldsymbol{x})\right) + \sigma_b \boldsymbol{b}^{(l+1)}$$

- **Output:** $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{z}^{(L)}(\boldsymbol{x})$

- **Initialization:** $W_{ij}^{(l)} \sim \mathcal{N}(0,1)$ and $b_i^{(l)} \sim \mathcal{N}(0,1)$

Notation

- Fully-connected L-layer network:

$$\boldsymbol{z}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{d_l}} \boldsymbol{W}^{(l+1)} \phi\left(\boldsymbol{z}^{(l)}(\boldsymbol{x})\right) + \sigma_b \boldsymbol{b}^{(l+1)}$$

- **Output:** $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{z}^{(L)}(\boldsymbol{x})$

- **Initialization:** $W_{ij}^{(l)} \sim \mathcal{N}(0,1)$ and $b_i^{(l)} \sim \mathcal{N}(0,1)$

- We can control **variances** σ_w, σ_b and study how the network behaves when varying these

Setup

Setup

- Data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d_0} \times \mathbb{R}$ for $i = 1, \ldots, n$

Setup

- Data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{d_0} \times \mathbb{R}$ for $i = 1, \ldots, n$
- Stack features and targets into matrices:

$$\boldsymbol{X} \in \mathbb{R}^{n \times d_0} \text{ and } \boldsymbol{y} \in \mathbb{R}^{n}$$

Recap: NNGP

Recap: NNGP

- At **initialization:**

$$f_{\boldsymbol{\theta}(0)}(\cdot) \xrightarrow{d_1,\ldots,d_L \to \infty} \mathcal{GP}(0, \Sigma^{(L)})$$

Recap: NNGP

- At **initialization:**

$$f_{\boldsymbol{\theta}(0)}(\cdot) \xrightarrow{d_1,\ldots,d_L \to \infty} \mathcal{GP}(0, \Sigma^{(L)})$$

where we have the recursion:

$$\Sigma^{(l)}\Big|_{x,x'} = \begin{pmatrix} \Sigma^{(l)}(x,x) & \Sigma^{(l)}(x,x') \\ \Sigma(x',x) & \Sigma^{(l)}(x',x') \end{pmatrix}$$

1) $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$

2) $\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \Sigma^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'})\right)}\Big[\phi(z_1)\phi(z_2)\Big] + \sigma_b^2$

Recap: NNGP

- At **initialization:**

$$f_{\boldsymbol{\theta}(0)}(\cdot) \xrightarrow{d_1,\ldots,d_L \to \infty} \mathcal{GP}(0, \Sigma^{(L)})$$

where we have the recursion:

$$\Sigma^{(l)}|_{x,x'} = \begin{pmatrix} \Sigma^{(l)}(x,x) & \Sigma^{(l)}(x,x') \\ \Sigma(x',x) & \Sigma^{(l)}(x',x') \end{pmatrix}$$

1) $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$

2) $\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \Sigma^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'})\right)} \left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2$

- If network has **multiple** outputs:

$$f_{\boldsymbol{\theta},i} \perp f_{\boldsymbol{\theta},j}$$

Recap: Neural Tangent Kernel (I)

Recap: Neural Tangent Kernel (I)

- Empirical **NTK**:

$$\hat{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x})\right)^T \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x}')$$

Recap: Neural Tangent Kernel (I)

- Empirical **NTK**:

$$\hat{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x})\right)^T \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x}')$$

- Governs the **dynamics**:

$$\partial_t f_{\boldsymbol{\theta}(t)}(\boldsymbol{x}) = -\hat{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})^T \frac{\partial L(f_{\boldsymbol{\theta}(t)}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_{\boldsymbol{\theta}(t)}(\boldsymbol{X})}$$

Recap: Neural Tangent Kernel (I)

- Empirical **NTK**:

$$\hat{\boldsymbol{\Theta}}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x})\right)^T \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}(t)}(\boldsymbol{x}')$$

- Governs the **dynamics**:

$$\partial_t f_{\boldsymbol{\theta}(t)}(\boldsymbol{x}) = -\hat{\boldsymbol{\Theta}}_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})^T \frac{\partial L(f_{\boldsymbol{\theta}(t)}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_{\boldsymbol{\theta}(t)}(\boldsymbol{X})}$$

- **Infinite-width** limit:

Deterministic and time-independent

$$\hat{\boldsymbol{\Theta}}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \xrightarrow{d_1, \ldots, d_L \to \infty} \boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$$

Recap: Neural Tangent Kernel (II)

Recap: Neural Tangent Kernel (II)

- Again recursive structure:

$\dot{\Sigma}^{(l+1)}(\underline{x}, \underline{x}') = \mathbb{E}_{z \sim \mathcal{N}(\underline{0}, \Sigma^{(l+1)}|_{\underline{x}, \underline{x}'})}[\dot{\phi}(z_1)\dot{\phi}(z_2)]$

1) $\mathbf{\Theta}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$

2) $\mathbf{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

Recap: Neural Tangent Kernel (II)

- Again recursive structure:

$\dot{\Sigma}^{(l+1)}(\underline{x}, \underline{x}') = \mathbb{E}_{z \sim \mathcal{N}(\underline{0}, \Sigma^{(l+1)}|_{\underline{x}, \underline{x}'})}[\dot{\phi}(z_1)\dot{\phi}(z_2)]$

1) $\mathbf{\Theta}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$
2) $\mathbf{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

- **Closed-form** expression for NN at $t = \infty$ and squared loss:

$$\begin{aligned} f_\infty(\boldsymbol{x}) &= \left(\mathbf{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\mathbf{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\ &= P(\mathbf{\Theta}^{(L)})\boldsymbol{y} \end{aligned}$$

Recap: Neural Tangent Kernel (II)

- Again recursive structure:

$$\dot{\Sigma}^{(l+1)}(\underline{x},\underline{x}') = \mathbb{E}_{z \sim \mathcal{N}(\underline{0}, \Sigma^{(l+1)}|_{\underline{x},\underline{x}'})}\left[\dot{\phi}(z_1)\dot{\phi}(z_2)\right]$$

  1) $\boldsymbol{\Theta}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$
  2) $\boldsymbol{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

- **Closed-form** expression for NN at $t = \infty$ and squared loss:

$$\begin{aligned} f_\infty(\boldsymbol{x}) &= \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\ &= P(\boldsymbol{\Theta}^{(L)})\boldsymbol{y} \end{aligned}$$

- We are assuming $f_0(\boldsymbol{x}) = 0$ here

What Have We Considered So Far?

- We considered very **overparametrized** networks:

$$\#params >> \#samples$$

What Have We Considered So Far?

- We considered very **overparametrized** networks:

$$\#params >> \#samples$$

- Overparametrization achieved with **huge widths**

$$d_1, \ldots, d_L \to \infty$$

What Have We Considered So Far?

- We considered very **overparametrized** networks:

$$\#params >> \#samples$$

- Overparametrization achieved with **huge widths**

$$d_1, \ldots, d_L \to \infty$$

- What happens if we let **depth** L go to infinity as well?

Large-Depth Limit

Large-Depth Limit

Two kinds of limits:

Large-Depth Limit

Two kinds of limits:

- Take width and depth to infinity **simultaneously**

Large-Depth Limit

Two kinds of limits:

- Take width and depth to infinity **simultaneously**
- Take width limit **first** and **then** consider depth limit

Large-Depth Limit

Two kinds of limits:

- Take width and depth to infinity **simultaneously**
- Take width limit **first** and **then** consider depth limit
- Two **different** results!

Simultaneous Limit

Simultaneous Limit

- Very **difficult** to analyze!

Simultaneous Limit

- Very **difficult** to analyze!
- "*Finite Depth and Width Corrections to the Neural Tangent Kernel*" analyzes this setting

Simultaneous Limit

- Very **difficult** to analyze!
- "*Finite Depth and Width Corrections to the Neural Tangent Kernel*" analyzes this setting
- They are only able to analyze **diagonal** entries $\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x})$

Simultaneous Limit

- Very **difficult** to analyze!
- "*Finite Depth and Width Corrections to the Neural Tangent Kernel*" analyzes this setting
- They are only able to analyze **diagonal** entries $\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x})$
- They prove that **SGD** induces change $\Delta\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x}')$ with

$$\text{var}(\Delta\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x}')) > 0$$

Simultaneous Limit

- Very **difficult** to analyze!
- "*Finite Depth and Width Corrections to the Neural Tangent Kernel*" analyzes this setting
- They are only able to analyze **diagonal** entries $\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x})$
- They prove that **SGD** induces change $\Delta\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x}')$ with

$$\text{var}(\Delta\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{x}, \boldsymbol{x}')) > 0$$

$\implies$ **Feature-learning** happening!

Sequential Limit

Sequential Limit

- **Easier** to analyze!

Sequential Limit

- **Easier** to analyze!
- Study the **dynamical systems**:

1. $\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \Sigma^{(l)}|_{\boldsymbol{x},\boldsymbol{x}'})\right)} \left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2$
2. $\boldsymbol{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

Sequential Limit

- **Easier** to analyze!
- Study the **dynamical systems**:

1. $\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \Sigma^{(l)}|_{\boldsymbol{x}, \boldsymbol{x}'})\right)}\left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2$
2. $\boldsymbol{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

- How do they behave as $l \to \infty$ as a function of σ_w and σ_b?

Sequential Limit

- **Easier** to analyze!
- Study the **dynamical systems**:

  1. $\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \Sigma^{(l)}|_{\boldsymbol{x}, \boldsymbol{x}'})\right)} \left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2$
  2. $\boldsymbol{\Theta}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

- How do they behave as $l \to \infty$ as a function of σ_w and σ_b?
- What happens to

$$P(\boldsymbol{\Theta}^{(L)}) = \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1}$$

This Work

This Work

- Consider the **sequential limit**

This Work

- Consider the **sequential limit**
- **Incremental**:
  - "*Exponential expressivity in deep neural networks through transient chaos*"
  - "*Deep Information Propagation*"

This Work

- Consider the **sequential limit**
- **Incremental**:
  - "*Exponential expressivity in deep neural networks through transient chaos*"
  - "*Deep Information Propagation*"
- Connect **chaotic** and **ordered** regimes with generalization and trainability

This Work

- Consider the **sequential limit**
- **Incremental**:
  - "*Exponential expressivity in deep neural networks through transient chaos*"
  - "*Deep Information Propagation*"
- Connect **chaotic** and **ordered** regimes with generalization and trainability
- How to characterize **generalization** and **trainability**?

Trainability (I)

Trainability (I)

- Recall that we have closed-form training dynamics:

$$f_t(\boldsymbol{x}) = \boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X}) \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \left(\mathbb{1} - e^{-t\eta \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})}\right) \boldsymbol{y}$$

Trainability (I)

- Recall that we have closed-form training dynamics:

$$f_t(\boldsymbol{x}) = \boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X}) \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \right)^{-1} \left(\mathbb{1} - e^{-t\eta \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})} \right) \boldsymbol{y}$$

- On the **training** set we have:

$$f_t(\boldsymbol{X}) = \left(\mathbb{1} - e^{-t\eta \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})} \right) \boldsymbol{y}$$

Trainability (I)

- Recall that we have closed-form training dynamics:

$$f_t(\boldsymbol{x}) = \boldsymbol{\Theta}^{(L)}(\boldsymbol{x}, \boldsymbol{X}) \left(\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \left(\mathbb{1} - e^{-t\eta\boldsymbol{\Theta}^{(L)}(\boldsymbol{X},\boldsymbol{X})}\right) \boldsymbol{y}$$

- On the **training** set we have:

$$f_t(\boldsymbol{X}) = \left(\mathbb{1} - e^{-t\eta\boldsymbol{\Theta}^{(L)}(\boldsymbol{X},\boldsymbol{X})}\right) \boldsymbol{y}$$

- **Diagonalize** NTK: $\boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{U}^T \boldsymbol{\Lambda} \boldsymbol{U}$ which leads to

rotated labels

$$\tilde{f}_{t,i}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-t\eta\lambda_i}\right) \tilde{y}_i$$

rotated predictions

Trainability (II)

Trainability (II)

- **Maximum** feasible learning rate $\eta \sim \frac{2}{\lambda_1}$ (Recall Seyed's presentation)

Trainability (II)

- **Maximum** feasible learning rate $\eta \sim \frac{2}{\lambda_1}$ (Recall Seyed's presentation)

- Plugging this into the dynamics:

$$\tilde{f}_{t,i}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_i}{\lambda_1}}\right)\tilde{y}_i$$

Trainability (II)

- **Maximum** feasible learning rate $\eta \sim \frac{2}{\lambda_1}$ (Recall Seyed's presentation)

- Plugging this into the dynamics:

$$\tilde{f}_{t,i}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_i}{\lambda_1}}\right)\tilde{y}_i$$

- **Hardest** component to learn follows

$\kappa = \frac{\lambda_1}{\lambda_m}$

$$\tilde{f}_{t,m}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_m}{\lambda_1}}\right)\tilde{y}_m = \left(\mathbb{1} - e^{-2t\kappa^{-1}}\right)\tilde{y}_m$$

Trainability (II)

- **Maximum** feasible learning rate $\eta \sim \frac{2}{\lambda_1}$ (Recall Seyed's presentation)

- Plugging this into the dynamics:

$$\tilde{f}_{t,i}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_i}{\lambda_1}}\right)\tilde{y}_i$$

- **Hardest** component to learn follows

$\kappa = \frac{\lambda_1}{\lambda_m}$

$$\tilde{f}_{t,m}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_m}{\lambda_1}}\right)\tilde{y}_m = \left(\mathbb{1} - e^{-2t\kappa^{-1}}\right)\tilde{y}_m$$

- **Hence:** If $\kappa \xrightarrow{L\to\infty} \infty$:

$$\tilde{f}^{(\infty)}_{t,m}(\boldsymbol{X}) = 0 \neq \tilde{y}_m \quad \forall t$$

Trainability (II)

- **Maximum** feasible learning rate $\eta \sim \frac{2}{\lambda_1}$ (Recall Seyed's presentation)

- Plugging this into the dynamics:

$$\tilde{f}_{t,i}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_i}{\lambda_1}}\right)\tilde{y}_i$$

- **Hardest** component to learn follows

$\kappa = \frac{\lambda_1}{\lambda_m}$

$$\tilde{f}_{t,m}(\boldsymbol{X}) = \left(\mathbb{1} - e^{-2t\frac{\lambda_m}{\lambda_1}}\right)\tilde{y}_m = \left(\mathbb{1} - e^{-2t\kappa^{-1}}\right)\tilde{y}_m$$

- **Hence:** If $\kappa \xrightarrow{L\to\infty} \infty$:

$$\tilde{f}^{(\infty)}_{t,m}(\boldsymbol{X}) = 0 \neq \tilde{y}_m \quad \forall t$$

$\implies$ **Cannot** learn training set

Generalization

Generalization

- To study the **generalization** ability, look at $P(\boldsymbol{\Theta}^{(\infty)})$

Generalization

- To study the **generalization** ability, look at $P(\boldsymbol{\Theta}^{(\infty)})$
- The limit will be a **function** of $(\sigma_{w}, \sigma_{\boldsymbol{b}})$

Generalization

- To study the **generalization** ability, look at $P(\mathbf{\Theta}^{(\infty)})$
- The limit will be a **function** of (σ_w, σ_b)
- Different regimes (chaotic and ordered) in the (σ_w, σ_b) plane where $P(\mathbf{\Theta}^{(\infty)})$ is **data-dependent** or **data-independent**

Generalization

- To study the **generalization** ability, look at $P(\boldsymbol{\Theta}^{(\infty)})$
- The limit will be a **function** of (σ_w, σ_b)
- Different regimes (chaotic and ordered) in the (σ_w, σ_b) plane where $P(\boldsymbol{\Theta}^{(\infty)})$ is **data-dependent** or **data-independent**
- Data-dependence of course **not** best measure to study generalization performance

Goals

Goals

- Characterize evolution of $\kappa^{(l)}$ to study trainability

$\implies$ We hence need to control the spectrum of $\boldsymbol{\Theta}^{(l)}$

Goals

- Characterize evolution of $\kappa^{(l)}$ to study trainability

  $\implies$ We hence need to control the spectrum of $\mathbf{\Theta}^{(l)}$

- Study trajectory of $P(\mathbf{\Theta}^{(l)})$ to understand generalization

NNGP and its Fixed Points

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

  Approximate $q_{aa}^{(l)} \approx q^* = \lim_{l \to \infty} q_{aa}^{(l)}$ in all calculations to follow.

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

  Approximate $q_{aa}^{(l)} \approx q^* = \lim_{l \to \infty} q_{aa}^{(l)}$ in all calculations to follow.

  Correlation between representations

- Introduce $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{\sqrt{q_{aa}^{(l)} q_{bb}^{(l)}}} = \frac{q_{ab}^{(l)}}{q^*}$ and write off-diagonal terms:

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

  Approximate $q_{aa}^{(l)} \approx q^* = \lim_{l \to \infty} q_{aa}^{(l)}$ in all calculations to follow.

  Correlation between representations

- Introduce $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{\sqrt{q_{aa}^{(l)} q_{bb}^{(l)}}} = \frac{q_{ab}^{(l)}}{q^*}$ and write off-diagonal terms:

$$q_{ab}^{(l+1)} = \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \Sigma^{(l)}\big|_{\boldsymbol{x}, \boldsymbol{x}'}\right)\right)} \left[\phi(z_1)\phi(z_2)\right] + \sigma_b^2$$

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

  Approximate $q_{aa}^{(l)} \approx q^* = \lim_{l \to \infty} q_{aa}^{(l)}$ in all calculations to follow.

Correlation between representations

- Introduce $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{\sqrt{q_{aa}^{(l)} q_{bb}^{(l)}}} = \frac{q_{ab}^{(l)}}{q^*}$ and write off-diagonal terms:

Gaussian measure: $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z_1^2} dz_1$

$$
\begin{aligned}
q_{ab}^{(l+1)} &= \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{0}, \Sigma^{(l)}\big|_{\boldsymbol{x},\boldsymbol{x}'}\right)\right)} \Big[\phi(z_1)\phi(z_2)\Big] + \sigma_b^2 \\
&= \sigma_w^2 \int \phi\left(\sqrt{q^*} z_1\right) \phi\left(\sqrt{q^*}(c_{ab}^{(l)} z_1 + \sqrt{1 - (c_{ab}^{(l)})^2} z_2)\right) \mathcal{D}z_1 \mathcal{D}z_2 + \sigma_b^2
\end{aligned}
$$

NNGP and its Fixed Points

- **Assume :** Diagonal entries $q_{aa}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_a)$ converge way quicker than off-diagonal terms $q_{ab}^{(l)} := \Sigma^{(l)}(\boldsymbol{x}_a, \boldsymbol{x}_b)$

  Approximate $q_{aa}^{(l)} \approx q^* = \lim_{l \to \infty} q_{aa}^{(l)}$ in all calculations to follow.

Correlation between representations

- Introduce $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{\sqrt{q_{aa}^{(l)} q_{bb}^{(l)}}} = \frac{q_{ab}^{(l)}}{q^*}$ and write off-diagonal terms:

Gaussian measure: $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_1^2} dz_1$

$$
\begin{aligned}
q_{ab}^{(l+1)} &= \sigma_w^2 \mathbb{E}_{\left(\boldsymbol{z} \sim \mathcal{N}\left(\mathbf{0}, \Sigma^{(l)}\big|_{\boldsymbol{x}, \boldsymbol{x}'}\right)\right)} \Big[\phi(z_1)\phi(z_2)\Big] + \sigma_b^2 \\
&= \sigma_w^2 \int \phi\left(\sqrt{q^*} z_1\right) \phi\left(\sqrt{q^*}(c_{ab}^{(l)} z_1 + \sqrt{1 - (c_{ab}^{(l)})^2} z_2)\right) \mathcal{D}z_1 \mathcal{D}z_2 + \sigma_b^2 \\
&= \sigma_w^2 \int \phi\left(u_1(z_1)\right) \phi\left(u_2(z_1, z_2, c_{ab}^{(l)})\right) \mathcal{D}z_1 \mathcal{D}z_2 + \sigma_b^2
\end{aligned}
$$

Recursive Correlation

Recursive Correlation

- Define the function

$$f(c) = \frac{1}{q^*}\left(\sigma_w^2 \int \phi\left(u_1(z_1)\right)\phi\left(u_2(z_1, z_2, c)\right)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2\right)$$

Recursive Correlation

- Define the function

$$f(c) = \frac{1}{q^*}\left(\sigma_w^2 \int \phi\left(u_1(z_1)\right)\phi\left(u_2(z_1, z_2, c)\right)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2\right)$$

- **Observe:** $f(1) = 1 \implies c^* = 1$ is a fixed point of f

Recursive Correlation

- Define the function

$$f(c) = \frac{1}{q^*}\left(\sigma_w^2 \int \phi\left(u_1(z_1)\right)\phi\left(u_2(z_1, z_2, c)\right)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2\right)$$

- **Observe:** $f(1) = 1 \implies c^* = 1$ is a fixed point of f

- **Attracting** or **repulsive** fixed point?

Recursive Correlation

- Define the function

$$f(c) = \frac{1}{q^*}\left(\sigma_w^2 \int \phi\left(u_1(z_1)\right)\phi\left(u_2(z_1, z_2, c)\right)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2\right)$$

- **Observe:** $f(1) = 1 \implies c^* = 1$ is a fixed point of f

- **Attracting** or **repulsive** fixed point?

- Calculus 101: Equivalent to $f'(1) < 1$ or $f'(1) > 1$?

Recursive Correlation

- Define the function

$$f(c) = \frac{1}{q^*}\left(\sigma_w^2 \int \phi\left(u_1(z_1)\right)\phi\left(u_2(z_1, z_2, c)\right)\mathcal{D}z_1\mathcal{D}z_2 + \sigma_b^2\right)$$

- **Observe:** $f(1) = 1 \implies c^* = 1$ is a fixed point of f

- **Attracting** or **repulsive** fixed point?

- Calculus 101: Equivalent to $f'(1) < 1$ or $f'(1) > 1$?

- Turns out: $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z$

  $\implies$ Depends on σ_w and σ_b

What are Chaotic and Ordered Regimes?

What are Chaotic and Ordered Regimes?

- **Ordered**: $f'(1) < 1 \implies$ Correlation $c_{ab}^{(l)}$ between hidden representations of inputs $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ converges to $c^* = 1$:

$$\lim_{l \to \infty} c_{ab}^{(l)} = 1$$

What are Chaotic and Ordered Regimes?

- **Ordered**: $f'(1) < 1 \implies$ Correlation $c_{ab}^{(l)}$ between hidden representations of inputs $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ converges to $c^* = 1$:

$$\lim_{l \to \infty} c_{ab}^{(l)} = 1$$

This means that two outputs always **collapse onto** each other in the infinite depth limit

What are Chaotic and Ordered Regimes?

- **Ordered**: $f'(1) < 1 \implies$ Correlation $c_{ab}^{(l)}$ between hidden representations of inputs $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ converges to $c^* = 1$:

  $$\lim_{l \to \infty} c_{ab}^{(l)} = 1$$

  This means that two outputs always **collapse onto** each other in the infinite depth limit

- **Chaotic**: $f'(1) > 1 \implies c^* = 1$ is repulsive

What are Chaotic and Ordered Regimes?

- **Ordered**: $f'(1) < 1 \implies$ Correlation $c_{ab}^{(l)}$ between hidden representations of inputs $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ converges to $c^* = 1$:

  $$\lim_{l \to \infty} c_{ab}^{(l)} = 1$$

  This means that two outputs always **collapse onto** each other in the infinite depth limit

- **Chaotic**: $f'(1) > 1 \implies c^* = 1$ is repulsive

  Since $c_{ab}^{(l)} < 1$:

  $$\lim_{l \to \infty} c_{ab}^{(l)} < 1$$

What are Chaotic and Ordered Regimes?

- **Ordered**: $f'(1) < 1 \implies$ Correlation $c_{ab}^{(l)}$ between hidden representations of inputs $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ converges to $c^* = 1$:

  $$\lim_{l \to \infty} c_{ab}^{(l)} = 1$$

  This means that two outputs always **collapse onto** each other in the infinite depth limit

- **Chaotic**: $f'(1) > 1 \implies c^* = 1$ is repulsive

  Since $c_{ab}^{(l)} < 1$:

  $$\lim_{l \to \infty} c_{ab}^{(l)} < 1$$

  Two outputs thus become more and more **dissimilar** to each other

Restate the Dynamics

Restate the Dynamics

- Dynamics in terms of new notation:

Restate the Dynamics

- Dynamics in terms of new notation:

1a) $q_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2$

1b) $q_{aa}^{(l+1)} = q^*$

$= \mathbb{E}_{z \sim \mathcal{N}\left(0, \begin{pmatrix} q^* & q_{ab}^{(l)} \\ q_{ab}^{(l)} & q^* \end{pmatrix}\right)} \left[\phi(z_1)\phi(z_2)\right]$

Restate the Dynamics

- Dynamics in terms of new notation:

1a) $q_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2$

$= \mathbb{E}_{z \sim \mathcal{N}\left(0, \begin{pmatrix} q^* & q_{ab}^{(l)} \\ q_{ab}^{(l)} & q^* \end{pmatrix}\right)}\left[\phi(z_1)\phi(z_2)\right]$

1b) $q_{aa}^{(l+1)} = q^*$

2a) $p_{ab}^{(l+1)} = q_{ab}^{(l+1)} + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^{(l)}) p_{ab}^{(l)}$

2b) $p_{aa}^{(l+1)} = q^* + \sigma_w^2 \dot{\mathcal{T}}(q^*) p_{aa}^{(l)} = q^* + f'(1) p_{aa}^{(l)}$

Restate the Dynamics

- Dynamics in terms of new notation:

1a) $q_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2$

1b) $q_{aa}^{(l+1)} = q^*$

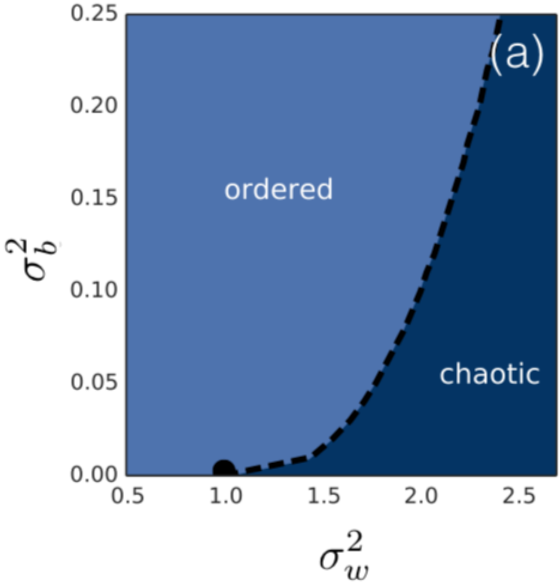
$= \mathbb{E}_{z \sim \mathcal{N}\left(0, \begin{pmatrix} q^* & q_{ab}^{(l)} \\ q_{ab}^{(l)} & q^* \end{pmatrix}\right)} \left[\phi(z_1)\phi(z_2)\right]$

2a) $p_{ab}^{(l+1)} = q_{ab}^{(l+1)} + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^{(l)}) p_{ab}^{(l)}$

2b) $p_{aa}^{(l+1)} = q^* + \sigma_w^2 \dot{\mathcal{T}}(q^*) p_{aa}^{(l)} = q^* + f'(1) p_{aa}^{(l)}$

- Analyze the limits in the two different regimes

An Example



0.25
0.20
0.15
0.10
0.05
0.00
0.5
1.0
1.5
2.0
2.5
(a)
ordered
chaotic
σ_b^2
σ_w^2

Chaotic Regime: Limits

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)}$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Hence: $p_{aa}^{(l+1)} = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Hence: $p_{aa}^{(l+1)} = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^* \approx f'(1)^l q^* \xrightarrow{l \to \infty} \infty$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \longrightarrow \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \longrightarrow \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Hence: $p_{aa}^{(l+1)} = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^* \approx f'(1)^l q^* \xrightarrow{l \longrightarrow \infty} \infty$

2.a) Taking the limit on both sides of **2a)**:

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Hence: $p_{aa}^{(l+1)} = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^* \approx f'(1)^l q^* \xrightarrow{l \to \infty} \infty$

2.a) Taking the limit on both sides of **2a)**:

$$p_{ab}^* = q_{ab}^* + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^*) p_{ab}^*$$

Chaotic Regime: Limits

1.b) $q_{aa}^{(l)} = q^*$ already given

1.a) **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* < 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} c^* q^*$

2.b) $p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^*$

Hence: $p_{aa}^{(l+1)} = \frac{f'(1)^{l+1} - 1}{f'(1) - 1} q^* \approx f'(1)^l q^* \xrightarrow{l \to \infty} \infty$

2.a) Taking the limit on both sides of **2a)**:

$$p_{ab}^* = q_{ab}^* + \sigma_w^2 \dot{\mathcal{T}}(q_{ab}^*) p_{ab}^* \iff p_{ab}^* = \frac{q_{ab}^*}{1 - f'(c^*)}$$

Chaotic Regime: Spectrum NNGP

Chaotic Regime: Spectrum NNGP

- We can summarize as follows:

Chaotic Regime: Spectrum NNGP

- We can summarize as follows:

$$\Sigma^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^*(c^*\mathbf{1}\mathbf{1}^T + (1-c^*)\mathbb{1}_{n\times n}) = \begin{pmatrix} q^* & c^*q^* & \dots & c^*q^* \\ c^*q^* & q^* & \dots & c^*q^* \\ \vdots & & & \\ c^*q^* & \dots & q^* & c^*q^* \\ c^*q^* & \dots & c^*q^* & q^* \end{pmatrix}$$

Chaotic Regime: Spectrum NNGP

- We can summarize as follows:

$$\Sigma^{(\infty)}(\boldsymbol{X},\boldsymbol{X}) = q^*(c^*\mathbf{1}\mathbf{1}^T + (1-c^*)\mathbb{1}_{n\times n}) = \begin{pmatrix} q^* & c^*q^* & \dots & c^*q^* \\ c^*q^* & q^* & \dots & c^*q^* \\ \vdots & & & \\ c^*q^* & \dots & q^* & c^*q^* \\ c^*q^* & \dots & c^*q^* & q^* \end{pmatrix}$$

- Has eigenvector $\boldsymbol{v}_1 = (1,\dots,1)$ with $\lambda_0 = q^*\left((n-1)c^* + 1\right)$

Chaotic Regime: Spectrum NNGP

- We can summarize as follows:

$$\Sigma^{(\infty)}(\boldsymbol{X},\boldsymbol{X}) = q^*(c^*\mathbf{1}\mathbf{1}^T + (1-c^*)\mathbb{1}_{n\times n}) = \begin{pmatrix} q^* & c^*q^* & \dots & c^*q^* \\ c^*q^* & q^* & \dots & c^*q^* \\ \vdots & & & \\ c^*q^* & \dots & q^* & c^*q^* \\ c^*q^* & \dots & c^*q^* & q^* \end{pmatrix}$$

- Has eigenvector $\boldsymbol{v}_1 = (1,\dots,1)$ with $\lambda_0 = q^*\left((n-1)c^* + 1\right)$

- Others are $\boldsymbol{v}_i = (\boldsymbol{e}_i - \boldsymbol{e}_{i+1})$ and $\lambda_i = \lambda_{\text{bulk}} = q^*(1-c^*)$

Chaotic Regime: Spectrum NNGP

- We can summarize as follows:

$$\Sigma^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^*(c^*\mathbf{1}\mathbf{1}^T + (1-c^*)\mathbb{1}_{n\times n}) = \begin{pmatrix} q^* & c^*q^* & \dots & c^*q^* \\ c^*q^* & q^* & \dots & c^*q^* \\ \vdots & & & \\ c^*q^* & \dots & q^* & c^*q^* \\ c^*q^* & \dots & c^*q^* & q^* \end{pmatrix}$$

- Has eigenvector $\boldsymbol{v}_1 = (1, \dots, 1)$ with $\lambda_0 = q^* \left((n-1)c^* + 1\right)$

- Others are $\boldsymbol{v}_i = (\boldsymbol{e}_i - \boldsymbol{e}_{i+1})$ and $\lambda_i = \lambda_{\text{bulk}} = q^*(1 - c^*)$

- $\kappa(\Sigma^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})) = \frac{\lambda_0}{\lambda_{\text{bulk}}} = \frac{(n-1)c^*+1}{1-c^*}$

Chaotic Regime: Spectrum NTK

Chaotic Regime: Spectrum NTK

- $\frac{1}{p^*_{aa}} \boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = \mathbb{1}_{n \times n}$

Chaotic Regime: Spectrum NTK

- $\frac{1}{p^*_{aa}} \boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = \mathbb{1}_{n \times n}$
- $\lambda_i \left(\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) \right) = \infty$

Chaotic Regime: Spectrum NTK

- $\frac{1}{p^*_{aa}}\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X},\boldsymbol{X}) = \mathbb{1}_{n\times n}$
- $\lambda_i\left(\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X},\boldsymbol{X})\right) = \infty$
- $\kappa\left(\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X},\boldsymbol{X})\right) = 1 \implies$ NTK **remains trainable** for infinite-depth in the chaotic regime

Ordered Regime: Limits

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)}$$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i$$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i \xrightarrow{l \to \infty} \sum_{i=0}^{\infty} q^* \left(f'(1)\right)^i$$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i \xrightarrow{l \to \infty} \sum_{i=0}^{\infty} q^* \left(f'(1)\right)^i$$

$$= \frac{q^*}{1 - f'(1)}$$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i \xrightarrow{l \to \infty} \sum_{i=0}^{\infty} q^* \left(f'(1)\right)^i$$

$$= \frac{q^*}{1 - f'(1)}$$

- Taking the limit on both sides of **2a)**:

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i \xrightarrow{l \to \infty} \sum_{i=0}^{\infty} q^* \left(f'(1)\right)^i$$
$$= \frac{q^*}{1 - f'(1)}$$

- Taking the limit on both sides of **2a)**:

$$p_{ab}^* = q_{ab}^* + \sigma_w^2 \dot{\mathcal{T}}(q^*) p_{ab}^*$$

Ordered Regime: Limits

- $q_{aa}^{(l)} = q^*$ already given
- **Recall**: $c_{ab}^{(l)} = \frac{q_{ab}^{(l)}}{q^*} \xrightarrow{l \to \infty} c^* = 1 \implies q_{ab}^{(l)} \xrightarrow{l \to \infty} q^*$
- Using $f'(1) < 1$, we can apply the geometric series

$$p_{aa}^{(l+1)} = q^* + f'(1) p_{aa}^{(l)} = \cdots = \sum_{i=0}^{l+1} q^* f'(1)^i \xrightarrow{l \to \infty} \sum_{i=0}^{\infty} q^* \left(f'(1)\right)^i$$

$$= \frac{q^*}{1 - f'(1)}$$

- Taking the limit on both sides of **2a)**:

$$p_{ab}^* = q_{ab}^* + \sigma_w^2 \dot{\mathcal{T}}(q^*) p_{ab}^* \iff p_{ab}^* = \frac{q^*}{1 - f'(1)} = p^*$$

Ordered Regime: Spectrum NNGP

Ordered Regime: Spectrum NNGP

- We can summarize the NNGP as

$$\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^* \mathbf{1}\mathbf{1}^T$$

Ordered Regime: Spectrum NNGP

- We can summarize the NNGP as

$$\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^* \mathbf{1}\mathbf{1}^T$$

- Again, $\boldsymbol{v}_1 = \mathbf{1}$ with $\lambda_1(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})) = nq^*$

Ordered Regime: Spectrum NNGP

- We can summarize the NNGP as

$$\mathbf{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^* \mathbf{1}\mathbf{1}^T$$

- Again, $\boldsymbol{v}_1 = \mathbf{1}$ with $\lambda_1(\mathbf{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})) = nq^*$

- Moreover, $\boldsymbol{v}_i = \boldsymbol{e}_i - \boldsymbol{e}_{i+1}$ with $\lambda_{\text{bulk}}(\mathbf{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = 0$

Ordered Regime: Spectrum NNGP

- We can summarize the NNGP as

$$\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = q^* \mathbf{1}\mathbf{1}^T$$

- Again, $\boldsymbol{v}_1 = \mathbf{1}$ with $\lambda_1(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})) = nq^*$
- Moreover, $\boldsymbol{v}_i = \boldsymbol{e}_i - \boldsymbol{e}_{i+1}$ with $\lambda_{\text{bulk}}(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = 0$
- We can calculate the condition number as

$$\kappa\left(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})\right) = \frac{\lambda_1\left(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})\right)}{\lambda_{\text{bulk}}\left(\boldsymbol{\Sigma}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})\right)} = \infty$$

Ordered Regime: Spectrum NTK

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\mathbf{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T$$

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T$$

- We get the same eigen-structure as for the NNGP

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T$$

- We get the same eigen-structure as for the NNGP

- $\lambda_1\left(\boldsymbol{\Theta}^{(\infty)}\right) = np^*$

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^{*}\mathbf{1}\mathbf{1}^{T}$$

- We get the same eigen-structure as for the NNGP
- $\lambda_1\left(\boldsymbol{\Theta}^{(\infty)}\right) = np^{*}$
- $\lambda_{\text{bulk}}\left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T$$

- We get the same eigen-structure as for the NNGP
- $\lambda_1 \left(\boldsymbol{\Theta}^{(\infty)}\right) = np^*$
- $\lambda_{\text{bulk}} \left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$
- $\kappa \left(\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty$

Ordered Regime: Spectrum NTK

- We can express the infinite-depth NTK compactly via

$$\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T$$

- We get the same eigen-structure as for the NNGP
- $\lambda_1 \left(\boldsymbol{\Theta}^{(\infty)}\right) = np^*$
- $\lambda_{\text{bulk}} \left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$
- $\kappa \left(\boldsymbol{\Theta}^{(\infty)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty$

$\implies$ Networks become **untrainable**!

Finite Depth Correction

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$
- Expanding and Taylor-approximating:

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^*$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^*$
- Expanding and Taylor-approximating:

$$q_{ab}^{(l+1)} = q_{ab}^* + \epsilon_{ab}^{(l+1)}$$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$

- Expanding and Taylor-approximating:

  $$q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2$$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$

- Expanding and Taylor-approximating:

$$q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2 = \sigma_w^2 \mathcal{T}\left(q_{ab}^{*} + \epsilon_{ab}^{(l)}\right) + \sigma_b^2$$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$
- Expanding and Taylor-approximating:

$$q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} = \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2 = \sigma_w^2 \mathcal{T}\left(q_{ab}^{*} + \epsilon_{ab}^{(l)}\right) + \sigma_b^2$$

$$\overset{\text{Taylor}}{=} \sigma_w^2 \mathcal{T}\left(q_{ab}^{*}\right) + \sigma_b^2 + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right)$$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$

- Expanding and Taylor-approximating:

$$
\begin{aligned}
q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} &= \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2 = \sigma_w^2 \mathcal{T}\left(q_{ab}^{*} + \epsilon_{ab}^{(l)}\right) + \sigma_b^2 \\
&\overset{\text{Taylor}}{=} \sigma_w^2 \mathcal{T}\left(q_{ab}^{*}\right) + \sigma_b^2 + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right) \\
&= q_{ab}^{*} + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right)
\end{aligned}
$$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$
- Expanding and Taylor-approximating:

$$
\begin{aligned}
q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} &= \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2 = \sigma_w^2 \mathcal{T}\left(q_{ab}^{*} + \epsilon_{ab}^{(l)}\right) + \sigma_b^2 \\
&\overset{\text{Taylor}}{=} \sigma_w^2 \mathcal{T}\left(q_{ab}^{*}\right) + \sigma_b^2 + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right) \\
&= q_{ab}^{*} + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right)
\end{aligned}
$$

- **Hence:** $\epsilon_{ab}^{(l+1)} \approx \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} = f'(c^*)^{l+1} \epsilon_{ab}^{(0)}$

Finite Depth Correction

- Define $\epsilon_{ab}^{(l)} = q_{ab}^{(l)} - q_{ab}^{*}$ and $\delta_{ab}^{(l)} = p_{ab}^{(l)} - p_{ab}^{*}$
- Expanding and Taylor-approximating:

$$
\begin{aligned}
q_{ab}^{(l+1)} = q_{ab}^{*} + \epsilon_{ab}^{(l+1)} &= \sigma_w^2 \mathcal{T}\left(q_{ab}^{(l)}\right) + \sigma_b^2 = \sigma_w^2 \mathcal{T}\left(q_{ab}^{*} + \epsilon_{ab}^{(l)}\right) + \sigma_b^2 \\
\text{Taylor} &= \sigma_w^2 \mathcal{T}\left(q_{ab}^{*}\right) + \sigma_b^2 + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right) \\
&= q_{ab}^{*} + \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} + \mathcal{O}\left(\left(\epsilon_{ab}^{(l)}\right)^2\right)
\end{aligned}
$$

- **Hence:** $\epsilon_{ab}^{(l+1)} \approx \sigma_w^2 \dot{\mathcal{T}}\left(q_{ab}^{*}\right) \epsilon_{ab}^{(l)} = f'(c^*)^{l+1} \epsilon_{ab}^{(0)}$
- **Similarly:** $\delta_{ab}^{(l+1)} \approx f'(c^*)^{l+1} \left(\delta_{ab}^{(0)} + l\left(1 + \frac{f''(c^*)}{f'(c^*)} p_{ab}^{*}\right) \epsilon_{ab}^{(0)}\right)$

Chaotic Regime: Generalization

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

$$P(\boldsymbol{\Theta}^{(l)})\boldsymbol{y} = \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y}$$

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(l)})\boldsymbol{y} &= \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\
&\approx (p^*_{ab}\mathbf{1})^T \left(p^*_{ab}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + p^{(l)}_{aa}\mathbb{1}\right)^{-1} \boldsymbol{y}
\end{aligned}
$$

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(l)})\boldsymbol{y} &= \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\
&\approx (p^*_{ab}\mathbf{1})^T \left(p^*_{ab}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + p^{(l)}_{aa}\mathbb{1}\right)^{-1} \boldsymbol{y} \\
&= \frac{1}{p^{(l)}_{aa}}(p^*_{ab}\mathbf{1})^T \left(\frac{p^*_{ab}}{p^{(l)}_{aa}}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + \mathbb{1}\right)^{-1} \boldsymbol{y}
\end{aligned}
$$

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(l)})\boldsymbol{y} &= \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\
&\approx (p^*_{ab}\mathbf{1})^T \left(p^*_{ab}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + p^{(l)}_{aa}\mathbb{1}\right)^{-1} \boldsymbol{y} \\
&= \frac{1}{\underbrace{p^{(l)}_{aa}}_{\infty}}(p^*_{ab}\mathbf{1})^T \left(\underbrace{\frac{p^*_{ab}}{p^{(l)}_{aa}}}_{0}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + \mathbb{1}\right)^{-1} \boldsymbol{y} \\
&\xrightarrow{l \to \infty} \mathbf{0}
\end{aligned}
$$

Chaotic Regime: Generalization

- Let's analyze the predictive function in the limit:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(l)})\boldsymbol{y} &= \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{y} \\
&\approx (p^*_{ab}\mathbf{1})^T \left(p^*_{ab}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + p^{(l)}_{aa}\mathbb{1}\right)^{-1} \boldsymbol{y} \\
&= \frac{1}{\underset{\infty}{p^{(l)}_{aa}}}(p^*_{ab}\mathbf{1})^T \left(\frac{p^*_{ab}}{\underset{0}{p^{(l)}_{aa}}}(\mathbf{1}\mathbf{1}^T - \mathbb{1}) + \mathbb{1}\right)^{-1} \boldsymbol{y} \\
&\xrightarrow{l\to\infty} \mathbf{0}
\end{aligned}
$$

- We hence get the trivial prediction **independent** of the data!

Chaotic Regime: Summary

Chaotic Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z > 1$

Chaotic Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z > 1$
- $\lim_{l \to \infty} \lambda_i \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \; \forall i$

Chaotic Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z > 1$
- $\lim_{l \to \infty} \lambda_i \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \ \forall i$
- $\lim_{l \to \infty} \kappa \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 1 \implies$ Remains **trainable**

Chaotic Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z > 1$
- $\lim_{l \to \infty} \lambda_i \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \ \forall i$
- $\lim_{l \to \infty} \kappa \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 1 \implies$ Remains **trainable**
- $\lim_{l \to \infty} f_\infty(\boldsymbol{x}) = \lim_{l \to \infty} P\left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{x}, \boldsymbol{X})\right) \boldsymbol{y} = \boldsymbol{0} \quad \forall \boldsymbol{x} \notin \boldsymbol{X}$

Chaotic Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z > 1$
- $\lim_{l \to \infty} \lambda_i \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \ \forall i$
- $\lim_{l \to \infty} \kappa \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 1 \implies$ Remains **trainable**
- $\lim_{l \to \infty} f_\infty(\boldsymbol{x}) = \lim_{l \to \infty} P \left(\boldsymbol{\Theta}^{(l)} (\boldsymbol{x}, \boldsymbol{X})\right) \boldsymbol{y} = \boldsymbol{0} \quad \forall \boldsymbol{x} \notin \boldsymbol{X}$

  $\implies$ Network **fails** to generalize

Ordered Regime: Generalization (I)

Ordered Regime: Generalization (I)

- More **subtle**: Can't use continuity of map

$$f : GL_n(\mathbb{R}) \to GL_n(\mathbb{R}), \;\; \boldsymbol{A} \mapsto f(\boldsymbol{A}) = \boldsymbol{A}^{-1}$$

because $\det\left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$ (can't just plugin the limit)

Ordered Regime: Generalization (I)

- More **subtle**: Can't use continuity of map

$$f : GL_n(\mathbb{R}) \to GL_n(\mathbb{R}), \ \ \boldsymbol{A} \mapsto f(\boldsymbol{A}) = \boldsymbol{A}^{-1}$$

because $\det\left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$ (can't just plugin the limit)

- Have to work with the finite depth approximation to argue like

$$\lim_{\lambda \to \infty} (\boldsymbol{A} + \lambda \mathbb{1})^{-1} = \mathbf{0}$$

to obtain the correct limit

Ordered Regime: Generalization (I)

- More **subtle**: Can't use continuity of map

$$f : GL_n(\mathbb{R}) \to GL_n(\mathbb{R}), \;\; \boldsymbol{A} \mapsto f(\boldsymbol{A}) = \boldsymbol{A}^{-1}$$

because $\det\left(\boldsymbol{\Theta}^{(\infty)}\right) = 0$ (can't just plugin the limit)

- Have to work with the finite depth approximation to argue like

$$\lim_{\lambda \to \infty} (\boldsymbol{A} + \lambda \mathbb{1})^{-1} = \boldsymbol{0}$$

to obtain the correct limit

- width correction matrix that depends on the data
- $A^{(l)}(X,X) \longrightarrow A^{(\infty)}(X,X$

- Write $\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X}) = p^* \mathbf{1}\mathbf{1}^T + l\,(f'(1))^l\,\boldsymbol{A}^{(l)}(\boldsymbol{X}, \boldsymbol{X})$

Ordered Regime: Generalization (II)

Ordered Regime: Generalization (II)

Let $\lambda = jf'(1)^j$, $\boldsymbol{A}^{(j)}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{A}_{\boldsymbol{X}}^{(j)}$. Let's express the finite depth inverse:

Ordered Regime: Generalization (II)

Let $\lambda = jf'(1)^j$, $\boldsymbol{A}^{(j)}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{A}_{\boldsymbol{X}}^{(j)}$. Let's express the finite depth inverse:

$$\left(\boldsymbol{\Theta}^{(j)}\left(\boldsymbol{X}, \boldsymbol{X}\right)\right)^{-1} = \left(p^* \mathbf{1}\mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}$$

Ordered Regime: Generalization (II)

Let $\lambda = jf'(1)^j$, $\boldsymbol{A}^{(j)}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{A}_{\boldsymbol{X}}^{(j)}$. Let's express the finite depth inverse:

$$\left(\boldsymbol{\Theta}^{(j)}\left(\boldsymbol{X}, \boldsymbol{X}\right)\right)^{-1} = \left(p^* \boldsymbol{1}\boldsymbol{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}$$

$(A+uv^T)^{-1} = A^{-1} + \dfrac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$

$$= \lambda^{-1}\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \frac{\lambda^{-2}\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\boldsymbol{1}\boldsymbol{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}}{\frac{1}{p^*} + \lambda^{-1}\boldsymbol{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\boldsymbol{1}}$$

Ordered Regime: Generalization (II)

Let $\lambda = jf'(1)^j$, $\boldsymbol{A}^{(j)}(\boldsymbol{X}, \boldsymbol{X}) = \boldsymbol{A}^{(j)}_{\boldsymbol{X}}$. Let's express the finite depth inverse:

$$\left(\boldsymbol{\Theta}^{(j)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} = \left(p^* \mathbf{1}\mathbf{1}^T + \lambda \boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}$$

$$(A+uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$$

$$= \lambda^{-1}\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1} - \frac{\lambda^{-2}\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}\mathbf{1}\mathbf{1}^T\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}}{\frac{1}{p^*} + \lambda^{-1}\mathbf{1}^T\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}\mathbf{1}}$$

$$= \lambda^{-1}\left(\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1} - \frac{\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}\mathbf{1}\mathbf{1}^T\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}}{\lambda\left(\frac{1}{p^*} + \lambda^{-1}\mathbf{1}^T\left(\boldsymbol{A}^{(j)}_{\boldsymbol{X}}\right)^{-1}\mathbf{1}\right)}\right)$$

Ordered Regime: Generalization (II)

Let $\lambda = jf'(1)^j$, $\boldsymbol{A}^{(j)}(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{A}_{\boldsymbol{X}}^{(j)}$. Let's express the finite depth inverse:

$$\left(\boldsymbol{\Theta}^{(j)}\left(\boldsymbol{X},\boldsymbol{X}\right)\right)^{-1} = \left(p^*\mathbf{1}\mathbf{1}^T + \lambda\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}$$

$(A+uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$

$$= \lambda^{-1}\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \frac{\lambda^{-2}\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\mathbf{1}\mathbf{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}}{\frac{1}{p^*} + \lambda^{-1}\mathbf{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\mathbf{1}}$$

$$= \lambda^{-1}\left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \frac{\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\mathbf{1}\mathbf{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}}{\lambda\left(\frac{1}{p^*} + \lambda^{-1}\mathbf{1}^T\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1}\mathbf{1}\right)}\right)$$

$$= \lambda^{-1}\left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p}\boldsymbol{a}\boldsymbol{a}^T\right)$$

$a = (A_X^{(j)})^{-1}\underline{1}$

Ordered Regime: Generalization (III)

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \Theta^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \Theta^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^{*}\mathbf{1}^{T} + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$P(\boldsymbol{\Theta}^{(j)}) = \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{a}\boldsymbol{a}^T\right)$$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(j)}) &= \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{a} \boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a} \boldsymbol{a}^T + \lambda^{-1} p^* \left(\boldsymbol{a}^T - \hat{p} \mathbf{1}^T \boldsymbol{a} \boldsymbol{a}^T\right)
\end{aligned}
$$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(j)}) &= \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{a}\boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \lambda^{-1} p^* \left(\boldsymbol{a}^T - \hat{p} \mathbf{1}^T \boldsymbol{a}\boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \hat{p} \boldsymbol{a}^T
\end{aligned}
$$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(j)}) &= \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)} \right)^{-1} - \hat{p} \boldsymbol{a} \boldsymbol{a}^T \right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)} \right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a} \boldsymbol{a}^T + \lambda^{-1} p^* \left(\boldsymbol{a}^T - \hat{p} \mathbf{1}^T \boldsymbol{a} \boldsymbol{a}^T \right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)} \right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a} \boldsymbol{a}^T + \hat{p} \boldsymbol{a}^T \\
&\xrightarrow{j \to \infty} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)} \right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \boldsymbol{a}^{(\infty)} \left(a^{(\infty)} \right)^T + \hat{p} \left(\boldsymbol{a}^{(\infty)} \right)^T
\end{aligned}
$$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(j)}) &= \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{a}\boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \lambda^{-1} p^* \left(\boldsymbol{a}^T - \hat{p} \mathbf{1}^T \boldsymbol{a}\boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \hat{p} \boldsymbol{a}^T \\
&\xrightarrow{j \to \infty} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \boldsymbol{a}^{(\infty)} \left(a^{(\infty)}\right)^T + \hat{p} \left(\boldsymbol{a}^{(\infty)}\right)^T
\end{aligned}
$$

- **Thus:** $f_\infty(\boldsymbol{z}) = \left(\boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} + \hat{\boldsymbol{A}}\right) \boldsymbol{y}$

Ordered Regime: Generalization (III)

- Take a new input $\boldsymbol{z} \implies \boldsymbol{\Theta}^{(j)}(\boldsymbol{z}, \boldsymbol{X}) = p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}$

- Let's express the predictive function:

$$
\begin{aligned}
P(\boldsymbol{\Theta}^{(j)}) &= \left(p^* \mathbf{1}^T + \lambda \boldsymbol{A}_{\boldsymbol{z}}^{(j)}\right) \lambda^{-1} \left(\left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{a}\boldsymbol{a}^T \right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \lambda^{-1} p^* \left(\boldsymbol{a}^T - \hat{p} \mathbf{1}^T \boldsymbol{a}\boldsymbol{a}^T\right) \\
&= \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(j)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(j)} \boldsymbol{a}\boldsymbol{a}^T + \hat{p} \boldsymbol{a}^T \\
&\xrightarrow{j \to \infty} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} - \hat{p} \boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \boldsymbol{a}^{(\infty)} \left(a^{(\infty)}\right)^T + \hat{p} \left(\boldsymbol{a}^{(\infty)}\right)^T
\end{aligned}
$$

- **Thus:** $f_\infty(\boldsymbol{z}) = \left(\boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} + \hat{\boldsymbol{A}} \right) \boldsymbol{y}$

  Non-trivial generalization possible!

Ordered Regime: Summary

Ordered Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$

Ordered Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$
- $\lim_{l \to \infty} \lambda_1 \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = np^*$

Ordered Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$
- $\lim_{l \to \infty} \lambda_1 \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = np^*$
- $\lim_{l \to \infty} \lambda_{\text{bulk}} \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 0$

Ordered Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$
- $\lim_{l \to \infty} \lambda_1 \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = np^*$
- $\lim_{l \to \infty} \lambda_{\text{bulk}} \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 0$
- $\lim_{l \to \infty} \kappa \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \implies$ Becomes **untrainable**

Ordered Regime: Summary

- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$
- $\lim_{l \to \infty} \lambda_1 \left(\mathbf{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = np^*$
- $\lim_{l \to \infty} \lambda_{\text{bulk}} \left(\mathbf{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 0$
- $\lim_{l \to \infty} \kappa \left(\mathbf{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \implies$ Becomes **untrainable**
- $\lim_{l \to \infty} f_{\infty}^{(l)}(\boldsymbol{x}) = \left(\boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} + \hat{\boldsymbol{A}}\right) \boldsymbol{y}$

Ordered Regime: Summary

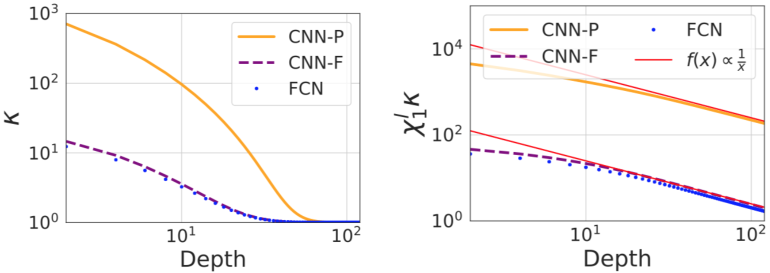
- Characterized by $f'(1) = \sigma_w^2 \int \left[\phi'(\sqrt{q^*}z)\right]^2 \mathcal{D}z < 1$
- $\lim_{l \to \infty} \lambda_1 \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = np^*$
- $\lim_{l \to \infty} \lambda_{\text{bulk}} \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = 0$
- $\lim_{l \to \infty} \kappa \left(\boldsymbol{\Theta}^{(l)}(\boldsymbol{X}, \boldsymbol{X})\right) = \infty \implies$ Becomes **untrainable**
- $\lim_{l \to \infty} f_\infty^{(l)}(\boldsymbol{x}) = \left(\boldsymbol{A}_{\boldsymbol{z}}^{(\infty)} \left(\boldsymbol{A}_{\boldsymbol{X}}^{(\infty)}\right)^{-1} + \hat{\boldsymbol{A}}\right) \boldsymbol{y}$

  $\implies$ Network might be **able** to generalize

Experiments (I)

Chaotic: $\kappa \to 1$

ordered: $\kappa \simeq (\chi_1^l \cdot c)^{-1} \to \infty$

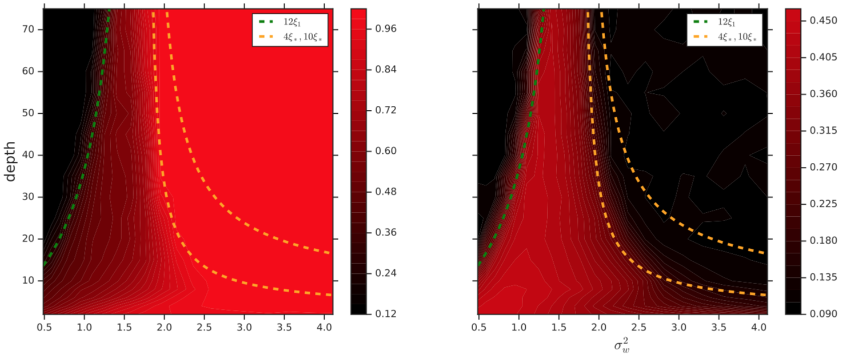


Condition number in the chaotic and ordered regime

Experiments (II)

Train accuracy

Test accuracy



Test and Training Accuracy for Finite Depth and Width Networks

Discussion

- Important to also study **infinite depth**, after all seems to be more important in practice than big widths
- Interesting **tradeoff** between generalization and trainability
- Characterization of generalization a bit **weak** but seems to be true empirically
- Approximation of the diagonal $\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) = q^*$ **very strong** assumption, unclear how this affects the obtained limits
- Very **messy** paper, lots of referencing to prior work and some statements are a bit unclear