Neural Tangent Kernel: Convergence and Generalization in Neural Networks

Arthur Jacot, Franck Gabriel, Clément Hongler

First Approach

- I will try to give **high-level** overview first

First Approach

- I will try to give **high-level** overview first
- Different notation to not enter into all mathematical details (*dual space*, *functional kernel regression* etc)

First Approach

- I will try to give **high-level** overview first
- Different notation to not enter into all mathematical details (*dual space*, *functional kernel regression* etc)
- If time allows we can dwelve deeper into the matter

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:

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

Notation

Notation

- Only consider **fully-connected** networks of depth L:

Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(L)}(\boldsymbol{x}) \in \mathbb{R}^{d_L}$ is the **output** of the network

Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(L)}(\boldsymbol{x}) \in \mathbb{R}^{d_L}$ is the **output** of the network

- Number of parameters $P = \sum_{l=1}^{L} d_{l-1} d_l$

Notation

- Only consider **fully-connected** networks of depth L:

  - $\tilde{\alpha}^{(l+1)}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \boldsymbol{W}^{(l)} \alpha^{(l)}(\boldsymbol{x}) + \beta \boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$
  - $\alpha^{(l+1)}(\boldsymbol{x}) = \sigma\left(\tilde{\alpha}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{d_{l+1}}$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d_{l+1} \times d_l}$, $\boldsymbol{b}^{(l)} \in \mathbb{R}^{d_{l+1}}$, σ is a coordinate-wise activation function and $\beta > 0$

- $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \tilde{\alpha}^{(L)}(\boldsymbol{x}) \in \mathbb{R}^{d_L}$ is the **output** of the network

- Number of parameters $P = \sum_{l=1}^{L} d_{l-1} d_l$

- **Same** notation as last time :)

Random Initialization

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Made **width dependence** of random initialization explicit in the architecture

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Made **width dependence** of random initialization explicit in the architecture
- **Effectively:** $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{d_l})$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \beta^2)$

  $\implies$ Usual Lecun initialization

Random Initialization

- Initialize as $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- Made **width dependence** of random initialization explicit in the architecture
- **Effectively:** $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{d_l})$ and $b_i^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \beta^2)$

  $\implies$ Usual Lecun initialization
- Again **identical** to first presentation

Loss

Loss

- This time we will **train** the neural network

Loss

- This time we will **train** the neural network
- We take some general **loss** function

$$L : \mathbb{R}^P \times \mathbb{R}^{n \times d_0} \times \mathbb{R}^n \to \mathbb{R}, \quad (\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) \mapsto L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

Loss

- This time we will **train** the neural network
- We take some general **loss** function

$$L : \mathbb{R}^P \times \mathbb{R}^{n \times d_0} \times \mathbb{R}^n \to \mathbb{R}, \quad (\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y}) \mapsto L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

- Assume that the loss decomposes over the samples:

$$L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y}) = \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

Gradient Flow (I)

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

- Let's take a continuous view point: $\boldsymbol{\theta}_k$ are samples at multiples of η from a function $\boldsymbol{\theta} : \mathbb{R}_+ \to \mathbb{R}^{d_0}$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}(k\eta)$$

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

- Let's take a continuous view point: $\boldsymbol{\theta}_k$ are samples at multiples of η from a function $\boldsymbol{\theta} : \mathbb{R}_+ \to \mathbb{R}^{d_0}$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}(k\eta)$$

- Let $t = k\eta$ and let's look at

$$\boldsymbol{\theta}(t + \eta)$$

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

- Let's take a continuous view point: $\boldsymbol{\theta}_k$ are samples at multiples of η from a function $\boldsymbol{\theta} : \mathbb{R}_+ \to \mathbb{R}^{d_0}$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}(k\eta)$$

- Let $t = k\eta$ and let's look at

$$\boldsymbol{\theta}(t + \eta) = \boldsymbol{\theta}_{k+1}$$

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

- Let's take a continuous view point: $\boldsymbol{\theta}_k$ are samples at multiples of η from a function $\boldsymbol{\theta} : \mathbb{R}_+ \to \mathbb{R}^{d_0}$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}(k\eta)$$

- Let $t = k\eta$ and let's look at

$$\boldsymbol{\theta}(t+\eta) = \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

Gradient Flow (I)

- We usually do **gradient descent**

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k}$$

- Let's take a continuous view point: $\boldsymbol{\theta}_k$ are samples at multiples of η from a function $\boldsymbol{\theta} : \mathbb{R}_+ \to \mathbb{R}^{d_0}$

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}(k\eta)$$

- Let $t = k\eta$ and let's look at

$$\begin{aligned}\boldsymbol{\theta}(t+\eta) = \boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_k} \\ &= \boldsymbol{\theta}(t) - \eta \nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}\end{aligned}$$

Gradient Flow (II)

Gradient Flow (II)

- Rearranging gives

$$\frac{1}{\eta}\left(\boldsymbol{\theta}(t+\eta)-\boldsymbol{\theta}(t)\right)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

Gradient Flow (II)

- Rearranging gives

$$\frac{1}{\eta}\left(\boldsymbol{\theta}(t+\eta)-\boldsymbol{\theta}(t)\right)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Letting $\eta \to 0$ thus gives the gradient flow equation:

$$\dot{\boldsymbol{\theta}}(t)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

Gradient Flow (II)

- Rearranging gives

$$\frac{1}{\eta}\left(\boldsymbol{\theta}(t+\eta)-\boldsymbol{\theta}(t)\right)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Letting $\eta \to 0$ thus gives the gradient flow equation:

$$\dot{\boldsymbol{\theta}}(t)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- **Side note**: Gradient Descent is the Euler discretization of the above differential equation

Gradient Flow (II)

- Rearranging gives

$$\frac{1}{\eta}\left(\boldsymbol{\theta}(t+\eta)-\boldsymbol{\theta}(t)\right)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Letting $\eta \to 0$ thus gives the gradient flow equation:

$$\dot{\boldsymbol{\theta}}(t)=-\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}),\boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- **Side note**: Gradient Descent is the Euler discretization of the above differential equation

- **Non-linear** ODE, hard to solve... If we could solve it, no need to train NNs since the solution describes the trajectory through time

Neural Tangent Kernel

Let us now introduce the quantity of interest:

Definition 1:

Neural Tangent Kernel

Let us now introduce the quantity of interest:

Definition 1:
The **Neural Tangent Kernel** is given by

$$\Theta_t^{(L)} = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}} \otimes \partial_{\theta_p} f_{\boldsymbol{\theta}}$$

Neural Tangent Kernel

Let us now introduce the quantity of interest:

Definition 1:
The **Neural Tangent Kernel** is given by

$$\Theta_t^{(L)} = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}} \otimes \partial_{\theta_p} f_{\boldsymbol{\theta}}$$

This means in particular that for $k, k' \in \{1, \dots, d_L\}$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$

$$\Theta_{kk'}^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \left\langle \partial_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}), \partial_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \right\rangle$$

Neural Tangent Kernel

Let us now introduce the quantity of interest:

Definition 1:
The **Neural Tangent Kernel** is given by

$$\Theta_t^{(L)} = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}} \otimes \partial_{\theta_p} f_{\boldsymbol{\theta}}$$

This means in particular that for $k, k' \in \{1, \dots, d_L\}$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{d_0}$

$$\Theta_{kk'}^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \left\langle \partial_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}), \partial_{\boldsymbol{\theta}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \right\rangle$$

and thus $\Theta^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{R}^{d_L \times d_L}$ (multi-dimensional kernel)

An Easier Setting

An Easier Setting

- Let's assume that $d_L = 1$ (e.g. binary classification problem or least squared regression)

An Easier Setting

- Let's assume that $d_L = 1$ (e.g. binary classification problem or least squared regression)

- Then the NTK simplifies to a scalar:

$$\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$$

An Easier Setting

- Let's assume that $d_L = 1$ (e.g. binary classification problem or least squared regression)
- Then the NTK simplifies to a scalar:

$$\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$$

- $[\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})]_i = \Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[\Theta_t^{(L)}(\boldsymbol{X}, \boldsymbol{X})]_{ij} = \Theta_t^{(L)}(\boldsymbol{x}_i, \boldsymbol{x}_j)$

An Easier Setting

- Let's assume that $d_L = 1$ (e.g. binary classification problem or least squared regression)
- Then the NTK simplifies to a scalar:

$$\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$$

- $[\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})]_i = \Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}_i)$ and $[\Theta_t^{(L)}(\boldsymbol{X}, \boldsymbol{X})]_{ij} = \Theta_t^{(L)}(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- **Question:** Why should we care about this object???

NTK and Gradient Flow

NTK and Gradient Flow

- Let us "train" f_{θ} with gradient flow:

NTK and Gradient Flow

- Let us "train" $f_{\boldsymbol{\theta}}$ with gradient flow:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

NTK and Gradient Flow

- Let us "train" $f_{\boldsymbol{\theta}}$ with gradient flow:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Component-wise: $\frac{\partial \theta_p}{\partial t} = -\frac{\partial}{\partial \theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$

NTK and Gradient Flow

- Let us "train" $f_{\boldsymbol{\theta}}$ with gradient flow:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Component-wise: $\frac{\partial \theta_p}{\partial t} = -\frac{\partial}{\partial \theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$

- This equation describes the dynamics in **weight space**

NTK and Gradient Flow

- Let us "train" $f_{\boldsymbol{\theta}}$ with gradient flow:

$$\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$$

- Component-wise: $\frac{\partial \theta_p}{\partial t} = -\frac{\partial}{\partial \theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(t)}$

- This equation describes the dynamics in **weight space**

- What can we say about the **function space** dynamics?

Function Space Dynamics (I)

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x})$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t}$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p}$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p}$$

$$= -\sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p}$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\begin{aligned}
\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y}) \\
&= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i) \\
&= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p} \\
&= -\sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p} \\
&= -\sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}_i)
\end{aligned}$$

Function Space Dynamics (I)

Let's take a new input $\boldsymbol{x} \in \mathbb{R}^{d_0}$:

$$\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial \theta_p}{\partial t} = -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \partial_{\theta_p} \sum_{i=1}^{n} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)$$

$$= -\sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p}$$

$$= -\sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \sum_{p=1}^{P} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \theta_p} \frac{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \theta_p}$$

$$= -\sum_{i=1}^{n} \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i, y_i)}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)} \boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}_i) = -\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})^T \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{X})}$$

Function Space Dynamics (II)

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

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

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

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

- **Notice:** This result so far holds for **any** neural network and **any** loss L, no infinite-width limit taken yet!

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

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

- **Notice:** This result so far holds for **any** neural network and **any** loss L, no infinite-width limit taken yet!

- $\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is a **random variable** (Depends on initialization)

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

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

- **Notice:** This result so far holds for **any** neural network and **any** loss L, no infinite-width limit taken yet!

- $\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is a **random variable** (Depends on initialization)

- $\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is **time-dependent** (It is a function of the parameters $\boldsymbol{\theta}(t)$)

Function Space Dynamics (II)

- The function space dynamics is hence governed by the NTK:

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

- **Notice:** This result so far holds for **any** neural network and **any** loss L, no infinite-width limit taken yet!

- $\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is a **random variable** (Depends on initialization)

- $\boldsymbol{\Theta}_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}')$ is **time-dependent** (It is a function of the parameters $\boldsymbol{\theta}(t)$)

- ODE hence **not** analytically solvable (again if it were, no need to train NNs at all)

Infinite-Width Limit

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

We recover the following properties:

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

We recover the following properties:

- $\boldsymbol{\Theta}_t^{(L)}$ converges to a **deterministic** kernel $\boldsymbol{\Theta}_{\infty,t}^{(L)}$

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

We recover the following properties:

- $\boldsymbol{\Theta}_t^{(L)}$ converges to a **deterministic** kernel $\boldsymbol{\Theta}_{\infty,t}^{(L)}$
- $\boldsymbol{\Theta}_{\infty,t}^{(L)}$ has a **closed-form** expression

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

We recover the following properties:

- $\boldsymbol{\Theta}_t^{(L)}$ converges to a **deterministic** kernel $\boldsymbol{\Theta}_{\infty,t}^{(L)}$
- $\boldsymbol{\Theta}_{\infty,t}^{(L)}$ has a **closed-form** expression
- $\boldsymbol{\Theta}_{\infty,t}^{(L)}$ does **not** depend on time: $\boldsymbol{\Theta}_{\infty,t}^{(L)} = \boldsymbol{\Theta}_{\infty}^{(L)}$

Infinite-Width Limit

Question: What happens if we let $d_1, \ldots, d_{L-1} \to \infty$?

We recover the following properties:

- $\boldsymbol{\Theta}_t^{(L)}$ converges to a **deterministic** kernel $\boldsymbol{\Theta}_{\infty,t}^{(L)}$
- $\boldsymbol{\Theta}_{\infty,t}^{(L)}$ has a **closed-form** expression
- $\boldsymbol{\Theta}_{\infty,t}^{(L)}$ does **not** depend on time: $\boldsymbol{\Theta}_{\infty,t}^{(L)} = \boldsymbol{\Theta}_{\infty}^{(L)}$

We hence don't have to worry about the initialization and time-dependence anymore!

Consequences

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \frac{\partial L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_t^{\infty}(\boldsymbol{X})}$$

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \frac{\partial L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_t^{\infty}(\boldsymbol{X})}$$

- Also assume we use the **squared loss**:

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \frac{\partial L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_t^{\infty}(\boldsymbol{X})}$$

- Also assume we use the **squared loss**:

$$L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y}) = \frac{1}{2} \sum_{i=1}^{n} (f_t^{\infty}(\boldsymbol{x}_i) - y_i)^2$$

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \frac{\partial L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_t^{\infty}(\boldsymbol{X})}$$

- Also assume we use the **squared loss**:

$$L(f_t^{\infty}(\boldsymbol{X}), \boldsymbol{y}) = \frac{1}{2} \sum_{i=1}^{n} (f_t^{\infty}(\boldsymbol{x}_i) - y_i)^2$$

- This gives the following expression for the gradient

Consequences

- Let us **assume** for now that all of this holds. What are the consequences?

- We have the following dynamics:

$$\partial_t f_t^\infty(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \frac{\partial L(f_t^\infty(\boldsymbol{X}), \boldsymbol{y})}{\partial f_t^\infty(\boldsymbol{X})}$$

- Also assume we use the **squared loss**:

$$L(f_t^\infty(\boldsymbol{X}), \boldsymbol{y}) = \frac{1}{2}\sum_{i=1}^{n}(f_t^\infty(\boldsymbol{x}_i) - y_i)^2$$

- This gives the following expression for the gradient

$$\partial_{f_t^\infty(\boldsymbol{x}_i)} L(f_t^\infty(\boldsymbol{X}), \boldsymbol{y}) = f_t^\infty(\boldsymbol{x}_i) - y_i$$

Closed Form Solution (I)

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

That's a system of **linear first order** differential equations!

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^\infty(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^\infty(\boldsymbol{X}) - \boldsymbol{y})$$

  That's a system of **linear first order** differential equations!

- Careful, every $\boldsymbol{x}$ induces a **different** function $f_t^\infty(\boldsymbol{x})$ in t!

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

  That's a system of **linear first order** differential equations!

- Careful, every $\boldsymbol{x}$ induces a **different** function $f_t^{\infty}(\boldsymbol{x})$ in t!

- After some **tricks** we find a closed form solution:

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^\infty(\boldsymbol{x}) = -\left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^\infty(\boldsymbol{X}) - \boldsymbol{y})$$

  That's a system of **linear first order** differential equations!

- Careful, every $\boldsymbol{x}$ induces a **different** function $f_t^\infty(\boldsymbol{x})$ in t!

- After some **tricks** we find a closed form solution:

$$f_t^\infty(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})^{-1} (e^{-\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})t} - \mathbb{1})\right) (f_0^\infty(\boldsymbol{X}) - \boldsymbol{y}) + f_0^\infty(\boldsymbol{x})$$

Training an Infinitely-Wide Network Infinitely Long

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

$$f_\infty^\infty(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} (\boldsymbol{y} - f_0^\infty(\boldsymbol{X})) + f_0^\infty(\boldsymbol{x})$$

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

$$f_{\infty}^{\infty}(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^{T} \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} (\boldsymbol{y} - f_{0}^{\infty}(\boldsymbol{X})) + f_{0}^{\infty}(\boldsymbol{x})$$

- We have a **perfect fit** on the training data:

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

$$f_{\infty}^{\infty}(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^{T} \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} (\boldsymbol{y} - f_{0}^{\infty}(\boldsymbol{X})) + f_{0}^{\infty}(\boldsymbol{x})$$

- We have a **perfect fit** on the training data:

$$f_{\infty}^{\infty}(\boldsymbol{X}) = \boldsymbol{y}$$

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

$$f_\infty^\infty(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} (\boldsymbol{y} - f_0^\infty(\boldsymbol{X})) + f_0^\infty(\boldsymbol{x})$$

- We have a **perfect fit** on the training data:

$$f_\infty^\infty(\boldsymbol{X}) = \boldsymbol{y}$$

(But one has to show that $\boldsymbol{\Theta}_\infty^{(L)}$ is positive-definite...)

Training an Infinitely-Wide Network Infinitely Long

- We can "read off" the **final state** of the infinitely-wide network:

$$f_\infty^\infty(\boldsymbol{x}) = \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} (\boldsymbol{y} - f_0^\infty(\boldsymbol{X})) + f_0^\infty(\boldsymbol{x})$$

- We have a **perfect fit** on the training data:

$$f_\infty^\infty(\boldsymbol{X}) = \boldsymbol{y}$$

(But one has to show that $\boldsymbol{\Theta}_\infty^{(L)}$ is positive-definite...)

- Gradient flow hence converges to the **global minimum** in the infinite regime.

Kernel Ridge Regression

Kernel Ridge Regression

- If we assume $f_0^{\infty}(\boldsymbol{x}) = 0$, we have

Kernel Ridge Regression

- If we assume $f_0^{\infty}(\boldsymbol{x}) = 0$, we have

$$f_{\infty}^{\infty}(\boldsymbol{x}) = \boldsymbol{y}^T \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \right)^{-1} \boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})$$

Kernel Ridge Regression

- If we assume $f_0^\infty(\boldsymbol{x}) = 0$, we have

$$f_\infty^\infty(\boldsymbol{x}) = \boldsymbol{y}^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})$$

- That's exactly the same solution as in unregularized **kernel regression** with a positive-definite kernel K:

Kernel Ridge Regression

- If we assume $f_0^{\infty}(\boldsymbol{x}) = 0$, we have

$$f_{\infty}^{\infty}(\boldsymbol{x}) = \boldsymbol{y}^T \left(\boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \right)^{-1} \boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})$$

- That's exactly the same solution as in unregularized **kernel regression** with a positive-definite kernel K:

$$\hat{y} = \boldsymbol{y}^T K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{x}, \boldsymbol{X})$$

Kernel Ridge Regression

- If we assume $f_0^\infty(\boldsymbol{x}) = 0$, we have

$$f_\infty^\infty(\boldsymbol{x}) = \boldsymbol{y}^T \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})\right)^{-1} \boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})$$

- That's exactly the same solution as in unregularized **kernel regression** with a positive-definite kernel K:

$$\hat{y} = \boldsymbol{y}^T K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{x}, \boldsymbol{X})$$

- Learning an infinitely-wide neural network **is the same** as kernel regression with $\boldsymbol{\Theta}_\infty^{(L)}$

Let's Reiterate

Let's Reiterate

- $\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$

Let's Reiterate

- $\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$

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

Let's Reiterate

- $\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta}}(\boldsymbol{x}')$
- $\partial_t f_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\Theta_t^{(L)}(\boldsymbol{x}, \boldsymbol{X})^T \frac{\partial L(f_{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{y})}{\partial f_{\boldsymbol{\theta}}(\boldsymbol{X})}$
- $\Theta_\infty^{(L)}$ closed-form, deterministic and time-independent

Let's Reiterate

- $\Theta_t^{(L)}(x, x') = \sum_{p=1}^{P} \partial_{\theta_p} f_\theta(x) \partial_{\theta_p} f_\theta(x')$
- $\partial_t f_\theta(x) = -\Theta_t^{(L)}(x, X)^T \frac{\partial L(f_\theta(X), y)}{\partial f_\theta(X)}$
- $\Theta_\infty^{(L)}$ closed-form, deterministic and time-independent
- Recover final state of the infinite model

$$f_\infty^\infty(x) = \left(\Theta_\infty^{(L)}(x, X)\right)^T \left(\Theta_\infty^{(L)}(X, X)\right)^{-1} \left(y - f_0^\infty(X)\right) + f_0^\infty(x)$$

Let's Reiterate

- $\Theta_t^{(L)}(x, x') = \sum_{p=1}^{P} \partial_{\theta_p} f_\theta(x) \partial_{\theta_p} f_\theta(x')$
- $\partial_t f_\theta(x) = -\Theta_t^{(L)}(x, X)^T \frac{\partial L(f_\theta(X), y)}{\partial f_\theta(X)}$
- $\Theta_\infty^{(L)}$ closed-form, deterministic and time-independent
- Recover final state of the infinite model

$$f_\infty^\infty(x) = \left(\Theta_\infty^{(L)}(x, X)\right)^T \left(\Theta_\infty^{(L)}(X, X)\right)^{-1} \left(y - f_0^\infty(X)\right) + f_0^\infty(x)$$

Recap: NNs as GPs at Initialization

Theorem 2:

Recap: NNs as GPs at Initialization

Theorem 2:
Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:

$$f^{\infty}_{\boldsymbol{\theta},i}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(L)})$$

Recap: NNs as GPs at Initialization

Theorem 2:
Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:

$$f^{\infty}_{\boldsymbol{\theta},i}(\cdot) \sim \mathcal{GP}(0, \Sigma^{(L)})$$

i.i.d. in i, where $\Sigma^{(L)}$ has the recursive structure

- $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \frac{1}{\sqrt{d_0}} \boldsymbol{x}^T \boldsymbol{x'} + \beta^2$
- $\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l-1)})} \left[\sigma(z_1) \sigma(z_2) \right] + \beta^2$

Recap: NNs as GPs at Initialization

Theorem 2:
Given a NN $f_{\boldsymbol{\theta}}$, it holds that for $d_1, \ldots, d_{L-1} \to \infty$:

$$f^{\infty}_{\boldsymbol{\theta},i}(\cdot) \sim \mathcal{GP}\big(0, \Sigma^{(L)}\big)$$

i.i.d. in i, where $\Sigma^{(L)}$ has the recursive structure

- $\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x'}) = \frac{1}{\sqrt{d_0}} \boldsymbol{x}^T \boldsymbol{x'} + \beta^2$
- $\Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l-1)})}\big[\sigma(z_1)\sigma(z_2)\big] + \beta^2$

for $\tilde{\boldsymbol{\Sigma}}^{(l-1)} = \begin{pmatrix} \Sigma^{(l-1)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(l-1)}(\boldsymbol{x}, \boldsymbol{x'}) \\ \Sigma^{(l-1)}(\boldsymbol{x'}, \boldsymbol{x}) & \Sigma^{(l-1)}(\boldsymbol{x'}, \boldsymbol{x'}) \end{pmatrix}$

NTK at Initialization

Theorem 3:

NTK at Initialization

Theorem 3:
Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

NTK at Initialization

Theorem 3:
Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

NTK at Initialization

Theorem 3:

Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

where $\boldsymbol{\Theta}_{0,\infty}^{(L)} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ follows the recursion

NTK at Initialization

Theorem 3:
Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

where $\boldsymbol{\Theta}_{0,\infty}^{(L)} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ follows the recursion

- $\boldsymbol{\Theta}_{0,\infty}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$
- $\boldsymbol{\Theta}_{0,\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}_{0,\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

NTK at Initialization

Theorem 3:

Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

where $\boldsymbol{\Theta}_{0,\infty}^{(L)} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ follows the recursion

- $\boldsymbol{\Theta}_{0,\infty}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$
- $\boldsymbol{\Theta}_{0,\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}_{0,\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

with $\dot{\Sigma}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\dot{\sigma}\left(z_1\right)\dot{\sigma}\left(z_2\right)\right]$

NTK at Initialization

Theorem 3:

Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

where $\boldsymbol{\Theta}_{0,\infty}^{(L)} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ follows the recursion

- $\boldsymbol{\Theta}_{0,\infty}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$
- $\boldsymbol{\Theta}_{0,\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}_{0,\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

with $\dot{\Sigma}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\dot{\sigma}\left(z_1\right)\dot{\sigma}\left(z_2\right)\right]$

Writing the expression on the scalar level gives

NTK at Initialization

Theorem 3:

Take a NN $f_{\boldsymbol{\theta}}$ of depth L with Lipschitz σ. Then, for $d_1, \ldots, d_L \to \infty$, it holds

$$\boldsymbol{\Theta}_0^{(L)} \xrightarrow{\mathbb{P}} \boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$$

where $\boldsymbol{\Theta}_{0,\infty}^{(L)} : \mathbb{R}^{d_0} \times \mathbb{R}^{d_0} \to \mathbb{R}$ follows the recursion

- $\boldsymbol{\Theta}_{0,\infty}^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$
- $\boldsymbol{\Theta}_{0,\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}_{0,\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$

with $\dot{\Sigma}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\dot{\sigma}\left(z_1\right)\dot{\sigma}\left(z_2\right)\right]$

Writing the expression on the scalar level gives

$$\left(\boldsymbol{\Theta}_{0,\infty}^{(L)} \otimes \mathbb{1}_{d_L \times d_L}\right)_{kk'}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\Theta}_{0,\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$$

Proof

Proof

- Proof is based on **induction** on the depth L

Proof

- Proof is based on **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ again **sequentially** taken

Proof

- Proof is based on **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ again **sequentially** taken
- For ease of notation, assume $\beta = 0$, so **no bias**

Proof

- Proof is based on **induction** on the depth L
- Limits $d_1, \ldots, d_{L-1} \to \infty$ again **sequentially** taken
- For ease of notation, assume $\beta = 0$, so **no bias**
- Can't do it for just $d_L = 1$, otherwise induction doesn't work. Have to do it for **general** d_L...

Base Case: $L = 1$ (I)

Base Case: $L = 1$ (I)

In this case the network is given by

Base Case: $L = 1$ (I)

In this case the network is given by

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \sum_{l=1}^{d_0} W_{kl}^{(0)} x_l$$

Base Case: $L = 1$ (I)

In this case the network is given by

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \sum_{l=1}^{d_0} W_{kl}^{(0)} x_l$$

Gradients can hence be written as

Base Case: $L = 1$ (I)

In this case the network is given by

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \sum_{l=1}^{d_0} W_{kl}^{(0)} x_l$$

Gradients can hence be written as

$$\partial_{W_{ij}^{(0)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} x_j \mathbb{1}_{\{k=i\}}$$

Base Case: $L = 1$ (I)

In this case the network is given by

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} \sum_{l=1}^{d_0} W_{kl}^{(0)} x_l$$

Gradients can hence be written as

$$\partial_{W_{ij}^{(0)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_0}} x_j \mathbb{1}_{\{k=i\}}$$

Remember: $\boldsymbol{\Theta}_{kk'}^{(L)}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{p=1}^{P} \partial_{\theta_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\theta_p} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$

Base Case: $L = 1$ (II)

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\boldsymbol{\Theta}^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}')$$

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\Theta^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$$

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\begin{aligned}\Theta^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{i=1}^{d_0}\sum_{j=1}^{d_1} \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \\ &= \sum_{i=1}^{d_0}\sum_{j=1}^{d_1} \frac{1}{d_0} x_j \mathbb{1}_{\{k=i\}} x'_j \mathbb{1}_{\{k'=i\}}\end{aligned}$$

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\begin{aligned}
\Theta^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \\
&= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \frac{1}{d_0} x_j \mathbb{1}_{\{k=i\}} x'_j \mathbb{1}_{\{k'=i\}} \\
&= \frac{1}{d_0} \boldsymbol{x}^T \boldsymbol{x}' \mathbb{1}_{\{k'=k\}}
\end{aligned}$$

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\begin{aligned}
\Theta^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \\
&= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \frac{1}{d_0} x_j \mathbb{1}_{\{k=i\}} x'_j \mathbb{1}_{\{k'=i\}} \\
&= \frac{1}{d_0} \boldsymbol{x}^T \boldsymbol{x}' \mathbb{1}_{\{k'=k\}} \\
&= \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{k'=k\}}
\end{aligned}$$

Notice: We didn't take any limit because input dimension d_0 of course remains fixed.

Base Case: $L = 1$ (II)

Plugging in the definition of the NTK leads to:

$$\begin{aligned}
\Theta^{(1)}_{k,k'}(\boldsymbol{x}, \boldsymbol{x}') &= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W^{(0)}_{ij}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') \\
&= \sum_{i=1}^{d_0} \sum_{j=1}^{d_1} \frac{1}{d_0} x_j \mathbb{1}_{\{k=i\}} x'_j \mathbb{1}_{\{k'=i\}} \\
&= \frac{1}{d_0} \boldsymbol{x}^T \boldsymbol{x}' \mathbb{1}_{\{k'=k\}} \\
&= \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{k'=k\}}
\end{aligned}$$

Notice: We didn't take any limit because input dimension d_0 of course remains fixed.

This **proves** the base case.

Induction Step

Induction Step

- Take a network of depth $l + 1$:

Induction Step

- Take a network of depth $l + 1$:

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{j=1}^{d_l} W_{kj}^{(l)} \sigma\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right)$$

Induction Step

- Take a network of depth $l + 1$:

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{j=1}^{d_l} W_{kj}^{(l)} \sigma\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right)$$

- Split the parameters into $\boldsymbol{\theta} = \left(\tilde{\boldsymbol{\theta}}, \boldsymbol{W}^{(l)}\right)$

Induction Step

- Take a network of depth $l+1$:

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{j=1}^{d_l} W_{kj}^{(l)} \sigma\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right)$$

- Split the parameters into $\boldsymbol{\theta} = \left(\tilde{\boldsymbol{\theta}}, \boldsymbol{W}^{(l)}\right)$

- We know that $\boldsymbol{\Theta}_{kk'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$

Induction Step

- Take a network of depth $l+1$:

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{j=1}^{d_l} W_{kj}^{(l)} \sigma\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right)$$

- Split the parameters into $\boldsymbol{\theta} = \left(\tilde{\boldsymbol{\theta}}, \boldsymbol{W}^{(l)}\right)$

- We know that $\boldsymbol{\Theta}_{kk'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$

We can now **split** the NTK:

Induction Step

- Take a network of depth $l+1$:

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \sum_{j=1}^{d_l} W_{kj}^{(l)} \sigma\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right)$$

- Split the parameters into $\boldsymbol{\theta} = \left(\tilde{\boldsymbol{\theta}}, \boldsymbol{W}^{(l)}\right)$

- We know that $\boldsymbol{\Theta}_{kk'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \xrightarrow{d_1,\ldots,d_{l-1}\to\infty} \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$

We can now **split** the NTK:

$$\boldsymbol{\Theta}_{kk'}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \underbrace{\sum_{l=1}^{\tilde{P}} \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k}(\boldsymbol{x}), \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')}_{A} + \underbrace{\sum_{i,j=1}^{d_l,d_{l+1}} \partial_{W_{ij}^{(L)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}), \partial_{W_{ij}^{(L)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')}_{B}$$

Terms in A (I)

Terms in A (I)

- Write out the derivatives as

Terms in A (I)

- Write out the derivatives as

$$\partial_{\tilde{\theta}_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_L}} \sum_{j=1}^{d_L} W_{kj}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j(\boldsymbol{x})$$

Terms in A (I)

- Write out the derivatives as

$$\partial_{\tilde{\theta}_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_L}} \sum_{j=1}^{d_L} W_{kj}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j(\boldsymbol{x})$$

Let us study the terms $A = \sum_{l=1}^{\tilde{P}} \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$ due to $\tilde{\boldsymbol{\theta}}$:

Terms in A (I)

- Write out the derivatives as

$$\partial_{\tilde{\theta}_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_L}} \sum_{j=1}^{d_L} W_{kj}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j(\boldsymbol{x})$$

Let us study the terms $A = \sum_{l=1}^{\tilde{P}} \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$ due to $\tilde{\boldsymbol{\theta}}$:

$$\frac{1}{d_l} \sum_{p=1}^{\tilde{P}} \left(\sum_{j=1}^{d_l} W_{kj}^{(L)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j^{(l)}(\boldsymbol{x}) \right) \left(\sum_{j'=1}^{d_l} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \partial_{\tilde{\theta}_p} \alpha_{j'}^{(l)}(\boldsymbol{x}') \right)$$

Terms in A (I)

- Write out the derivatives as

$$\partial_{\tilde{\theta}_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_L}} \sum_{j=1}^{d_L} W_{kj}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j(\boldsymbol{x})$$

Let us study the terms $A = \sum_{l=1}^{\tilde{P}} \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$ due to $\tilde{\boldsymbol{\theta}}$:

$$\frac{1}{d_l} \sum_{p=1}^{\tilde{P}} \left(\sum_{j=1}^{d_l} W_{kj}^{(L)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j^{(l)}(\boldsymbol{x}) \right) \left(\sum_{j'=1}^{d_l} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \partial_{\tilde{\theta}_p} \alpha_{j'}^{(l)}(\boldsymbol{x}') \right)$$

$$= \frac{1}{d_l} \sum_{j,j'=1}^{d_L} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \sum_{p=1}^{\tilde{P}} \partial_{\tilde{\theta}_p} \tilde{\alpha}_j^{(l)}(\boldsymbol{x}) \partial_{\tilde{\theta}_p} \tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')$$

Terms in A (I)

- Write out the derivatives as

$$\partial_{\tilde{\theta}_p} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_L}} \sum_{j=1}^{d_L} W_{kj}^{(l)} \dot{\sigma}\Big(\tilde{\alpha}_j(\boldsymbol{x})\Big) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j(\boldsymbol{x})$$

Let us study the terms $A = \sum_{l=1}^{\tilde{P}} \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{\tilde{\theta}_l} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$ due to $\tilde{\boldsymbol{\theta}}$:

$$\frac{1}{d_l} \sum_{p=1}^{\tilde{P}} \left(\sum_{j=1}^{d_l} W_{kj}^{(L)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \partial_{\tilde{\theta}_p} \tilde{\alpha}_j^{(l)}(\boldsymbol{x}) \right) \left(\sum_{j'=1}^{d_l} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \partial_{\tilde{\theta}_p} \alpha_{j'}^{(l)}(\boldsymbol{x}') \right)$$

$$= \frac{1}{d_l} \sum_{j,j'=1}^{d_L} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \sum_{p=1}^{\tilde{P}} \partial_{\tilde{\theta}_p} \tilde{\alpha}_j^{(l)}(\boldsymbol{x}) \partial_{\tilde{\theta}_p} \tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')$$

$$= \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

Terms in A (II)

Terms in A (II)

A

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_1, \ldots d_{l-1} \to \infty} \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j',\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{j=j'\}}$$

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_1, \ldots d_{l-1} \to \infty} \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j',\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{j=j'\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_1, \dots d_{l-1} \to \infty} \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j',\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{j=j'\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_l \to \infty} \mathbb{E}\left[W_{kj}^{(L)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\right]$$

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_1,\dots d_{l-1}\to\infty} \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j',\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{j=j'\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_l\to\infty} \mathbb{E}\left[W_{kj}^{(L)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\right]$$

$$= \boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{E}\left[W_{kj}^{(l)} W_{k'j}^{(l)}\right] \mathbb{E}\left[\dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right)\right]$$

Terms in A (II)

$$A = \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j'}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{j,j'}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_1,\dots d_{l-1}\to\infty} \frac{1}{d_l} \sum_{j,j'=1}^{d_l} W_{kj}^{(l)} W_{k'j'}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j',\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{j=j'\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\xrightarrow{d_l\to\infty} \mathbb{E}\left[W_{kj}^{(L)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\right]$$

$$= \boldsymbol{\Theta}_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{E}\left[W_{kj}^{(l)} W_{k'j}^{(l)}\right] \mathbb{E}\left[\dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{\theta}})\right) \dot{\sigma}\left(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}')\right)\right]$$

$$= \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{k=k'\}} \dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}')$$

Terms in B

Terms in B

Let's calculate the gradients w.r.t. last layer weights

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

B

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}')$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \frac{1}{d_l} \sum_{i,j=1}^{d_l, d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}} \alpha_j(\boldsymbol{x}') \mathbb{1}_{\{k'=i\}}$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \frac{1}{d_l} \sum_{i,j=1}^{d_l, d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}} \alpha_j(\boldsymbol{x}') \mathbb{1}_{\{k'=i\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}}$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \frac{1}{d_l} \sum_{i,j=1}^{d_l, d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}} \alpha_j(\boldsymbol{x}') \mathbb{1}_{\{k'=i\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}} \xrightarrow{d_l \to \infty} \mathbb{E}\left[\sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})) \sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}'))\right] \mathbb{1}_{\{k=k\}}$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \frac{1}{d_l} \sum_{i,j=1}^{d_l, d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}} \alpha_j(\boldsymbol{x}') \mathbb{1}_{\{k'=i\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}} \xrightarrow{d_l \to \infty} \mathbb{E}\left[\sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})) \sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}'))\right] \mathbb{1}_{\{k=k\}}$$

$$= \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})} \left[\sigma(z_1)\, \sigma(z_2)\right] \mathbb{1}_{\{k=k'\}}$$

Terms in B

Let's calculate the gradients w.r.t. last layer weights

$$\partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) = \frac{1}{\sqrt{d_l}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}}$$

$$B = \sum_{i,j=1}^{d_l, d_{l+1}} \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k}(\boldsymbol{x}) \partial_{W_{ij}^{(l)}} f_{\boldsymbol{\theta},k'}(\boldsymbol{x}') = \frac{1}{d_l} \sum_{i,j=1}^{d_l, d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \mathbb{1}_{\{k=i\}} \alpha_j(\boldsymbol{x}') \mathbb{1}_{\{k'=i\}}$$

$$= \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}} \xrightarrow{d_l \to \infty} \mathbb{E}\left[\sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x})) \sigma(\tilde{\alpha}_{j,\infty}^{(l)}(\boldsymbol{x}'))\right] \mathbb{1}_{\{k=k\}}$$

$$= \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(l)})}\left[\sigma(z_1)\, \sigma(z_2)\right] \mathbb{1}_{\{k=k'\}} = \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') \mathbb{1}_{\{k=k'\}}$$

All in All

All in All

- Summing up the two quantities gives:

All in All

- Summing up the two quantities gives:

$$\Theta_{kk'}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \Theta_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}\dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}')+\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$$

All in All

- Summing up the two quantities gives:

$$\Theta_{kk'}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \Theta_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}\dot{\Sigma}^{(L+1)}(\boldsymbol{x}, \boldsymbol{x}')+\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')\mathbb{1}_{\{k=k'\}}$$

- This concludes the proof!

Time-independence

Theorem 4:

Time-independence

Let us state the time-independence theorem for completeness:

Theorem 4:

Time-independence

Let us state the time-independence theorem for completeness:

Theorem 4:
Assume σ is Lipschitz, twice differentiable and has bounded second derivative. Fix some time horizon T and assume that
$\int_0^T \sqrt{\sum_{i=1}^n \frac{\partial L(f(\boldsymbol{x}),y_i)}{\partial f(\boldsymbol{x})}\Big|_{\boldsymbol{x}=\boldsymbol{x}_i}}\, dt$ *is bounded.*

Time-independence

Let us state the time-independence theorem for completeness:

Theorem 4:
Assume σ is Lipschitz, twice differentiable and has bounded second derivative. Fix some time horizon T and assume that
$\int_0^T \sqrt{\sum_{i=1}^n \frac{\partial L(f(\boldsymbol{x}),y_i)}{\partial f(\boldsymbol{x})}\Big|_{\boldsymbol{x}=\boldsymbol{x}_i}}\, dt$ *is bounded.*

Then: $\boldsymbol{\Theta}_t^{(L)} \xrightarrow{d_1,\ldots,d_{L-1}\to\infty} \boldsymbol{\Theta}_\infty^{(L)} \otimes \mathbb{1}_{d_L \times d_L}$ *for all* $t \in [0, T]$

Proof Sketch

Proof Sketch

- Again relies on **induction** over the depth:

Proof Sketch

- Again relies on **induction** over the depth:

$$\boldsymbol{\Theta}_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$$

Proof Sketch

- Again relies on **induction** over the depth:

$$\Theta_{\infty}^{(l+1)}(x, x') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x)\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x')\right) \Theta_{\infty}^{(l)}(x, x')$$

$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(x) \alpha_j^{(l)}(x') \mathbb{1}_{\{k=k'\}}$$

Proof Sketch

- Again relies on **induction** over the depth:

$$\Theta_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x}')\right) \Theta_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$
$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}}$$

- The time derivative $\partial_t \Theta_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$ asympotically looks like:

Proof Sketch

- Again relies on **induction** over the depth:

$$\boldsymbol{\Theta}_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(\boldsymbol{x}')\right) \boldsymbol{\Theta}_{\infty}^{(l)}(\boldsymbol{x}, \boldsymbol{x}')$$
$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(\boldsymbol{x}) \alpha_j^{(l)}(\boldsymbol{x}') \mathbb{1}_{\{k=k'\}}$$

- The time derivative $\partial_t \boldsymbol{\Theta}_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$ asympotically looks like:

$$\partial_t \boldsymbol{\Theta}_{\infty}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$$

Proof Sketch

- Again relies on **induction** over the depth:

$$\Theta_\infty^{(l+1)}(x, x') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x)\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x')\right) \Theta_\infty^{(l)}(x, x')$$

$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(x) \alpha_j^{(l)}(x') \mathbb{1}_{\{k=k'\}}$$

- The time derivative $\partial_t \Theta_\infty^{(l+1)}(x, x')$ asympotically looks like:

$$\partial_t \Theta_\infty^{(l+1)}(x, x') \sim \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} \mathcal{O}(\partial_t W_{kj}^{(l)}) + \mathcal{O}(\partial_t \tilde{\alpha}_j^{(l)}) + \mathcal{O}(\partial_t \alpha_j^{(l)})$$

Proof Sketch

- Again relies on **induction** over the depth:

$$\Theta_{\infty}^{(l+1)}(x, x') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x)\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x')\right) \Theta_{\infty}^{(l)}(x, x')$$

$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(x) \alpha_j^{(l)}(x') \mathbb{1}_{\{k=k'\}}$$

- The time derivative $\partial_t \Theta_{\infty}^{(l+1)}(x, x')$ asympotically looks like:

$$\partial_t \Theta_{\infty}^{(l+1)}(x, x') \sim \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} \mathcal{O}(\partial_t W_{kj}^{(l)}) + \mathcal{O}(\partial_t \tilde{\alpha}_j^{(l)}) + \mathcal{O}(\partial_t \alpha_j^{(l)})$$

$$\sim \mathcal{O}(\partial_t W_{kj}^{(l)}) + \mathcal{O}(\partial_t \tilde{\alpha}_j^{(l)}) + \mathcal{O}(\partial_t \alpha_j^{(l)})$$

Proof Sketch

- Again relies on **induction** over the depth:

$$\Theta_{\infty}^{(l+1)}(x, x') = \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} W_{kj}^{(l)} W_{k'j}^{(l)} \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x)\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(l)}(x')\right) \Theta_{\infty}^{(l)}(x, x')$$

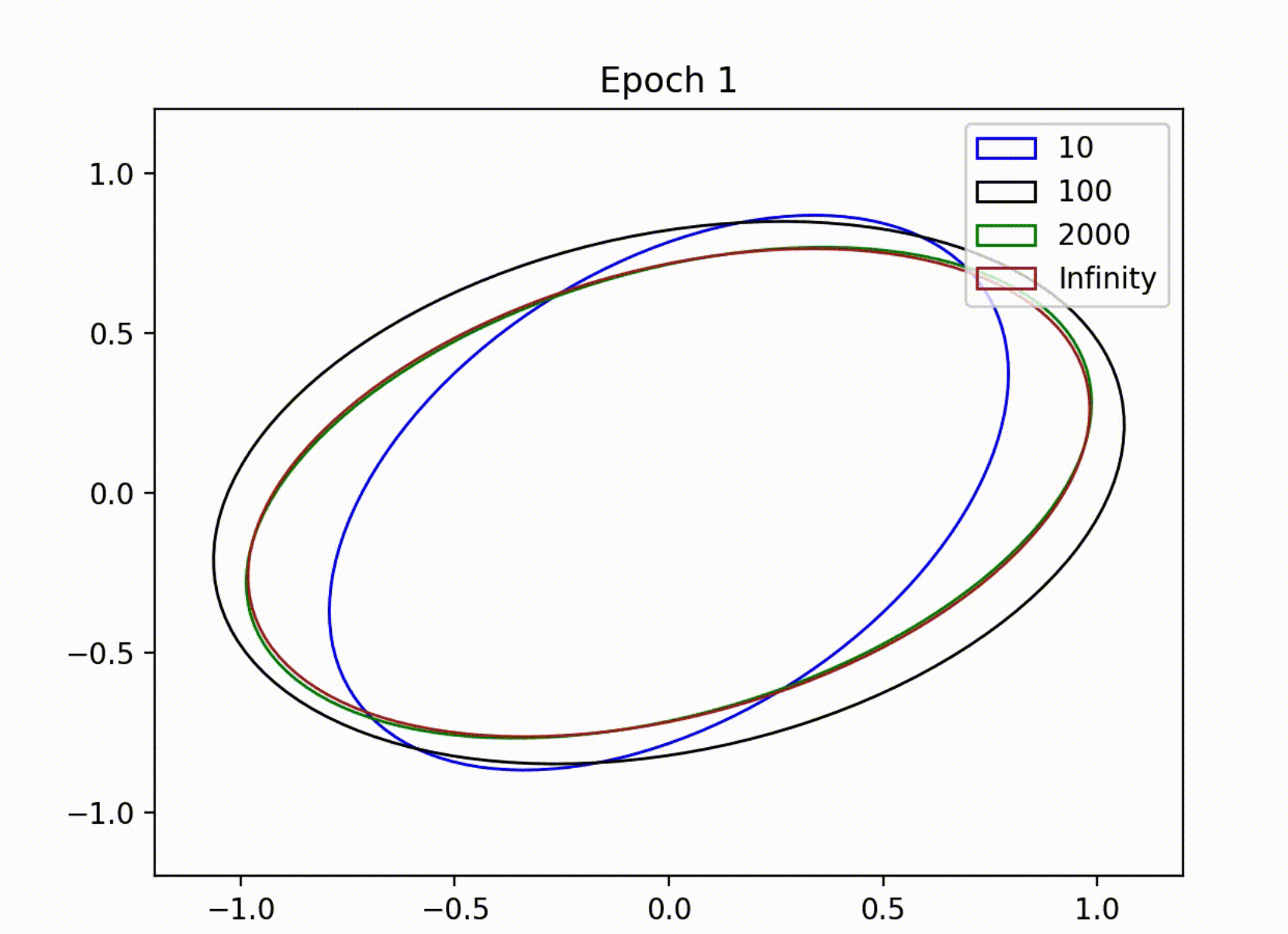
$$+ \frac{1}{d_l} \sum_{j=1}^{d_{l+1}} \alpha_j^{(l)}(x) \alpha_j^{(l)}(x') \mathbb{1}_{\{k=k'\}}$$

- The time derivative $\partial_t \Theta_{\infty}^{(l+1)}(x, x')$ asympotically looks like:

$$\partial_t \Theta_{\infty}^{(l+1)}(x, x') \sim \lim_{d_l \to \infty} \frac{1}{d_l} \sum_{j=1}^{d_l} \mathcal{O}(\partial_t W_{kj}^{(l)}) + \mathcal{O}(\partial_t \tilde{\alpha}_j^{(l)}) + \mathcal{O}(\partial_t \alpha_j^{(l)})$$

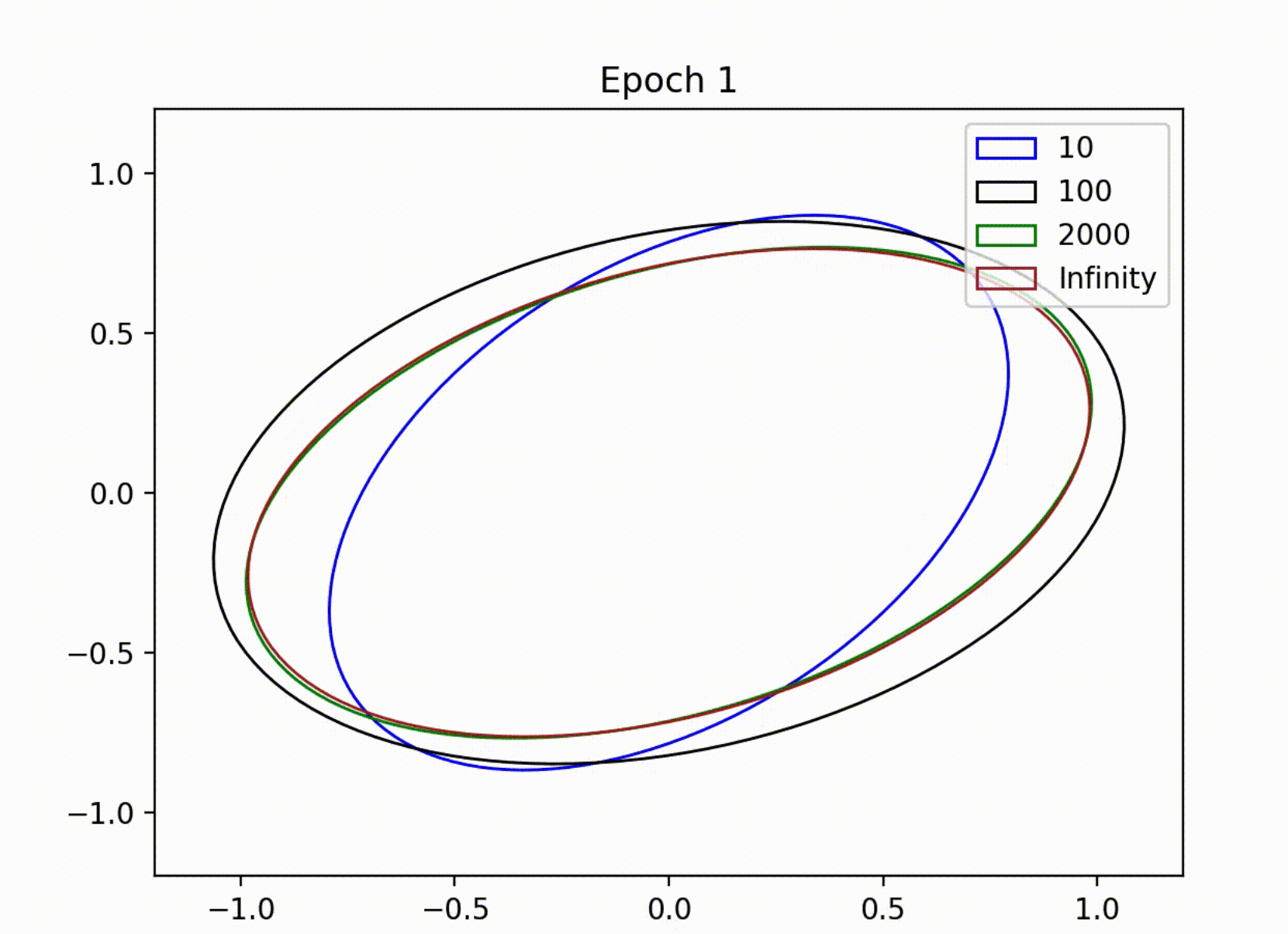
$$\sim \mathcal{O}(\partial_t W_{kj}^{(l)}) + \mathcal{O}(\partial_t \tilde{\alpha}_j^{(l)}) + \mathcal{O}(\partial_t \alpha_j^{(l)}) = \mathcal{O}\left(\frac{1}{\sqrt{d_l}}\right)$$

NTK Animation



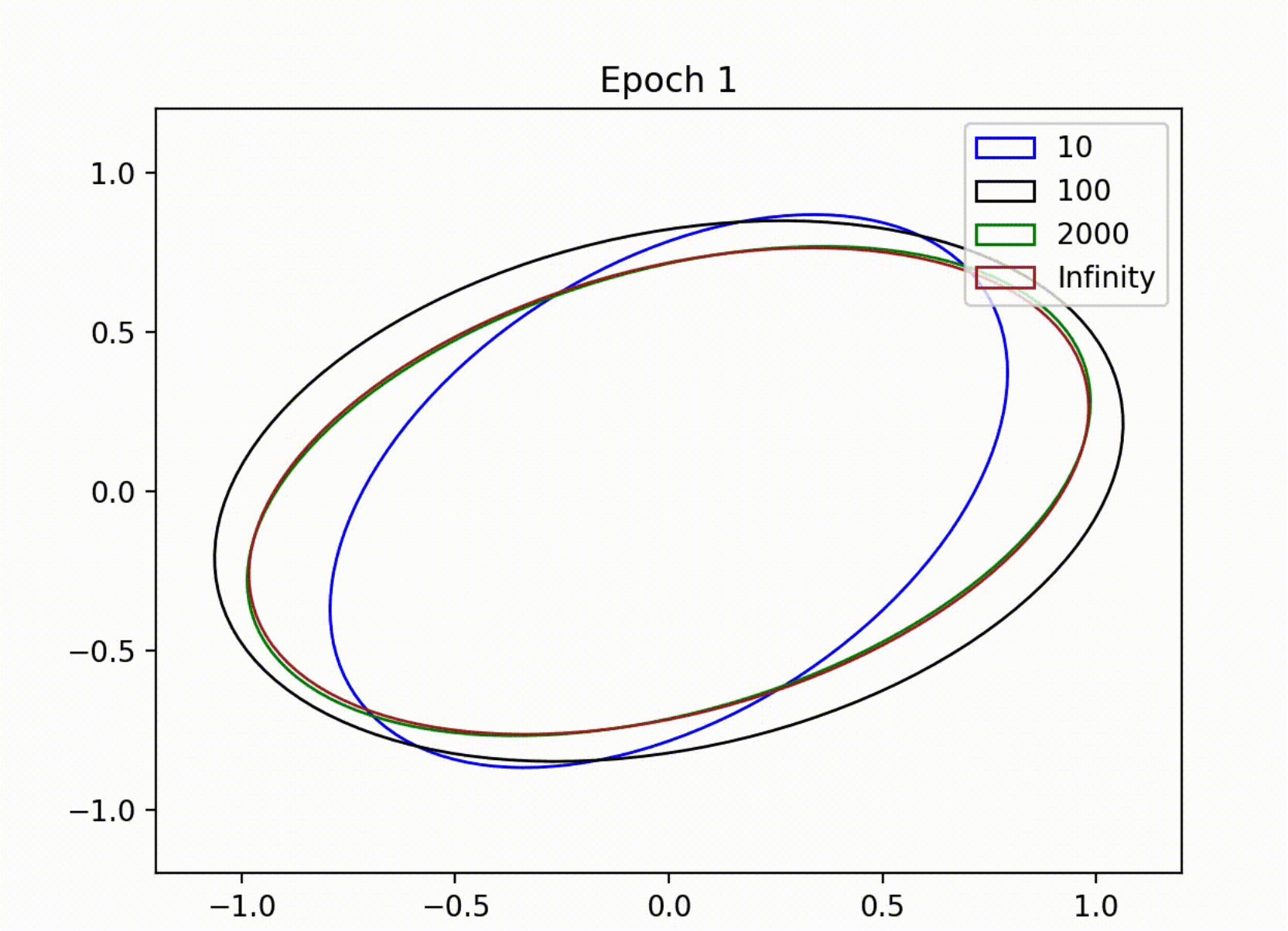
NTK Animation

- Train on just two images of *MNIST* $\implies \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{2\times 2}$



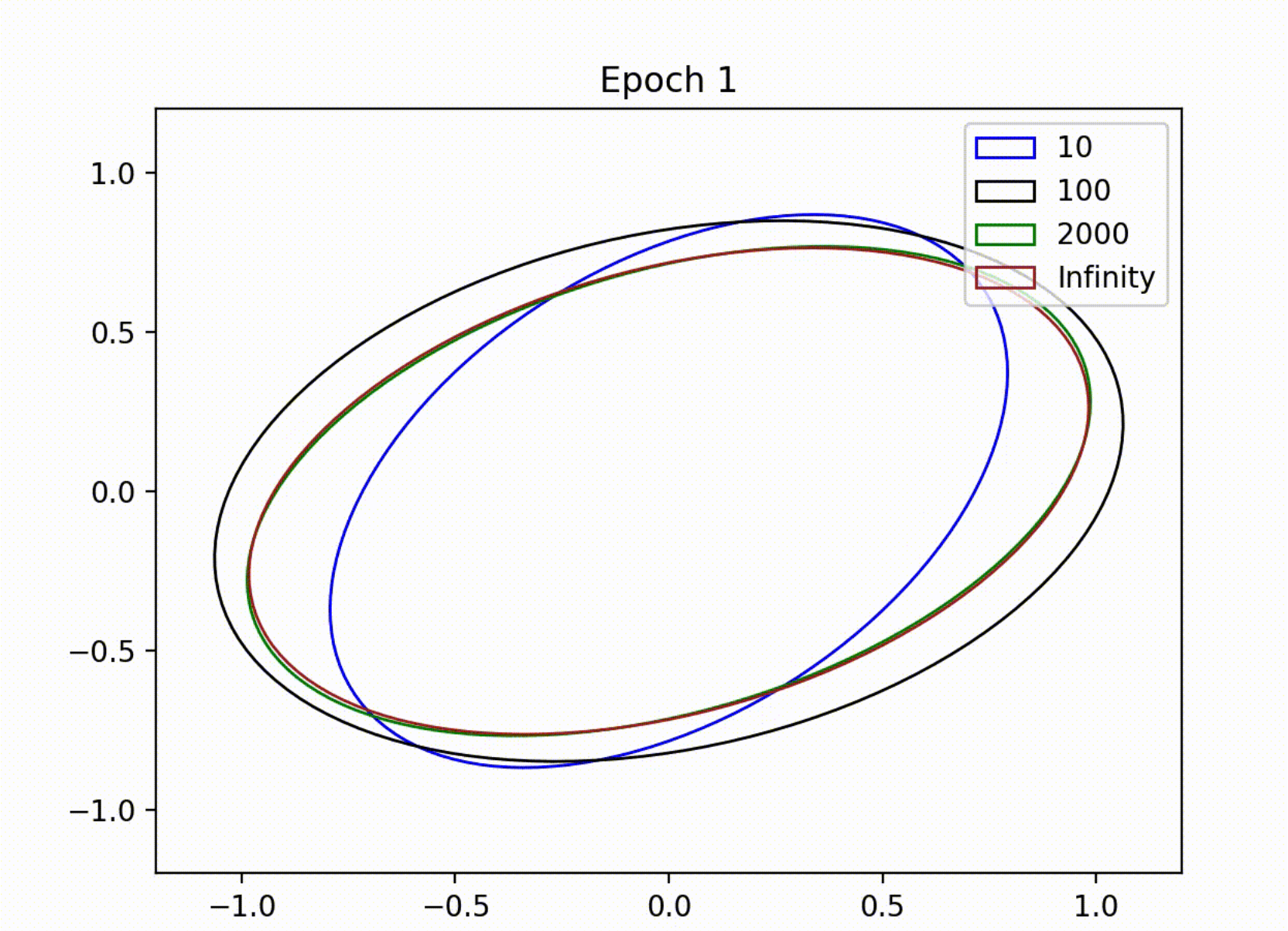
NTK Animation

- Train on just two images of *MNIST* $\implies \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{2\times 2}$
- Plot corresponding ellipse through training for different widths



NTK Animation

- Train on just two images of *MNIST* $\implies \boldsymbol{\Theta}^{(L)}(\boldsymbol{X}, \boldsymbol{X}) \in \mathbb{R}^{2\times 2}$
- Plot corresponding ellipse through training for different widths



Discussion

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc
- Only gradient flow, Seyed will present a paper that also talks about **gradient descent**

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc
- Only gradient flow, Seyed will present a paper that also talks about **gradient descent**
- Concerning a generalization bound on loss:

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc
- Only gradient flow, Seyed will present a paper that also talks about **gradient descent**
- Concerning a generalization bound on loss:

Bound $\geq$ observed kernel performance $\geq$ observed NN performance

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc
- Only gradient flow, Seyed will present a paper that also talks about **gradient descent**
- Concerning a generalization bound on loss:

Bound $\geq$ observed kernel performance $\geq$ observed NN performance

- Can hence probably not explain why neural networks work **so** well if we take the kernel route but it may be a good start

Discussion

- Training NNs $\approx$ kernel regression with NTK $\implies$ The corresponding kernel literature can be connected to NNs concerning generalization, optimization etc
- Only gradient flow, Seyed will present a paper that also talks about **gradient descent**
- Concerning a generalization bound on loss:

Bound $\geq$ observed kernel performance $\geq$ observed NN performance

- Can hence probably not explain why neural networks work **so** well if we take the kernel route but it may be a good start
- We observe that NNs work better with increasing width. This is a bit conflicting with infinite width as there performance decreases (Aurelien will present a paper that shows this empirically)

Backup Slides

Closed Form Solution (I)

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\Theta_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^{\infty}(\boldsymbol{x}) = -\left(\Theta_{\infty}^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

That's a system of **linear first order** differential equations!

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^\infty(\boldsymbol{x}) = -\left(\Theta_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^\infty(\boldsymbol{X}) - \boldsymbol{y})$$

  That's a system of **linear first order** differential equations!

- Careful, every $\boldsymbol{x}$ induces a **different** function $f_t^\infty(\boldsymbol{x})$ in t!

Closed Form Solution (I)

- Plugging this gradient into our ODE gives:

$$\partial_t f_t^\infty(\boldsymbol{x}) = -\left(\Theta_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T (f_t^\infty(\boldsymbol{X}) - \boldsymbol{y})$$

  That's a system of **linear first order** differential equations!

- Careful, every $\boldsymbol{x}$ induces a **different** function $f_t^\infty(\boldsymbol{x})$ in t!

- We first need to solve it on the training set $\boldsymbol{X}$:

$$\partial_t f_t^\infty(\boldsymbol{X}) = -\boldsymbol{\Theta}_\infty(\boldsymbol{X}, \boldsymbol{X})(f_t^\infty(\boldsymbol{X}) - \boldsymbol{y})$$

Closed Form Solution (II)

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$
- We hence get the following solution on the data

$$f_t^{\infty}(\boldsymbol{X}) = \boldsymbol{y} + e^{-t\Theta_{\infty}^{(L)}(\boldsymbol{X},\boldsymbol{X})}(f_0^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$
- We hence get the following solution on the data

$$f_t^{\infty}(\boldsymbol{X}) = \boldsymbol{y} + e^{-t\Theta_{\infty}^{(L)}(\boldsymbol{X},\boldsymbol{X})}(f_0^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

- Getting the function value at an arbitrary point $\boldsymbol{x}$ is now just a matter of **integrating** the original ODE

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$
- We hence get the following solution on the data

$$f_t^{\infty}(\boldsymbol{X}) = \boldsymbol{y} + e^{-t\Theta_{\infty}^{(L)}(\boldsymbol{X},\boldsymbol{X})}(f_0^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

- Getting the function value at an arbitrary point $\boldsymbol{x}$ is now just a matter of **integrating** the original ODE
- Some more tricks:

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$
- We hence get the following solution on the data

$$f_t^{\infty}(\boldsymbol{X}) = \boldsymbol{y} + e^{-t\Theta_{\infty}^{(L)}(\boldsymbol{X},\boldsymbol{X})}(f_0^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

- Getting the function value at an arbitrary point $\boldsymbol{x}$ is now just a matter of **integrating** the original ODE
- Some more tricks:
  1) $\int \boldsymbol{a}^T \boldsymbol{X}(t)\boldsymbol{b}\, dt = \boldsymbol{a}^T \left(\int \boldsymbol{X}(t)dt\right) \boldsymbol{b}$

Closed Form Solution (II)

- **Recall:** $\dot{\boldsymbol{x}}(t) = A(\boldsymbol{x}(t) - \boldsymbol{x}^*) \implies \boldsymbol{x}(t) = \boldsymbol{x}^* + e^{At}(\boldsymbol{x}(0) - \boldsymbol{x}^*)$
- We hence get the following solution on the data

$$f_t^{\infty}(\boldsymbol{X}) = \boldsymbol{y} + e^{-t\Theta_{\infty}^{(L)}(\boldsymbol{X},\boldsymbol{X})}(f_0^{\infty}(\boldsymbol{X}) - \boldsymbol{y})$$

- Getting the function value at an arbitrary point $\boldsymbol{x}$ is now just a matter of **integrating** the original ODE
- Some more tricks:
  1) $\int \boldsymbol{a}^T \boldsymbol{X}(t)\boldsymbol{b}\, dt = \boldsymbol{a}^T \left(\int \boldsymbol{X}(t)dt\right) \boldsymbol{b}$
  2) $\int e^{\boldsymbol{A}t} dt = A^{-1}\left(e^{\boldsymbol{A}t} - \mathbb{1}\right) + C$

Closed Form Solution (III)

Performing this integral gives

$$
\begin{aligned}
f_t^\infty(\boldsymbol{x}) &= \int - \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(f_t^\infty(\boldsymbol{X}) - \boldsymbol{y}\right)\, dt \\
&= \int - \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T e^{-\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})t} \left(f_0^\infty(\boldsymbol{X}) - \boldsymbol{y}\right)\, dt \\
&\overset{1)}{=} - \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\int e^{-\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})t}\, dt\right) \left(f_0^\infty(\boldsymbol{X}) - \boldsymbol{y}\right) \\
&\overset{2)}{=} \left(\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{x}, \boldsymbol{X})\right)^T \left(\left(\boldsymbol{\Theta}_\infty^{(L)}\right)^{-1} \left(e^{-\boldsymbol{\Theta}_\infty^{(L)}(\boldsymbol{X}, \boldsymbol{X})t} - \mathbb{1}\right)\right) \left(f_0^\infty(\boldsymbol{X}) - \boldsymbol{y}\right) \\
&\quad + f_0^\infty(\boldsymbol{x})
\end{aligned}
$$