Kernel-Based Smoothness Analysis of Residual Networks

Tom Tirer, Joan Bruna, Raja Giryes,

What's it about?

What's it about?

- Very recent submission (25 September)

What's it about?

- Very recent submission (25 September)
- NTK for **Residual** neural networks
  $\implies$ Form of NNGP and NTK are derived

What's it about?

- Very recent submission (25 September)
- NTK for **Residual** neural networks
  $\implies$ Form of NNGP and NTK are derived
- Proof for **time-invariance** of kernel for gradient descent (not flow)

What's it about?

- Very recent submission (25 September)
- NTK for **Residual** neural networks
  $\Longrightarrow$ Form of NNGP and NTK are derived
- Proof for **time-invariance** of kernel for gradient descent (not flow)
- ResNets are **superior** to MLPs in most tasks. Structural difference visible in the two NTKs?

What's it about?

- Very recent submission (25 September)
- NTK for **Residual** neural networks
  $\Longrightarrow$ Form of NNGP and NTK are derived
- Proof for **time-invariance** of kernel for gradient descent (not flow)
- ResNets are **superior** to MLPs in most tasks. Structural difference visible in the two NTKs?
- Provides **smoother** interpolations

Residual Networks

Residual Networks

- One of the **state-of-the-art** neural networks for computer vision

Residual Networks

- One of the **state-of-the-art** neural networks for computer vision
- Previous representation gets added to new representation:

$$\boldsymbol{x}^{(l+1)} = \boldsymbol{x}^{(l)} + F\left(\boldsymbol{x}^{(l)}\right)$$

Residual Networks

- One of the **state-of-the-art** neural networks for computer vision
- Previous representation gets added to new representation:

$$\boldsymbol{x}^{(l+1)} = \boldsymbol{x}^{(l)} + F\left(\boldsymbol{x}^{(l)}\right)$$

- Only very **heuristic** understanding (gradient highway) in optimization, basically **no work** on generalization

$$\underline{x}^{(l)} \longrightarrow W^{(l)}\underline{x}^{(l)} \longrightarrow V^{(l)}\phi\left(W^{(l)}\underline{x}^{(l)}\right) \longrightarrow \underline{x}^{(l)} + V^{(l)}\phi\left(W^{(l)}\underline{x}^{(l)}\right)$$

Skip connection

Notation: Residual Networks

Notation: Residual Networks

Fully-connected residual neural network with square weights:

Notation: Residual Networks

Fully-connected residual neural network with square weights:

- $\boldsymbol{x}^{(0)}(\boldsymbol{x}) = \frac{1}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x} \in \mathbb{R}^{n}$

Notation: Residual Networks

Fully-connected residual neural network with square weights:

- $\boldsymbol{x}^{(0)}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \boldsymbol{U} \boldsymbol{x} \in \mathbb{R}^{n}$
- $\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\boldsymbol{g}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^{n}$

Notation: Residual Networks

Fully-connected residual neural network with square weights:

- $\boldsymbol{x}^{(0)}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \boldsymbol{U}\boldsymbol{x} \in \mathbb{R}^n$
- $\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\boldsymbol{g}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^n$
- $\boldsymbol{g}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}$

Notation: Residual Networks

Fully-connected residual neural network with square weights:

- $\boldsymbol{x}^{(0)}(\boldsymbol{x}) = \frac{1}{\sqrt{d}} \boldsymbol{U}\boldsymbol{x} \in \mathbb{R}^n$
- $\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\boldsymbol{g}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^n$
- $\boldsymbol{g}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}$
- **Output:** $f(\boldsymbol{x}) = g^{(L+1)}(\boldsymbol{x}) = \left(\boldsymbol{w}^{(L+1)}\right)^T \boldsymbol{x}^{(L)} \in \mathbb{R}$

Notation: Residual Networks

Fully-connected residual neural network with square weights:

- $\boldsymbol{x}^{(0)}(\boldsymbol{x}) = \frac{1}{\sqrt{d}}\boldsymbol{U}\boldsymbol{x} \in \mathbb{R}^n$
- $\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha\frac{\sigma_v}{\sqrt{n}}\boldsymbol{V}^{(l+1)}\phi\left(\boldsymbol{g}^{(l+1)}(\boldsymbol{x})\right) \in \mathbb{R}^n$
- $\boldsymbol{g}^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}}\boldsymbol{W}^{(l+1)}\boldsymbol{x}^{(l)}$
- **Output:** $f(\boldsymbol{x}) = g^{(L+1)}(\boldsymbol{x}) = \left(\boldsymbol{w}^{(L+1)}\right)^T\boldsymbol{x}^{(L)} \in \mathbb{R}$

where $\boldsymbol{U} \in \mathbb{R}^{d\times n}$, $\boldsymbol{W}^{(l)} \in \mathbb{R}^{n\times n}$, $\boldsymbol{V}^{(l)} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{w}^{(L+1)} \in \mathbb{R}^n$

More Notation

More Notation

- Dataset $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$

More Notation

- Dataset $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$
- For convenience introduce the function

$$\mathcal{T} : \mathbb{R}^{n \times n} \to \mathbb{R},\ \boldsymbol{\Sigma} \mapsto \mathcal{T}(\boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})}[\phi(z_1)\phi(z_2)]$$

More Notation

- Dataset $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$
- For convenience introduce the function

$$\mathcal{T} : \mathbb{R}^{n \times n} \to \mathbb{R},\ \boldsymbol{\Sigma} \mapsto \mathcal{T}(\boldsymbol{\Sigma}) = \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})}[\phi(z_1)\phi(z_2)]$$

- This function has **closed form** for *ReLU* and *erf* activation

NNGP for Fully Connected Networks

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Sigma^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sigma_w^2 \mathcal{T}\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}\right)$

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Sigma^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sigma_w^2 \mathcal{T}\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}\right)$

Neural networks are linked in **two ways** to this kernel:

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Sigma^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sigma_w^2 \mathcal{T} \left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}} \right)$

Neural networks are linked in **two ways** to this kernel:

- $f(\cdot) \xrightarrow{(d)} \mathcal{GP}(0, \Sigma^{(L+1)})$ and kernel regression with $\Sigma^{(L+1)}$ corresponds to **Bayesian inference** with $\mathcal{GP}(0, \Sigma^{(L+1)})$ as a prior

NNGP for Fully Connected Networks

Recall the NNGP for fully-connected networks

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Sigma^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \sigma_w^2 \mathcal{T}\left(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}\right)$

Neural networks are linked in **two ways** to this kernel:

- $f(\cdot) \xrightarrow{(d)} \mathcal{GP}(0, \Sigma^{(L+1)})$ and kernel regression with $\Sigma^{(L+1)}$ corresponds to **Bayesian inference** with $\mathcal{GP}(0, \Sigma^{(L+1)})$ as a prior
- $\Sigma^{(L+1)}$ also arises as the **NTK** of $f(\cdot)$ if only the top layer of f is trained

NNGP for Residual Networks

NNGP for ResNet

NNGP for Residual Networks

NNGP for ResNet

Denote by f again the output of above residual network. For width $n \to \infty$ we have:

$$f(\cdot) \xrightarrow{(d)} \mathcal{GP}(0, \Sigma_{\text{res}}^{(L+1)})$$

where we have the recursion

NNGP for Residual Networks

NNGP for ResNet

Denote by f again the output of above residual network. For width $n \to \infty$ we have:

$$f(\cdot) \xrightarrow{(d)} \mathcal{GP}(0, \Sigma_{\text{res}}^{(L+1)})$$

where we have the recursion

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$

NNGP for Residual Networks

NNGP for ResNet

Denote by f again the output of above residual network. For width $n \to \infty$ we have:

$$f(\cdot) \xrightarrow{(d)} \mathcal{GP}(0, \Sigma_{\text{res}}^{(L+1)})$$

where we have the recursion

- $\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Sigma_{\text{res}}^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \Sigma_{\text{res}}^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \alpha^2 \sigma_v^2 \sigma_w^2 \mathcal{T}\left(\boldsymbol{\Sigma}_{\text{res}}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}\right)$

Comments

Comments

- Structure very similar to standard NNGP, the ResNet NNGP however inherits the **additive structure**

Comments

- Structure very similar to standard NNGP, the ResNet NNGP however inherits the **additive structure**
- We will look at a **sequential limit** proof for presentation purposes (but bit confusing because all widths are n...)

Comments

- Structure very similar to standard NNGP, the ResNet NNGP however inherits the **additive structure**
- We will look at a **sequential limit** proof for presentation purposes (but bit confusing because all widths are n...)
- Proof hence via **induction**

Comments

- Structure very similar to standard NNGP, the ResNet NNGP however inherits the **additive structure**
- We will look at a **sequential limit** proof for presentation purposes (but bit confusing because all widths are n...)
- Proof hence via **induction**
- There are proofs **independent** of the order of limits

Proof for NNGP: Base Case

Proof for NNGP: Base Case

- $L = 1$:

$$f(\boldsymbol{x}) = x_i^{(0)} = \frac{\sigma_w}{\sqrt{d}}(\boldsymbol{U}\boldsymbol{x})_i = \frac{\sigma_w}{\sqrt{d}}\sum_{k=1}^{d} U_{ik}x_k \sim \mathcal{N}(0, \sigma_w||\boldsymbol{x}||_2^2)$$

Proof for NNGP: Base Case

- $L = 1$:

$$f(\boldsymbol{x}) = x_i^{(0)} = \frac{\sigma_w}{\sqrt{d}}(\boldsymbol{U}\boldsymbol{x})_i = \frac{\sigma_w}{\sqrt{d}}\sum_{k=1}^{d} U_{ik}x_k \sim \mathcal{N}(0, \sigma_w||\boldsymbol{x}||_2^2)$$

- For the covariance we have that

$$\mathbb{E}[f(\boldsymbol{x})f(\tilde{\boldsymbol{x}})] = \frac{\sigma_w^2}{d}\sum_{k,l=1}^{d}\mathbb{E}[U_{ik}U_{il}]x_k\tilde{x}_l = \frac{\sigma_w^2}{d}\boldsymbol{x}^T\tilde{\boldsymbol{x}} = \Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

Proof for NNGP: Base Case

- $L = 1$:

$$f(\boldsymbol{x}) = x_i^{(0)} = \frac{\sigma_w}{\sqrt{d}}(\boldsymbol{U}\boldsymbol{x})_i = \frac{\sigma_w}{\sqrt{d}}\sum_{k=1}^{d} U_{ik}x_k \sim \mathcal{N}(0, \sigma_w ||\boldsymbol{x}||_2^2)$$

- For the covariance we have that

$$\mathbb{E}[f(\boldsymbol{x})f(\tilde{\boldsymbol{x}})] = \frac{\sigma_w^2}{d}\sum_{k,l=1}^{d} \mathbb{E}[U_{ik}U_{il}]x_k\tilde{x}_l = \frac{\sigma_w^2}{d}\boldsymbol{x}^T\tilde{\boldsymbol{x}} = \Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

- One can easily check that Gaussianity also holds for the vector

$$(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{(1)})$$

Proof for NNGP: Inductive Step

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$
- First look at the the **inner** representation of the residual block:

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$g_j^{(l+1)}(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x})$$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\left[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2 \right]\right)
\end{aligned}
$$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\Big[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2 \Big]\right) \\
&= \mathcal{N}\left(0, \sigma_w^2 \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$
- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\left[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2\right]\right) \\
&= \mathcal{N}\left(0, \sigma_w^2 \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

- Due to multi-dimensional **CLT** it suffices to consider:

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\Big[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2\Big]\right) \\
&= \mathcal{N}\left(0, \sigma_w^2 \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

- Due to multi-dimensional **CLT** it suffices to consider:

$$
\mathbb{E}\left[W_{js}^{(l+1)} x_s^{(l)}(\boldsymbol{x}) W_{jt}^{(l+1)} x_t^{(l)}(\tilde{\boldsymbol{x}})\right] = \mathbb{1}_{\{s=t\}} \mathbb{E}[x_s^{(l)}(\boldsymbol{x}) x_t^{(l)}(\tilde{\boldsymbol{x}})]
$$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\Big[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2 \Big]\right) \\
&= \mathcal{N}\left(0, \sigma_w^2 \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

- Due to multi-dimensional **CLT** it suffices to consider:

$$
\begin{aligned}
\mathbb{E}\left[W_{js}^{(l+1)} x_s^{(l)}(\boldsymbol{x}) W_{jt}^{(l+1)} x_t^{(l)}(\tilde{\boldsymbol{x}})\right] &= \mathbb{1}_{\{s=t\}} \mathbb{E}[x_s^{(l)}(\boldsymbol{x}) x_t^{(l)}(\tilde{\boldsymbol{x}})] \\
&= \mathbb{E}[x_t^{(l)}(\boldsymbol{x}) x_t^{(l)}(\tilde{\boldsymbol{x}})]
\end{aligned}
$$

Proof for NNGP: Inductive Step

- **Assume:** $x_i^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

- First look at the the **inner** representation of the residual block:

$$
\begin{aligned}
g_j^{(l+1)}(\boldsymbol{x}) &= \frac{\sigma_w}{\sqrt{n}} \sum_{k=1}^{n} W_{jk}^{(l+1)} x_k^{(l)}(\boldsymbol{x}) \\
&\xrightarrow{(d)} \mathcal{N}\left(0, \sigma_w^2 \mathbb{E}\left[\left(W_{jk}^{(l+1)}\right)^2 \left(x_k^{(l)}\right)^2 \right]\right) \\
&= \mathcal{N}\left(0, \sigma_w^2 \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

- Due to multi-dimensional **CLT** it suffices to consider:

$$
\begin{aligned}
\mathbb{E}\left[W_{js}^{(l+1)} x_s^{(l)}(\boldsymbol{x}) W_{jt}^{(l+1)} x_t^{(l)}(\tilde{\boldsymbol{x}})\right] &= \mathbb{1}_{\{s=t\}} \mathbb{E}[x_s^{(l)}(\boldsymbol{x}) x_t^{(l)}(\tilde{\boldsymbol{x}})] \\
&= \mathbb{E}[x_t^{(l)}(\boldsymbol{x}) x_t^{(l)}(\tilde{\boldsymbol{x}})] \\
&= \Sigma^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})
\end{aligned}
$$

Proof for NNGP: Inductive Step

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$
- Now we can take the **next** width to infinity:

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$

- Now we can take the **next** width to infinity:

$$x_i^{(l+1)}(\boldsymbol{x}) = x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha \sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right)$$

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$
- Now we can take the **next** width to infinity:

$$x_i^{(l+1)}(\boldsymbol{x}) = x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha\sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right)$$

$$\xrightarrow{(d)} \mathcal{N}(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \mathcal{N}\left(0, \alpha^2\sigma_v^2 \mathbb{E}\left[\left(V_{ij}^{(l+1)}\right)^2 \phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\right]\right)$$

Actually needs more work: $X_n \xrightarrow{(d)} X$ and $Y_n \xrightarrow{(d)} Y$ does not imply that

$X_n + Y_n \xrightarrow{(d)} X_n + Y_n$

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$

- Now we can take the **next** width to infinity:

$$x_i^{(l+1)}(\boldsymbol{x}) = x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha \sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right)$$

$$\xrightarrow{(d)} \mathcal{N}(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \mathcal{N}\left(0, \alpha^2 \sigma_v^2 \mathbb{E}\Big[\left(V_{ij}^{(l+1)}\right)^2 \phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\overset{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2 \sigma_v^2 \mathbb{E}\Big[\phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$
- Now we can take the **next** width to infinity:

$$x_i^{(l+1)}(\boldsymbol{x}) = x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha \sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right)$$

$$\xrightarrow{(d)} \mathcal{N}(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \mathcal{N}\left(0, \alpha^2 \sigma_v^2 \mathbb{E}\Big[\left(V_{ij}^{(l+1)}\right)^2 \phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\overset{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2 \sigma_v^2 \mathbb{E}\Big[\phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\overset{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2 \sigma_w^2 \sigma_v^2 \mathbb{E}\Big[\phi^2\left(\sigma_w^{-1} g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$
- Now we can take the **next** width to infinity:

$$
\begin{aligned}
x_i^{(l+1)}(\boldsymbol{x}) &= x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha\sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right) \\
&\xrightarrow{(d)} \mathcal{N}(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \mathcal{N}\left(0, \alpha^2\sigma_v^2 \mathbb{E}\left[\left(V_{ij}^{(l+1)}\right)^2 \phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\right]\right) \\
&\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_v^2 \mathbb{E}\left[\phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\right]\right) \\
&\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_w^2\sigma_v^2 \mathbb{E}\left[\phi^2\left(\sigma_w^{-1} g_j^{(l+1)}(\boldsymbol{x})\right)\right]\right) \\
&\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_w^2\sigma_v^2 \mathcal{T}(\Sigma(\boldsymbol{x}, \boldsymbol{x})\right)
\end{aligned}
$$

Seems like we need 1-homogeneous activations?

Proof for NNGP: Inductive Step

- **Hence:** $g_i^{(l+1)} \sim \mathcal{GP}(0, \sigma_w^2 \Sigma^{(l)})$
- Now we can take the **next** width to infinity:

$$x_i^{(l+1)}(\boldsymbol{x}) = x_i^{(l)}(\boldsymbol{x}) + \frac{\alpha\sigma_v}{\sqrt{n}} \sum_{j=1}^{n} V_{ij}^{(l+1)} \phi\left(g_j^{(l+1)}(\boldsymbol{x})\right)$$

$$\xrightarrow{(d)} \mathcal{N}(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \mathcal{N}\left(0, \alpha^2\sigma_v^2 \mathbb{E}\Big[\left(V_{ij}^{(l+1)}\right)^2 \phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_v^2 \mathbb{E}\Big[\phi^2\left(g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_w^2\sigma_v^2 \mathbb{E}\Big[\phi^2\left(\sigma_w^{-1} g_j^{(l+1)}(\boldsymbol{x})\right)\Big]\right)$$

$$\stackrel{(d)}{=} \mathcal{N}\left(0, \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) + \alpha^2\sigma_w^2\sigma_v^2 \mathcal{T}(\Sigma(\boldsymbol{x}, \boldsymbol{x})\right)$$

- **Off-diagonal** terms again follow from multidimensional CLT

NTK for MLPs

Recall: For $f(\cdot)$ an MLP, we have that

$$\hat{\Theta}_{\mathrm{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta_{\mathrm{MLP}}^{(L+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

NTK for MLPs

Recall: For $f(\cdot)$ an MLP, we have that

$$\hat{\Theta}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta^{(L+1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

NTK for MLPs

Recall: For $f(\cdot)$ an MLP, we have that

$$\hat{\Theta}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta^{(L+1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

- $\Theta^{(1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$

NTK for MLPs

Recall: For $f(\cdot)$ an MLP, we have that

$$\hat{\Theta}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta^{(L+1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

- $\Theta^{(1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Theta^{(l+1)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \Theta^{(l)}_{\text{MLP}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \dot{\mathcal{T}}(\Sigma^{(l+1)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}) + \mathcal{T}(\Sigma^{(l+1)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}})$

NTK for ResNet

NTK for ResNet

For width $n \to \infty$ we have:

$$\hat{\Theta}_{\text{res}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta_{\text{res}}^{(L+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

NTK for ResNet

For width $n \to \infty$ we have:

$$\hat{\Theta}_{\text{res}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta_{\text{res}}^{(L+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

- $\Theta_{\text{res}}^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$

NTK for ResNet

For width $n \to \infty$ we have:

$$\hat{\Theta}_{\text{res}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta_{\text{res}}^{(L+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

- $\Theta_{\text{res}}^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $\Theta_{\text{res}}^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \Sigma_{\text{res}}^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \Pi^{(0)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$
  $+ \alpha^2 \sum_{m=1}^{l} \Pi^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\sigma_v^2\sigma_w^2 \mathcal{T}(\Sigma^{(m)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}})$
  $+ \alpha^2 \sum_{m=1}^{l} \Pi^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\Sigma^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})\dot{\mathcal{T}}(\Sigma^{(m)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}})$

NTK for ResNet

For width $n \to \infty$ we have:

$$\hat{\Theta}_{\text{res}}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle \xrightarrow{(d)} \Theta_{\text{res}}^{(L+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$$

where we have the recursion

- $\Theta_{\text{res}}^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \frac{\sigma_w^2}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}}$
- $$\begin{aligned}\Theta_{\text{res}}^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \Sigma_{\text{res}}^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \Pi^{(0)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \Sigma^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \\ + \alpha^2 \textstyle\sum_{m=1}^{l} \Pi^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \sigma_v^2 \sigma_w^2 \mathcal{T}(\Sigma^{(m)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}) \\ + \alpha^2 \textstyle\sum_{m=1}^{l} \Pi^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \Sigma^{(m)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \dot{\mathcal{T}}(\Sigma^{(m)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}})\end{aligned}$$

where

$$\Pi^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = \Pi^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \left(1 + \alpha^2 \sigma_v^2 \sigma_w^2 \dot{\mathcal{T}}(\Sigma^{(l+1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}))\right)$$

$$\Pi^{(L)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) = 1$$

Proof Sketch (1)

Proof Sketch (1)

- We need to calculate all the derivatives

Proof Sketch (1)

- We need to calculate all the derivatives

$$\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{l=1}^{L} \left(\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle + \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{V}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{V}^{(l)}} \right\rangle \right)$$
$$+ \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{U}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{U}} \right\rangle + \left\langle \frac{\partial f(\underline{x})}{\partial \underline{W}^{(L+1)}}, \frac{\partial f(\tilde{\underline{x}})}{\partial \underline{W}^{(L+1)}} \right\rangle$$

Proof Sketch (1)

- We need to calculate all the derivatives

$$\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{l=1}^{L} \left(\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle + \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{V}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{V}^{(l)}} \right\rangle \right)$$
$$+ \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{U}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{U}} \right\rangle + \left\langle \frac{\partial f(\underline{x})}{\partial \underline{w}^{(L+1)}}, \frac{\partial f(\tilde{\underline{x}})}{\partial \underline{w}^{(L+1)}} \right\rangle$$

- **Recall:** $f(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \left(\boldsymbol{w}^{(L+1)} \right)^T \boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(0)} = \frac{1}{\sqrt{d}} \boldsymbol{U}\boldsymbol{x}$

Proof Sketch (1)

- We need to calculate all the derivatives

$$\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{l=1}^{L} \left(\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle + \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{V}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{V}^{(l)}} \right\rangle \right)$$
$$+ \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{U}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{U}} \right\rangle + \left\langle \frac{\partial f(\underline{x})}{\partial \underline{w}^{(L+1)}}, \frac{\partial f(\underline{\tilde{x}})}{\partial \underline{w}^{(L+1)}} \right\rangle$$

- **Recall:** $f(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \left(\boldsymbol{w}^{(L+1)} \right)^T \boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(0)} = \frac{1}{\sqrt{d}} \boldsymbol{U}\boldsymbol{x}$

- $\frac{\partial f}{\partial \boldsymbol{w}^{(L+1)}} = \frac{\sigma_w}{\sqrt{n}} \boldsymbol{x}^{(L)}$

Proof Sketch (1)

- We need to calculate all the derivatives

$$\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{\theta}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{\theta}} \right\rangle = \sum_{l=1}^{L} \left(\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle + \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{V}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{V}^{(l)}} \right\rangle \right)$$
$$+ \left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{U}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{U}} \right\rangle + \left\langle \frac{\partial f(\underline{x})}{\partial \underline{w}^{(L+1)}}, \frac{\partial f(\tilde{\underline{x}})}{\partial \underline{w}^{(L+1)}} \right\rangle$$

- **Recall:** $f(\boldsymbol{x}) = \frac{\sigma_w}{\sqrt{n}} \left(\boldsymbol{w}^{(L+1)} \right)^T \boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(0)} = \frac{1}{\sqrt{d}} \boldsymbol{U}\boldsymbol{x}$

- $\frac{\partial f}{\partial \boldsymbol{w}^{(L+1)}} = \frac{\sigma_w}{\sqrt{n}} \boldsymbol{x}^{(L)}$

- $\frac{\partial f}{\partial U_{kl}} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial U_{kl}} = \frac{1}{\sqrt{d}} \frac{\partial f}{\partial x_k^{(0)}} x_l = \frac{1}{\sqrt{d}} \left(\frac{\partial f}{\partial \boldsymbol{x}^{(0)}} \boldsymbol{x}^T \right)_{kl}$

Proof Sketch (2)

Proof Sketch (2)

- **Recall** the form of the residual network

$$\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}(\boldsymbol{x})\right) \in \mathbb{R}^n$$

Proof Sketch (2)

- **Recall** the form of the residual network

$$\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}(\boldsymbol{x})\right) \in \mathbb{R}^n$$

- $\frac{\partial f}{\partial \boldsymbol{x}^{(l)}} = \frac{\sigma_w}{\sqrt{n}} \left(\prod_{k=l+1}^{L} \frac{\partial \boldsymbol{x}^{(k)}}{\partial \boldsymbol{x}^{(k-1)}}\right)^T \boldsymbol{w}^{(L+1)} = \boldsymbol{\delta}^{(l)}$

Proof Sketch (2)

- **Recall** the form of the residual network

$$\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}(\boldsymbol{x})\right) \in \mathbb{R}^n$$

- $\frac{\partial f}{\partial \boldsymbol{x}^{(l)}} = \frac{\sigma_w}{\sqrt{n}} \left(\prod_{k=l+1}^{L} \frac{\partial \boldsymbol{x}^{(k)}}{\partial \boldsymbol{x}^{(k-1)}}\right)^T \boldsymbol{w}^{(L+1)} = \boldsymbol{\delta}^{(l)}$

- $\frac{\partial f}{\partial W_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial W_{ij}^{(l)}}$ and $\frac{\partial f}{\partial V_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial V_{ij}^{(l)}}$

Proof Sketch (2)

- **Recall** the form of the residual network

$$\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}(\boldsymbol{x})\right) \in \mathbb{R}^n$$

- $\frac{\partial f}{\partial \boldsymbol{x}^{(l)}} = \frac{\sigma_w}{\sqrt{n}} \left(\prod_{k=l+1}^{L} \frac{\partial \boldsymbol{x}^{(k)}}{\partial \boldsymbol{x}^{(k-1)}}\right)^T \boldsymbol{w}^{(L+1)} = \boldsymbol{\delta}^{(l)}$

- $\frac{\partial f}{\partial W_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial W_{ij}^{(l)}}$ and $\frac{\partial f}{\partial V_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial V_{ij}^{(l)}}$

- $\frac{\partial x_k^{(l)}}{\partial W_{ij}^{(l)}} = \alpha \frac{\sigma_v \sigma_w}{n} V_{ki}^{(l)} \phi'\left(g_i^{(l)}\right) x_j^{(l-1)}$

Proof Sketch (2)

- **Recall** the form of the residual network

$$\boldsymbol{x}^{(l+1)}(\boldsymbol{x}) = \boldsymbol{x}^{(l)}(\boldsymbol{x}) + \alpha \frac{\sigma_v}{\sqrt{n}} \boldsymbol{V}^{(l+1)} \phi\left(\frac{\sigma_w}{\sqrt{n}} \boldsymbol{W}^{(l+1)} \boldsymbol{x}^{(l)}(\boldsymbol{x})\right) \in \mathbb{R}^n$$

- $\frac{\partial f}{\partial \boldsymbol{x}^{(l)}} = \frac{\sigma_w}{\sqrt{n}} \left(\prod_{k=l+1}^{L} \frac{\partial \boldsymbol{x}^{(k)}}{\partial \boldsymbol{x}^{(k-1)}}\right)^T \boldsymbol{w}^{(L+1)} = \boldsymbol{\delta}^{(l)}$

- $\frac{\partial f}{\partial W_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial W_{ij}^{(l)}}$ and $\frac{\partial f}{\partial V_{ij}^{(l)}} = \sum_{k=1}^{n} \delta_k^{(l)} \frac{\partial x_k^{(l)}}{\partial V_{ij}^{(l)}}$

- $\frac{\partial x_k^{(l)}}{\partial W_{ij}^{(l)}} = \alpha \frac{\sigma_v \sigma_w}{n} V_{ki}^{(l)} \phi'\left(g_i^{(l)}\right) x_j^{(l-1)}$

- $\frac{\partial x_k^{(l)}}{\partial V_{ij}^{(l)}} = \alpha \frac{\sigma_v}{\sqrt{n}} \phi\left(g_j^{(l)}\right) \mathbb{1}_{\{k=i\}}$

Proof Sketch (3)

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$

$$= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$

Proof Sketch (3)

- $$\begin{aligned}\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] &= \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right] \\ &= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right] \\ &= \frac{K^{(1)}(\boldsymbol{x},\tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]\end{aligned}$$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d}\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$
$$= \frac{1}{d}\boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$
$$= \frac{K^{(1)}(\boldsymbol{x},\tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]$$

- $\mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial V_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial V_{ij}^{(l)}}\right] = \frac{\alpha^2\sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right)\right]$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$
$$= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$
$$= \frac{K^{(1)}(\boldsymbol{x},\tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]$$

- $\mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial V_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial V_{ij}^{(l)}}\right] = \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right)\right]$
$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathbb{E}\left[\phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right) \middle| \delta_i^{(l)}\right]\right]$$

Tower property: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$
$$= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$
$$= \frac{K^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]$$

- $\mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial V_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial V_{ij}^{(l)}}\right] = \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right)\right]$
$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathbb{E}\left[\phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right) \middle| \delta_i^{(l)}\right]\right]$$
$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}})\right]$$

$g_j^{(l)} \sim \mathcal{GP}(0, \Sigma^{(l)})$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$
$$= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$
$$= \frac{K^{(1)}(\boldsymbol{x},\tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]$$

- $\mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial V_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial V_{ij}^{(l)}}\right] = \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right)\right]$
$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathbb{E}\left[\phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right) \middle| \delta_i^{(l)}\right]\right]$$
$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}})\right]$$
$$= \frac{\alpha^2 \sigma_v^2}{n} \underbrace{\mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}})}_{\text{deterministic}} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)}\right]$$

Proof Sketch (3)

- $\sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial U_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial U_{ij}^{(l)}}\right] = \frac{1}{d} \sum_{i,j=1}^{d,n} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} x_j \frac{\partial \tilde{f}}{\partial x_i^{(0)}} \tilde{x}_j\right]$

$$= \frac{1}{d} \boldsymbol{x}^T \tilde{\boldsymbol{x}} \mathbb{E}\left[\frac{\partial f}{\partial x_i^{(0)}} \frac{\partial \tilde{f}}{\partial x_i^{(0)}}\right]$$

$$= \frac{K^{(1)}(\boldsymbol{x}, \tilde{\boldsymbol{x}})}{\sigma_w^2} \mathbb{E}\left[\left(\boldsymbol{\delta}^{(0)}(\boldsymbol{x})\right)^T \boldsymbol{\delta}^{(0)}(\tilde{\boldsymbol{x}})\right]$$

- $\mathbb{E}\left[\frac{\partial f(\boldsymbol{x})}{\partial V_{ij}^{(l)}} \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial V_{ij}^{(l)}}\right] = \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right)\right]$

$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathbb{E}\left[\phi\left(g_j^{(l)}\right) \phi\left(\tilde{g}_j^{(l)}\right) \middle| \delta_i^{(l)}\right]\right]$$

$$= \frac{\alpha^2 \sigma_v^2}{n} \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)} \mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}})\right]$$

$$= \frac{\alpha^2 \sigma_v^2}{n} \mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}) \mathbb{E}\left[\delta_i^{(l)} \tilde{\delta}_i^{(l)}\right]$$

- **So:** $\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{V}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{V}^{(l)}} \right\rangle = \frac{\alpha^2 \sigma_v^2}{n} \mathcal{T}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x}, \tilde{\boldsymbol{x}}}) \mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$

Proof Sketch (4)

Proof Sketch (4)

- A similar sequence of steps leads to

$$\mathbb{E}\left[\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle\right] = \alpha^2 \sigma_v^2 \Sigma^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \dot{\mathcal{T}}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}}) \mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$$

Proof Sketch (4)

- A similar sequence of steps leads to

$$\mathbb{E}\left[\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle\right] = \alpha^2 \sigma_v^2 \Sigma^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \dot{\mathcal{T}}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}}) \mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$$

- Remains to deal with

$$\mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$$

Proof Sketch (4)

- A similar sequence of steps leads to

$$\mathbb{E}\left[\left\langle \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\tilde{\boldsymbol{x}})}{\partial \boldsymbol{W}^{(l)}} \right\rangle\right] = \alpha^2 \sigma_v^2 \Sigma^{(l)}(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \dot{\mathcal{T}}(\boldsymbol{\Sigma}^{(l)}|_{\boldsymbol{x},\tilde{\boldsymbol{x}}}) \mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$$

- Remains to deal with

$$\mathbb{E}\left[\left(\boldsymbol{\delta}^{(l)}\right)^T \tilde{\boldsymbol{\delta}}^{(l)}\right]$$

- This is given exactly by $\Pi(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ in the theorem

Time-invariance of Θ_{res}

Time-invariance of Θ_{res}

- One can now prove that the NTK remains **constant** through training with GD

Time-invariance of Θ_{res}

- One can now prove that the NTK remains **constant** through training with GD

The paper that Seyed presented

- Proof very similar to *"Wide neural networks of any depth evolve as linear models"*

Time-invariance of Θ_{res}

- One can now prove that the NTK remains **constant** through training with GD
- Proof very similar to ”*Wide neural networks of any depth evolve as linear models*”
- Interesting decomposition of proof into two steps, one **dependent** on the architecture and the other **independent** of particular architecture

First Step

First Step

- Proving the following lemma depends on the architecture

First Step

- Proving the following lemma depends on the architecture

Lemma: *Fix two weights* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : ||\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}||_2\}$. *There exists* $K > 0$ *such that* $\forall C > 0$ *with* $N >> C^2$ *we have that*

First Step

- Proving the following lemma depends on the architecture

Lemma: *Fix two weights* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : ||\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}||_2\}$. *There exists* $K > 0$ *such that* $\forall C > 0$ *with* $N >> C^2$ *we have that*

$$||J(\boldsymbol{\theta})||_F \leq K$$

First Step

- Proving the following lemma depends on the architecture

Lemma: *Fix two weights* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : ||\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}||_2\}$*. There exists* $K > 0$ *such that* $\forall C > 0$ *with* $N >> C^2$ *we have that*

$$||J(\boldsymbol{\theta})||_F \leq K$$

$$||J(\boldsymbol{\theta}) - J(\tilde{\boldsymbol{\theta}})||_F \leq K||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2$$

First Step

- Proving the following lemma depends on the architecture

Lemma: *Fix two weights* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : ||\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}||_2\}$*. There exists* $K > 0$ *such that* $\forall C > 0$ *with* $N >> C^2$ *we have that*

$$||J(\boldsymbol{\theta})||_F \leq K$$

$$||J(\boldsymbol{\theta}) - J(\tilde{\boldsymbol{\theta}})||_F \leq K||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2$$

where $J(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{|\boldsymbol{X}| \times n}$

First Step

- Proving the following lemma depends on the architecture

Lemma: *Fix two weights* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \{\boldsymbol{\theta} : ||\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}||_2\}$. *There exists* $K > 0$ *such that* $\forall C > 0$ *with* $N >> C^2$ *we have that*

$$||J(\boldsymbol{\theta})||_F \leq K$$

$$||J(\boldsymbol{\theta}) - J(\tilde{\boldsymbol{\theta}})||_F \leq K||\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}||_2$$

where $J(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{|\boldsymbol{X}| \times n}$

- In essence, one needs to control the norm of the Jacobian and show the Lipschitzness

Second Step

Second Step

Time-Invariance for ResNet

Denote the training loss by $L(\boldsymbol{\theta}_t) = ||\boldsymbol{f}(\boldsymbol{\theta}_t) - \boldsymbol{y}||_2^2$ and by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\text{res}}(\boldsymbol{X}, \boldsymbol{X})$ the kernel evaluated on the training set. Fix a learning rate $\eta_0 \leq 2(\lambda_{\min}(\boldsymbol{\Theta}) + \lambda_{\max}(\boldsymbol{\Theta}))^{-1}$. Let $n > C^2$. Then with high probability there exists $R_0, K > 0$ such that:

Second Step

Time-Invariance for ResNet

Denote the training loss by $L(\boldsymbol{\theta}_t) = ||\boldsymbol{f}(\boldsymbol{\theta}_t) - \boldsymbol{y}||_2^2$ and by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\text{res}}(\boldsymbol{X}, \boldsymbol{X})$ the kernel evaluated on the training set. Fix a learning rate $\eta_0 \leq 2(\lambda_{\text{min}}(\boldsymbol{\Theta}) + \lambda_{\text{max}}(\boldsymbol{\Theta}))^{-1}$. Let $n > C^2$. Then with high probability there exists $R_0, K > 0$ such that:

- $L(\boldsymbol{\theta}_t) \leq \left(1 - \frac{\eta_0}{3}\lambda_{\text{min}}(\boldsymbol{\Theta})\right)^t R_0$ (Zero Training Loss)

Second Step

Time-Invariance for ResNet

Denote the training loss by $L(\boldsymbol{\theta}_t) = ||\boldsymbol{f}(\boldsymbol{\theta}_t) - \boldsymbol{y}||_2^2$ and by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\text{res}}(\boldsymbol{X}, \boldsymbol{X})$ the kernel evaluated on the training set. Fix a learning rate $\eta_0 \leq 2(\lambda_{\min}(\boldsymbol{\Theta}) + \lambda_{\max}(\boldsymbol{\Theta}))^{-1}$. Let $n > C^2$. Then with high probability there exists $R_0, K > 0$ such that:

- $L(\boldsymbol{\theta}_t) \leq \left(1 - \frac{\eta_0}{3}\lambda_{\min}(\boldsymbol{\Theta})\right)^t R_0$ (Zero Training Loss)
- $\sum_{j=1}^{t} ||\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}||_2 \leq \frac{3KR_0}{\lambda_{\min}(\boldsymbol{\Theta})}$ (Weight stability)

Second Step

Time-Invariance for ResNet

Denote the training loss by $L(\boldsymbol{\theta}_t) = ||\boldsymbol{f}(\boldsymbol{\theta}_t) - \boldsymbol{y}||_2^2$ and by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\text{res}}(\boldsymbol{X}, \boldsymbol{X})$ the kernel evaluated on the training set. Fix a learning rate $\eta_0 \leq 2(\lambda_{\min}(\boldsymbol{\Theta}) + \lambda_{\max}(\boldsymbol{\Theta}))^{-1}$. Let $n > C^2$. Then with high probability there exists $R_0, K > 0$ such that:

- $L(\boldsymbol{\theta}_t) \leq \left(1 - \frac{\eta_0}{3}\lambda_{\min}(\boldsymbol{\Theta})\right)^t R_0$ (Zero Training Loss)
- $\sum_{j=1}^t ||\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}||_2 \leq \frac{3KR_0}{\lambda_{\min}(\boldsymbol{\Theta})}$ (Weight stability)
- $\sup_{t>0} ||\hat{\Theta}_t - \hat{\Theta}_0||_F = \frac{6K^3R_0}{\lambda_{\min}(\boldsymbol{\Theta})} n^{-\frac{1}{2}}$ (NTK stability)

Second Step

Time-Invariance for ResNet

Denote the training loss by $L(\boldsymbol{\theta}_t) = ||\boldsymbol{f}(\boldsymbol{\theta}_t) - \boldsymbol{y}||_2^2$ and by $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\text{res}}(\boldsymbol{X}, \boldsymbol{X})$ the kernel evaluated on the training set. Fix a learning rate $\eta_0 \leq 2(\lambda_{\min}(\boldsymbol{\Theta}) + \lambda_{\max}(\boldsymbol{\Theta}))^{-1}$. Let $n > C^2$. Then with high probability there exists $R_0, K > 0$ such that:

- $L(\boldsymbol{\theta}_t) \leq \left(1 - \frac{\eta_0}{3}\lambda_{\min}(\boldsymbol{\Theta})\right)^t R_0$ (Zero Training Loss)
- $\sum_{j=1}^{t} ||\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j-1}||_2 \leq \frac{3KR_0}{\lambda_{\min}(\boldsymbol{\Theta})}$ (Weight stability)
- $\sup_{t>0} ||\hat{\Theta}_t - \hat{\Theta}_0||_F = \frac{6K^3R_0}{\lambda_{\min}(\boldsymbol{\Theta})} n^{-\frac{1}{2}}$ (NTK stability)

Follows from Lemma **without** any architecture specific arguments!

Smoothness of Interpolant (I)

Smoothness of Interpolant (I)

- Let's look at the derivative of the **input-output** map:

Smoothness of Interpolant (I)

Reflects smoothness of f_{res} to some degree

- Let's look at the derivative of the **input-output** map:

$$\frac{\partial f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{x}^{(L)}} \left(\prod_{l=1}^{L} \frac{\partial \boldsymbol{x}^{(l)}}{\partial \boldsymbol{x}^{(l-1)}} \right) \frac{\partial \boldsymbol{x}^{(0)}}{\partial \boldsymbol{x}}$$

Smoothness of Interpolant (I)

- Let's look at the derivative of the **input-output** map:

$$\frac{\partial f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{x}^{(L)}} \left(\prod_{l=1}^{L} \frac{\partial \boldsymbol{x}^{(l)}}{\partial \boldsymbol{x}^{(l-1)}} \right) \frac{\partial \boldsymbol{x}^{(0)}}{\partial \boldsymbol{x}}$$

$$\propto \left(\boldsymbol{w}^{(L+1)} \right)^T \left(\prod_{l=1}^{L} \left(\mathbb{1} + \frac{\alpha \sigma_v}{n} \boldsymbol{V}^{(l)} diag \left(\phi' \left(\boldsymbol{g}^{(l)} \right) \right) \boldsymbol{W}^{(l)} \right) \right) \frac{\boldsymbol{U}}{\sqrt{nd}}$$

Smoothness of Interpolant (I)

- Let's look at the derivative of the **input-output** map:

$$\frac{\partial f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial f}{\partial \boldsymbol{x}^{(L)}} \left(\prod_{l=1}^{L} \frac{\partial \boldsymbol{x}^{(l)}}{\partial \boldsymbol{x}^{(l-1)}} \right) \frac{\partial \boldsymbol{x}^{(0)}}{\partial \boldsymbol{x}}$$

$$\propto \left(\boldsymbol{w}^{(L+1)} \right)^{T} \left(\prod_{l=1}^{L} \left(\mathbb{1} + \frac{\alpha \sigma_v}{n} \boldsymbol{V}^{(l)} diag \left(\phi' \left(\boldsymbol{g}^{(l)} \right) \right) \boldsymbol{W}^{(l)} \right) \right) \frac{\boldsymbol{U}}{\sqrt{nd}}$$

- Bounding the norm of the derivative leads to

$$\left|\left| \frac{f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right|\right|_2 \leq \frac{||\boldsymbol{w}^{(L+1)}||_2}{\sqrt{nd}} ||\boldsymbol{U}||_2 \prod_{l=1}^{L} \left(1 + \frac{C_\phi \alpha \sigma_v}{n} ||\boldsymbol{V}^{(l)}||_2 ||\boldsymbol{W}^{(l)}||_2 \right)$$

Singular Values of Gaussian Random Matrices

Singular Values of Gaussian Random Matrices

Lemma: *Take $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ with $W_{ij} \sim \mathcal{N}(0,1)$. Then with probability $\geq 1 - 2\exp(-\frac{t^2}{2})$ it holds:*

$$\sqrt{m} - \sqrt{n} - t \leq \sigma_{min}(\boldsymbol{W}) \leq \sigma_{max}(\boldsymbol{W}) \leq \sqrt{m} + \sqrt{n} + t$$

Singular Values of Gaussian Random Matrices

Lemma: *Take* $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ *with* $W_{ij} \sim \mathcal{N}(0,1)$. *Then with probability* $\geq 1 - 2\exp(-\frac{t^2}{2})$ *it holds:*

$$\sqrt{m} - \sqrt{n} - t \leq \sigma_{min}(\boldsymbol{W}) \leq \sigma_{max}(\boldsymbol{W}) \leq \sqrt{m} + \sqrt{n} + t$$

- Let $C = \frac{3KR_0}{\lambda_{\min}(\boldsymbol{\Theta})}$. It holds (in NTK regime) that $\boldsymbol{\theta}_t \in B(\boldsymbol{\theta}_0, C)$. We can hence arrive at

$$\begin{aligned} ||\boldsymbol{W}_t^{(l)}||_2 &\leq ||\boldsymbol{W}_0^{(l)}||_2 + ||\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}||_2 \\ &\leq 2\sqrt{n} + t + C \leq 3\sqrt{n} \end{aligned}$$

Singular Values of Gaussian Random Matrices

Lemma: *Take* $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ *with* $W_{ij} \sim \mathcal{N}(0,1)$. *Then with probability* $\geq 1 - 2\exp(-\frac{t^2}{2})$ *it holds:*

$$\sqrt{m} - \sqrt{n} - t \leq \sigma_{min}(\boldsymbol{W}) \leq \sigma_{max}(\boldsymbol{W}) \leq \sqrt{m} + \sqrt{n} + t$$

- Let $C = \frac{3KR_0}{\lambda_{\min}(\boldsymbol{\Theta})}$. It holds (in NTK regime) that $\boldsymbol{\theta}_t \in B(\boldsymbol{\theta}_0, C)$. We can hence arrive at

$$\begin{aligned} ||\boldsymbol{W}_t^{(l)}||_2 &\leq ||\boldsymbol{W}_0^{(l)}||_2 + ||\boldsymbol{W}_t^{(l)} - \boldsymbol{W}_0^{(l)}||_2 \\ &\leq 2\sqrt{n} + t + C \leq 3\sqrt{n} \end{aligned}$$

- Similary we find $||\boldsymbol{U}||_2 \leq \sqrt{d} + 2\sqrt{n}$ and $||\boldsymbol{V}^{(l)}||_2 \leq 3\sqrt{n}$

Smoothness of Interpolant (II)

Smoothness of Interpolant (II)

- We can combine these bounds:

Smoothness of Interpolant (II)

- We can combine these bounds:

$$\left|\left|\frac{f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|\right|_2 \leq \frac{||\boldsymbol{w}^{(L+1)}||_2}{\sqrt{nd}} ||\boldsymbol{U}||_2 \prod_{l=1}^{L} \left(1 + \frac{C_\phi \alpha \sigma_v}{n} ||\boldsymbol{V}^{(l)}||_2 ||\boldsymbol{W}^{(l)}||_2\right)$$

Smoothness of Interpolant (II)

- We can combine these bounds:

$$
\begin{aligned}
\left|\left|\frac{f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|\right|_2 &\leq \frac{||\boldsymbol{w}^{(L+1)}||_2}{\sqrt{nd}}||\boldsymbol{U}||_2 \prod_{l=1}^{L}\left(1 + \frac{C_\phi \alpha \sigma_v}{n}||\boldsymbol{V}^{(l)}||_2||\boldsymbol{W}^{(l)}||_2\right) \\
&\leq 2\frac{\sqrt{n}}{\sqrt{nd}}(\sqrt{d} + 2\sqrt{n})\prod_{l=1}^{L}\left(1 + \frac{C_\phi \alpha \sigma_v}{n}3\sqrt{n}\sqrt{n}\right)
\end{aligned}
$$

Smoothness of Interpolant (II)

- We can combine these bounds:

$$
\begin{aligned}
\left|\left|\frac{f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|\right|_2 &\leq \frac{||\boldsymbol{w}^{(L+1)}||_2}{\sqrt{nd}} ||\boldsymbol{U}||_2 \prod_{l=1}^{L} \left(1 + \frac{C_\phi \alpha \sigma_v}{n} ||\boldsymbol{V}^{(l)}||_2 ||\boldsymbol{W}^{(l)}||_2\right) \\
&\leq 2\frac{\sqrt{n}}{\sqrt{nd}}(\sqrt{d} + 2\sqrt{n}) \prod_{l=1}^{L} \left(1 + \frac{C_\phi \alpha \sigma_v}{n} 3\sqrt{n}\sqrt{n}\right) \\
&\leq 2(1 + 2\sqrt{\frac{n}{d}})(1 + 9\alpha C_\phi)^L = B_{\text{res}}
\end{aligned}
$$

Smoothness of Interpolant (II)

- We can combine these bounds:

$$\left|\left|\frac{f_{\text{res}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|\right|_2 \leq \frac{||\boldsymbol{w}^{(L+1)}||_2}{\sqrt{nd}}||\boldsymbol{U}||_2 \prod_{l=1}^{L}\left(1 + \frac{C_\phi \alpha \sigma_v}{n}||\boldsymbol{V}^{(l)}||_2||\boldsymbol{W}^{(l)}||_2\right)$$

$$\leq 2\frac{\sqrt{n}}{\sqrt{nd}}(\sqrt{d} + 2\sqrt{n})\prod_{l=1}^{L}\left(1 + \frac{C_\phi \alpha \sigma_v}{n}3\sqrt{n}\sqrt{n}\right)$$

$$\leq 2(1 + 2\sqrt{\frac{n}{d}})(1 + 9\alpha C_\phi)^L = B_{\text{res}}$$

- **Observe:** Bound decreases with α. Better empirical performance has been observed for smaller α!

Smoothness of Interpolant MLP

Smoothness of Interpolant MLP

- We can make a similar derivation for the MLP:

$$\left\| \frac{\partial f_{\mathsf{MLP}}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\|_2 \leq 2C_\phi \left(1 + 2\sqrt{\frac{n}{d}}\right) (3C_\phi)^{L-1} = B_{\mathsf{MLP}}$$

Smoothness of Interpolant MLP

- We can make a similar derivation for the MLP:

$$\left\|\frac{\partial f_{\mathrm{MLP}}(\boldsymbol{x})}{\partial \boldsymbol{x}}\right\|_2 \leq 2C_\phi \left(1 + 2\sqrt{\frac{n}{d}}\right) (3C_\phi)^{L-1} = B_{\mathrm{MLP}}$$

- Comparing the two bounds shows:

$$\frac{B_{\mathrm{res}}}{B_{\mathrm{MLP}}} = \frac{(1+9\alpha)^L}{3^{L-1}} \leq 1 \iff \alpha \leq \frac{3^{1-L^{-1}} - 1}{9}$$

Smoothness of Interpolant MLP

- We can make a similar derivation for the MLP:

$$\left\| \frac{\partial f_{\mathsf{MLP}}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\|_2 \leq 2C_\phi \left(1 + 2\sqrt{\frac{n}{d}}\right) (3C_\phi)^{L-1} = B_{\mathsf{MLP}}$$

- Comparing the two bounds shows:

$$\frac{B_{\mathsf{res}}}{B_{\mathsf{MLP}}} = \frac{(1+9\alpha)^L}{3^{L-1}} \leq 1 \iff \alpha \leq \frac{3^{1-L^{-1}} - 1}{9}$$

- We can hence obtain **smoother** interpolations by using $\alpha = 0.1$ for any depth. Apparently values on this order also work best **in practice**!

Smoothness of Interpolant (III)

Smoothness of Interpolant (III)

- **Fact:** Gaussian kernel promotes very smooth functions.

Smoothness of Interpolant (III)

- **Fact:** Gaussian kernel promotes very smooth functions.

- The induced RKHS norm $||\cdot||_{\mathcal{H}_{\text{Gauss}}}$ should hence be a good **measure** for smoothness:

$$\mu(f) = \frac{||f_{\text{Gauss}}||_{\mathcal{H}_{\text{Gauss}}}}{||f||_{\mathcal{H}_{\text{Gauss}}}}$$

  where f_{Gauss} is the fit with the Gaussian kernel and f is the NTK fit.

Smoothness of Interpolant (III)

- **Fact:** Gaussian kernel promotes very smooth functions.
- The induced RKHS norm $||\cdot||_{\mathcal{H}_{\text{Gauss}}}$ should hence be a good **measure** for smoothness:

$$\mu(f) = \frac{||f_{\text{Gauss}}||_{\mathcal{H}_{\text{Gauss}}}}{||f||_{\mathcal{H}_{\text{Gauss}}}}$$

  where f_{Gauss} is the fit with the Gaussian kernel and f is the NTK fit.

- If $\mu(f_1) < \mu(f_2)$ will (in some sense) be interpreted that f_2 is smoother than f_1. It is also scale-invariant.

Smoothness of Interpolant (III)

- **Fact:** Gaussian kernel promotes very smooth functions.

- The induced RKHS norm $||\cdot||_{\mathcal{H}_{\text{Gauss}}}$ should hence be a good **measure** for smoothness:

$$\mu(f) = \frac{||f_{\text{Gauss}}||_{\mathcal{H}_{\text{Gauss}}}}{||f||_{\mathcal{H}_{\text{Gauss}}}}$$

  where f_{Gauss} is the fit with the Gaussian kernel and f is the NTK fit.

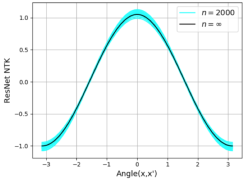
- If $\mu(f_1) < \mu(f_2)$ will (in some sense) be interpreted that f_2 is smoother than f_1. It is also scale-invariant.

- Calculate norm as

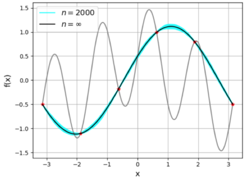
Fourier transform

$$||f||_{\text{Gauss}} = \frac{1}{(2\phi)^{\frac{1}{d}}} \int \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[K_{\text{Gauss}}](\omega)} d\omega$$

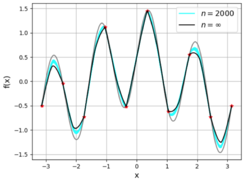
Experiments (I)



(a) Empirical and asymptotic NTK



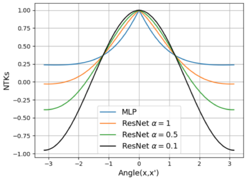
(b) Interpolation with 6 samples



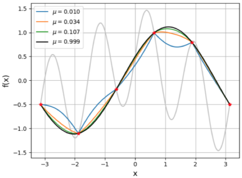
(c) Interpolation with 10 samples

Finite (black) versus infinite width ResNet (blue). In a) the two kernels are compared, in b) and c) the predictive functions with different learning rates. Empirical results perfectly **agree** with the predictions of theory.

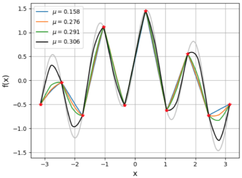
Experiments (II)



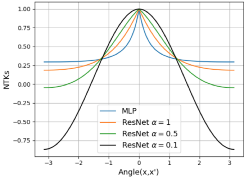
(a) NTKs (normalized to unit peak)
$L = 5$



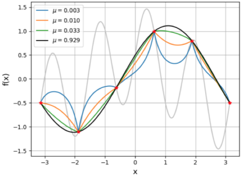
(b) Interpolation with 6 samples
$L = 5$



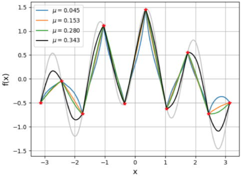
(c) Interpolation with 10 samples
$L = 5$



(d) NTKs (normalized to unit peak)
$L = 15$



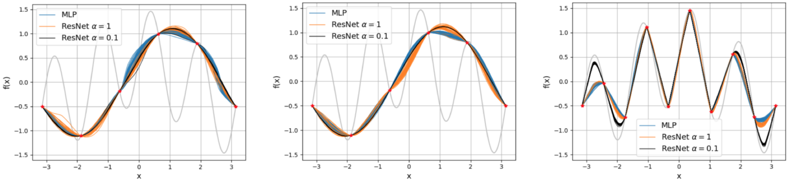
(e) Interpolation with 6 samples
$L = 15$



(f) Interpolation with 10 samples
$L = 15$

MLP (black) and Resnets for different α's and different number of layers. Visually, fits become smoother with **smaller** α. μ-measure agrees with that observation.

Experiments (III)



Same visualizations **out of** the NTK regime (lazy regime), meaning **smaller** width $n = 500$, **Xavier** initialization (no $\frac{1}{\sqrt{n}}$) and different (**adaptive**) optimizers. Same observations roughly hold in this regime, ResNet is smoother than MLP and decreasing α promotes smoothness.

Discussion

- Notions of smoothness seem interesting and the bounds are rather easy. Maybe one can do something in the spirit of **norm-based generalization bounds** but for trained networks instead of all networks (avoiding the union bound because we have roughly a bound on Lipschitz constant of trained network)
- As always in the NTK works, experiments are rather **unsatisfying**...
- They test their theory on 2d data from circle with 10 (**???**) samples
- Nevertheless, smoothness of induced predictive function could be interesting to compare between NNGP and NTK as well

Discussion

- In general very cool if NTK can reflect properties encountered in the **empirical** world

- Notions of smoothness seem interesting and the bounds are rather easy. Maybe one can do something in the spirit of **norm-based generalization bounds** but for trained networks instead of all networks (avoiding the union bound because we have roughly a bound on Lipschitz constant of trained network)
- As always in the NTK works, experiments are rather **unsatisfying**...

- They test their theory on 2d data from circle with 10 (**???**) samples

- Nevertheless, smoothness of induced predictive function could be interesting to compare between NNGP and NTK as well