On the linearity of large non-linear models: when and why the tangent kernel is constant

Chaoyue Liu, Libin Zhu, Mikhail Belkin,

What's it about?

What's it about?

- NeurIPS paper this year

What's it about?

- NeurIPS paper this year
- Revisiting of the proof that neural tangent kernel remains **constant** during training

What's it about?

- NeurIPS paper this year
- Revisiting of the proof that neural tangent kernel remains **constant** during training
- Conceptually **easier** reasoning when and why NTK is constant

What’s it about?

- NeurIPS paper this year
- Revisiting of the proof that neural tangent kernel remains **constant** during training
- Conceptually **easier** reasoning when and why NTK is constant
- **Lazy training** is not the origin of constancy, it’s a consequence of the architecture!

Setup

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}}\boldsymbol{W}^{(l)}\boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}}\boldsymbol{W}^{(l)}\boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{m_{l-1}\times m_l}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ some **non-linearity**

Setup

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}}\boldsymbol{W}^{(l)}\boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{m_{l-1}\times m_l}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ some **non-linearity**
- The final output is given by

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{m}}\boldsymbol{v}^T\boldsymbol{\alpha}^{(L)}(\boldsymbol{x})$$

  where $\boldsymbol{v} \in \mathbb{R}^m$ is **not** trainaible

Initialization and Conditions

Initialization and Conditions

Initialize all the parameters in the network as follows:

Initialization and Conditions

Initialize all the parameters in the network as follows:

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

Initialization and Conditions

Initialize all the parameters in the network as follows:

- $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $v_i \sim \mathcal{U}(\{-1,1\})$

Initialization and Conditions

Initialize all the parameters in the network as follows:

- $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $v_i \sim \mathcal{U}(\{-1,1\})$

I guess could be any almost surely **bounded** initialization for $\boldsymbol{v}$ as we can see later.

Initialization and Conditions

Initialize all the parameters in the network as follows:

- $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $v_i \sim \mathcal{U}(\{-1,1\})$

I guess could be any almost surely **bounded** initialization for $\boldsymbol{v}$ as we can see later.

Conditions on non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$:

Initialization and Conditions

Initialize all the parameters in the network as follows:

- $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$
- $v_i \sim \mathcal{U}(\{-1, 1\})$

I guess could be any almost surely **bounded** initialization for $\boldsymbol{v}$ as we can see later.

Conditions on non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$:

- β-smooth: $||\sigma(\boldsymbol{w}) - \sigma(\boldsymbol{v}) - \nabla\sigma(\boldsymbol{v})(\boldsymbol{w} - \boldsymbol{v})||_2 \leq \frac{\beta}{2}||\boldsymbol{w} - \boldsymbol{v}||_2^2$

Initialization and Conditions

Initialize all the parameters in the network as follows:

- $W_{ij}^{(l)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- $v_i \sim \mathcal{U}(\{-1,1\})$

I guess could be any almost surely **bounded** initialization for $\boldsymbol{v}$ as we can see later.

Conditions on non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$:

- β-smooth: $||\sigma(\boldsymbol{w}) - \sigma(\boldsymbol{v}) - \nabla\sigma(\boldsymbol{v})(\boldsymbol{w} - \boldsymbol{v})||_2 \leq \frac{\beta}{2}||\boldsymbol{w} - \boldsymbol{v}||_2^2$
- Lipschitz: $||\sigma(\boldsymbol{w}) - \sigma(\boldsymbol{v})||_2 \leq \gamma||\boldsymbol{v} - \boldsymbol{w}||_2$

Notations

Notations

- We will mostly only deal with the "empirical" NTK (meaning finite width):

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x})\right)^T \nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x}')$$

Notations

- We will mostly only deal with the "empirical" NTK (meaning finite width):

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x})\right)^T \nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x}')$$

- Will view this kernel mostly as a function of $\boldsymbol{w}$

Notations

- We will mostly only deal with the "empirical" NTK (meaning finite width):

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') = \left(\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x})\right)^T \nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x}')$$

- Will view this kernel mostly as a function of $\boldsymbol{w}$

- Of particular interest are diagonal entries:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2$$

Observation I

Observation I

A rather "intuitive" but very important fact is the following:

Observation I

A rather "intuitive" but very important fact is the following:

Constancy $\iff$ Linear

The tangent kernel $\boldsymbol{w} \mapsto K_{\boldsymbol{w}}$ of a differentiable model $f(\boldsymbol{w}; \boldsymbol{x})$ is **constant** if and only if $f(\boldsymbol{w}; \boldsymbol{x})$ is **linear** in $\boldsymbol{w}$.

Observation I

A rather "intuitive" but very important fact is the following:

Constancy $\iff$ Linear

The tangent kernel $\boldsymbol{w} \mapsto K_{\boldsymbol{w}}$ of a differentiable model $f(\boldsymbol{w}; \boldsymbol{x})$ is **constant** if and only if $f(\boldsymbol{w}; \boldsymbol{x})$ is **linear** in $\boldsymbol{w}$.

Proof is surprisingly long but uses very elementary tools from calculus.

Observation II

Observation II

A new proof strategy for the constancy of NTK:

Observation II

A new proof strategy for the constancy of NTK:

- **Originally:** $\overset{\text{Jacot}}{\Longrightarrow}$ $K_{\boldsymbol{w}}$ constant $\overset{\text{Lee et al.}}{\Longrightarrow}$ $f(\boldsymbol{w};\boldsymbol{x})$ linear

Observation II

A new proof strategy for the constancy of NTK:

- **Originally:** $\overset{\text{Jacot}}{\Longrightarrow}$ $K_{\boldsymbol{w}}$ constant $\overset{\text{Lee et al.}}{\Longrightarrow}$ $f(\boldsymbol{w};\boldsymbol{x})$ linear
- **Now:** $\Longrightarrow$ $f(\boldsymbol{w},\boldsymbol{x})$ linear $\Longrightarrow$ $K_{\boldsymbol{w}}$ constant

Observation II

A new proof strategy for the constancy of NTK:

- **Originally:** $\overset{\text{Jacot}}{\Longrightarrow}$ $K_{\boldsymbol{w}}$ constant $\overset{\text{Lee et al.}}{\Longrightarrow}$ $f(\boldsymbol{w}; \boldsymbol{x})$ linear
- **Now:** $\Longrightarrow$ $f(\boldsymbol{w}, \boldsymbol{x})$ linear $\Longrightarrow$ $K_{\boldsymbol{w}}$ constant

Let us hence first show that $f(\boldsymbol{w}; \boldsymbol{x})$ becomes linear.

Observation III

Observation III

We have the following well-known characterization of linearity:

Observation III

We have the following well-known characterization of linearity:

Linearity vs Hessian

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$,$\boldsymbol{w} \mapsto f(\boldsymbol{w})$ is linear if and only if the Hessian $\boldsymbol{H}_f = \frac{\partial f}{\partial \boldsymbol{w}^2}$ vanishes $\forall \boldsymbol{w} \in \mathbb{R}^d$.

Observation III

We have the following well-known characterization of linearity:

Linearity vs Hessian

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$,$\boldsymbol{w} \mapsto f(\boldsymbol{w})$ is linear if and only if the Hessian $\boldsymbol{H}_f = \frac{\partial f}{\partial \boldsymbol{w}^2}$ vanishes $\forall \boldsymbol{w} \in \mathbb{R}^d$.

It hence suffices to show that $\boldsymbol{H}_f = 0$ for $m \to \infty$ or equivalently $||\boldsymbol{H}_f||_2 = 0$

Observation IV

Observation IV

We can convert above asymptotic facts to the **non-asymptotic** (finite-width) setting:

Observation IV

We can convert above asymptotic facts to the **non-asymptotic** (finite-width) setting:

Small Hessian $\implies$ Small change in $K_{\boldsymbol{w}}$

Observation IV

We can convert above asymptotic facts to the **non-asymptotic** (finite-width) setting:

Small Hessian $\implies$ Small change in $K_{\boldsymbol{w}}$

Fix $\boldsymbol{w}_0 \in \mathbb{R}^d$ and $R > 0$ and consider $\mathcal{B}(\boldsymbol{w}_0, R)$. If it holds that $||\boldsymbol{H}_f(\boldsymbol{w})||_2 < \epsilon \; \forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$, then we have that

Observation IV

We can convert above asymptotic facts to the **non-asymptotic** (finite-width) setting:

Small Hessian $\implies$ Small change in $K_{\boldsymbol{w}}$

Fix $\boldsymbol{w}_0 \in \mathbb{R}^d$ and $R > 0$ and consider $\mathcal{B}(\boldsymbol{w}_0, R)$. If it holds that $||\boldsymbol{H}_f(\boldsymbol{w})||_2 < \epsilon \ \forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$, then we have that

$$\left| K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') - K_{\boldsymbol{w}_0}(\boldsymbol{x}, \boldsymbol{x}') \right| = \mathcal{O}(R\epsilon)$$

$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{m_0}$ and $\forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$

Observation IV

We can convert above asymptotic facts to the **non-asymptotic** (finite-width) setting:

Small Hessian $\implies$ Small change in $K_{\boldsymbol{w}}$

Fix $\boldsymbol{w}_0 \in \mathbb{R}^d$ and $R > 0$ and consider $\mathcal{B}(\boldsymbol{w}_0, R)$. If it holds that $||\boldsymbol{H}_f(\boldsymbol{w})||_2 < \epsilon \; \forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$, then we have that

$$\left| K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') - K_{\boldsymbol{w}_0}(\boldsymbol{x}, \boldsymbol{x}') \right| = \mathcal{O}(R\epsilon)$$

$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^{m_0}$ and $\forall \boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$

Hence if $\epsilon = \epsilon(m) \xrightarrow{m \to \infty} 0$ then also

$$\left| K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') - K_{\boldsymbol{w}_0}(\boldsymbol{x}, \boldsymbol{x}') \right| \xrightarrow{m \to \infty} 0$$

Observation V

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$

- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$
- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Exploration of GD

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$
- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Exploration of GD

Take NN $f(\boldsymbol{w}; \boldsymbol{x})$ and consider β-smooth and $\mu - PL^*$ loss $L(\boldsymbol{w})$. Fix $R = \frac{2\sqrt{2\beta L(\boldsymbol{w}_0)}}{\mu}$. Then it holds that:

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$
- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Exploration of GD

Take NN $f(\boldsymbol{w}; \boldsymbol{x})$ and consider β-smooth and $\mu - PL^*$ loss $L(\boldsymbol{w})$. Fix $R = \frac{2\sqrt{2\beta L(\boldsymbol{w}_0)}}{\mu}$. Then it holds that:

- $\exists \boldsymbol{w}^* \in \mathcal{B}(\boldsymbol{w}_0, R)$ such that $L(\boldsymbol{w}^*) = 0$

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$
- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Exploration of GD

Take NN $f(\boldsymbol{w}; \boldsymbol{x})$ and consider β-smooth and $\mu - PL^*$ loss $L(\boldsymbol{w})$. Fix $R = \frac{2\sqrt{2\beta L(\boldsymbol{w}_0)}}{\mu}$. Then it holds that:

- $\exists \boldsymbol{w}^* \in \mathcal{B}(\boldsymbol{w}_0, R)$ such that $L(\boldsymbol{w}^*) = 0$
- GD with $\eta = \frac{1}{\sup_{\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)}} \lambda_{\max}(\boldsymbol{H}_L)$ converges in $\mathcal{B}(\boldsymbol{w}_0, R)$

Observation V

- **Notice:** We only have a valid result in $\mathcal{B}(\boldsymbol{w}_0, R)$
- We don't want constancy $\forall \boldsymbol{w} \in \mathbb{R}^d$, we only want it along the **training trajectory** of GD!

Exploration of GD

Take NN $f(\boldsymbol{w}; \boldsymbol{x})$ and consider β-smooth and $\mu - PL^*$ loss $L(\boldsymbol{w})$. Fix $R = \frac{2\sqrt{2\beta L(\boldsymbol{w}_0)}}{\mu}$. Then it holds that:

- $\exists \boldsymbol{w}^* \in \mathcal{B}(\boldsymbol{w}_0, R)$ such that $L(\boldsymbol{w}^*) = 0$
- GD with $\eta = \frac{1}{\sup_{\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)}} \lambda_{\max}(\boldsymbol{H}_L)$ converges in $\mathcal{B}(\boldsymbol{w}_0, R)$

Thus: Fixing $R = \frac{2\sqrt{2\beta L(\boldsymbol{w}_0)}}{\mu}$ above (independent of m) **suffices**!

Warm-up: 1 hidden layer

Warm-up: 1 hidden layer

Consider 1 hidden layer network with one dimensional input $x \in \mathbb{R}$:

Warm-up: 1 hidden layer

Consider 1 hidden layer network with one dimensional input $x \in \mathbb{R}$:

- $f(\boldsymbol{w}; x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \alpha_i(x)$

Warm-up: 1 hidden layer

Consider 1 hidden layer network with one dimensional input $x \in \mathbb{R}$:

- $f(\boldsymbol{w}; x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} v_i \alpha_i(x)$
- $\alpha_i(x) = \sigma(w_i x)$

where $\boldsymbol{w} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^m$.

Hessian H_f

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j}$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x)$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$

$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H})$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H}) = \max_{j=1}^{m} |H_{jj}|$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H}) = \max_{j=1}^{m} |H_{jj}| = \max_{j=1}^{m} \frac{1}{\sqrt{m}} x^2 |v_j| |\sigma''(w_j x)|$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$

$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H}) = \max_{j=1}^{m} |H_{jj}| = \max_{j=1}^{m} \frac{1}{\sqrt{m}} x^2 |v_j| |\sigma''(w_j x)|$$

$$\leq \frac{1}{\sqrt{m}} C^2 \beta_\sigma$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H}) = \max_{j=1}^{m} |H_{jj}| = \max_{j=1}^{m} \frac{1}{\sqrt{m}} x^2 |v_j| |\sigma''(w_j x)|$$
$$\leq \frac{1}{\sqrt{m}} C^2 \beta_\sigma = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$$

Hessian H_f

Let's calculate the **Hessian** of $f(\boldsymbol{w}; x)$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} v_j \sigma'(w_j x) x$$
$$= \frac{1}{\sqrt{m}} v_j x^2 \sigma''(w_j x) \mathbb{1}_{\{i=j\}}$$

Since $\boldsymbol{H}$ is diagonal:

$$||\boldsymbol{H}||_2 = \sigma_{\max}(\boldsymbol{H}) = \max_{j=1}^{m} |H_{jj}| = \max_{j=1}^{m} \frac{1}{\sqrt{m}} x^2 |v_j| |\sigma''(w_j x)|$$
$$\leq \frac{1}{\sqrt{m}} C^2 \beta_\sigma = \mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$$

Thus: $||\boldsymbol{H}||_2 \xrightarrow{m \to \infty} 0$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x)$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{1}{\sqrt{m}} v_i \sigma'(w_i x) x$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{1}{\sqrt{m}} v_i \sigma'(w_i x) x$$

Thus:

$$||\nabla f(\boldsymbol{w}, x)||_2^2 = \frac{x^2}{m} \sum_{i=1}^{m} v_i^2 \sigma'(w_i x)$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{1}{\sqrt{m}} v_i \sigma'(w_i x) x$$

Thus:

$$||\nabla f(\boldsymbol{w}, x)||_2^2 = \frac{x^2}{m} \sum_{i=1}^{m} v_i^2 \sigma'(w_i x) = \frac{x^2}{m} \sum_{i=1}^{m} \sigma'(w_i x)$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{1}{\sqrt{m}} v_i \sigma'(w_i x) x$$

Thus:

$$||\nabla f(\boldsymbol{w}, x)||_2^2 = \frac{x^2}{m} \sum_{i=1}^{m} v_i^2 \sigma'(w_i x) = \frac{x^2}{m} \sum_{i=1}^{m} \sigma'(w_i x) = \Theta(1)$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i}\frac{1}{\sqrt{m}}\sum_{k=1}^{m} v_k\sigma(w_k x) = \frac{1}{\sqrt{m}} v_i\sigma'(w_i x)x$$

Thus:

$$||\nabla f(\boldsymbol{w}, x)||_2^2 = \frac{x^2}{m}\sum_{i=1}^{m} v_i^2\sigma'(w_i x) = \frac{x^2}{m}\sum_{i=1}^{m}\sigma'(w_i x) = \Theta(1)$$

Moreover:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = x^2\frac{1}{m}\sum_{i=1}^{m}\sigma'(w_i x)$$

Diagonal Kernel Entry $K_{\boldsymbol{w}}(x, x)$

On the other hand we have that:

$$\frac{\partial f}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{\sqrt{m}} \sum_{k=1}^{m} v_k \sigma(w_k x) = \frac{1}{\sqrt{m}} v_i \sigma'(w_i x) x$$

Thus:

$$||\nabla f(\boldsymbol{w}, x)||_2^2 = \frac{x^2}{m} \sum_{i=1}^{m} v_i^2 \sigma'(w_i x) = \frac{x^2}{m} \sum_{i=1}^{m} \sigma'(w_i x) = \Theta(1)$$

Moreover:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = x^2 \frac{1}{m} \sum_{i=1}^{m} \sigma'(w_i x) \xrightarrow{m \to \infty} x^2 \mathbb{E}_{w \sim \mathcal{N}(0,1)}[\sigma'(w_i x)]$$

Why Different Scalings?

Why Different Scalings?

- **Question:** How come that different scalings in m arise in **Hessian** norm and **gradient** norm??

Why Different Scalings?

- **Question:** How come that different scalings in m arise in **Hessian** norm and **gradient** norm??

- Let's study the order 3 tensor $\boldsymbol{T} = \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}^2}$

Why Different Scalings?

- **Question:** How come that different scalings in m arise in **Hessian** norm and **gradient** norm??

- Let's study the order 3 tensor $\boldsymbol{T} = \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}^2}$

- Equip it with the **norm**

$$||\boldsymbol{T}||_{2,2,1} = \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_k \Big| \sum_i \sum_j T_{ijk} a_i b_j \Big|$$

Why Different Scalings?

- **Question:** How come that different scalings in m arise in **Hessian** norm and **gradient** norm??

- Let's study the order 3 tensor $\boldsymbol{T} = \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}^2}$

- Equip it with the **norm**

$$||\boldsymbol{T}||_{2,2,1} = \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_k \Big| \sum_i \sum_j T_{ijk} a_i b_j \Big|$$

- We have the following easy consequence of Hölder inequality:

Why Different Scalings?

- **Question:** How come that different scalings in m arise in **Hessian** norm and **gradient** norm??
- Let's study the order 3 tensor $\boldsymbol{T} = \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}^2}$
- Equip it with the **norm**

$$||\boldsymbol{T}||_{2,2,1} = \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_k \left| \sum_i \sum_j T_{ijk} a_i b_j \right|$$

- We have the following easy consequence of Hölder inequality:

Given $\boldsymbol{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ and $\boldsymbol{v} \in \mathbb{R}^{d_3}$, consider $A_{ij} = \sum_k T_{ijk} v_k$.
It holds that

$$||\boldsymbol{A}||_2 \leq ||\boldsymbol{T}||_{2,2,1} ||\boldsymbol{v}||_\infty$$

Hessian and $|| \cdot ||_\infty$ (I)

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j}$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j}$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

$$||\boldsymbol{H}||_2 \leq ||\,\boldsymbol{T}\,|||_{2,2,1} \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

$$||\boldsymbol{H}||_2 \leq ||\boldsymbol{T}|||_{2,2,1} \Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_\infty$$

- Observe that

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

$$||\boldsymbol{H}||_2 \leq ||\,\boldsymbol{T}\,|||_{2,2,1} \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty$$

- Observe that

$$T_{ijk} = \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j}$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

$$||\boldsymbol{H}||_2 \leq ||\,\boldsymbol{T}\,|||_{2,2,1} \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty$$

- Observe that

$$T_{ijk} = \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \sigma(w_k x)$$

Hessian and $||\cdot||_\infty$ (I)

- This time we express $\boldsymbol{H}$ in terms of $\boldsymbol{T}$:

$$H_{ij} = \frac{\partial^2 f}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \sum_{k=1}^{m} \frac{\partial f}{\partial \alpha_k} T_{ijk}$$

- From previous statement:

$$||\boldsymbol{H}||_2 \leq ||\boldsymbol{T}|||_{2,2,1} \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty$$

- Observe that

$$T_{ijk} = \frac{\partial^2 \alpha_k}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \sigma(w_k x) = \sigma''(w_k x) x^2 \mathbb{1}_{\{i=j=k\}}$$

Hessian and $||\cdot||_\infty$ (II)

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \left| \sum_i \sum_j T_{ijl} a_i b_i \right|$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$

$$\leq x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i|$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$

$$\leq x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i| \leq x^2 L_\sigma$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$
\begin{aligned}
||\boldsymbol{T}||_{2,2,1} &= \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)| \\
&\le x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i| \le x^2 L_\sigma = \mathcal{O}(1)
\end{aligned}
$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$
$$\leq x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i| \leq x^2 L_\sigma = \mathcal{O}(1)$$

- This in turn implies that

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$
$$\leq x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i| \leq x^2 L_\sigma = \mathcal{O}(1)$$

- This in turn implies that

$$||\boldsymbol{H}||_2 = \mathcal{O}\left(\Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty \right)$$

Hessian and $||\cdot||_\infty$ (II)

- We can hence calculate the norm as:

$$||\boldsymbol{T}||_{2,2,1} = \sup \sum_k \Big| \sum_i \sum_j T_{ijl} a_i b_i \Big| = x^2 \sup \sum_{k=1}^{m} |\sigma''(w_k x)|$$

$$\leq x^2 L_\sigma \sup_{||\boldsymbol{a}||_2 = ||\boldsymbol{b}||_2 = 1} \sum_{k=1} |a_i b_i| \leq x^2 L_\sigma = \mathcal{O}(1)$$

- This in turn implies that

$$||\boldsymbol{H}||_2 = \mathcal{O}\left(\Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}} \Big|\Big|_\infty \right)$$

- The l_∞ - norm thus governs the scaling of the Hessian

Kernel and $||\cdot||_2$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\mathrm{max}}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\mathrm{max}}(\boldsymbol{K})$$

Kernel and $|| \cdot ||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n} \sum_{i=1}^{n} \lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n} \sum_{i=1}^{n} \lambda_i(\boldsymbol{K})$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$
\begin{aligned}
||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) &= \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K}) \\
&= \frac{1}{n}\,\mathrm{tr}(\boldsymbol{K})
\end{aligned}
$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2 \approx \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\Big|\Big|_2^2 \Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2$$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2 \approx \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\Big|\Big|_2^2 \Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j}$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2 \approx \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\right|\right|_2^2 \left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j} = \sigma'(w_i x)x\mathbb{1}_{\{i=j\}}$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2 \approx \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\Big|\Big|_2^2 \Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j} = \sigma'(w_i x)x\mathbb{1}_{\{i=j\}} \leq L_\sigma x$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n\times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}\lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n}\lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n}K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2 \approx \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\right|\right|_2^2 \left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j} = \sigma'(w_i x)x\mathbb{1}_{\{i=j\}} \leq L_\sigma x \implies \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\right|\right|_2^2$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}} \frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2 \approx \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\right|\right|_2^2 \left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j} = \sigma'(w_i x) x \mathbb{1}_{\{i=j\}} \leq L_\sigma x \implies \left|\left|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\right|\right|_2^2 = \mathcal{O}(1)$

Kernel and $||\cdot||_2$

- Observe the following for the kernel $\boldsymbol{K} \in \mathbb{R}^{n \times n}$:

$$||\boldsymbol{K}||_2 = \lambda_{\max}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_{\max}(\boldsymbol{K}) \geq \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\boldsymbol{K})$$
$$= \frac{1}{n}\operatorname{tr}(\boldsymbol{K}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{w}}(\boldsymbol{x}_i, \boldsymbol{x}_i)$$

- Consider thus a diagonal kernel entry:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = ||\nabla_{\boldsymbol{w}} f(\boldsymbol{w}; x)||_2^2 = \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2 \approx \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\Big|\Big|_2^2 \Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2$$

- **Moreover** $\frac{\partial \alpha_i}{\partial w_j} = \sigma'(w_i x) x \mathbb{1}_{\{i=j\}} \leq L_\sigma x \implies \Big|\Big|\frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{w}}\Big|\Big|_2^2 = \mathcal{O}(1)$

- **Hence:** $||\boldsymbol{K}||_2 = \mathcal{O}\left(\Big|\Big|\frac{\partial f}{\partial \boldsymbol{\alpha}}\Big|\Big|_2^2\right)$

$||\cdot||_\infty$ versus $||\cdot||_2$

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$$\frac{\partial f}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} v_j \alpha_j\right)$$

$|| \cdot ||_\infty$ versus $|| \cdot ||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$$\frac{\partial f}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} v_j \alpha_j\right) = \frac{1}{\sqrt{m}} v_i$$

$|| \cdot ||_\infty$ versus $|| \cdot ||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$$\frac{\partial f}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} v_j \alpha_j\right) = \frac{1}{\sqrt{m}} v_i$$

- **Thus:**

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$$\frac{\partial f}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} v_j \alpha_j\right) = \frac{1}{\sqrt{m}} v_i$$

- **Thus:**

  1) $\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty = \frac{1}{\sqrt{m}}$

$||\cdot||_\infty$ versus $||\cdot||_2$

- Discrepancy hence stems from the different norms:

  1) $||\boldsymbol{H}||_2 \leq \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty\right)$

  2) $||\boldsymbol{K}||_2 = \mathcal{O}\left(\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2\right)$

- Now let's calculate this vector:

$$\frac{\partial f}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i}\left(\frac{1}{\sqrt{m}}\sum_{j=1}^{m} v_j \alpha_j\right) = \frac{1}{\sqrt{m}} v_i$$

- **Thus:**

  1) $\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_\infty = \frac{1}{\sqrt{m}}$

  2) $\left|\left|\frac{\partial f}{\partial \boldsymbol{\alpha}}\right|\right|_2^2 = \frac{1}{m}\sum_{i=1}^{m} v_i^2 = 1$

Setup, Restated

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}}\boldsymbol{W}^{(l)}\boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}} \boldsymbol{W}^{(l)} \boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{m_{l-1} \times m_l}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ some **non-linearity**

Setup, Restated

- As always, fully-connected network of L layers mapping to $\mathbb{R}$:
  - $\boldsymbol{\alpha}^{(0)}(\boldsymbol{x}) = \boldsymbol{x}$
  - $\boldsymbol{\alpha}^{(l)}(\boldsymbol{x}) = \sigma\left(\tilde{\boldsymbol{\alpha}}^{(l-1)}(\boldsymbol{x})\right)$
  - $\tilde{\boldsymbol{\alpha}}^{(l)}(\boldsymbol{x}) = \frac{1}{\sqrt{m_{l-1}}} \boldsymbol{W}^{(l)} \boldsymbol{\alpha}^{(l-1)}(\boldsymbol{x})$

  where $\boldsymbol{W}^{(l)} \in \mathbb{R}^{m_{l-1} \times m_l}$ and $\sigma : \mathbb{R} \to \mathbb{R}$ some **non-linearity**
- The final output is given by

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{m}} \boldsymbol{v}^T \boldsymbol{\alpha}^{(L)}(\boldsymbol{x})$$

  where $\boldsymbol{v} \in \mathbb{R}^m$ is **not** trainable

Hessian for General Neural Network

Hessian for General Neural Network

- Turns out that the same behaviour is observed!

Hessian for General Neural Network

- Turns out that the same behaviour is observed!
- Define the following quantities:

Hessian for General Neural Network

- Turns out that the same behaviour is observed!

- Define the following quantities:

  1) $\mathcal{Q}_\infty = \max_{1 \leq i \leq L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_\infty$

Hessian for General Neural Network

- Turns out that the same behaviour is observed!
- Define the following quantities:
  1) $\mathcal{Q}_\infty = \max_{1 \leq i \leq L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_\infty$
  2) $\mathcal{Q}_2 = \max_{1 \leq i \leq L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2$

Hessian for General Neural Network

- Turns out that the same behaviour is observed!
- Define the following quantities:
  1) $\mathcal{Q}_{\infty} = \max_{1 \leq i \leq L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_{\infty}$
  2) $\mathcal{Q}_{2} = \max_{1 \leq i \leq L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_{2}$
  3) $\mathcal{Q}_{2,2,1} = \max_{1 \leq l_1 < l_2 < l_3 \leq L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l_1)}}{\partial \boldsymbol{w}^{(l_1)}} \right|\right|_{2} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l_2)}}{\partial \boldsymbol{w}^{(l_2)}} \right|\right|_{2} \left|\left| \frac{\partial^2 \boldsymbol{\alpha}^{(l_3)}}{\left(\partial \boldsymbol{\alpha}^{(l_3-1)}\right)^2} \right|\right|_{2,2,1}$

Hessian for General Neural Network

- Turns out that the same behaviour is observed!
- Define the following quantities:
  1) $\mathcal{Q}_\infty = \max_{1 \le i \le L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_\infty$
  2) $\mathcal{Q}_2 = \max_{1 \le i \le L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2$
  3) $\mathcal{Q}_{2,2,1} = \max_{1 \le l_1 < l_2 < l_3 \le L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l_1)}}{\partial \boldsymbol{w}^{(l_1)}} \right|\right|_2 \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l_2)}}{\partial \boldsymbol{w}^{(l_2)}} \right|\right|_2 \left|\left| \frac{\partial^2 \boldsymbol{\alpha}^{(l_3)}}{\left(\partial \boldsymbol{\alpha}^{(l_3-1)}\right)^2} \right|\right|_{2,2,1}$

Bound on Hessian

We have the following bound on the norm of the Hessian:

$$||\boldsymbol{H}||_2 \le C_1 \mathcal{Q}_{2,2,1} \mathcal{Q}_\infty + \frac{1}{\sqrt{m}} C_2 \mathcal{Q}_2$$

Kernel for General Neural Network

Kernel for General Neural Network

- We use the same argument:

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x})$$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2$$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Again since all is random at initialization (a bit shady argument):

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Again since all is random at initialization (a bit shady argument):

$$\left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2 \approx \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \right|\right|_2^2 \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2$$

- Again since all is random at initialization (a bit shady argument):

$$\Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2 \approx \Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \Big|\Big|_2^2 \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2$$

- Moreover we get the same asymptotics:

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2$$

- Again since all is random at initialization (a bit shady argument):

$$\Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2 \approx \Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \Big|\Big|_2^2 \Big|\Big| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \Big|\Big|_2^2$$

- Moreover we get the same asymptotics:

$$\Big|\Big| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \Big|\Big|_2^2 = \Theta(1)$$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Again since all is random at initialization (a bit shady argument):

$$\left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2 \approx \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \right|\right|_2^2 \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Moreover we get the same asymptotics:

$$\left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \right|\right|_2^2 = \Theta(1)$$

- **So:** $K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \mathcal{O}\left(\max_{i=1}^{L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2\right)$

Kernel for General Neural Network

- We use the same argument:

$$K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \sum_{l=1}^{L} ||\nabla_{\boldsymbol{W}^{(l)}} f(\boldsymbol{w}; \boldsymbol{x})||_2^2 = \sum_{l=1}^{L} \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Again since all is random at initialization (a bit shady argument):

$$\left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2 \approx \left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \right|\right|_2^2 \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2$$

- Moreover we get the same asymptotics:

$$\left|\left| \frac{\partial \boldsymbol{\alpha}^{(l)}}{\partial \boldsymbol{W}^{(l)}} \right|\right|_2^2 = \Theta(1)$$

- **So:** $K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \mathcal{O}\left(\max_{i=1}^{L} \left|\left| \frac{\partial f}{\partial \boldsymbol{\alpha}^{(l)}} \right|\right|_2^2 \right) = \mathcal{O}(\mathcal{Q}_2)$

Controlling $\mathcal{Q}$

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Asymptotics of $\mathcal{Q}$

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Asymptotics of $\mathcal{Q}$

Take fully connected network $f(\boldsymbol{w};\boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Asymptotics of $\mathcal{Q}$

Take fully connected network $f(\boldsymbol{w}; \boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

- $\mathcal{Q}_\infty = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Asymptotics of $\mathcal{Q}$

Take fully connected network $f(\boldsymbol{w};\boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

- $\mathcal{Q}_\infty = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$
- $\mathcal{Q}_{2,2,1} = \mathcal{O}(1)$

Controlling $\mathcal{Q}$

Finally we have the following asymptotic behaviours for the above quantities:

Asymptotics of $\mathcal{Q}$

Take fully connected network $f(\boldsymbol{w}; \boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

- $\mathcal{Q}_\infty = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$
- $\mathcal{Q}_{2,2,1} = \mathcal{O}(1)$
- $\mathcal{Q}_2 = \mathcal{O}(1)$

Main Theorem

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Asymptotics of Hessian

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Asymptotics of Hessian

Take fully connected network $f(\boldsymbol{w};\boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Asymptotics of Hessian

Take fully connected network $f(\boldsymbol{w}; \boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

$$||\boldsymbol{H}||_2 = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$$

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Asymptotics of Hessian

Take fully connected network $f(\boldsymbol{w}; \boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

$$||\boldsymbol{H}||_2 = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$$

- Fixing $R > 0$ big enough to include GD trajectory and applying the Lemma from before implies the **constancy** of the tangent kernel during **training**.

Main Theorem

Now we can put all together to obtain as a **direct** consequence:

Asymptotics of Hessian

Take fully connected network $f(\boldsymbol{w}; \boldsymbol{x})$ with initialization $\boldsymbol{w}_0$, $\boldsymbol{w} \in \mathcal{B}(\boldsymbol{w}_0, R)$ with $R > 0$. Then with high probability over initialization:

$$||\boldsymbol{H}||_2 = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{m}}\right)$$

- Fixing $R > 0$ big enough to include GD trajectory and applying the Lemma from before implies the **constancy** of the tangent kernel during **training**.
- $||\boldsymbol{K}||_2 = \Theta(1)$ since $\mathcal{Q}_2 = \mathcal{O}(1)$

Some Comments

Some Comments

- The role of the radius R is a bit unclear.

Some Comments

- The role of the radius R is a bit unclear.
- What dependence on R in the bound for $||\boldsymbol{H}||_2$?

Some Comments

- The role of the radius R is a bit unclear.
- What dependence on R in the bound for $||\boldsymbol{H}||_2$?
- Probably all hidden in $\mathcal{O}$ but for instance $R = \infty$ shouldn't work? Or only low probability bound?

Some Comments

- The role of the radius R is a bit unclear.
- What dependence on R in the bound for $||\boldsymbol{H}||_2$?
- Probably all hidden in $\mathcal{O}$ but for instance $R = \infty$ shouldn't work? Or only low probability bound?
- How does the choice of R as GD radius affect the convergence/probability guarantee?

Negative Result for Lazy Learning

Negative Result for Lazy Learning

- Lots of literature claims that constancy of NTK is a consequence of the setting (**lazy learning**):

Negative Result for Lazy Learning

- Lots of literature claims that constancy of NTK is a consequence of the setting (**lazy learning**):

Small Learning rate + Huge Widths + Small Parameter Change

Negative Result for Lazy Learning

- Lots of literature claims that constancy of NTK is a consequence of the setting (**lazy learning**):

Small Learning rate + Huge Widths + Small Parameter Change

- Authors show that using a non-linear output layer instead of a linear layer induces a non-constant NTK

Negative Result for Lazy Learning

- Lots of literature claims that constancy of NTK is a consequence of the setting (**lazy learning**):

Small Learning rate + Huge Widths + Small Parameter Change

- Authors show that using a non-linear output layer instead of a linear layer induces a non-constant NTK

- Namely, using the 1-dim 1-hidden layer network $f(\boldsymbol{w}; x)$ as

$$\tilde{f}(\boldsymbol{w}, x) = \phi(f(\boldsymbol{w}; x))$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is some twice-differentiable non-linearity

NTK is Not Constant (I)

NTK is Not Constant (I)

- Let's look at the Hessian:

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j}$$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j} \phi(f)$$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$

$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:**

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:** $\tilde{\boldsymbol{H}} = \phi''(f)\left(\nabla_{\boldsymbol{w}} f\right)\left(\nabla_{\boldsymbol{w}} f\right)^T + \phi'(f)\boldsymbol{H}$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:** $\tilde{\boldsymbol{H}} = \phi''(f)\left(\nabla_{\boldsymbol{w}} f\right)\left(\nabla_{\boldsymbol{w}} f\right)^T + \phi'(f)\boldsymbol{H}$

- From **Weyl's**-inequality we have that for all symmetric matrices $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$:

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:** $\tilde{\boldsymbol{H}} = \phi''(f)\left(\nabla_{\boldsymbol{w}} f\right)\left(\nabla_{\boldsymbol{w}} f\right)^T + \phi'(f)\boldsymbol{H}$

- From **Weyl's**-inequality we have that for all symmetric matrices $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$:

$$\lambda_1(\boldsymbol{A}) \geq \lambda_1(\boldsymbol{B}) + \lambda_n(\boldsymbol{C}) \iff ||\boldsymbol{A}||_2 \geq ||\boldsymbol{B}||_2 - ||\boldsymbol{C}||_2$$

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:** $\tilde{\boldsymbol{H}} = \phi''(f)\left(\nabla_{\boldsymbol{w}} f\right)\left(\nabla_{\boldsymbol{w}} f\right)^T + \phi'(f)\boldsymbol{H}$

- From **Weyl's**-inequality we have that for all symmetric matrices $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$:

$$\lambda_1(\boldsymbol{A}) \geq \lambda_1(\boldsymbol{B}) + \lambda_n(\boldsymbol{C}) \iff ||\boldsymbol{A}||_2 \geq ||\boldsymbol{B}||_2 - ||\boldsymbol{C}||_2$$

- Applying this to $\tilde{\boldsymbol{H}}$ leads to

NTK is Not Constant (I)

- Let's look at the Hessian:

$$\tilde{H}_{ij} = \frac{\partial^2 \tilde{f}}{\partial w_i \partial w_j} = \frac{\partial^2}{\partial w_i \partial w_j}\phi(f) = \frac{\partial}{\partial w_i}\left(\phi'(f)\frac{\partial f}{\partial w_j}\right)$$
$$= \phi''(f)\frac{\partial f}{\partial w_i}\frac{\partial f}{\partial w_j} + \phi'(f)\frac{\partial^2 f}{\partial w_j \partial w_i}$$

- **Thus:** $\tilde{\boldsymbol{H}} = \phi''(f)\left(\nabla_{\boldsymbol{w}} f\right)\left(\nabla_{\boldsymbol{w}} f\right)^T + \phi'(f)\boldsymbol{H}$

- From **Weyl's**-inequality we have that for all symmetric matrices $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$:

$$\lambda_1(\boldsymbol{A}) \geq \lambda_1(\boldsymbol{B}) + \lambda_n(\boldsymbol{C}) \iff ||\boldsymbol{A}||_2 \geq ||\boldsymbol{B}||_2 - ||\boldsymbol{C}||_2$$

- Applying this to $\tilde{\boldsymbol{H}}$ leads to

$$||\tilde{\boldsymbol{H}}||_2 \geq |\phi''(f)|\left|\left|\nabla_{\boldsymbol{w}} f\right|\right|_2^2 - |\phi'(f)|||\boldsymbol{H}||_2$$

NTK is Not Constant (II)

NTK is Not Constant (II)

- We found that

$$||\tilde{\boldsymbol{H}}||_2 \geq |\phi''(f)| \left|\left|\nabla_{\boldsymbol{w}} f\right|\right|_2^2 - |\phi'(f)| ||\boldsymbol{H}||_2$$

NTK is Not Constant (II)

- We found that

$$||\tilde{\boldsymbol{H}}||_2 \geq |\phi''(f)|\left|\left|\nabla_{\boldsymbol{w}} f\right|\right|_2^2 - |\phi'(f)|||\boldsymbol{H}||_2$$

- **Recall:** $||\boldsymbol{H}||_2 \xrightarrow{m \to \infty} 0$

NTK is Not Constant (II)

- We found that

$$||\tilde{\boldsymbol{H}}||_2 \geq |\phi''(f)| \left|\left| \nabla_{\boldsymbol{w}} f \right|\right|_2^2 - |\phi'(f)| ||\boldsymbol{H}||_2$$

- **Recall:** $||\boldsymbol{H}||_2 \xrightarrow{m \to \infty} 0$
- **However:** $\left|\left| \nabla_{\boldsymbol{w}} f \right|\right|_2^2 = K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \Theta(1)$

NTK is Not Constant (II)

- We found that

$$||\tilde{\boldsymbol{H}}||_2 \geq |\phi''(f)|\big|\big|\nabla_{\boldsymbol{w}} f\big|\big|_2^2 - |\phi'(f)|||\boldsymbol{H}||_2$$

- **Recall:** $||\boldsymbol{H}||_2 \xrightarrow{m \to \infty} 0$
- **However:** $\big|\big|\nabla_{\boldsymbol{w}} f\big|\big|_2^2 = K_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}) = \Theta(1)$
- **Thus**, as long as ϕ'' doesn't vanish (so if not linear) we have

$$||\tilde{\boldsymbol{H}}||_2 = \Omega(1)$$

Other Example

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture
- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

- **But then** a term in $\mathcal{Q}_{2,2,1}$ scales as

$$=$$

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

- **But then** a term in $\mathcal{Q}_{2,2,1}$ scales as

$$\Big|\Big|\frac{\partial \boldsymbol{\alpha}^{(l+1)}}{(\partial \alpha^{(l)})^2}\Big|\Big|_{2,2,1} =$$

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

- **But then** a term in $\mathcal{Q}_{2,2,1}$ scales as

$$\left|\left|\frac{\partial \boldsymbol{\alpha}^{(l+1)}}{(\partial \alpha^{(l)})^2}\right|\right|_{2,2,1} = \sum_{i=1}^{m} \left|\sigma''\left(w_i^{(l+1)}\alpha^{(l)}\right)\left(w_i^{(l+1)}\right)^2\right|$$

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

- **But then** a term in $\mathcal{Q}_{2,2,1}$ scales as

$$\left|\left|\frac{\partial \boldsymbol{\alpha}^{(l+1)}}{(\partial \alpha^{(l)})^2}\right|\right|_{2,2,1} = \sum_{i=1}^{m} \left|\sigma''\left(w_i^{(l+1)}\alpha^{(l)}\right)\left(w_i^{(l+1)}\right)^2\right| = \Theta(m)$$

Other Example

- Non-constancy of the NTK also for a **bottle-neck** architecture

- Specifically, they assume one hidden layer l does **not** go to infinity but $\alpha^{(l)}(\boldsymbol{x}) \in \mathbb{R}$, so $m_l = 1$ fixed and

$$\boldsymbol{\alpha}^{(l+1)} = \sigma(\boldsymbol{w}^{(l+1)}\alpha^{(l)})$$

- **But then** a term in $\mathcal{Q}_{2,2,1}$ scales as

$$\left|\left|\frac{\partial \boldsymbol{\alpha}^{(l+1)}}{(\partial \alpha^{(l)})^2}\right|\right|_{2,2,1} = \sum_{i=1}^{m} \left|\sigma''\left(w_i^{(l+1)}\alpha^{(l)}\right)\left(w_i^{(l+1)}\right)^2\right| = \Theta(m)$$

- This "kills" the $\frac{1}{\sqrt{m}}$ from before!

Overview

Overview

- In summary we found that

Overview

- In summary we found that

Model	$\left\Vert \frac{\partial f}{\partial \alpha^{(l)}} \right\Vert_\infty$	$(2, 2, 1)$-norms	Hessian norm	Trans. to linearity?
linear output layer	$\tilde{O}(1/\sqrt{m})$	$O(1)$	$\tilde{O}(1/\sqrt{m})$	**Yes**
non-linear output layer	$\tilde{O}(1)$	$O(1)$	$\tilde{O}(1)$	**No**
bottleneck	$\tilde{O}(1/\sqrt{m})$	$O(m)$	$\tilde{O}(1)$	**No**

Overview

- In summary we found that

Model	$\left\Vert \frac{\partial f}{\partial \alpha^{(l)}} \right\Vert_\infty$	$(2,2,1)$-norms	Hessian norm	Trans. to linearity?
linear output layer	$\tilde{O}(1/\sqrt{m})$	$O(1)$	$\tilde{O}(1/\sqrt{m})$	**Yes**
non-linear output layer	$\tilde{O}(1)$	$O(1)$	$\tilde{O}(1)$	**No**
bottleneck	$\tilde{O}(1/\sqrt{m})$	$O(m)$	$\tilde{O}(1)$	**No**

- One important side note. For GD to converge fast, it is not necessary that the tangent kernel becomes constant but that it remains **well-conditioned** along the optimization path.

Overview

- In summary we found that

Model	$\left\| \frac{\partial f}{\partial \alpha^{(l)}} \right\|_\infty$	$(2,2,1)$-norms	Hessian norm	Trans. to linearity?
linear output layer	$\tilde{O}(1/\sqrt{m})$	$O(1)$	$\tilde{O}(1/\sqrt{m})$	**Yes**
non-linear output layer	$\tilde{O}(1)$	$O(1)$	$\tilde{O}(1)$	**No**
bottleneck	$\tilde{O}(1/\sqrt{m})$	$O(m)$	$\tilde{O}(1)$	**No**

- One important side note. For GD to converge fast, it is not necessary that the tangent kernel becomes constant but that it remains **well-conditioned** along the optimization path.

- Well-conditionedness is somewhat **"passed"** on to the non-linear output model from the linear output model. That's why these models also converge fast in practice.

Experiments (I)

Experiments (I)

- Track the change of the tangent kernel by computing

Experiments (I)

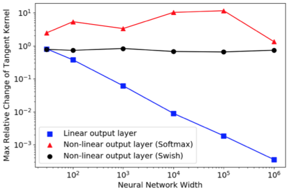
- Track the change of the tangent kernel by computing

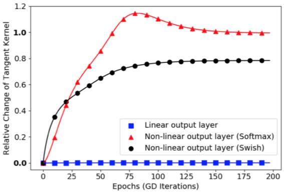
$$\Delta K = \sup_{t>0} \frac{||K(\boldsymbol{w}_t - K(\boldsymbol{w}_0)||_F}{||K(\boldsymbol{w}_0)||_F}$$

Experiments (I)

- Track the change of the tangent kernel by computing

$$\Delta K = \sup_{t>0} \frac{||K(\boldsymbol{w}_t - K(\boldsymbol{w}_0)||_F}{||K(\boldsymbol{w}_0)||_F}$$

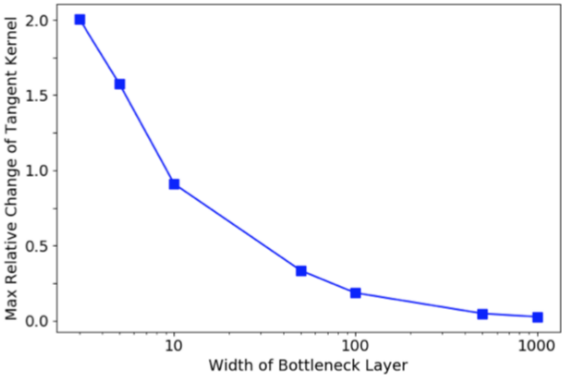




Change in NTK for **linear** versus **non-Linear** output activation. 1 hidden layer network with either linear, softmax or swish activation at output.

Experiments (II)

Experiments (II)



Width of the bottleneck layer and corresponding change in NTK.
3 hidden layer network of widths $m = 10^4$

Closing Thoughts

Closing Thoughts

- Overall very interesting and intuitive formulation

Closing Thoughts

- Overall very interesting and intuitive formulation
- Very elementary maths (Zero Hessian $\implies$ Linear etc)

Closing Thoughts

- Overall very interesting and intuitive formulation
- Very elementary maths (Zero Hessian $\implies$ Linear etc)
- Of course lots of math hidden in the statement that GD only explores $\mathcal{B}(\boldsymbol{w}_0, R)$. Some mistakes in the proofs but I think only little fixes needed.

Closing Thoughts

- Overall very interesting and intuitive formulation
- Very elementary maths (Zero Hessian $\implies$ Linear etc)
- Of course lots of math hidden in the statement that GD only explores $\mathcal{B}(\boldsymbol{w}_0, R)$. Some mistakes in the proofs but I think only little fixes needed.
- Very interesting that the optimization mechanism is clearly **separated**. I have not understood the paper in detail but seems like it suffices to show (for instance for SGD) that the iterates stay in $\mathcal{B}(\boldsymbol{w}_0, R)$ for R independent of width m.

Closing Thoughts

- Overall very interesting and intuitive formulation
- Very elementary maths (Zero Hessian $\implies$ Linear etc)
- Of course lots of math hidden in the statement that GD only explores $\mathcal{B}(\boldsymbol{w}_0, R)$. Some mistakes in the proofs but I think only little fixes needed.
- Very interesting that the optimization mechanism is clearly **separated**. I have not understood the paper in detail but seems like it suffices to show (for instance for SGD) that the iterates stay in $\mathcal{B}(\boldsymbol{w}_0, R)$ for R independent of width m.
- Are similar results known for SGD, ADAM, second order methods? What does it take to fall out of this ball (except for big learning rates)?