What Does It Mean To Be a Transformer?

Insights from a Theoretical Hessian Analysis

Weronika Ormaniec

ML in PL Conference, Warsaw, 16.10.2025

@wormaniec

Overview

Overview

1. What do we mean by neural network loss Hessian?

Overview

1. What do we mean by neural network loss Hessian?

2. Motivation:
   - Why do we care about the loss Hessian?
   - Why study the Hessian for Transformer loss?

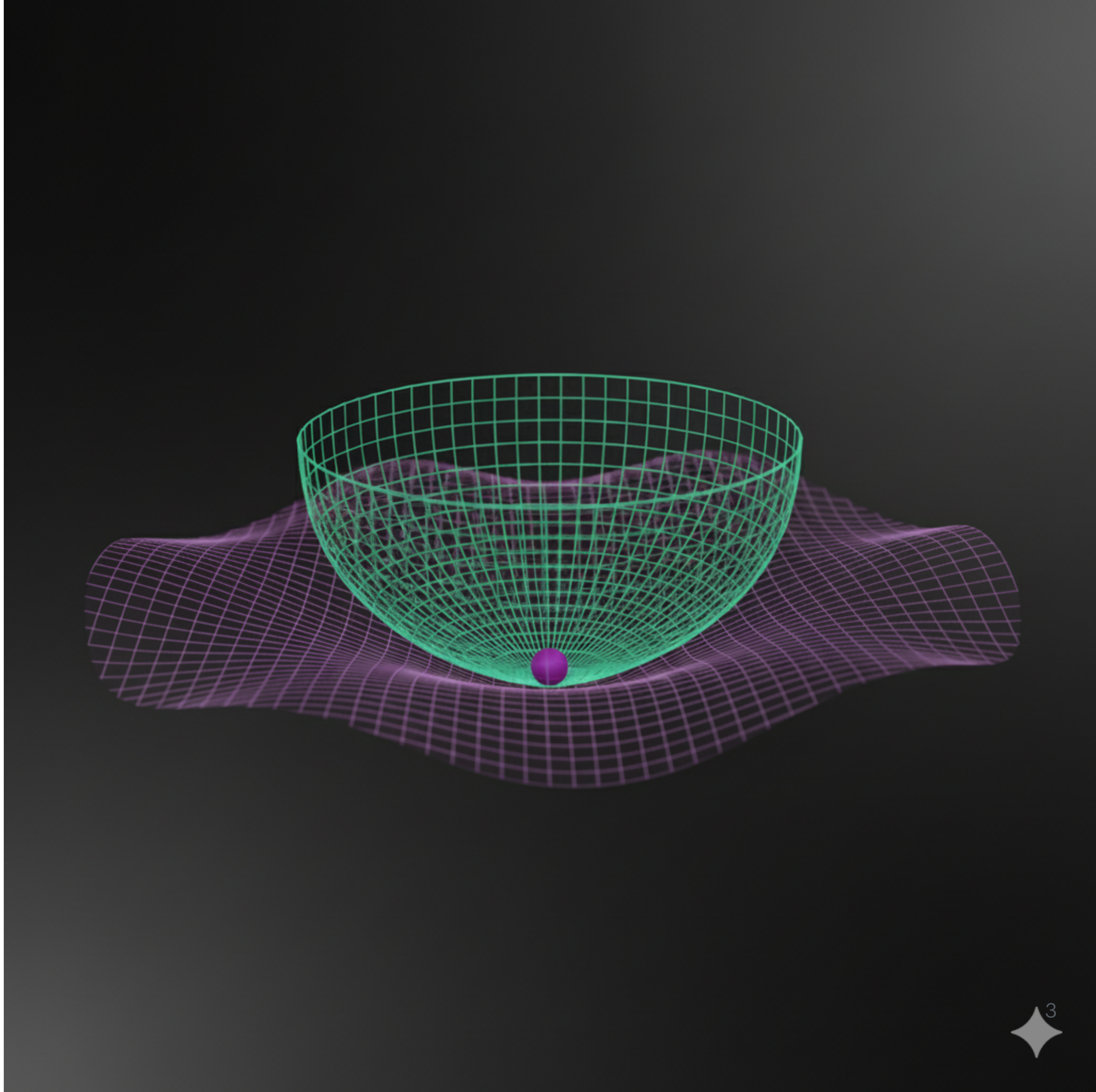
Overview

1. What do we mean by neural network loss Hessian?

2. Motivation:
   - Why do we care about the loss Hessian?
   - Why study the Hessian for Transformer loss?

3. Results:
   - Hessian block-heterogeneity
   - Comparison of loss Hessian for Transformers and MLPs
   - Outlook: Primer on Transformer loss Hessian spectra

Neural Network Loss Hessian



Neural Network Loss Hessian

Neural Network Loss Hessian

Neural network: $\mathbf{F} : \mathbb{R}^{d_1} \ni \theta \longrightarrow \mathbf{F}(\theta) \in \mathbb{R}^{d_2}$

Neural Network Loss Hessian

Neural network: $\mathbf{F} : \mathbb{R}^{d_1} \ni \theta \longrightarrow \mathbf{F}(\theta) \in \mathbb{R}^{d_2}$

Neural network loss: $\mathcal{L} : \mathbb{R}^{d_1} \ni \theta \longrightarrow (\ell \circ \mathbf{F})(\theta) \in \mathbb{R}$

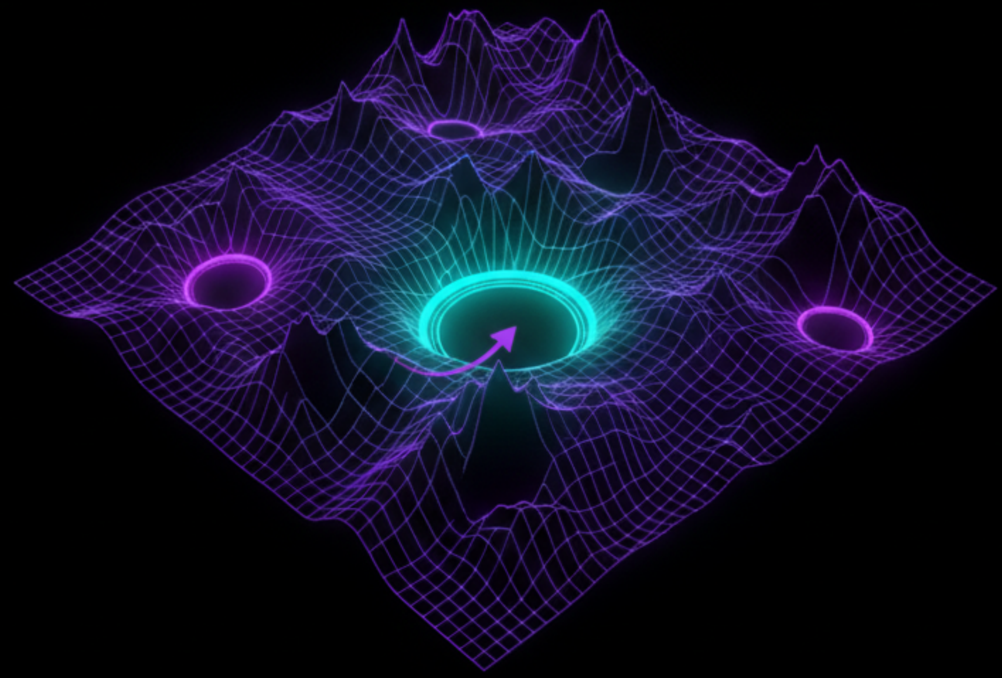
Neural Network Loss Hessian

Neural network: $\mathbf{F} : \mathbb{R}^{d_1} \ni \theta \longrightarrow \mathbf{F}(\theta) \in \mathbb{R}^{d_2}$

Neural network loss: $\mathcal{L} : \mathbb{R}^{d_1} \ni \theta \longrightarrow (\ell \circ \mathbf{F})(\theta) \in \mathbb{R}$

$$\mathbf{H} \in \mathbb{R}^{d_1 \times d_1} \qquad \mathbf{H}_{i,j} = \frac{\partial^2 \left(\ell \circ \mathbf{F}\left(\theta\right)\right)}{\partial \theta_i \partial \theta_j}$$

Motivation



Loss Approximation

Loss Approximation

We can approximate the loss around any point with a polynomial defined with gradient and Hessian

Loss Approximation

We can approximate the loss around any point with a polynomial defined with gradient and Hessian

Taylor expansion $$\mathcal{L}(\theta + \Delta\theta) \approx \mathcal{L}(\theta) + \Delta\theta^\top \nabla\mathcal{L}(\theta) + \frac{1}{2}\Delta\theta^\top \mathbf{H}(\theta)\Delta\theta$$

Loss Approximation

We can approximate the loss around any point with a polynomial defined with gradient and Hessian

Taylor expansion $$\mathcal{L}(\theta^* + \Delta\theta) \approx \mathcal{L}(\theta^*) + \frac{1}{2}\Delta\theta^\top \mathbf{H}(\theta^*)\Delta\theta$$

Loss Landscape Analysis

Loss Landscape Analysis

The Hessian...

Loss Landscape Analysis

The Hessian...

- Locally characterises the curvature of the loss

Loss Landscape Analysis

The Hessian...

- Locally characterises the curvature of the loss
- Eigenvectors and eigenvalues $\Longrightarrow$ in which directions the loss is locally flat, convex or concave

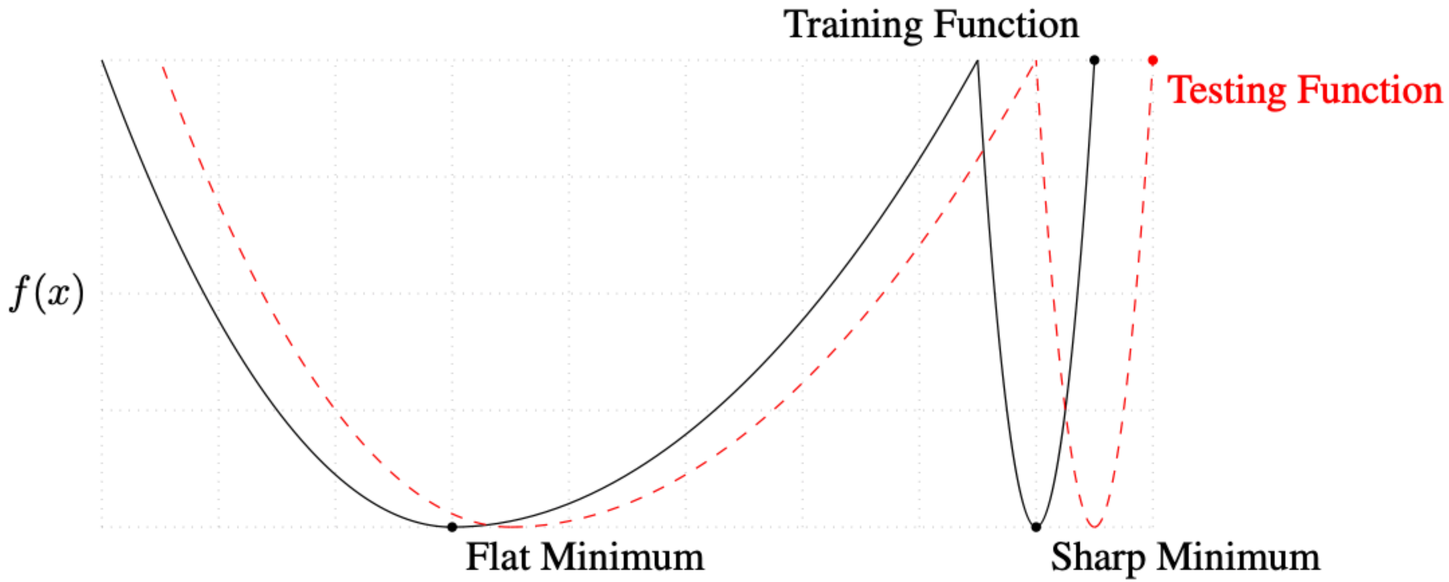
Loss Landscape Analysis

The Hessian...

- Locally characterises the curvature of the loss
- Eigenvectors and eigenvalues $\Longrightarrow$ in which directions the loss is locally flat, convex or concave
- Specifies the nature of critical points

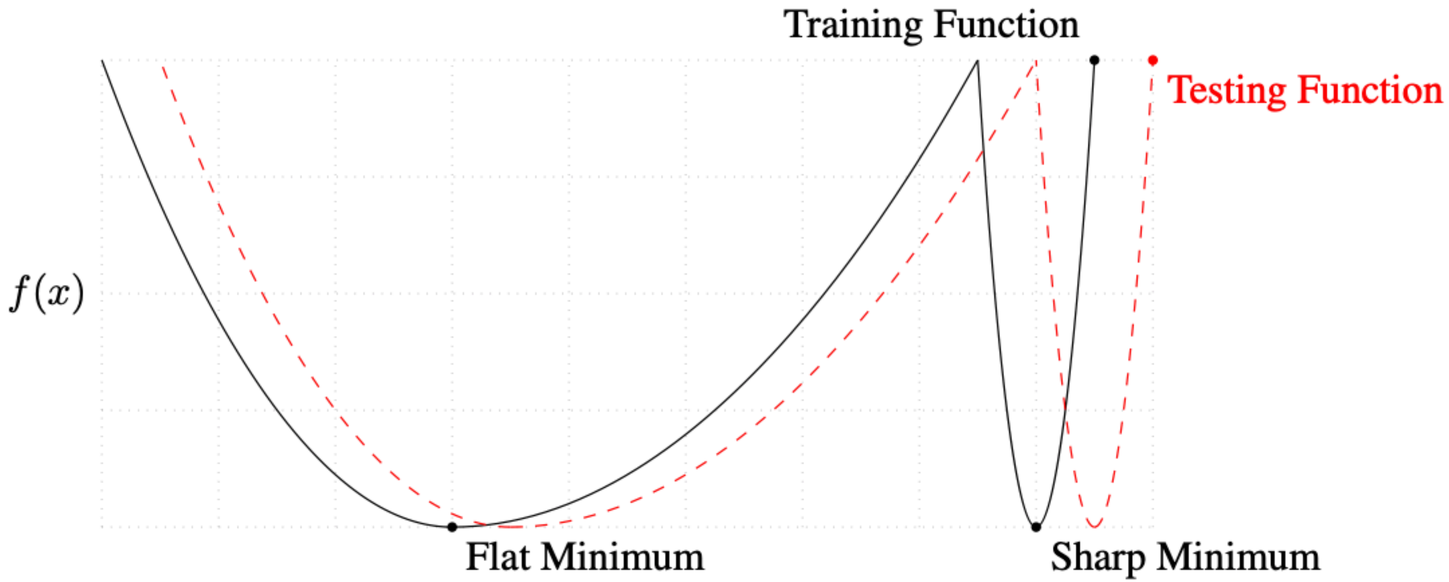
Generalisation

Generalisation



Source: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. Keskar et al. (2017)

Generalisation



Loss Hessian properties at minima are used as a measure of flatness

Source: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. Keskar et al. (2017)

Optimisation

Optimisation

Newton method $\theta_{t+1} = \theta_t - \eta \mathbf{H}^{-1}(\theta_t) \nabla \mathcal{L}(\theta_t)$

Optimisation

Newton method $\theta_{t+1} = \theta_t - \eta \mathbf{H}^{-1}(\theta_t) \nabla \mathcal{L}(\theta_t)$

For convex functions quadratic rate of convergence vs linear one for GD

Optimisation

Newton method $\theta_{t+1} = \theta_t - \eta \mathbf{H}^{-1}(\theta_t) \nabla \mathcal{L}(\theta_t)$

For convex functions quadratic rate of convergence vs linear one for GD

Adam can be viewed as a diagonal approximation of the inverse Hessian

Optimisation

Newton method $\theta_{t+1} = \theta_t - \eta \mathbf{H}^{-1}(\theta_t) \nabla \mathcal{L}(\theta_t)$

For convex functions quadratic rate of convergence vs linear one for GD

Adam can be viewed as a diagonal approximation of the inverse Hessian

Many other block-diagonal approximations: KFAC, Shampoo, PSGD, etc.

Optimisation

Optimisation

For gradient descent, the Hessian suggests the learning rate

Optimisation

For gradient descent, the Hessian suggests the learning rate

The loss should decrease for $\eta \in (0, \frac{2}{\lambda_{\max}(\mathbf{H}(\theta))})$

Pruning & Quantisation

Pruning & Quantisation

Optimal Brain Surgeon Framework

Pruning & Quantisation

Optimal Brain Surgeon Framework

$$\min_{q} \min_{\delta \mathbf{w}^\top} \frac{1}{2} \delta \mathbf{w}^\top \mathbf{H} \delta \mathbf{w}, \; s.t. \; \mathbf{e}_q^\top \delta \mathbf{w} + \mathbf{w}_q = 0$$

$$\delta \mathbf{w} = -\frac{w_q^2}{[\mathbf{H}_{qq}^{-1}]} \mathbf{H}^{-1} \mathbf{e}_q$$

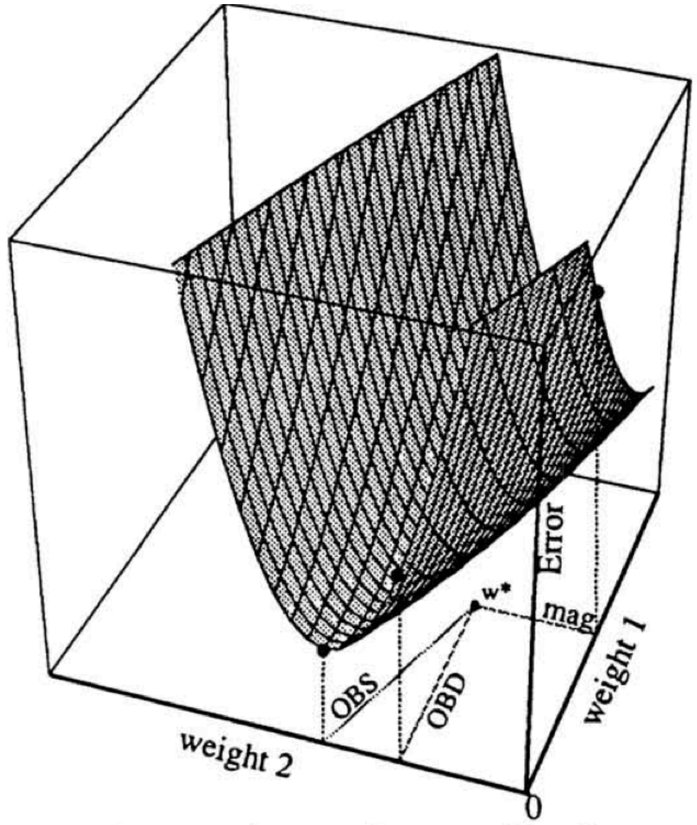


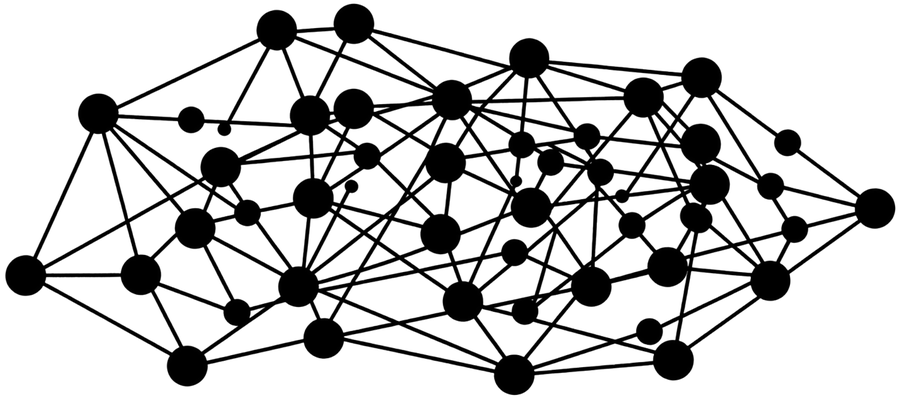
Figure 1: Error as a function of two weights in a network. The (local) minimum occurs at weight w*, found by gradient descent or other learning method. In this illustration, a magnitude based pruning technique (mag) then removes the smallest weight, weight 2; Optimal Brain Damage before retraining (OBD) removes weight 1. In contrast, our Optimal Brain Surgeon method (OBS) not only removes weight 1, but *also* automatically adjusts the value of weight 2 to minimize the error, without retraining.

Source:

Second order derivatives for network pruning: Optimal Brain Surgeon. Hassibi and Stork (1992)

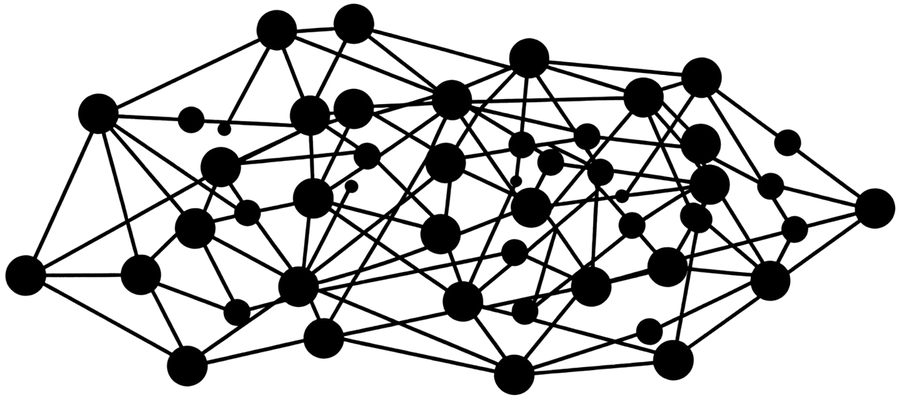
Why study the loss Hessian for Transformers?

Why study the loss Hessian for Transformers?

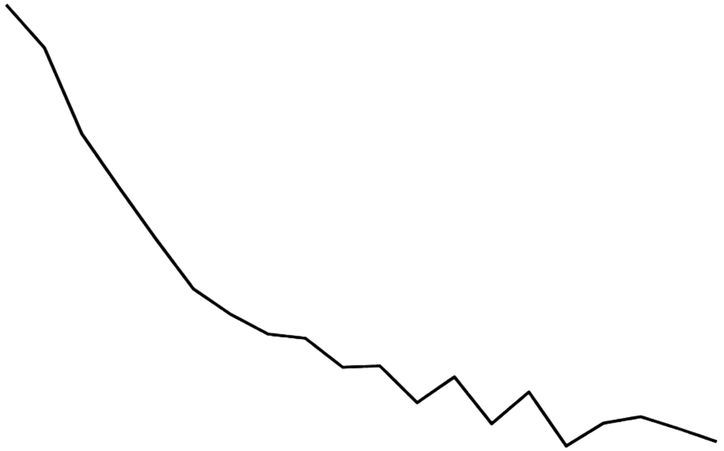


Unique architecture

Why study the loss Hessian for Transformers?

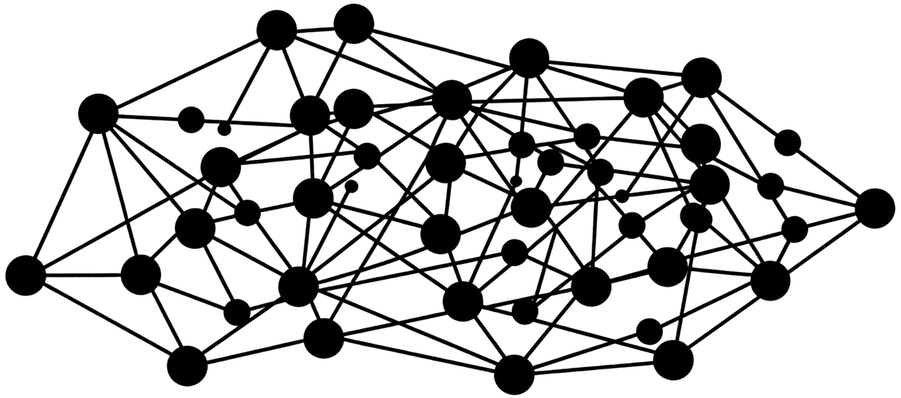


Unique architecture

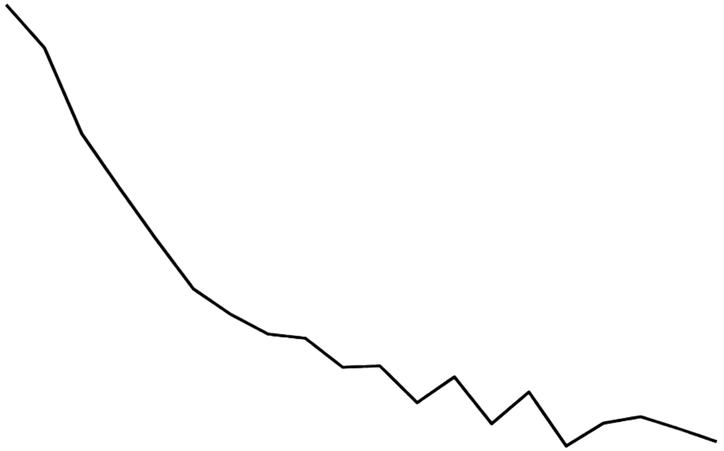


Tricky optimisation

Why study the loss Hessian for Transformers?



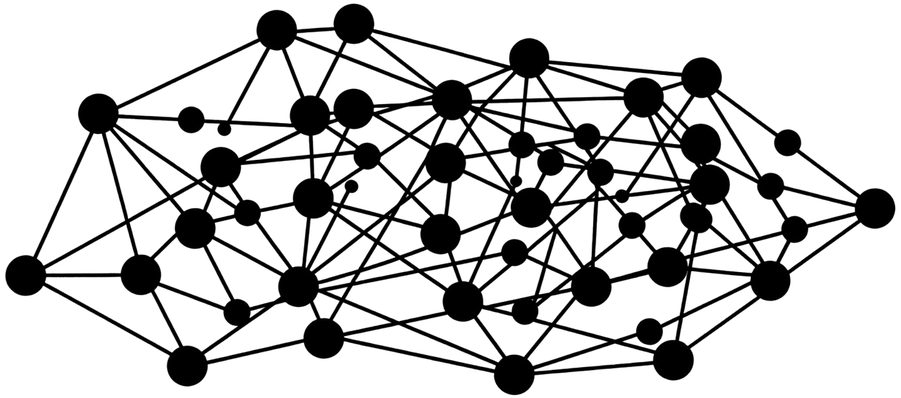
Unique architecture



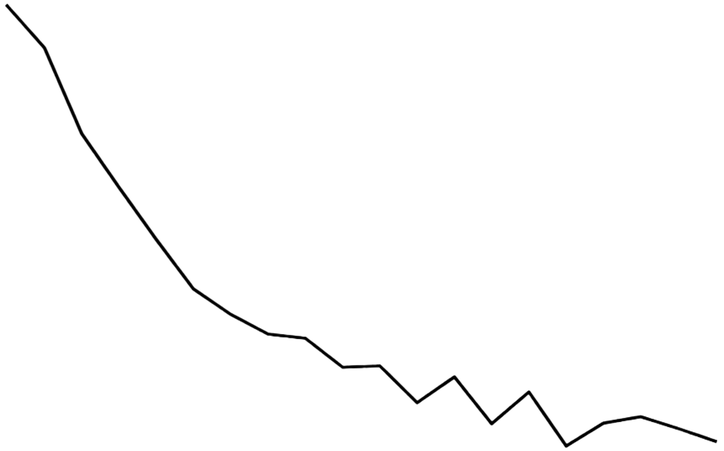
Tricky optimisation

Objective:

Why study the loss Hessian for Transformers?



Unique architecture



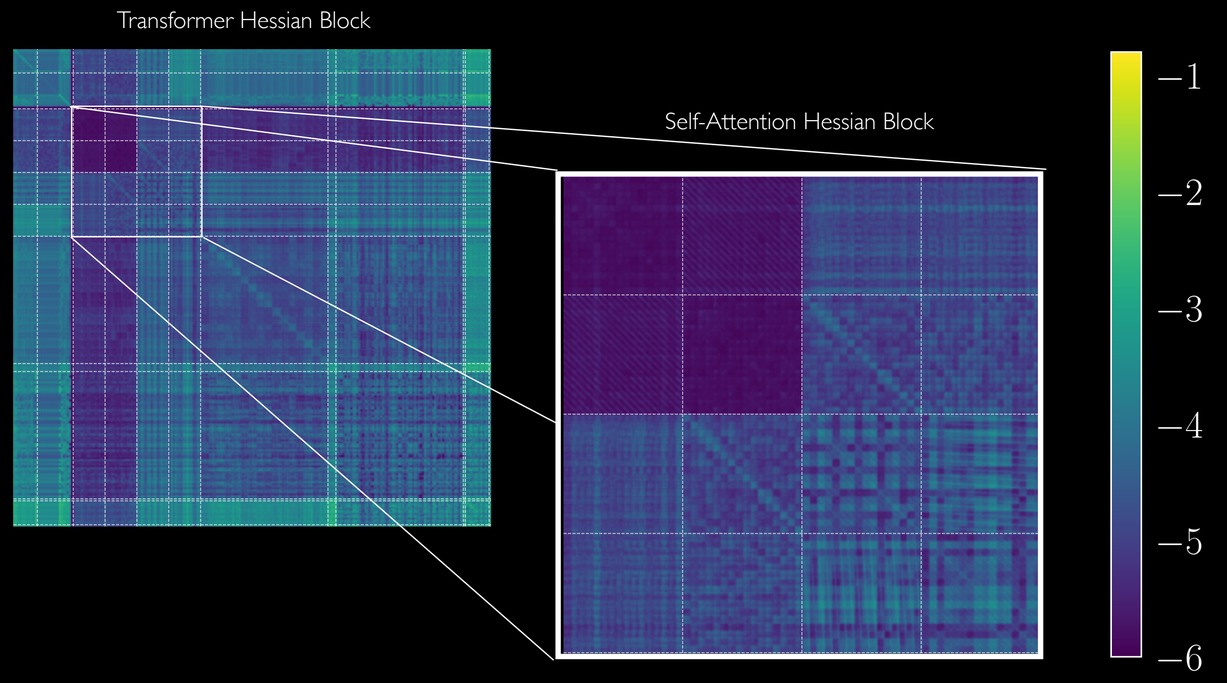
Tricky optimisation

Objective:

Analytically characterise the influence of self-attention on the loss Hessian and contrast it with classical architectures

What Does It Mean To Be a Transformer?

Insights from the Hessian



Collaborators

Felix Dangel

Sidak Pal Singh

Thomas Hofmann

Problem setup

Problem setup

Model consisting of a single self-attention layer

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}(\mathbf{X})\mathbf{X}\mathbf{W}_{\mathrm{V}}$$

where

$$\mathbf{A}(\mathbf{X}) = \mathrm{softmax}\left(\mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\top}\mathbf{X}^{\top} / \sqrt{d_K}\right)$$

Problem setup

Model consisting of a single self-attention layer

$$\mathbf{F}(\mathbf{X}) = \mathbf{A}(\mathbf{X})\mathbf{X}\mathbf{W}_{\mathrm{V}}$$

where

$$\mathbf{A}(\mathbf{X}) = \mathrm{softmax}\left(\mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\top}\mathbf{X}^{\top} / \sqrt{d_K}\right)$$

MSE loss

$$\ell(\mathbf{F}(\mathbf{X}), \mathbf{Y}) = \|\mathbf{F}(\mathbf{X}) - \mathbf{Y}\|_{\mathrm{F}}^2 / (Ld_V)$$

Problem setup

Problem setup

Split the Hessian according to so called Gauss-Newton decomposition

Problem setup

Split the Hessian according to so called Gauss-Newton decomposition

$$\frac{\partial^2 (\ell \circ \mathbf{F})}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot) = \underbrace{\frac{\partial \mathbf{F}}{\partial \mathbf{W}_i}(\cdot)^\top \frac{\partial^2 \ell}{\partial \mathbf{F}^2}(\mathbf{F}(\cdot)) \frac{\partial \mathbf{F}}{\partial \mathbf{W}_j}(\cdot)}_{\text{Outer-product Hessian}} + \underbrace{\left(\frac{\partial \ell}{\partial \mathbf{F}}(\mathbf{F}(\cdot)) \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot)}_{\text{Functional Hessian}}$$

Problem setup

Split the Hessian according to so called Gauss-Newton decomposition

$$\frac{\partial^2 (\ell \circ \mathbf{F})}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot) = \underbrace{\frac{\partial \mathbf{F}}{\partial \mathbf{W}_i}(\cdot)^\top \frac{\partial^2 \ell}{\partial \mathbf{F}^2}(\mathbf{F}(\cdot)) \frac{\partial \mathbf{F}}{\partial \mathbf{W}_j}(\cdot)}_{\text{Outer-product Hessian}} + \underbrace{\left(\frac{\partial \ell}{\partial \mathbf{F}}(\mathbf{F}(\cdot)) \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot)}_{\text{Functional Hessian}}$$

Analyse per-block for each of the weight matrices

Problem setup

Split the Hessian according to so called Gauss-Newton decomposition

$$\frac{\partial^2 (\ell \circ \mathbf{F})}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot) = \underbrace{\frac{\partial \mathbf{F}}{\partial \mathbf{W}_i}(\cdot)^\top \frac{\partial^2 \ell}{\partial \mathbf{F}^2}(\mathbf{F}(\cdot)) \frac{\partial \mathbf{F}}{\partial \mathbf{W}_j}(\cdot)}_{\text{Outer-product Hessian}} + \underbrace{\left(\frac{\partial \ell}{\partial \mathbf{F}}(\mathbf{F}(\cdot)) \otimes \mathbf{I}_{p_i q_i} \right) \frac{\partial^2 \mathbf{F}}{\partial \mathbf{W}_i \partial \mathbf{W}_j}(\cdot)}_{\text{Functional Hessian}}$$

Analyse per-block for each of the weight matrices

$$\mathbf{H}_\bullet = \begin{bmatrix} \mathbf{H}_\bullet(\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{Q}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{K}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{V}) \\ \mathbf{H}_\bullet(\mathbf{W}_\mathrm{K}, \mathbf{W}_\mathrm{Q}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{K}, \mathbf{W}_\mathrm{K}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{K}, \mathbf{W}_\mathrm{V}) \\ \mathbf{H}_\bullet(\mathbf{W}_\mathrm{V}, \mathbf{W}_\mathrm{Q}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{V}, \mathbf{W}_\mathrm{K}) & \mathbf{H}_\bullet(\mathbf{W}_\mathrm{V}, \mathbf{W}_\mathrm{V}) \end{bmatrix}$$

$$\bullet \in \{\mathrm{o}, \mathrm{f}\}$$

Main Result — Outer-Product Hessian

Theorem 3.1. Outer-product Hessian $\mathbf{H}_o$. For a single self-attention layer, the blocks of $\mathbf{H}_o$ are

$$\mathbf{H}_o(\mathbf{W}_V, \mathbf{W}_V) = \frac{2}{Ld_V}\mathbf{M}_1^\top \mathbf{M}_1 \otimes \mathbf{I}_{d_V},$$

$$\mathbf{H}_o(\mathbf{W}_Q, \mathbf{W}_Q) = \frac{2}{Ld_V d_K}\left(\mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top\right)\mathbf{Z}_1^\top\left(\mathbf{I}_L \otimes \mathbf{W}_V\mathbf{W}_V^\top\right)\mathbf{Z}_1\left(\mathbf{I}_{d_V} \otimes \mathbf{W}_K\right),$$

$$\mathbf{H}_o(\mathbf{W}_V, \mathbf{W}_Q) = \frac{2}{Ld_V\sqrt{d_K}}\left(\mathbf{M}_1^\top \otimes \mathbf{W}_V^\top\right)\mathbf{Z}_1\left(\mathbf{I}_{d_V} \otimes \mathbf{W}_K\right),$$

$$\mathbf{H}_o(\mathbf{W}_Q, \mathbf{W}_K) = \frac{2}{Ld_V d_K}\left(\mathbf{I}_{d_V} \otimes \mathbf{W}_K^\top\right)\mathbf{Z}_1^\top\left(\mathbf{I}_L \otimes \mathbf{W}_V\mathbf{W}_V^\top\right)\mathbf{Z}_1\left(\mathbf{W}_Q \otimes \mathbf{I}_{d_V}\right)\mathbf{K}_{d_K,d_V},$$

where $\mathbf{Z}_1 = \mathbf{X} * \mathbf{M}_2 \in \mathbb{R}^{Ld_V \times d_V^2}$ with $\mathbf{M}_1$ and $\mathbf{M}_2$ containing moments of the distributions given by self-attention scores.

Main Result — Functional Hessian

Theorem 3.2. Functional Hessian $\mathbf{H}_\mathrm{f}$. For a single self-attention layer, the functional Hessian w.r.t. the value weight matrix $\mathbf{H}_\mathrm{f}\left(\mathbf{W}_\mathrm{V}, \mathbf{W}_\mathrm{V}\right)$ is zero and the remaining blocks are given by

$$\mathbf{H}_\mathrm{f}\left(\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{Q}\right) = \frac{2}{L d_V d_K} \mathbf{R}_{d_V d_K} \left(\mathbf{I}_L \otimes \mathbf{W}_\mathrm{V}^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_\mathrm{K}^\top\right) \mathbf{Z}_2 \left(\mathbf{I}_{d_V} \otimes \mathbf{W}_\mathrm{K}\right),$$

$$\mathbf{H}_\mathrm{f}\left(\mathbf{W}_\mathrm{V}, \mathbf{W}_\mathrm{Q}\right) = \frac{2}{L d_V \sqrt{d_K}} \mathbf{R}_{d_V^2} \left(\mathbf{I}_L \otimes \mathbf{S}\right) \mathbf{Z}_1 \left(\mathbf{I}_{d_V} \otimes \mathbf{W}_\mathrm{K}\right),$$

$$\begin{aligned}\mathbf{H}_\mathrm{f}\left(\mathbf{W}_\mathrm{Q}, \mathbf{W}_\mathrm{K}\right) = &\frac{2}{L d_V d_K} \mathbf{R}_{d_V d_K} \left(\mathbf{I}_L \otimes \mathbf{W}_\mathrm{V}^\top \otimes \mathbf{I}_{d_V} \otimes \mathbf{W}_\mathrm{K}^\top\right) \mathbf{Z}_2 \left(\mathbf{W}_\mathrm{Q} \otimes \mathbf{I}_{d_V}\right) \mathbf{K}_{d_K, d_V} \\ &+ \frac{2}{L d_V \sqrt{d_K}} \mathbf{R}_{d_V} \left(\mathbf{I}_L \otimes \mathbf{W}_\mathrm{V}^\top \otimes \mathbf{I}_{d_V}\right) \left(\mathbf{Z}_1 \otimes \mathbf{I}_{d_V}\right) \mathbf{S} \otimes \mathbf{I}_{d_K},\end{aligned}$$

with the residual $\mathbf{R}_m := \mathrm{vec}_r\left(\mathbf{F}(\mathbf{X}) - \mathbf{Y}\right)^\top \otimes \mathbf{I}_m \in \mathbb{R}^{m \times m L d_V}$,
a shuffling matrix $\mathbf{S} := \left(\mathbf{I}_{d_V} \otimes \mathbf{K}_{d_V, d_V}\right)\left(\mathrm{vec}_r\, \mathbf{I}_{d_V} \otimes \mathbf{I}_{d_V}\right) \in \mathbb{R}^{d_V^3 \times d_V}$,
and $\mathbf{Z}_2 := \left(\mathbf{I}_L \otimes \mathbf{K}_{d_V, d_V} \otimes \mathbf{I}_{d_V}\right)\left(\mathbf{X} * \mathbf{X}^\top * \mathbf{M}_3\right) \in \mathbb{R}^{L d_V^3 \times d_V^2}$ with $\mathbf{M}_2$ and $\mathbf{M}_3$ containing moments of the distributions given by self-attention scores.

Let’s simplify with big O notation

Let's simplify with big O notation

We summarise the dependencies using big O notation

Let's simplify with big O notation

We summarise the dependencies using big O notation

Definition

For matrices $\mathbf{A}, \mathbf{B}$ *,* $f \in \mathcal{O}(\mathbf{AB})$ *means that all entries of* f *are* $\mathcal{O}(A_{i,j}B_{k,l})$ *for any* i, j, k, l

Let's simplify with big O notation

We summarise the dependencies using big O notation

Definition

For matrices $\mathbf{A}, \mathbf{B}$ *,* $f \in \mathcal{O}(\mathbf{AB})$ *means that all entries of* f *are* $\mathcal{O}(A_{i,j}B_{k,l})$ *for any* i, j, k, l

We omit the softmax expressions as they are bounded to [0,1]

Heterogeneity in big O notation

Data

Heterogeneity in big O notation

Data

- Self-attention Hessian shows heterogeneous dependence on data

Heterogeneity in big O notation

Data

- Self-attention Hessian shows heterogeneous dependence on data

$$\mathbf{H}_o \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

Data

- Self-attention Hessian shows heterogeneous dependence on data

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot (\mathbf{X}^5 + \mathbf{X}^3)) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

Data

Self-attention Hessian shows heterogeneous dependence on data

Predicted growth rates generalise to deeper Transformers

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot (\mathbf{X}^5 + \mathbf{X}^3)) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

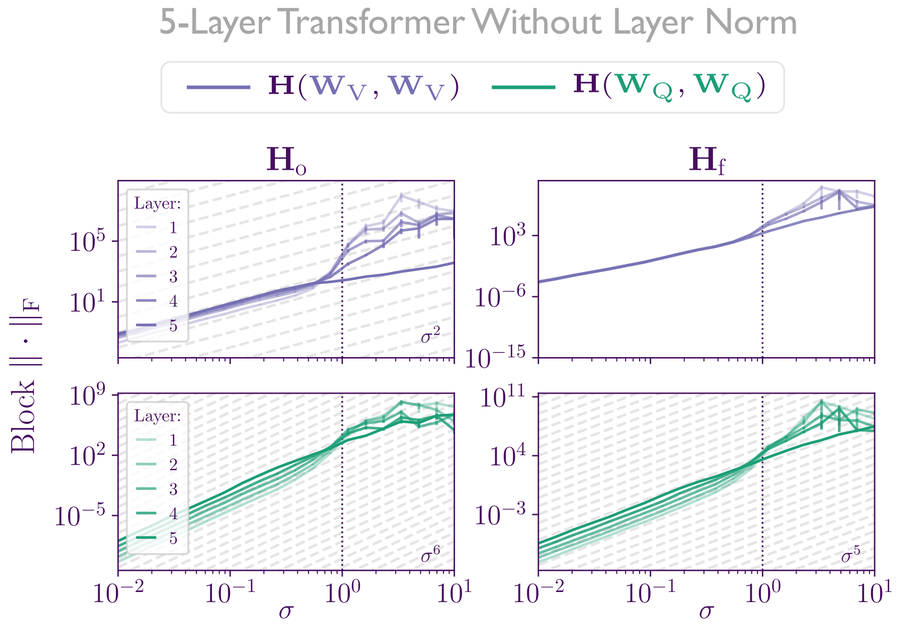
Data

Self-attention Hessian shows heterogeneous dependence on data

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot (\mathbf{X}^5 + \mathbf{X}^3)) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Predicted growth rates generalise to deeper Transformers



Heterogeneity in big O notation

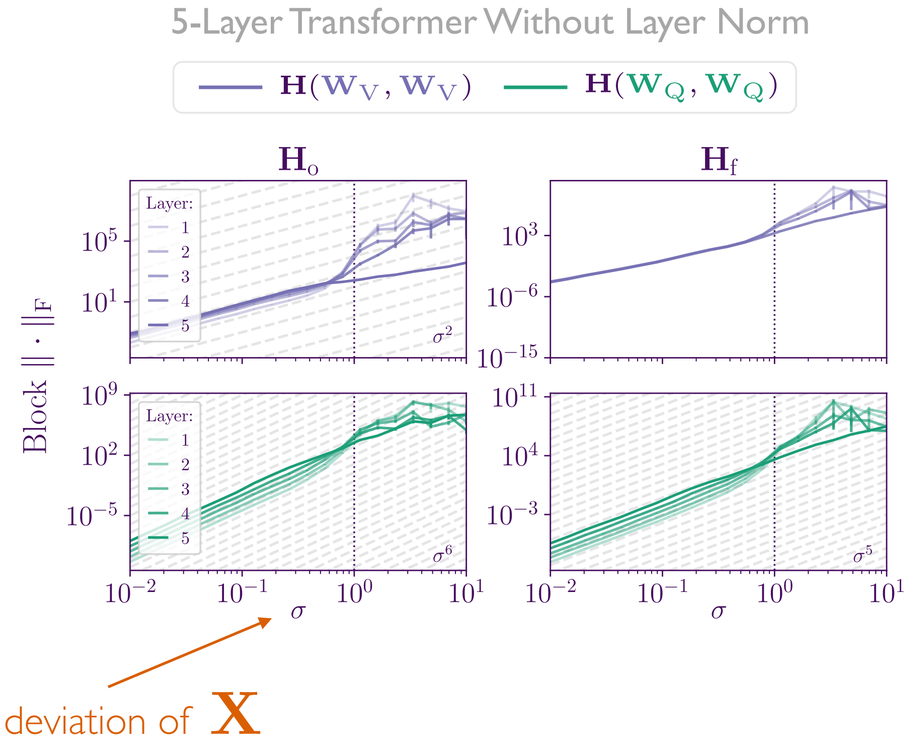
Data

Self-attention Hessian shows heterogeneous dependence on data

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot (\mathbf{X}^5 + \mathbf{X}^3)) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \mathcal{O}(\partial\ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial\ell \cdot \mathbf{X}^3) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Predicted growth rates generalise to deeper Transformers



Heterogeneity in big O notation

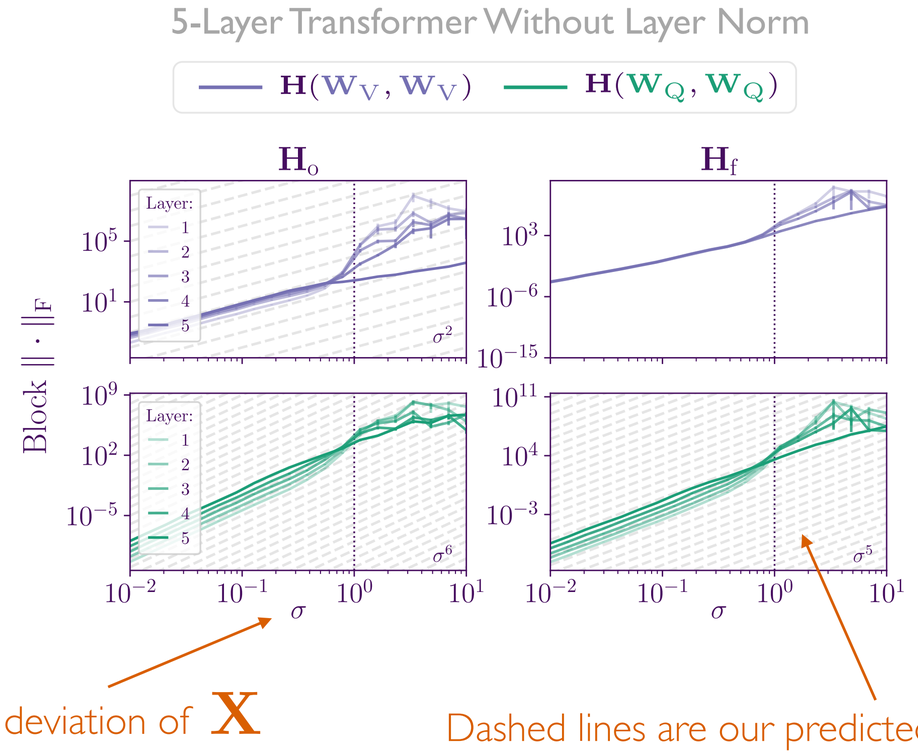
Data

Self-attention Hessian shows heterogeneous dependence on data

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

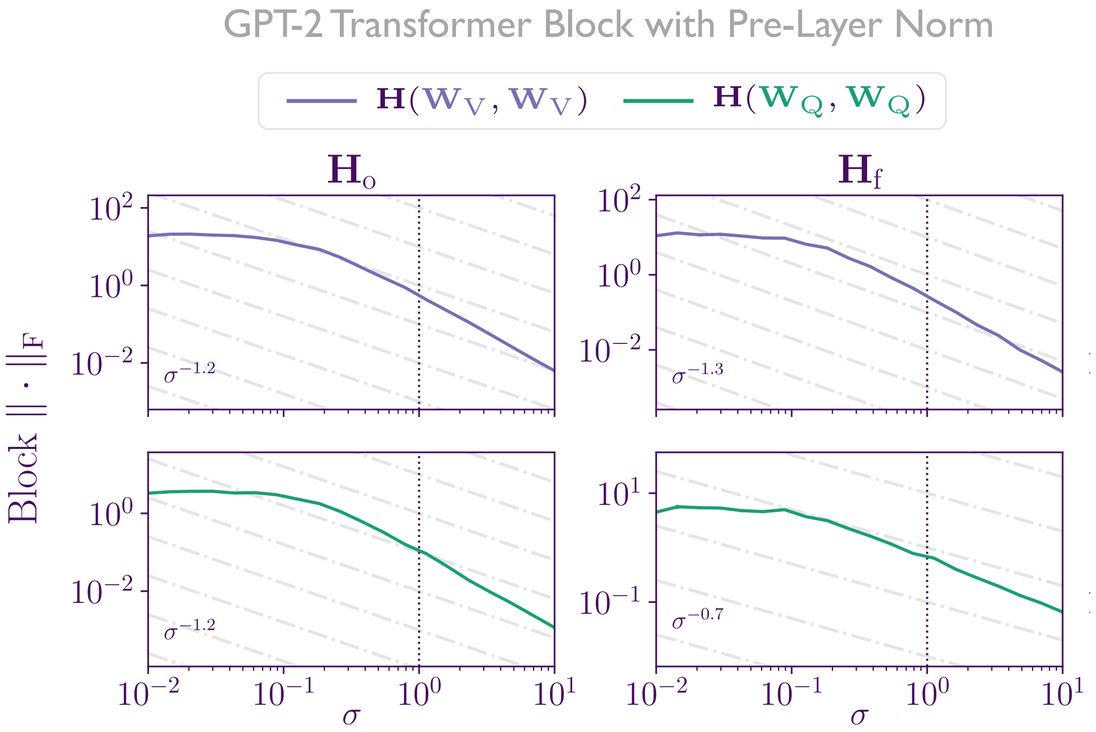
$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial \ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial \ell \cdot (\mathbf{X}^5 + \mathbf{X}^3)) & \mathcal{O}(\partial \ell \cdot \mathbf{X}^3) \\ \cdot & \mathcal{O}(\partial \ell \cdot \mathbf{X}^5) & \mathcal{O}(\partial \ell \cdot \mathbf{X}^3) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Predicted growth rates generalise to deeper Transformers

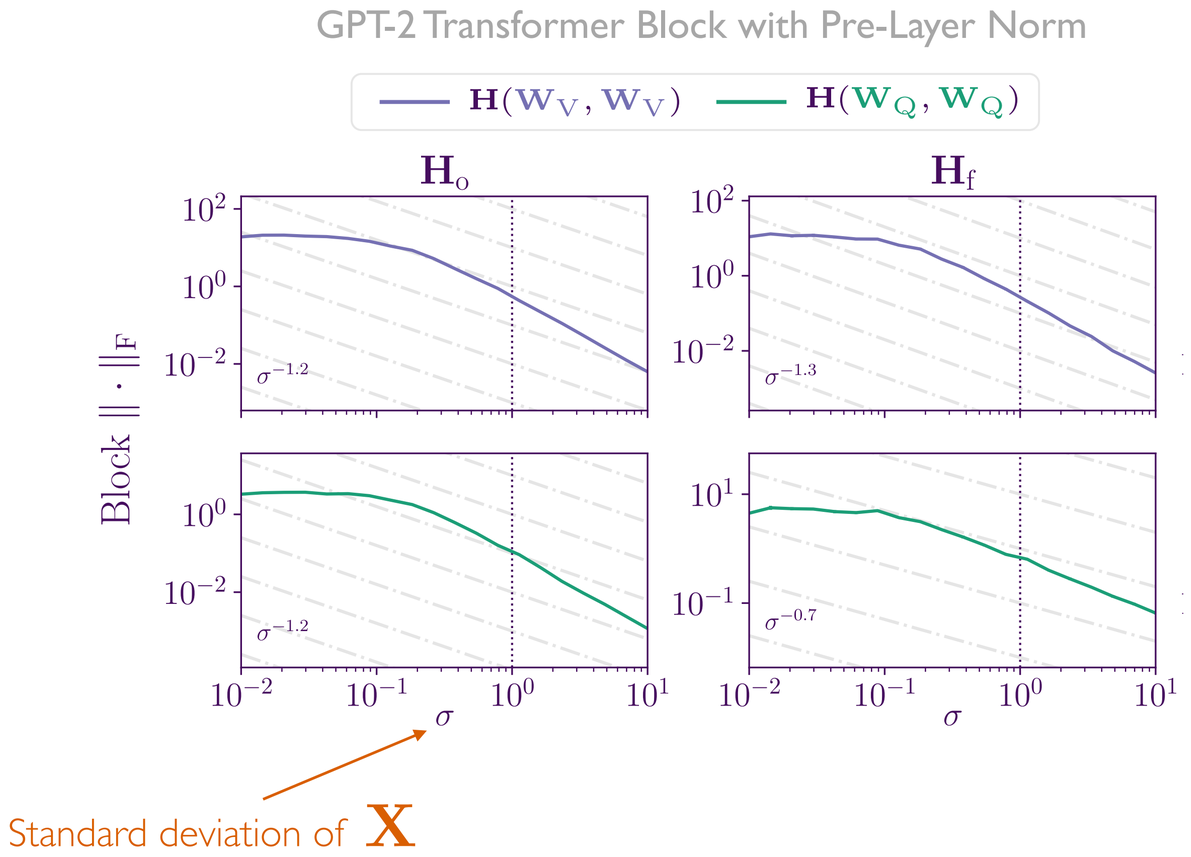


Partial Fix: Pre-Layer Norm

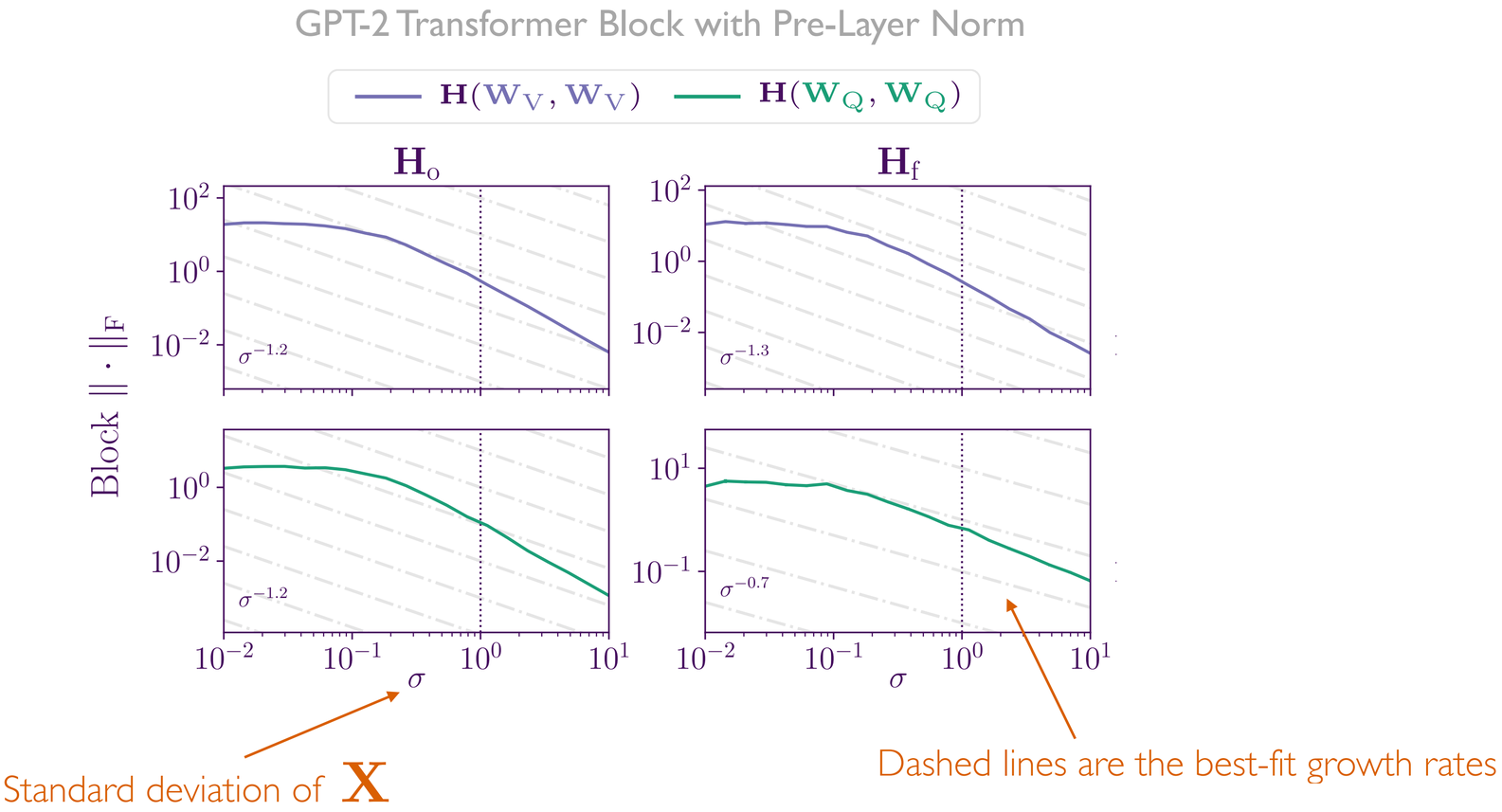
Partial Fix: Pre-Layer Norm



Partial Fix: Pre-Layer Norm



Partial Fix: Pre-Layer Norm



Heterogeneity in big O notation

Moments of attention scores distribution

Heterogeneity in big O notation

Moments of attention scores distribution

Definition

Every row of the attention matrix is a normalised distribution over input tokens. We can compute their (centred) moments:

$$\mathbf{M}_1 := \mathbf{A}\mathbf{X} = \left[\mathbf{A}_{i,:}^\top \mathbf{X}\right]_{1 \le i \le L}$$

$$\mathbf{M}_2 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j} \left(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}\right) \left(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}\right)^\top\right]_{1 \le i \le L}$$

$$\mathbf{M}_3 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j} (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}) \otimes (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}) (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})^\top\right]_{1 \le i \le L}$$

Heterogeneity in big O notation

Moments of attention scores distribution

Definition

Every row of the attention matrix is a normalised distribution over input tokens. We can compute their (centred) moments:

$$\mathbf{M}_1 := \mathbf{A}\mathbf{X} = \left[\mathbf{A}_{i,:}^\top \mathbf{X}\right]_{1 \le i \le L}$$

$$\mathbf{M}_2 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j} \left(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}\right) \left(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}\right)^\top\right]_{1 \le i \le L}$$

$$\mathbf{M}_3 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j} (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}) \otimes (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}) (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})^\top\right]_{1 \le i \le L}$$

$$\mathbf{H}_\text{o} \in \begin{array}{c} \\ \text{Q} \\ \text{K} \\ \text{V} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Q} & \text{K} & \text{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_1\mathbf{M}_2) \\ \cdot & \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_1\mathbf{M}_2) \\ \cdot & \cdot & \mathcal{O}(\mathbf{M}_1^2) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

Moments of attention scores distribution

Definition

Every row of the attention matrix is a normalised distribution over input tokens. We can compute their (centred) moments:

$$\mathbf{M}_1 := \mathbf{A}\mathbf{X} = \left[\mathbf{A}_{i,:}^\top \mathbf{X}\right]_{1\le i\le L}$$

$$\mathbf{M}_2 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j}\,(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})\,(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})^\top\right]_{1\le i\le L}$$

$$\mathbf{M}_3 := \left[\sum_{j=1}^{L} \mathbf{A}_{i,j}(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:}) \otimes (\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})(\mathbf{X}_{j,:} - [\mathbf{M}_1]_{i,:})^\top\right]_{1\le i\le L}$$

$$\mathbf{H}_\mathrm{o} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_1\mathbf{M}_2) \\ \cdot & \mathcal{O}(\mathbf{M}_2^2) & \mathcal{O}(\mathbf{M}_1\mathbf{M}_2) \\ \cdot & \cdot & \mathcal{O}(\mathbf{M}_1^2) \end{bmatrix} \end{array} \qquad \mathbf{H}_\mathrm{f} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial\ell \cdot \mathbf{M}_3) & \mathcal{O}(\partial\ell \cdot (\mathbf{M}_3 + \mathbf{M}_2)) & \mathcal{O}(\partial\ell \cdot \mathbf{M}_2 \\ \cdot & \mathcal{O}(\partial\ell \cdot \mathbf{M}_3) & \mathcal{O}(\partial\ell \cdot \mathbf{M}_2) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

Weights

Heterogeneity in big O notation

Weights

$$\mathbf{H}_o \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{W}_\mathrm{K}^2 \mathbf{W}_\mathrm{V}^2) & \mathcal{O}(\mathbf{W}_\mathrm{K} \mathbf{W}_\mathrm{Q} \mathbf{W}_\mathrm{V}^2) & \mathcal{O}(\mathbf{W}_\mathrm{K} \mathbf{W}_\mathrm{V}) \\ \cdot & \mathcal{O}(\mathbf{W}_\mathrm{Q}^2 \mathbf{W}_\mathrm{V}^2) & \mathcal{O}(\mathbf{W}_\mathrm{Q} \mathbf{W}_\mathrm{V}) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Heterogeneity in big O notation

Weights

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{W}_{\mathrm{K}}^2 \mathbf{W}_{\mathrm{V}}^2) & \mathcal{O}(\mathbf{W}_{\mathrm{K}} \mathbf{W}_{\mathrm{Q}} \mathbf{W}_{\mathrm{V}}^2) & \mathcal{O}(\mathbf{W}_{\mathrm{K}} \mathbf{W}_{\mathrm{V}}) \\ \cdot & \mathcal{O}(\mathbf{W}_{\mathrm{Q}}^2 \mathbf{W}_{\mathrm{V}}^2) & \mathcal{O}(\mathbf{W}_{\mathrm{Q}} \mathbf{W}_{\mathrm{V}}) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

$$\mathbf{H}_{\mathrm{f}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\partial \ell \cdot \mathbf{W}_{\mathrm{K}}^2 \mathbf{W}_{\mathrm{V}}) & \mathcal{O}(\partial \ell \cdot (\mathbf{W}_{\mathrm{K}} \mathbf{W}_{\mathrm{Q}} \mathbf{W}_{\mathrm{V}} + \mathbf{W}_{\mathrm{V}})) & \mathcal{O}(\partial \ell \cdot \mathbf{W}_{\mathrm{K}}) \\ \cdot & \mathcal{O}(\partial \ell \cdot \mathbf{W}_{\mathrm{Q}}^2 \mathbf{W}_{\mathrm{V}}) & \mathcal{O}(\partial \ell \cdot \mathbf{W}_{\mathrm{Q}}) \\ \cdot & \cdot & \mathcal{O}(1) \end{bmatrix} \end{array}$$

Softmax is the source of the heterogeneity

Softmax is the source of the heterogeneity

If there was no softmax in self-attention, so $\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\top}\mathbf{X}^{\top} / \sqrt{d_K}$, then

Softmax is the source of the heterogeneity

If there was no softmax in self-attention, so $\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\top}\mathbf{X}^{\top} / \sqrt{d_K}$, then

$$\mathbf{H}_{\mathrm{o}} \in \mathcal{O}(\boldsymbol{\Sigma}_{\mathbf{XX}}^{3})$$

where

$$\boldsymbol{\Sigma}_{\mathbf{XX}} := \mathbf{X}^{\top}\mathbf{X} / L$$

Softmax is the source of the heterogeneity

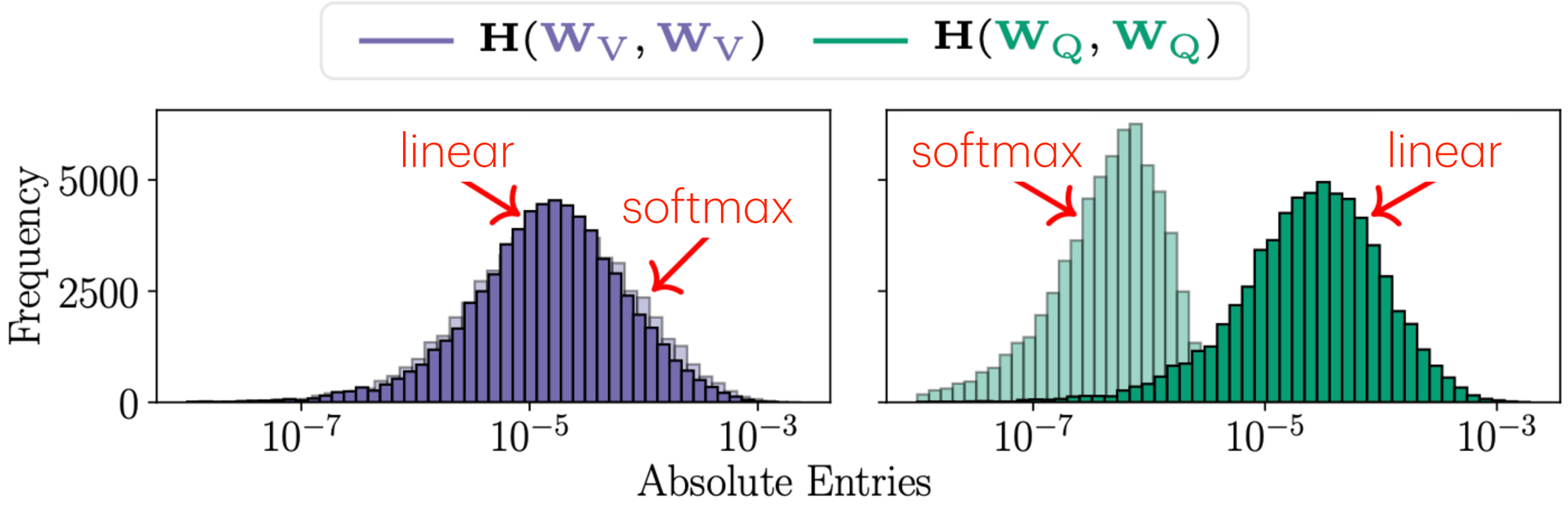
If there was no softmax in self-attention, so $\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathrm{Q}}\mathbf{W}_{\mathrm{K}}^{\top}\mathbf{X}^{\top} / \sqrt{d_K}$, then

$$\mathbf{H}_{\mathrm{o}} \in \mathcal{O}(\boldsymbol{\Sigma}_{\mathbf{XX}}^{3}) \quad \text{and} \quad \mathbf{H}_{\mathrm{f}} \in \mathcal{O}(\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{XX}})$$

where

$$\boldsymbol{\Sigma}_{\mathbf{XX}} := \mathbf{X}^{\top}\mathbf{X} / L \quad \text{and} \quad \boldsymbol{\Omega}_{\mathbf{xy}} := \mathbf{X}^{\top}(\mathbf{F}(\mathbf{X}) - \mathbf{Y}) / L$$

Softmax is the source of the heterogeneity



Hessian for self-attention is more data-dependent than that of MLP

Hessian for self-attention is more data-dependent than that of MLP

We showed that for self-attention

$$\mathbf{H}_{\mathrm{o}} \in \mathcal{O}(\boldsymbol{\Sigma}_{\mathbf{XX}}^{3}) \quad \text{and} \quad \mathbf{H}_{\mathrm{f}} \in \mathcal{O}(\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{XX}})$$

Hessian for self-attention is more data-dependent than that of MLP

We showed that for self-attention

$$\mathbf{H}_{\mathrm{o}} \in \mathcal{O}(\boldsymbol{\Sigma}_{\mathbf{XX}}^{3}) \quad \text{and} \quad \mathbf{H}_{\mathrm{f}} \in \mathcal{O}(\boldsymbol{\Omega}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{XX}})$$

From Singh et al. (2023) we know that for a linear MLP of any depth

$$\mathbf{H}_{\mathrm{o}} \in \mathcal{O}(\boldsymbol{\Sigma}_{\mathbf{XX}}) \quad \text{and} \quad \mathbf{H}_{\mathrm{f}} \in \mathcal{O}(\boldsymbol{\Omega}_{\mathbf{xy}})$$

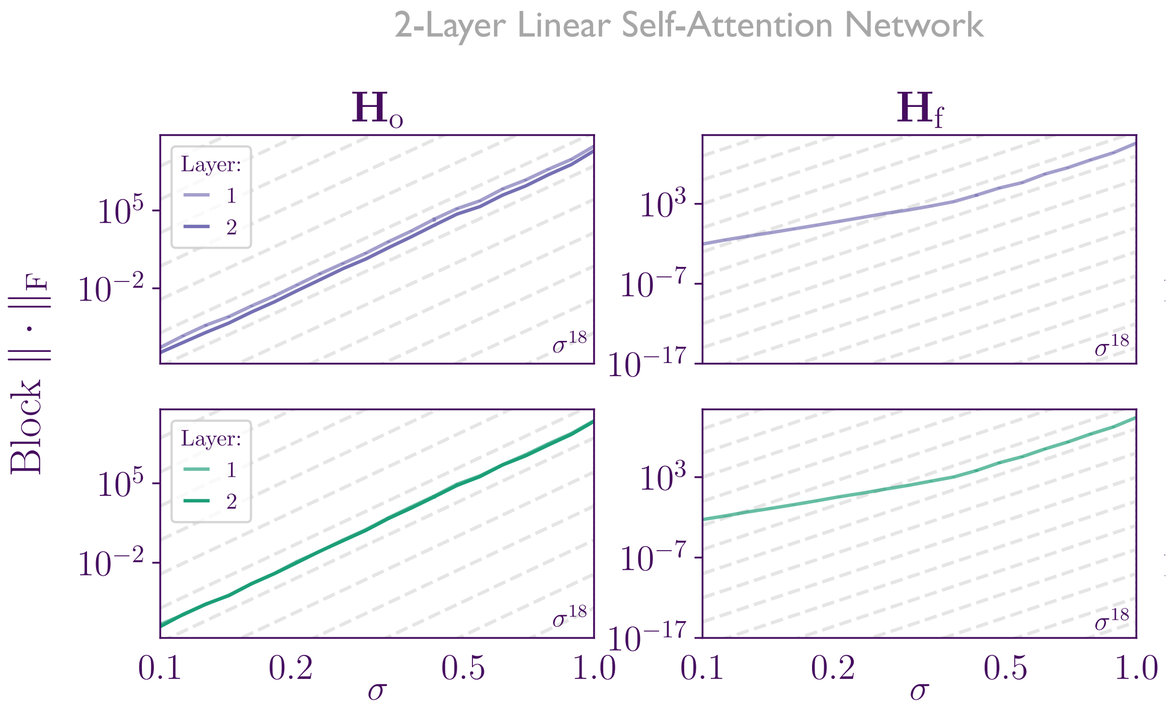
Hessian for self-attention is more data-dependent than that of MLP

Hessian for self-attention is more data-dependent than that of MLP

Contrary to MLPs, data dependence of the linear Transformer Hessian grows with depth

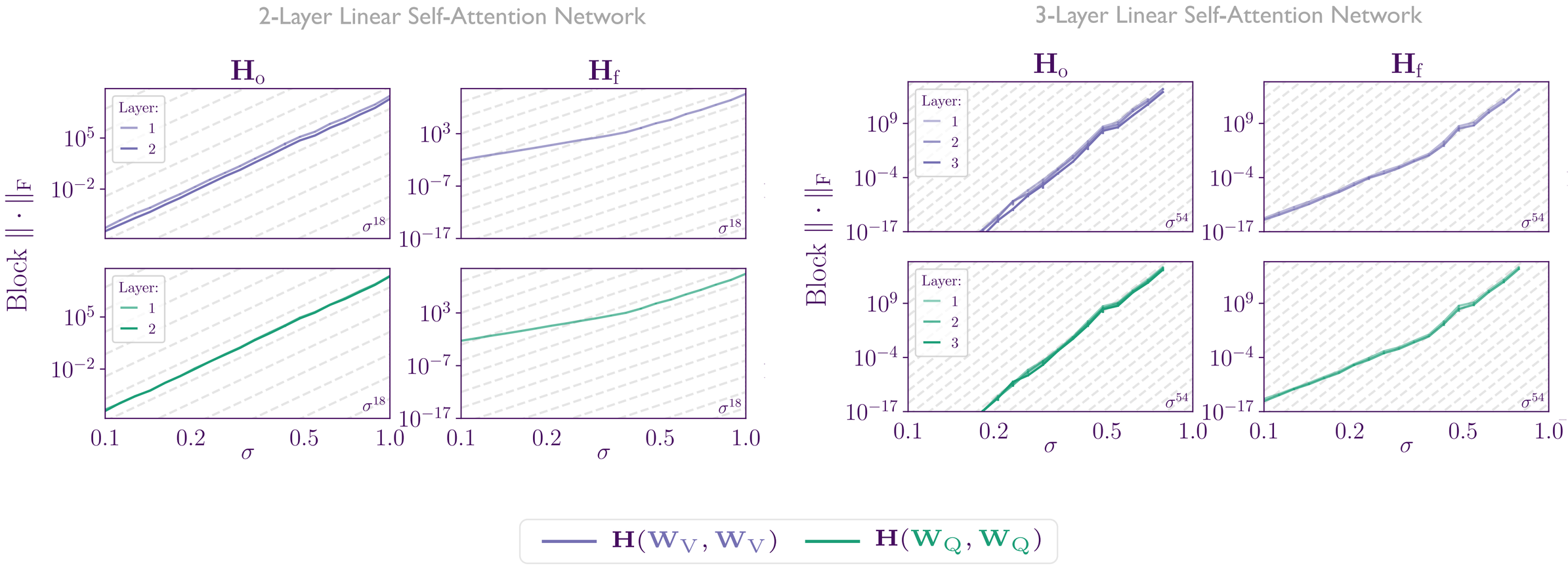
Hessian for self-attention is more data-dependent than that of MLP

Contrary to MLPs, data dependence of the linear Transformer Hessian grows with depth



Hessian for self-attention is more data-dependent than that of MLP

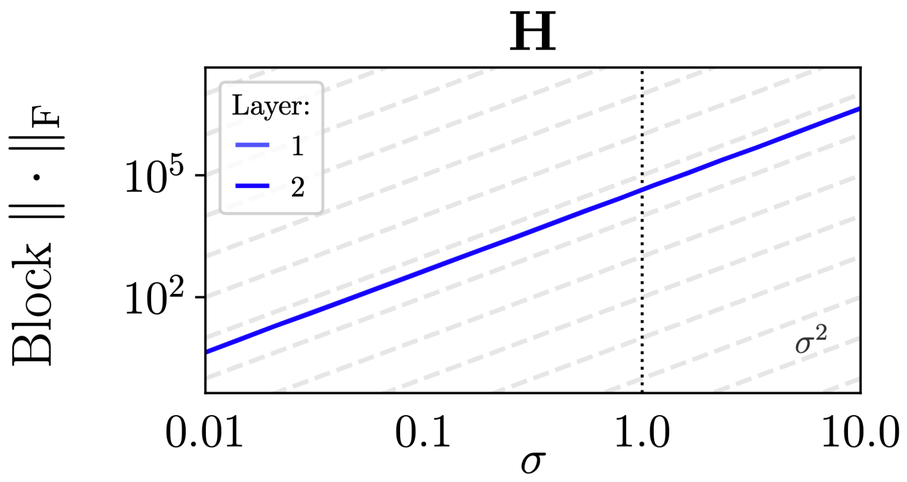
Contrary to MLPs, data dependence of the linear Transformer Hessian grows with depth



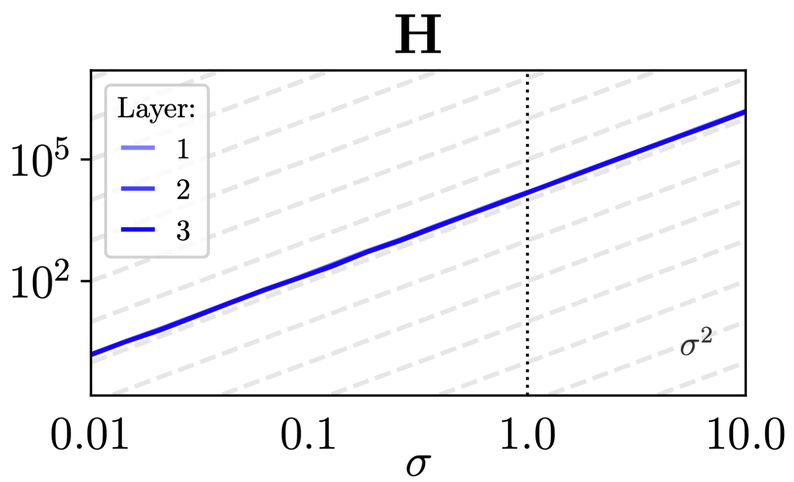
Hessian for self-attention is more data-dependent than that of MLP

Contrary to MLPs, data dependence of the linear Transformer Hessian grows with depth

2-Layer Linear MLP



3-Layer Linear MLP



What about Hessian eigenvalues?

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

What about Hessian eigenvalues?

Assume the following model:

$$f_{w_V, \mathbf{w}_Q, \mathbf{w}_K}(\mathbf{x}) := w_V \mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \text{softmax}\left(\frac{1}{\sqrt{d_K}} \mathbf{x}\mathbf{w}_Q{}^\top \mathbf{w}_K \mathbf{x}^\top\right)$$

$$\mathbf{w}_Q, \mathbf{w}_K \in \mathbb{R}^{d_K} \qquad w_V \in \mathbb{R}$$

In this model we can derive all loss Hessian eigenvalues in closed form!

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

The loss Hessian spectrum consists of:

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

The loss Hessian spectrum consists of:

- $d_{\mathrm{K}} - 1$ pairs of bulk eigenvalues which are symmetric around zero

What about Hessian eigenvalues?

Assume the following model:

$$f_{w_V,\mathbf{w}_Q,\mathbf{w}_K}(\mathbf{x}) := w_V \mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \text{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_Q{}^\top \mathbf{w}_K \mathbf{x}^\top\right)$$

$$\mathbf{w}_Q, \mathbf{w}_K \in \mathbb{R}^{d_K} \qquad w_V \in \mathbb{R}$$

The loss Hessian spectrum consists of:

- $d_K - 1$ pairs of bulk eigenvalues which are symmetric around zero
- 3 outlier eigenvalues with potentially much larger magnitude

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

We learn that:

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

We learn that:

- Hessian condition number is sensitive to the norm of the query and key weight vectors

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

We learn that:

- Hessian condition number is sensitive to the norm of the query and key weight vectors
- GGN-based approximations of the largest eigenvalue are worse for Transformers than for MLPs

What about Hessian eigenvalues?

Assume the following model:

$$\boldsymbol{f}_{w_{\mathrm{V}},\mathbf{w}_{\mathrm{Q}},\mathbf{w}_{\mathrm{K}}}(\mathbf{x}) := w_{\mathrm{V}}\mathbf{A}(\mathbf{x})\mathbf{x} \quad \text{where} \quad \mathbf{A}(\mathbf{x}) = \mathrm{softmax}\left(\frac{1}{\sqrt{d_K}}\mathbf{x}\mathbf{w}_{\mathrm{Q}}{}^{\top}\mathbf{w}_{\mathrm{K}}\mathbf{x}^{\top}\right)$$

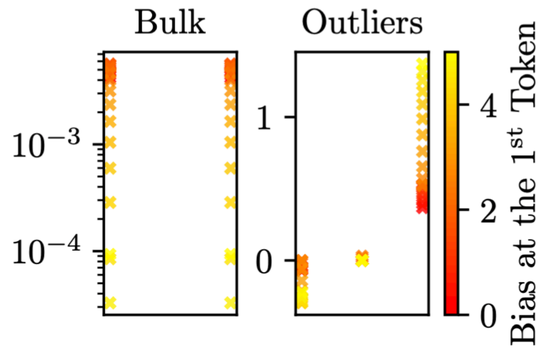
$$\mathbf{w}_{\mathrm{Q}}, \mathbf{w}_{\mathrm{K}} \in \mathbb{R}^{d_K} \qquad w_{\mathrm{V}} \in \mathbb{R}$$

We learn that:

- Hessian condition number is sensitive to the norm of the query and key weight vectors
- GGN-based approximations of the largest eigenvalue are worse for Transformers than for MLPs
- The Hessian spectrum is sensitive to the outlier activations and attention sinks

Outliers in activations and the spectrum

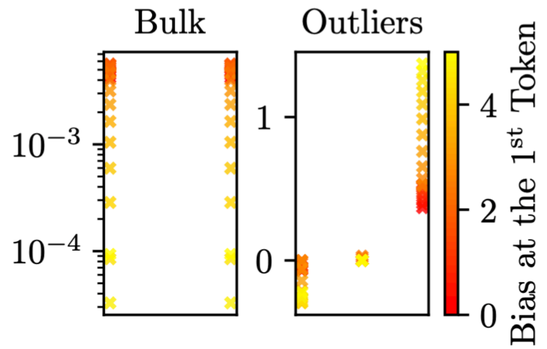
Outliers in activations and the spectrum



(c) Larger bias at the first token (an attention sink) shrinks the bulk and grows the outliers.

Source:
Sidak Pal Singh, Weronika Ormaniec, Thomas Hofmann. *Cracking the Hessian: Closed-Form Hessian Spectra for Fundamental Neural Networks*. 2025 (Preprint)

Outliers in activations and the spectrum



(c) Larger bias at the first token (an attention sink) shrinks the bulk and grows the outliers.

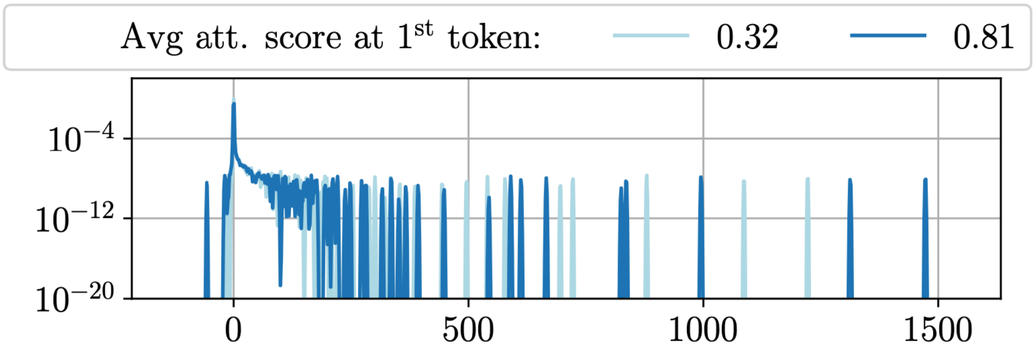


Figure 6: Hessian spectra in the GPT2 model for different magnitudes of attention sinks. Larger attention sinks stretch the spectrum.

Source:
Sidak Pal Singh, Weronika Ormaniec, Thomas Hofmann. *Cracking the Hessian: Closed-Form Hessian Spectra for Fundamental Neural Networks*. 2025 (Preprint)

Summary

Summary

1. Self-attention blocks of the Hessian exhibit heterogenous dependence on the input scale, weights and self-attention moments. The reason is softmax.

Summary

1. Self-attention blocks of the Hessian exhibit heterogenous dependence on the input scale, weights and self-attention moments. The reason is softmax.

2. Transformer Hessian is much more data-dependent than that of MLP. The effect strengthens with depth.

Summary

1. Self-attention blocks of the Hessian exhibit heterogenous dependence on the input scale, weights and self-attention moments. The reason is softmax.

2. Transformer Hessian is much more data-dependent than that of MLP. The effect strengthens with depth.

3. The outliers in self-attention scores have a direct impact on the self-attention Hessian spectrum.

ETH zürich

Thank you!

Questions?

@wormaniec