On the Special Hessian Structure of Neural Networks

Yushun Zhang, Jan, 2026

Presented at SJTU

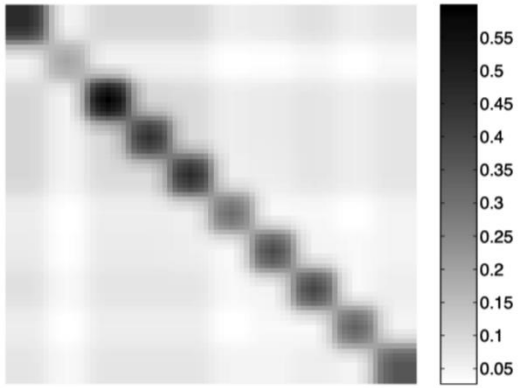
School of Data Science,

The Chinese University of Hong Kong, Shenzhen

Contents

Classical Results in 2004

- **Hessian of NNs are numerically observed to be near-block-diagonal**

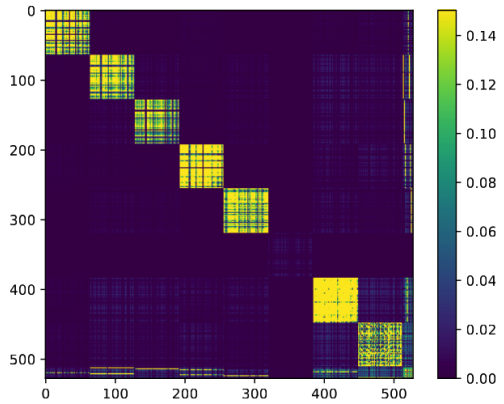


(a) Hessian of an MLP [18] after 1 step

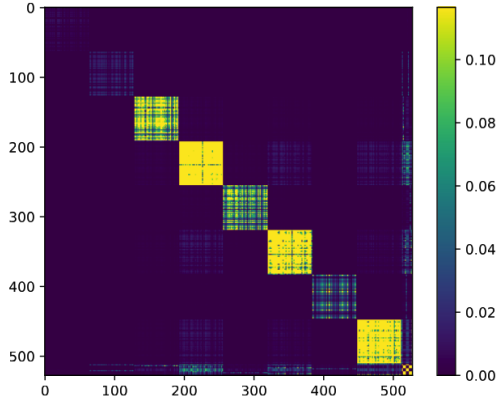
Hessian of a 1-hidden-layer NN

Figure from: Large Scale Machine Learning, Collobert, thesis, 2004

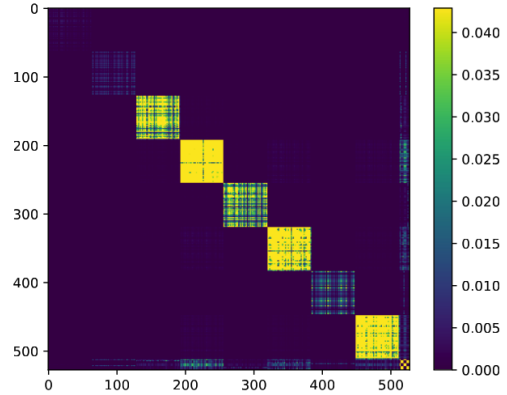
We reproduced the classical results recently



(b) Hessian of an MLP at 1% step



(c) Hessian of an MLP at 50% step



(d) Hessian of an MLP at 100% step

Hessian of 1-hidden-layer NNs

Figure (b,c,d): Why Transformers Need Adam: A Hessian Perspective, Zhang, Chen, Ding, Li, Sun, Luo, NeurIPS 2024

Hessian of Transformers?

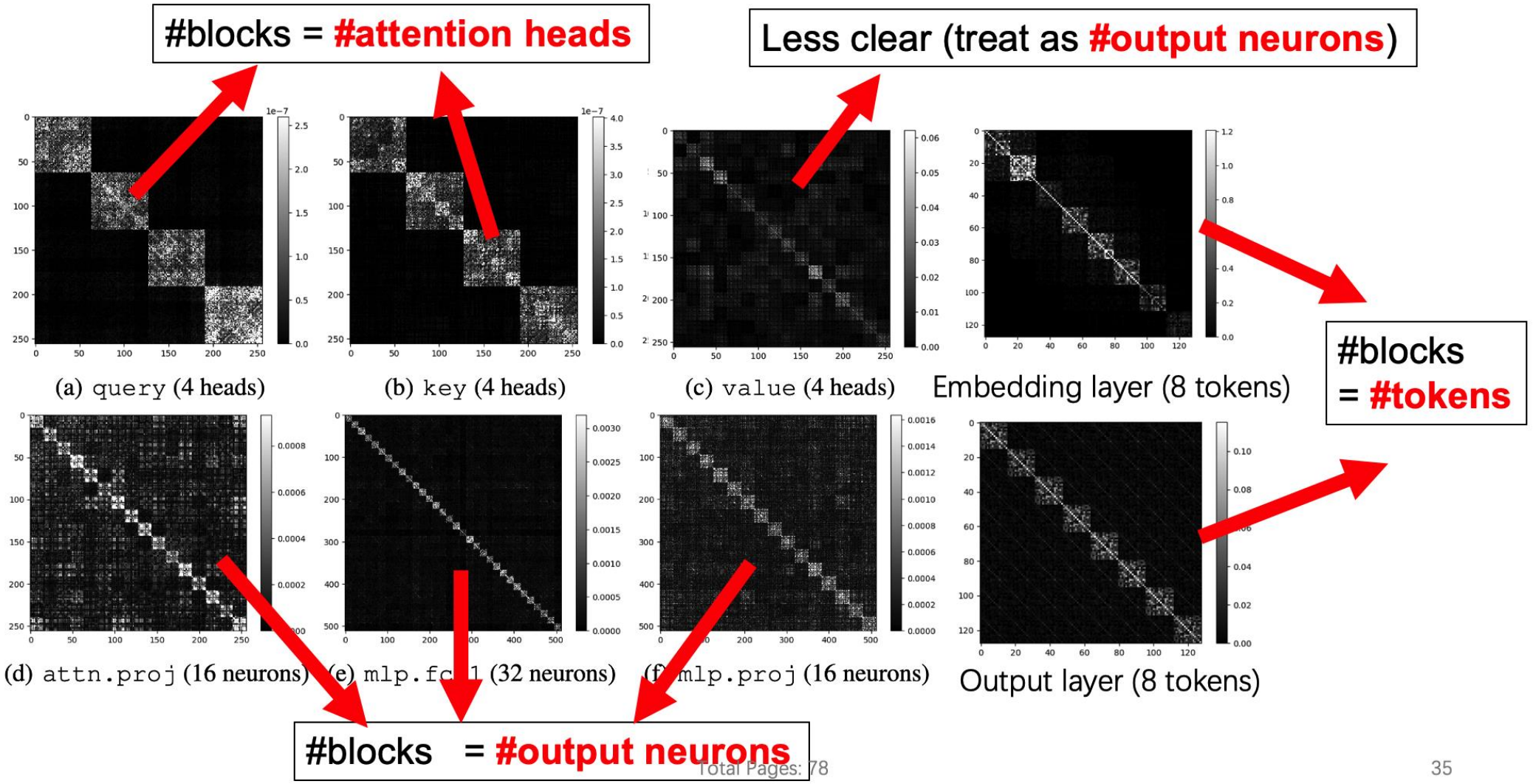
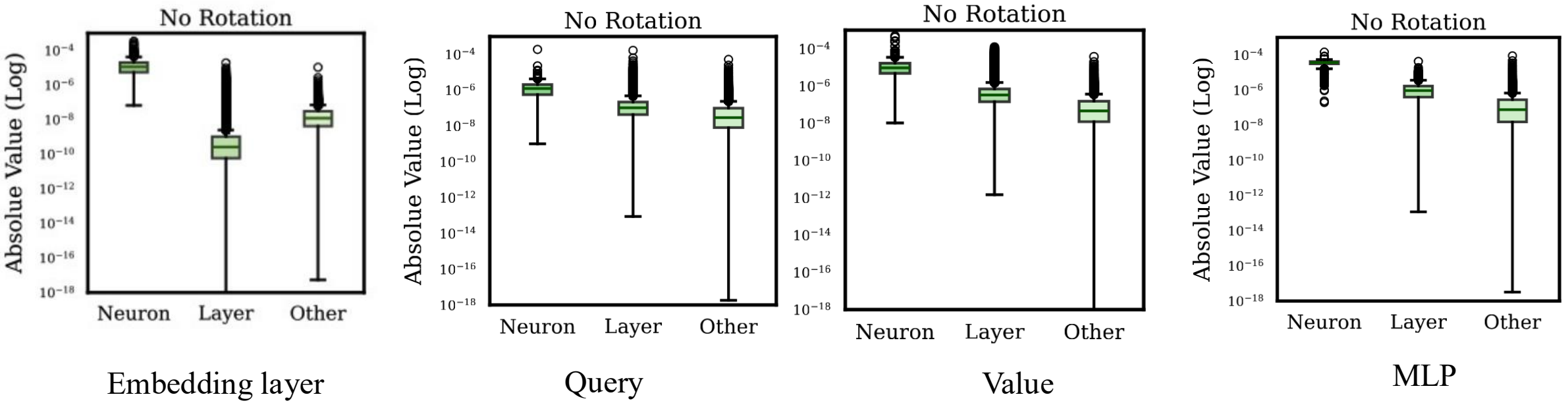


Figure from: Adam-mini: Use Fewer Learning Rates To Gain More, Zhang*, Chen*, …, Luo, Sun, ICLR 2025

Follow-up observations from community

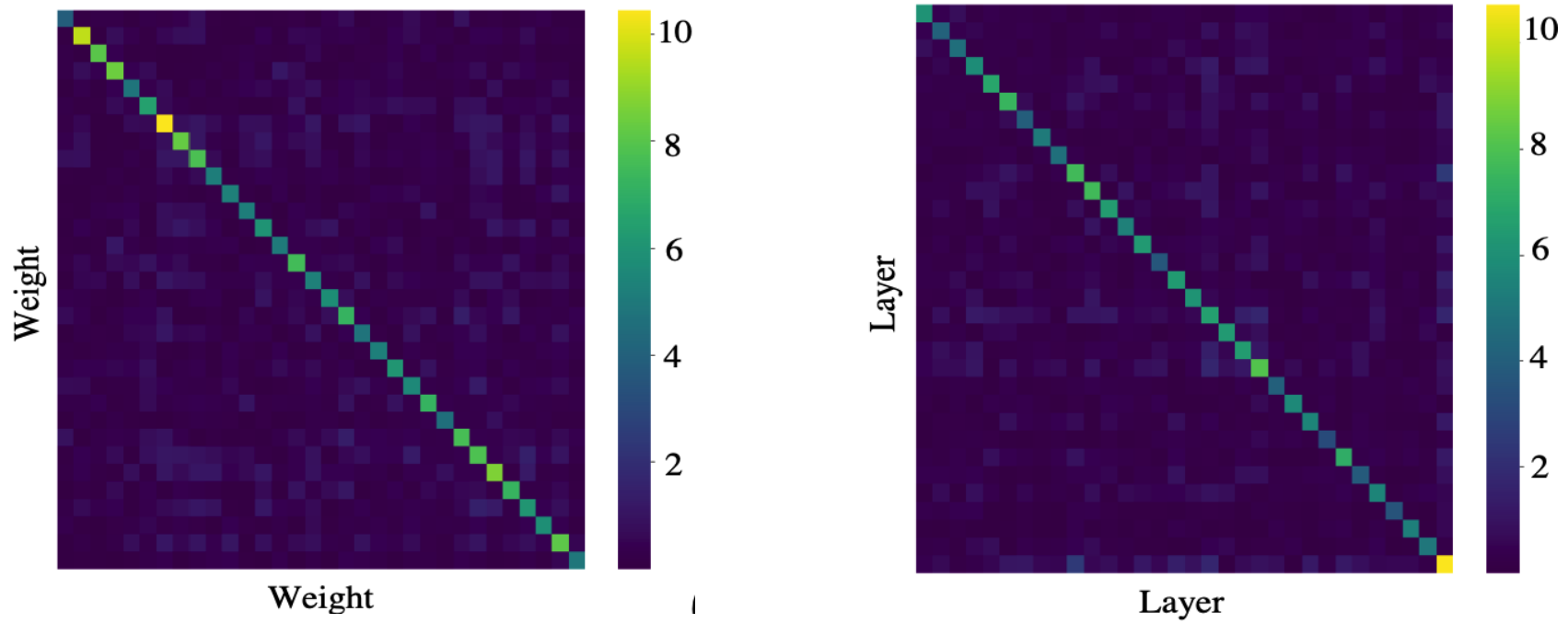


Embedding layer | Query | Value | MLP

Hessian sub-blocks sampled from GPT2-125M (diag-blocks > 10^4 off-diag-blocks)

Figure from: Understanding Adam Requires Better Rotation Dependent Assumptions, Maes, et al., 2024

More observations from community



Approximated Hessian of 1 layer in Llama-7B & 32 layers in Llama-7B

Figure from: CBQ: Cross-Block Quantization for Large Language Models, Ding, et al., ICLR 2025

Motivation: Why Studying Hessian Structure?

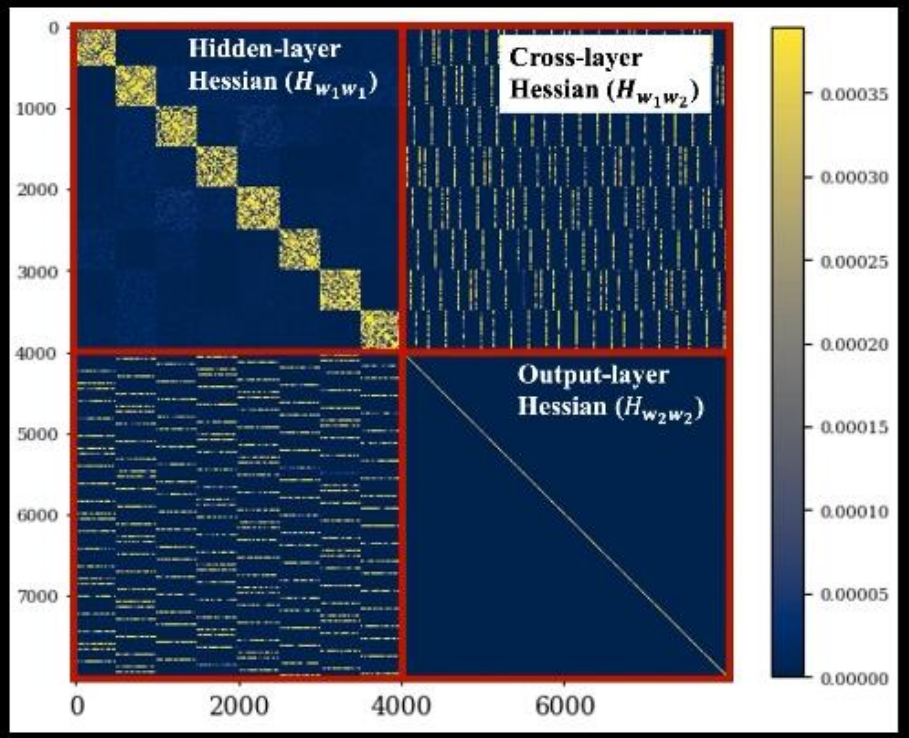
- **1. Hessian structure is crucial for understanding NN training**
  - The effectiveness of Adam
    (Zhang et al 24a, Kunstner et al. 24)
  - The effectiveness of general diagonal-preconditioned methods
    (Sun and Ye, 21, Qu et al. 22, Das et al. 24)
  - The effectiveness of recent block-diagonal-preconditioned methods
    (Shampoo, Muon)
- **2. Hessian structure can help design new training methods for NNs**
- **3. Offering a new function class for optimization community**
  - Typical problems do NOT have such structure:
    In classical non-linear programming dataset (Lavezzi et al 22), all problems have non-block-diag Hessian
  - Motivate new study into this specific class of problems

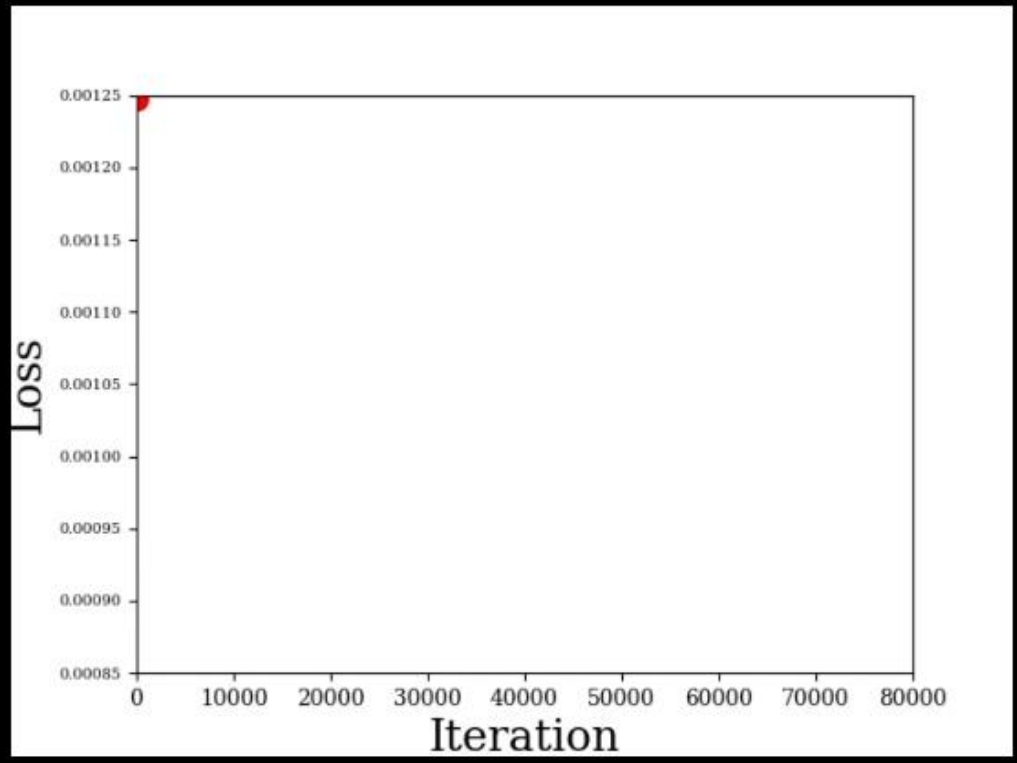
Today, we focus on…

- **Why do Hessian matrices look like this? Is it trivial?**
- **What does one block correspond to?**
- **What is the fundamental reason for this structure?**
  - Does it always hold for arbitrary NNs?
  - If not, is there common factor holds in all above, but we overlooked?
  - Is it a local property or global?
- **Any more structure missed in the previous experiments?**

Hessian throughout training

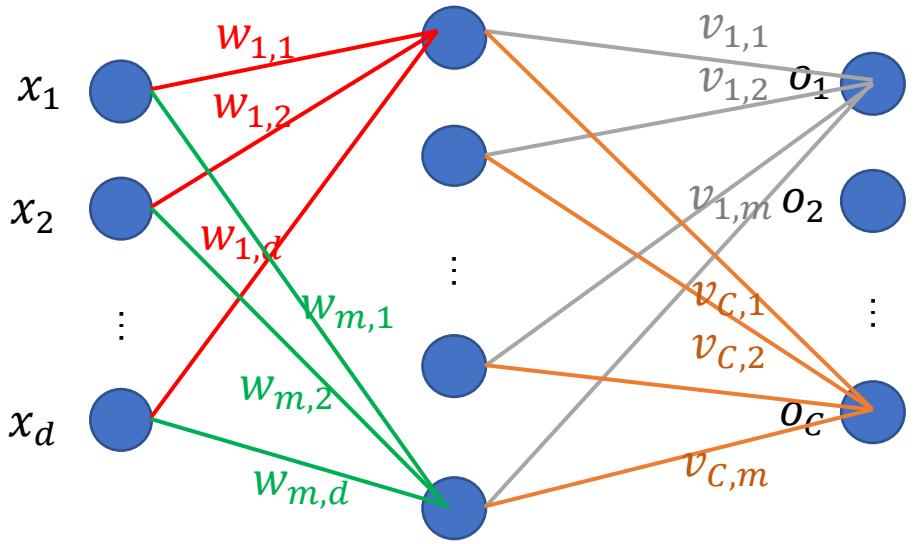
Hessian of a 2-layer relu NN, input dim = # classes = 500, width = 8,
CE loss +Adam, Gaussian data + random label, sample size = 5000





We will go back to 1-hidden-layer NN

Data: $x \in R^d, y \in R^C$ one-hot



Calculate Loss $\mathcal{L}$

Label

0

1

0

$$\min_{W \in R^{m \times d}, V \in R^{m \times C}} \frac{1}{N} \sum_{n=1} \ell(f(x_n), y_n)$$

$$W = \begin{bmatrix} w_1^T \\ \vdots \\ w_m^T \end{bmatrix} \in R^{m \times d}, V = \begin{bmatrix} v_1^T \\ \vdots \\ v_C^T \end{bmatrix} \in R^{C \times m}$$

$$f(x_n) = \begin{bmatrix} v_1^T \sigma(W x_n) \\ \vdots \\ v_C^T \sigma(W x_n) \end{bmatrix} \in R^C$$

$$\min_{W,V} \ell_{\text{MSE}}(W, V) := \frac{1}{N} \sum_{n=1}^{N} \|V\sigma(Wx) - \mathcal{Y}_n\|_2^2,$$

$$\min_{W,V} \ell_{\text{CE}}(W, V) := -\frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{\exp(v_{y_n}^\top \sigma(Wx))}{\sum_{k=1}^{c} \exp(v_k^\top \sigma(Wx))} \right)$$

Review: What is Hessian Matrix for NNs

	d	d	d	m	m	m
d	$H_{w_1w_1}$	$\cdots$	$H_{w_1w_m}$	$H_{w_1v_1}$	$\cdots$	$H_{w_1v_C}$
d		$\ddots$				
d	$H_{w_mw_1}$	$\cdots$	$H_{w_mw_m}$	$H_{w_mv_1}$	$\cdots$	$H_{w_mv_C}$
m	$H_{v_1w_1}$	$\cdots$	$H_{v_1w_m}$	$H_{v_1v_1}$	$\cdots$	$H_{v_1v_C}$
m		$\ddots$			$\ddots$	
m	$H_{v_Cw_1}$	$\cdots$	$H_{v_Cw_1}$	$H_{v_Cv_1}$	$\cdots$	$H_{v_Cv_C}$

Size of Hessian $= (md + Cm) * (md + Cm)$

$$W = \begin{bmatrix} w_1^T \\ \vdots \\ w_m^T \end{bmatrix} \in R^{m\times d}, V = \begin{bmatrix} v_1^T \\ \vdots \\ v_C^T \end{bmatrix} \in R^{C\times m}$$

$$H_{w_iw_i} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial w_i^T} \in R^{d\times d}$$

$$H_{w_iw_j} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial w_j^T} \in R^{d\times d}$$

$$H_{v_iv_i} = \frac{\partial^2 \mathcal{L}}{\partial v_i \partial v_i^T} \in R^{m\times m}$$

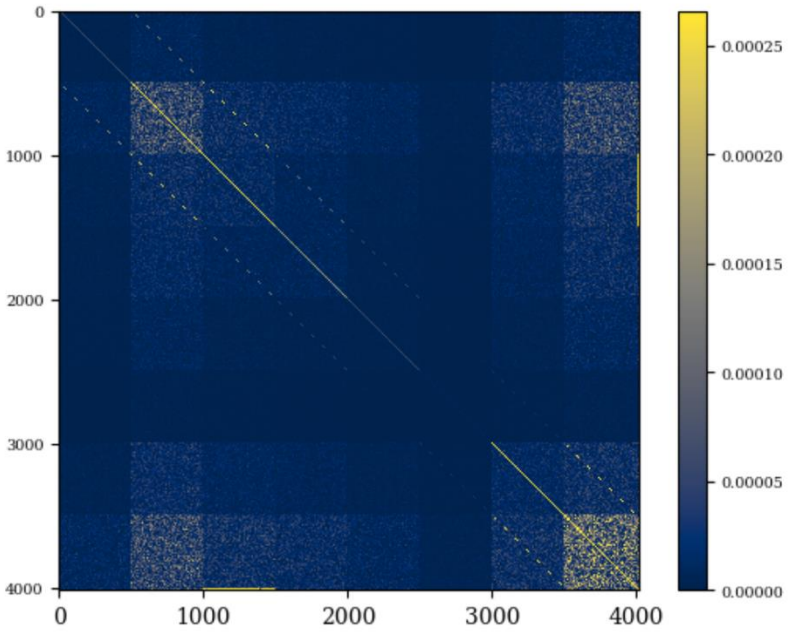
$$H_{v_iv_j} = \frac{\partial^2 \mathcal{L}}{\partial v_i \partial v_j^T} \in R^{m\times m}$$

$$H_{w_iv_i} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial v_i^T} \in R^{d\times m}$$

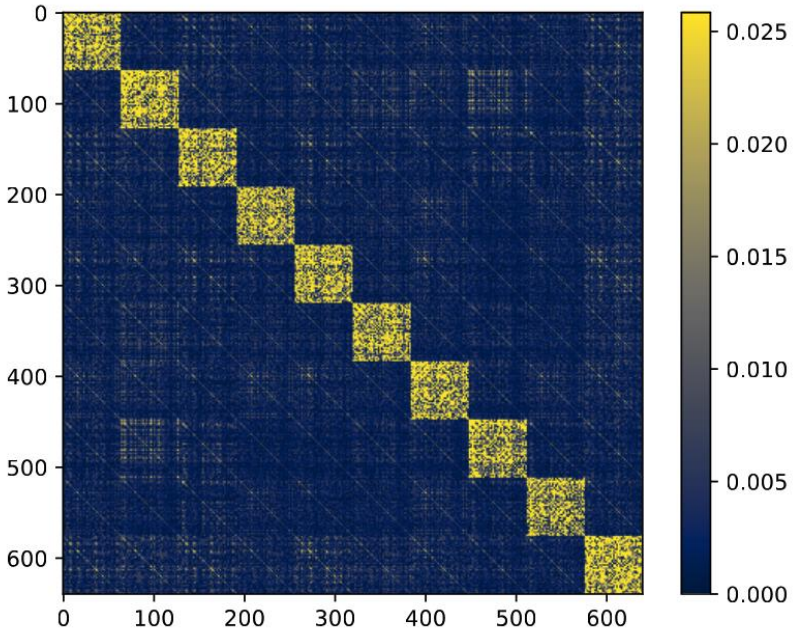
$$H_{w_iv_j} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial v_j^T} \in R^{d\times m}$$

We find a phase transition as # class C → ∞

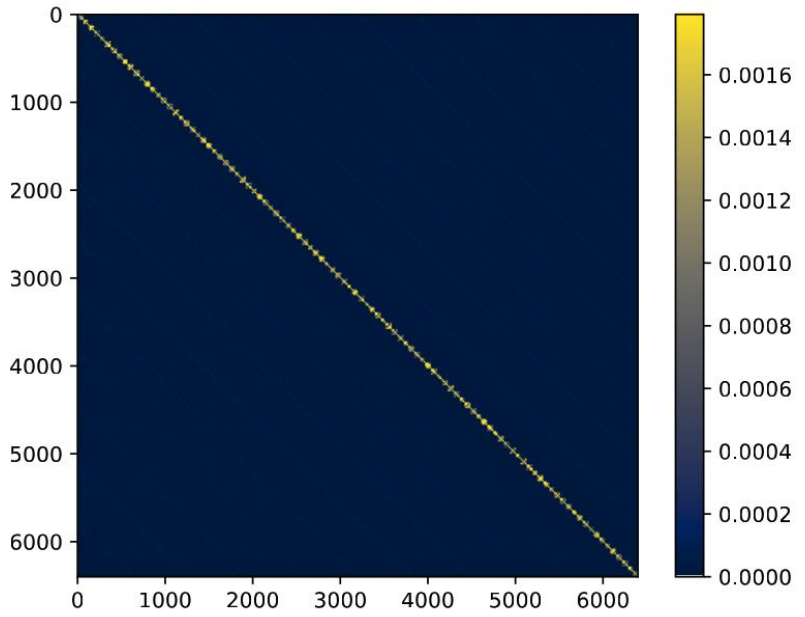
- **Simple setting:** Linear model + CE loss, #C class classification
- **Initial trial:** binary classification C = 2
- **Cannot see special Hessian structures. Why?**



C = 2



C = 10



C = 100

🤔 It seems that **large #class C** is important

We reveal two forces that shape the Hessian structure:

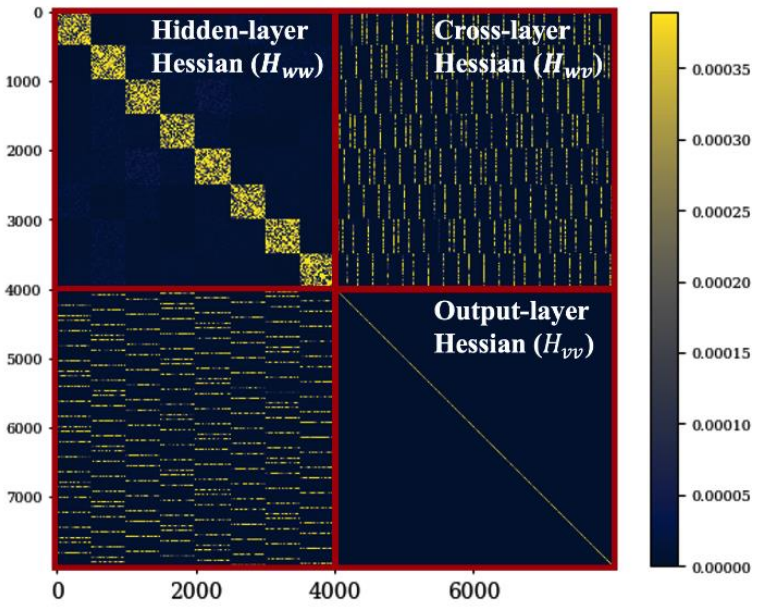
(a) Hessian at initialization

(f) Hessian at 100% steps

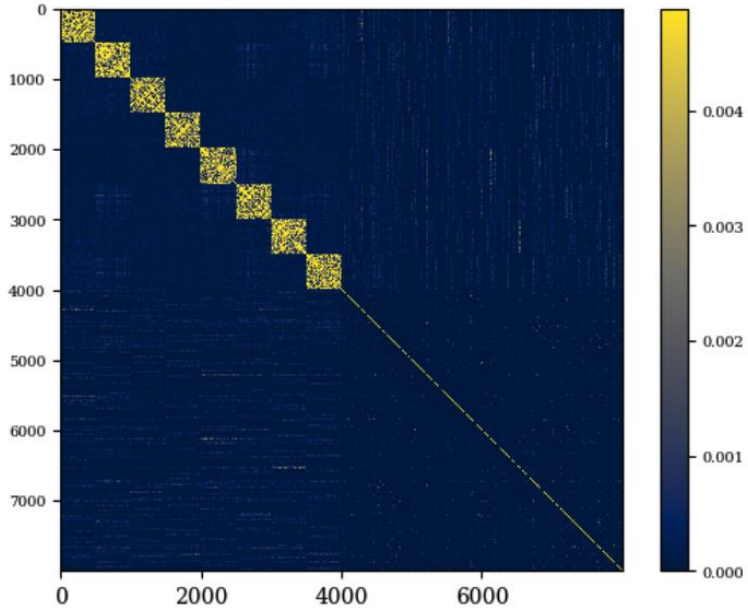
Force I: a ``static force'' rooted in the architecture design;
Force II: and a ``dynamic force'' arisen from training.

- In the following:
  - 1. We first provide intuitions on the structure. **Key cause:** the definition of matrix product
  - 2. a simple explanation on the ``dynamic force''
  - 3. rigorous theory on the ``static force'' at random initialization

“dynamic force”?



(a) Hessian at initialization

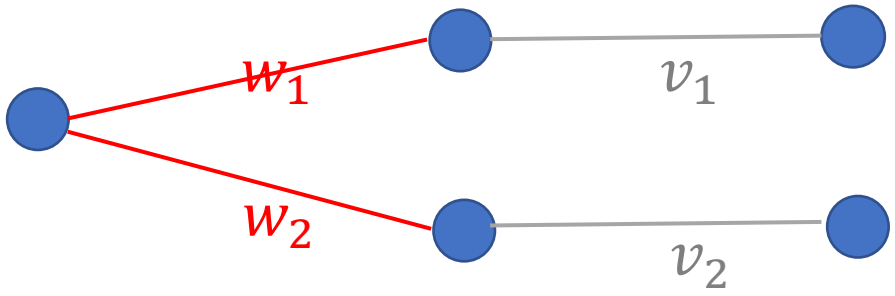


(f) Hessian at 100% steps

Training eliminates the block-circulant structure in H_{wv}. Why?

Warm-up: Example 1

- **Example 1: Single-input-multi-output (SIMO):**
  (this is not a standard NN, but is good for understanding)



check the links!

Multiplicative relation

Gives non-zero Hessian entry

Input data x = 1. No activation, label = 0, MSE loss: $\ell(w_1, w_2, v_1, v_2) = \frac{1}{2}(w_1 v_1)^2 + \frac{1}{2}(w_2 v_2)^2$

Hessian:

	w_1	w_2	v_1	v_2
w_1		0		0
w_2	0		0	
v_1		0		0
v_2	0		0	

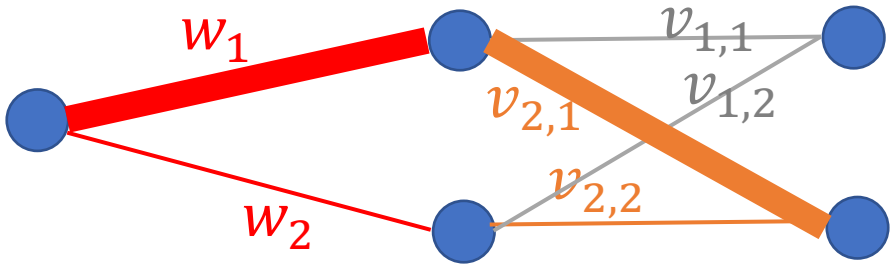
$\frac{\partial^2 \ell}{\partial w_1 \partial w_1} = v_1^2$ $\frac{\partial^2 \ell}{\partial w_1 \partial w_2} = 0$ $\frac{\partial^2 \ell}{\partial w_1 \partial v_1} = 2 w_1 v_1$ $\frac{\partial^2 \ell}{\partial w_1 \partial v_2} = 0$

This is a most simple block-circulant-block-diagonal matrix

Some Hessian entries are naturally 0

Warm-up: Example 2

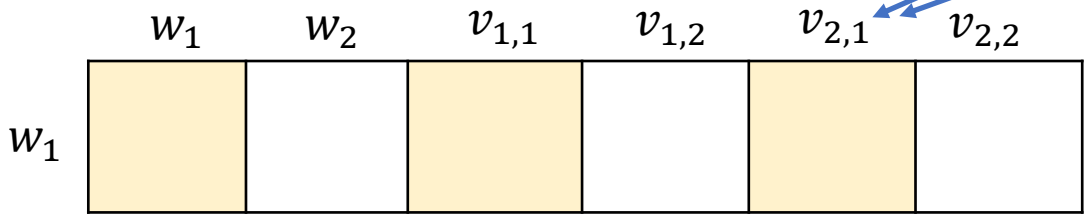
- **Example 2-2: Single-input-multi-output (SIMO):**



Denote $w = (w_1,\ w_2)$,
$v_1 = (v_{1,1}, v_{1,2})$,
$v_2 = (v_{2,1}, v_{2,2})$

$$\ell(w, v_1, v_2) = \frac{1}{2}(v_1^T w)^2 + \frac{1}{2}(v_2^T w)^2 = \frac{1}{2}\left(v_{1,1}w_1 + v_{1,2}w_2\right)^2 + \frac{1}{2}\left(v_{2,1}w_1 + v_{2,2}w_2\right)^2$$

Hessian (1st row):



check the links!

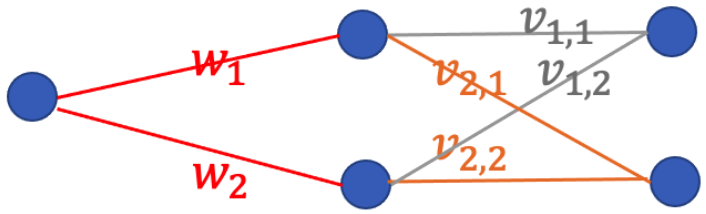
Remark: The white box might not be 0 due to the cross-term,
Need more detailed calculation, but usually the signal can be rather weak
(due to indirect muliplicative relation)

Warm-up: Example 2 (Continued)

$$\ell(w, v_1, v_2) = \frac{1}{2}(v_1^T w)^2 + \frac{1}{2}(v_2^T w)^2 = \frac{1}{2}\left(v_{1,1}w_1 + v_{1,2}w_2\right)^2 + \frac{1}{2}\left(v_{2,1}w_1 + v_{2,2}w_2\right)^2$$

Hessian:

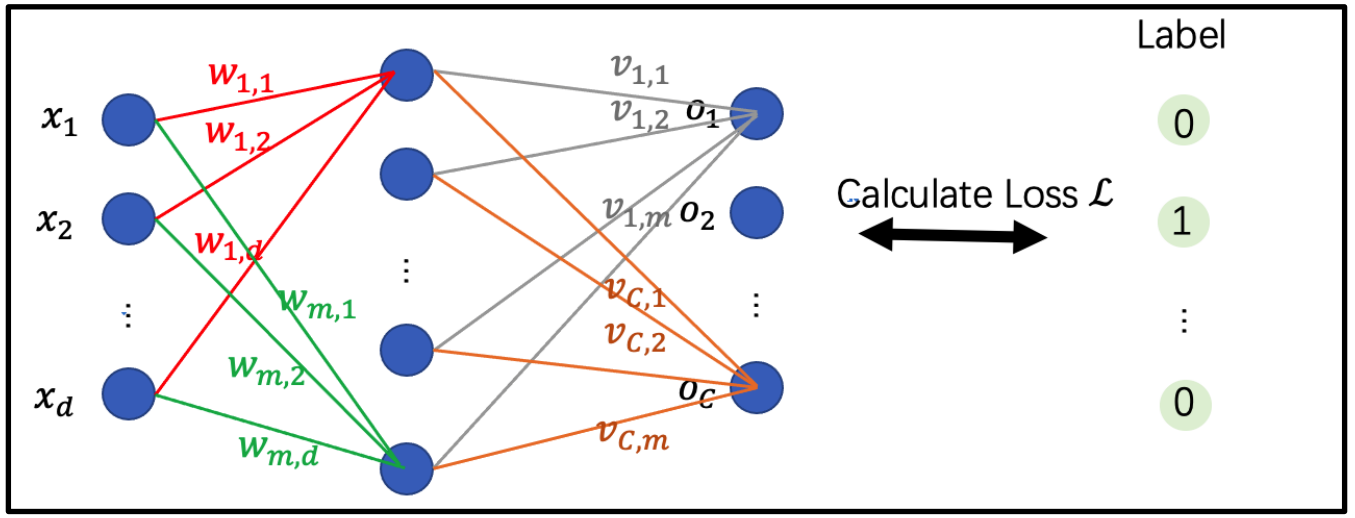
	w_1	w_2	$v_{1,1}$	$v_{1,2}$	$v_{2,1}$	$v_{2,2}$
w_1						
w_2						
$v_{1,1}$					0	0
$v_{1,2}$					0	0
$v_{2,1}$			0	0		
$v_{2,2}$			0	0		



No correlation between v_1 and v_2
(Check the graph:
NO link between them!)

Now we are ready to explain the
"dynamic force"

“dynamic force”

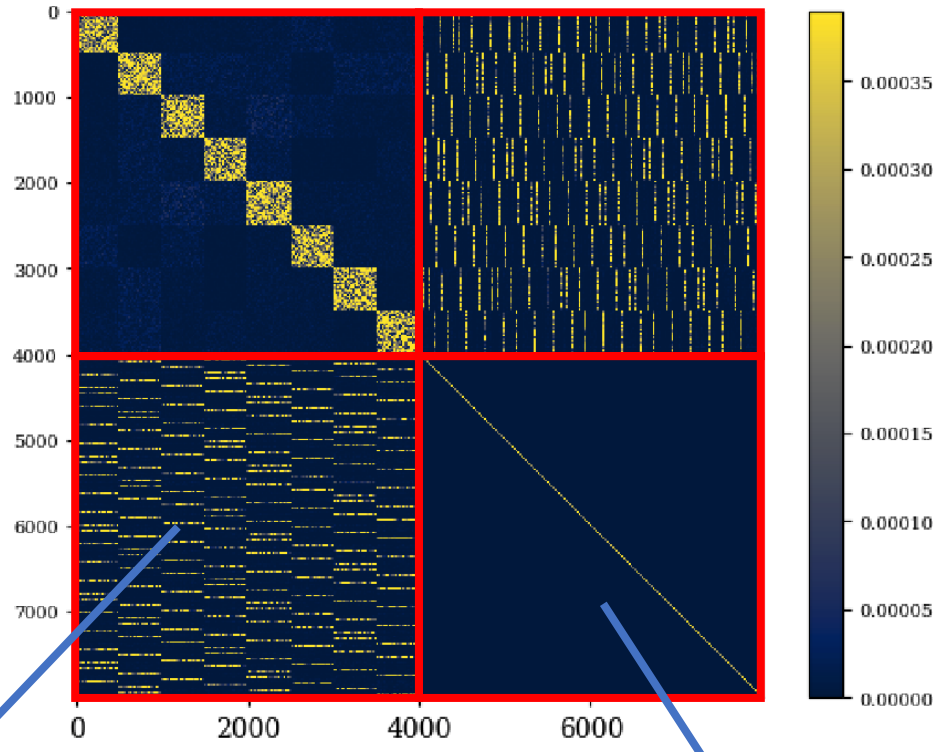


$$H_{w_i v_j} = \frac{\partial^2 \ell}{\partial w_i \, \partial v_j^T} = \begin{bmatrix} 0 & \cdots & a_{i,1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & a_{i,d} & 0 & \cdots & 0 \end{bmatrix} + O\left(\frac{1}{c}\right) \in R^{d \times m}, \text{ where } a_{i,d'} = -\frac{1}{N} \sum_n \sum_c \left(\delta_{y_n,c} - p_{n,c}\right) v_{c,i} \mathbb{I}\left(w_i^T x_n \geq 0\right) x_{n,d'}$$

- **Linear algebra perspective (like previous part):**
  from computation graph, only $v_{1,i}, \cdots, v_{C,i}$ are linked to w_i. So only i-th column in $H_{w_i v_j}$ is non-zero

Key take-away: $H_{wv} \approx O(optimality\ gap)$, which are expected to vanish (experiments: vanishes quickly as training begins)

What about the "static force"?



Hessian at initialization with CE loss

Why "block-circulant" in H_{wv}
and why vanish with training **(previous parts)**

H_{ww} and H_{vv} seems always "block-diag":
(Need large C here)

Why need large C ? An intuition

Hessian of hidden weights H_{ww}:

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_i^\top} = \frac{1}{N}\sum_{n=1}^N \left(\sum_{c=1}^C p_{n,c} v_{c,i}^2 - \left(\sum_{c=1}^C p_{n,c} v_{c,i}\right)^2\right) \mathbf{1}(w_i^\top x_n > 0) x_n x_n^\top \\ \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_j^\top} = \frac{1}{N}\sum_{n=1}^N \left(\sum_{c=1}^C p_{n,c} v_{c,i} v_{c,j} - \left(\sum_{c=1}^C p_{n,c} v_{c,i}\right)\left(\sum_{c=1}^C p_{n,c} v_{c,j}\right)\right) \mathbf{1}(w_i^\top x_n > 0)\mathbf{1}(w_j^\top x_n > 0) x_n x_n^\top \end{cases} \tag{12}$$

Hessian of output weights H_{vv}:

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top} = \frac{1}{N}\sum_{n=1}^N p_{n,i}(1-p_{n,i})\sigma(Wx_n)\sigma(Wx_n)^\top \\ \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_j^\top} = -\frac{1}{N}\sum_{n=1}^N p_{n,i}p_{n,j}\sigma(Wx_n)\sigma(Wx_n)^\top. \end{cases}$$

Intuitive understanding: at random initialization, suppose entries in W, V follows i.i.d. zero-mean Gaussian distribution, we have $p_{n,i} \approx \frac{1}{C}$ for all $n \in [N], i \in [C]$. As such:

$$\frac{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\mathrm{F}}}{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\mathrm{F}}} \approx \frac{\left(\sum_{c=1}^C v_{c,i}v_{c,j} - \left(\sum_{c=1}^C v_{c,i}\right)\left(\sum_{c=1}^C v_{c,j}\right)\right)/C}{\left(\sum_{c=1}^C v_{c,i}^2 - \left(\sum_{c=1}^C v_{c,i}\right)^2\right)/C} \overset{C\to\infty}{=} \frac{\mathrm{Cov}(v_{i,i}, v_{i,j})}{\mathrm{Var}(v_{i,i})}. \tag{14}$$

Since $v_{i,i}, v_{i,j}$ are independent, $\mathrm{Cov}(v_{i,i}, v_{i,j}) = 0$ and thus the block-diagonal structure occurs as $C \to \infty$. Similarly, we have

$$\frac{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_j^\top}\right\|_{\mathrm{F}}}{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top}\right\|_{\mathrm{F}}} \approx \frac{\sum_{n=1}^N p_{n,i}p_{n,j}}{\sum_{n=1}^N p_{n,i}(1-p_{n,i})} \approx \frac{\frac{1}{C^2}}{\frac{1}{C}\left(1-\frac{1}{C}\right)} = \frac{1}{C-1}, \tag{15}$$

and thus the block-diagonal structure arises as $C \to \infty$.

- **But how to prove rigorously?**
- **Need tools from Random Matrix Theory (RMT)**

This is why large C (=NN output dim) helps!

Rigorous Theory

Assumption 1 *The entries of the data matrix* $X_N = (x_1, \cdots, x_N) \in \mathbb{R}^{d\times N}$ *are i.i.d.* $\mathcal{N}(0,1)$.

Assumption 2 *The model weights in* W *and* V *are initialized by LeCun initialization. That is: for the linear model,* $V_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$, $i \in [C], j \in [d]$; *for 1-hidden-layer network,* $W_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$, $i \in [m], j \in [d]$, $V_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{m})$, $i \in [C], j \in [m]$.

- A1 is restricted
- A2 is widely used

Theorem 2 **(1-hidden-layer networks.)** *Consider the Hessian expressions in (8) to (13), and assume Assumptions 1 and 2 hold. Then for any fixed* $m \geq 3$, *suppose* $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$, *it holds that*

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\text{MSE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\text{F}}^2\right]}{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\text{MSE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\text{F}}^2\right]}, \quad \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\text{CE}}(W,V)}{\partial v_i \partial v_j^\top}\right\|_{\text{F}}^2\right]}{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\text{CE}}(W,V)}{\partial v_i \partial v_i^\top}\right\|_{\text{F}}^2\right]} \tag{28}$$

vanish at the rate $\mathcal{O}(1/C)$, $\mathcal{O}(1/C^2)$, respectively, and the block-diagonal structure also emerges as C increases.

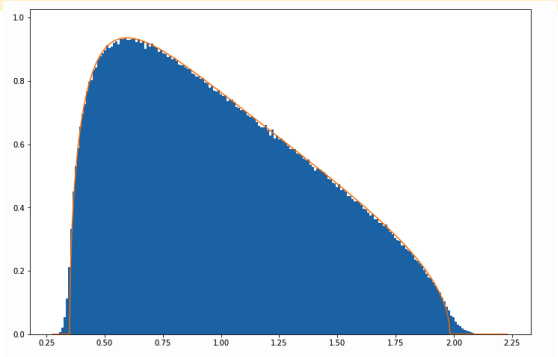
Key messages:
the block-diagonal structure arises when
classes C (i.e., NN output dim) → ∞

Key Challenges in the Proof

Diagonal Hessian block: $$\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top} \overset{(5)}{=} \frac{1}{N}\sum_{n=1}^{N} p_{n,i}(1-p_{n,i}) x_n x_n^\top := \frac{1}{N} X_N \Lambda_N X_N^\top \in \mathbb{R}^{d\times d},$$

Q: How to characterize the Hessian block $\mathrm{E}||\frac{1}{N} X_N \Lambda_N X_N^T||_F$?

Just 2nd-order moment of the eigenvalue distribution



Eigenvalue histogram of $A_n = \frac{1}{n} X\Lambda X^T \in R^{d\times d}, \Lambda = I, n = 50, d = 300$, 1000 samples of A_n

We will use Random Matrix Theory (RMT), but classical methods cannot be directly applied:

- If X_N, Λ_N are **independent**: can use **Generalized Marchenko–Pastur Theorem (1967)**
- In our $\frac{1}{N} X_N \Lambda_N X_N^T$, X_N, Λ_N are **clearly NOT independent**, so MP theorem cannot be applied
- Dependent matrix product is a difficult topic in RMT
- **Our solution:** a new method built upon **the Lindeberg principle** (originally proposed to prove CLT)

Our Proof Strategies (Overview)

Key challenge: Need $||\frac{1}{N}X_N\Lambda_N X_N^T||_F$, But X_N and Λ_N **are dependent**

Our solution: a new method built upon **the Lindeberg principle** (originally proposed to prove CLT)

Our "decouple" Strategy:

Step 1 (Important): "indept. copy $\tilde{X}_N$ + interpolation": we introduce the following $X_N(t)$

$$X_N(t) = \sqrt{t}\,X_N + \sqrt{1-t}\,\widetilde{X_N},\ t \in [0,1].\ \text{Note that } X_N(0) = X_N, X_N(1) = \tilde{X}_N$$

Goal: Wish to show that: for any $z \in C^+$, $\delta_N(z) = \mathrm{E}s_{\widetilde{H_{ii}}}(z) - \mathrm{E}\,s_{H_{ii}}(z)$ vanishes as N increases

Step 2 (Important): Fundamental theorem of calculus

$$\delta_N(z) = \int_0^1 E\left[\frac{d}{dt}\,s_{H_{ii}(t)}\right]dt$$

Step 3 (Important): Using Cauchy Integral Formular, we prove that $\delta_N(z) \le Const.\,\mathrm{E}[Z_1 f(Z_1) - Z_2 f(Z_2)]$, where $\mathrm{Z}_i \sim N(0,1)$

This step needs NN structure: Λ_N is made by relu + CE loss

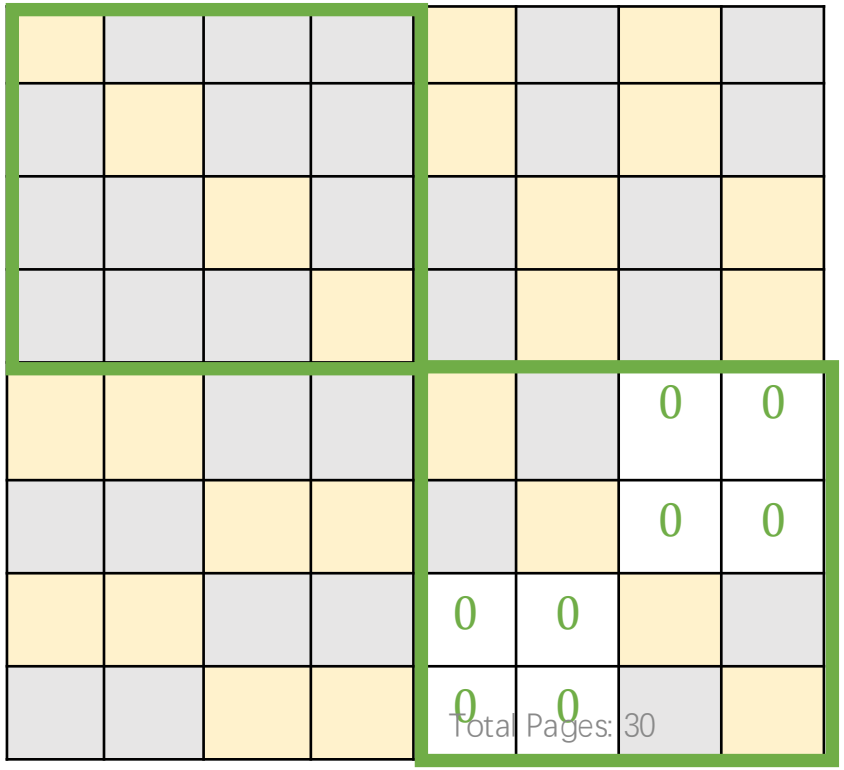
Step 4 (Important): Using Stein's Lemma, we prove that:

$$\mathrm{E}[Z_1 f(Z_1) - Z_2 f(Z_2)] = \mathrm{E}[\,f'(Z_1) - f'(Z_2)] = O\left(\frac{1}{\sqrt{N}}\right)$$

Step 5 (Standard): Apply GMP to recover $\mathbf{S}_{\widetilde{\mathbf{H_{ii}}}}(\mathbf{z})$, $\boldsymbol{\mu}_{\widetilde{H_{ii}}}$, and $\boldsymbol{\mu}_{H_{ii}}$

Summary: 3-level sources of block-diag structure

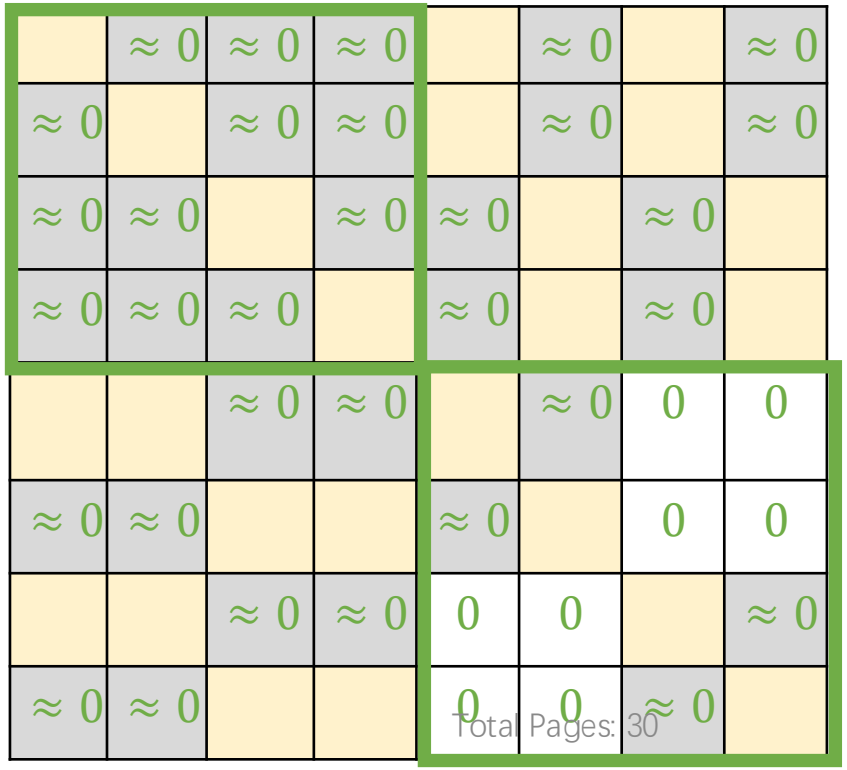
- **Level 1: definition of matrix product: some zeros, no links**



Summary: 3-level sources of block-diag structure

- **Level 1: definition of matrix product: some zeros, no links**
- **Level 2: #Class C goes to infinity: weaken many links in H_{ww}, H_{vv}**

Static force
(Proved by RMT at random initialization)



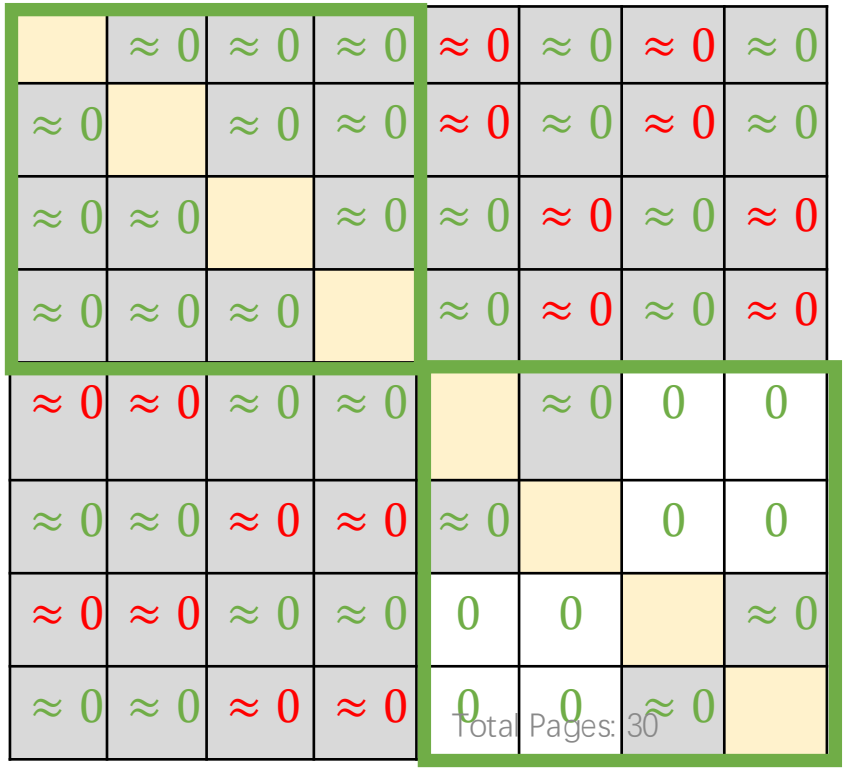
Summary: 3-level sources of block-diag structure

- **Level 1: definition of matrix product: some zeros, no links**
- **Level 2: #Class C goes to infinity: weaken many links in H_{ww}, H_{vv}**

} **Static force**
(Proved by RMT at random initialization)

- **Level 3: Training: eliminates strong links in H_{wv}**

} **Dynamic force**
(See from direct calculation)

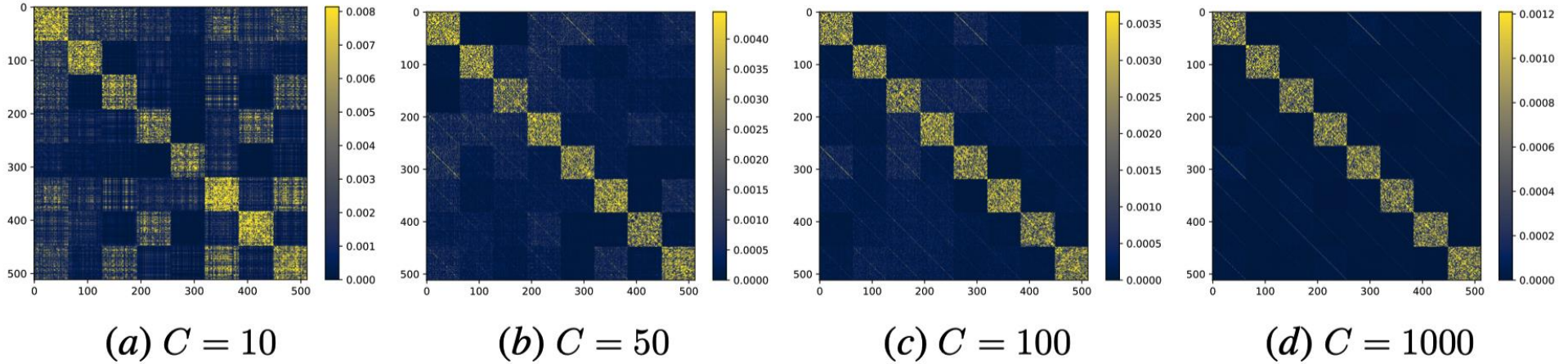


Limitations:
Current RMT does not consider training dynamics

Still incomplete, has a long way to go…

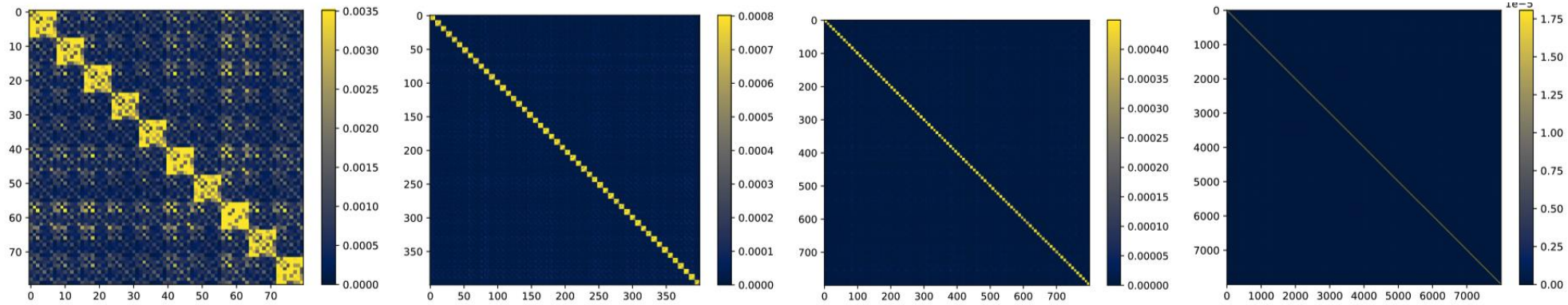
Experiments: The Effect of Increasing # classes C

Hessian of hidden weights



$(a)\ C = 10$ $(b)\ C = 50$ $(c)\ C = 100$ $(d)\ C = 1000$

Hessian of output weights



$(a)\ C = 10$ $(b)\ C = 50$ $(c)\ C = 100$ $(d)\ C = 1000$

- The Hessian blocks of **1-hidden-laye NN** with 8 hidden neurons + #class C at random init.
- The block-diag structure **becomes clearer as C increases**

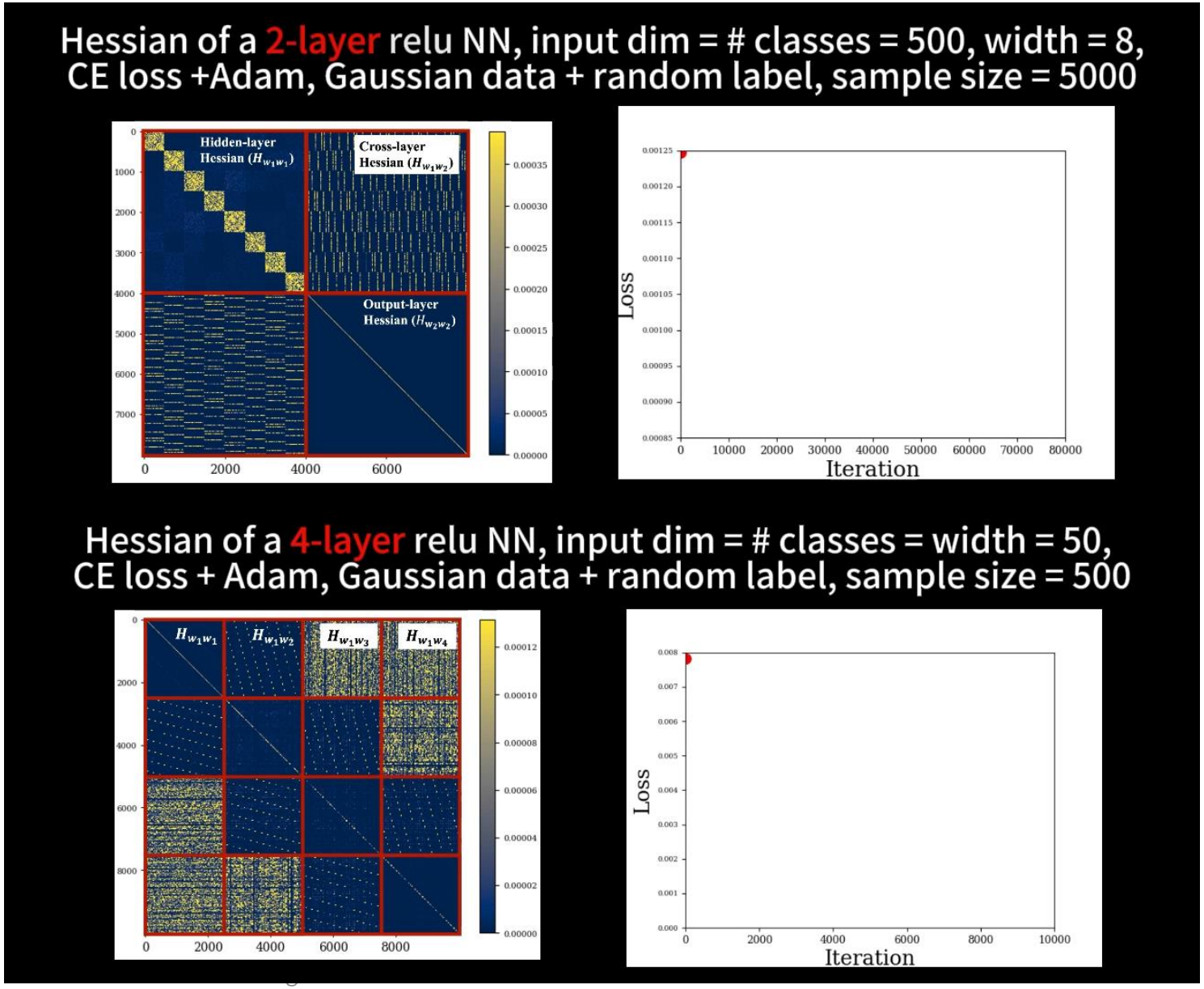
Hessian for Deep NNs?

For a rough estimate:
just check the links in the computational graph:

Check if there exists
a connected path between two links

Source of special Hessian:
consecutive multiplication of
large and well-shaped matrix variables

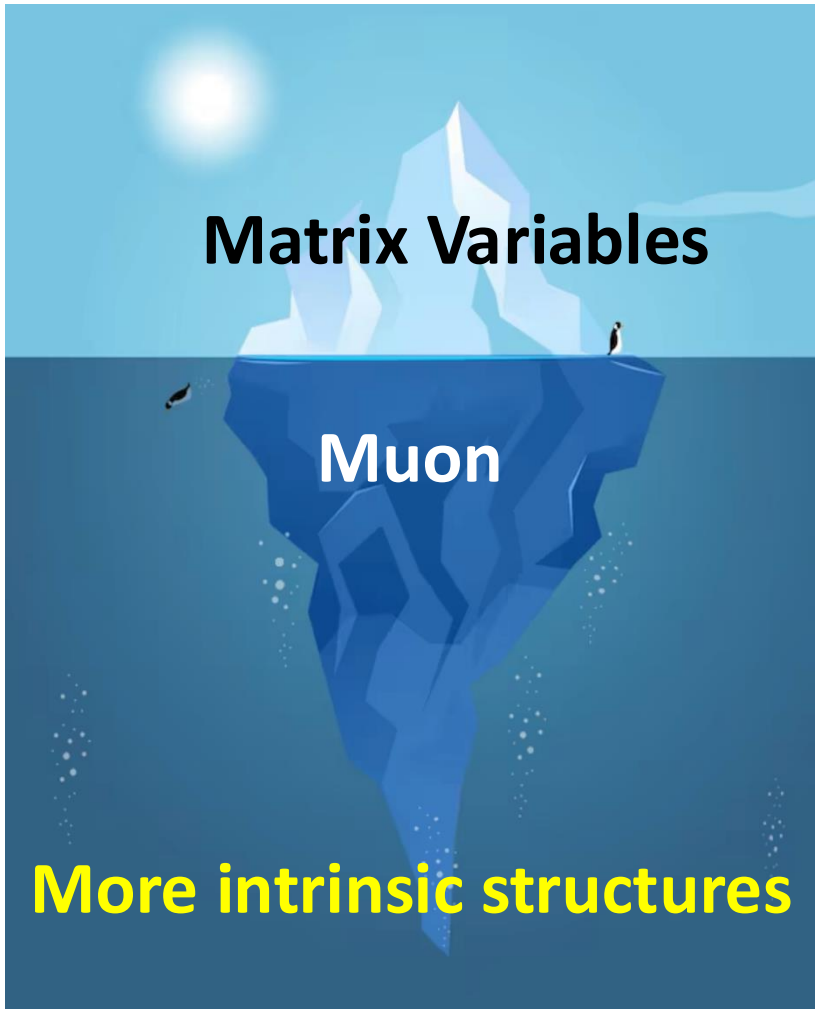
$\ldots W_1 W_2 W_3 W_4 \ldots$



Summary

- **Source of special Hessian:**
  consecutive multiplication of **large** and **well-shaped** matrix variables

$$y = W_3 \; W_2 \; W_1 \; x$$



Thanks for listening!