Towards Quantifying the Hessian Structure of Neural Networks

Zhaorui Dong[*] ZHAORUIDONG@LINK.CUHK.EDU.CN
Yushun Zhang[*] YUSHUNZHANG@LINK.CUHK.EDU.CN
Jianfeng Yao[†] JEFFYAO@CUHK.EDU.CN
Ruoyu Sun[†] SUNRUOYU@CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen

Abstract

Empirical studies reported that the Hessian matrix of neural networks (NNs) exhibits a near-block-diagonal structure, yet its theoretical foundation remains unclear. In this work, we reveal that the reported Hessian structure comes from a mixture of two forces: a "static force" rooted in the architecture design, and a "dynamic force" arisen from training. We then provide a rigorous theoretical analysis of "static force" at random initialization. We study linear models and 1-hidden-layer networks for classification tasks with C classes. By leveraging random matrix theory, we compare the limit distributions of the diagonal and off-diagonal Hessian blocks and find that the block-diagonal structure arises as $C \to \infty$. Our findings reveal that C is one primary driver of the near-block-diagonal structure. These results may shed new light on the Hessian structure of large language models (LLMs), which typically operate with a large C exceeding 10^4.

Keywords: Hessian, Neural Networks, Random Matrix Theory

1. Introduction

The Hessian matrix of neural networks (NNs) is crucial for understanding training dynamics, as well as motivating better algorithm designs. A classical work [11] first empirically reported that the Hessian of NNs is highly structured: the Hessian is observed to be *near-block-diagonal*. We reproduce this result in Figure 1. Unfortunately, no rigorous theory has been established in the past two decades to explain this phenomenon.

Very recently, the near-block diagonal structure of Hessian has drawn renewed attention in the machine learning community as it helps understanding the training of large language models (LLMs) [34, 85, 86]. Again, these works primarily focus on empirical observations, and there is no rigorous theoretical results on the underlying source of the special structure. The following fundamental question remains largely open:

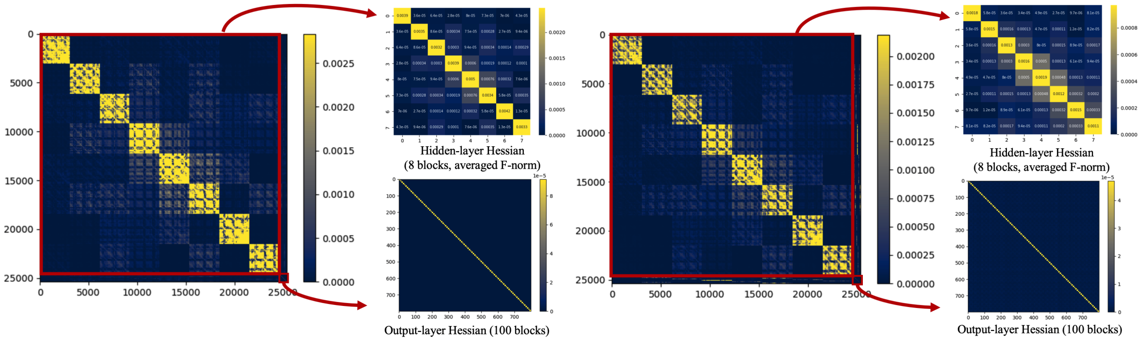
When and why does the Hessian of NNs exhibit near-block-diagonal structure?

Before delving into this question, we first list some of its important implications.

- **I.** Understanding Hessian structure can help understand NN training. For instance, the effectiveness of diagonal preconditioned methods such as Adam [33] is usually strongly

. $*$: Equal contribution. These authors are listed in alphabetical order.

. †:Correspondence authors: Ruoyu Sun, Email: `sunruoyu@cuhk.edu.cn`; Jianfeng Yao, Email: `jeffyao@cuhk.edu.cn`.



(*a*) 1-hidden-layer network with MSE loss (*b*) 1-hidden-layer network with CE loss

Figure 1: **(a, b):** The Hessian matrix of a 1-hidden-layer network with 8 hidden neurons at random initialization on CIFAR-100 dataset (# hidden neuron $m = 8$ and # classes or output neuron $C = 100$). For clearer visualization, we report the absolute value of each Hessian entry, and this applies to all Hessian matrices reported in this work. We observe near-block-diagonal structures under both MSE and CE loss with $m + C = 108$ blocks in total.

related to the Hessian structure; see, e.g., Das et al. [13], Qu et al. [61], Sun and Ye [68]. Recently, the near-block-diagonal Hessian is observed along the training process and such structure is shown to be related to the effectiveness of Adam on LLMs [34, 85]. Besides Adam, the near-block-diagonal Hessian structure may play a crucial role in the effectiveness of block-diagonal preconditioned methods (e.g., [20, 24, 31, 47, 72]). Among these methods, Muon optimizer [31] is used for training Moonlight [43], Kimi-K2 [50], and GLM-4.5 [82] very recently.

- **II.** Understanding Hessian structure can help design new training methods for NNs. For instance, Adam-mini [86], a recently proposed optimizer, utilizes the near-block-diagonal Hessian structure to cut down 50% memory consumption in Adam. We believe the special Hessian structure can inspire more new optimizers.
- **III.** The near-block-diagonal Hessian structure can offer a new class of problems for the optimization community to study. For the optimization community, it is rare to analyze (near-) block-diagonal Hessian structure since typical problems do *not* have such structure. For instance, in the classical non-linear programming dataset [35], all problems have non-block-diagonal Hessians. Understanding the special Hessian structure of NN can draw attention from the optimization community, motivating further study into this specialized class of problems.

In this work, we explore the Hessian structure of NNs both numerically and theoretically. First, we report more fine-grained numerical findings on Hessian: we observe *"block-circulant-block-diagonal"* structure at the random initialization and *"block-diagonal"* structure after training starts (presented later in Section 2). In particular, the "dynamic force" compresses cross-layer Hessian components during training; and the "static force" compresses the cross-neuron component in each layer for both initialization stage and training stage. Our findings suggest that the previously reported block-diagonal structure actually comes from a mixture of two forces: a "static force" rooted in the architecture design, and a "dynamic force" arisen from training.

Then, we provide a rigorous theoretical analysis of "static force" at random initialization. We focus on linear models and 1-hidden-layer networks for standard classification tasks with C classes. Leveraging tools from random matrix theory, we characterize the limit distributions of diagonal and off-diagonal Hessian blocks as the sample size N and input dimension d grow proportionally to infinity. Our theory shows that the off-diagonal blocks will be pushed to 0 as the number of classes C increases, suggesting that C is a primary driver of the near-block-diagonal Hessian structure. Our theory may shed new light on the Hessian structures of LLMs since they usually have large C (more than 10^4 or 10^5) [1].

Our main contributions are summarized as follows.

- We numerically investigate the source of the near-block-diagonal Hessian structure. We reveal two forces that shape such structure: a "static force" rooted in the architecture, and a "dynamic force" arisen from training. In particular, the "dynamic force" compresses cross-layer Hessian components during training; and the "static force" compresses the cross-neuron component in each layer for both initialization stage and training stage.

- We provide rigorous theory on the Hessian of linear models at random initialization. As the sample size N and input dimension d grow proportionally to infinity, we calculate the limit distribution for the Frobenius norm of the diagonal and off-diagonal blocks of Hessian. Specifically, the diagonal blocks correspond to the Hessian of weights associated with the same class, while the off-diagonal blocks represent the Hessian of weights from different classes. We find that: the ratio between the off-diagonal and diagonal blocks decays to zero at the rate of $O(1/C)$, where C is the number of classes. This demonstrates that the Hessian becomes block-diagonal as $C \to \infty$, and the number of blocks equals C.

- We extend the above analysis to 1-hidden-layer networks. We focus on two sub-matrices in Hessian: the hidden-layer Hessian and the output-layer Hessian, which are highlighted with red boxes in Figure 1. For the hidden-layer Hessian, the ratios between their off-diagonal and diagonal blocks decay to zero at the rate of $O(1/\sqrt{C})$. For the output-layer Hessian, the decay rate is $O(1/C)$. This demonstrates that these sub-matrices will become block-diagonal as $C \to \infty$. In this case, the total number of blocks in these sub-matrices equals $(m + C)$, where m denotes the number of hidden neurons.

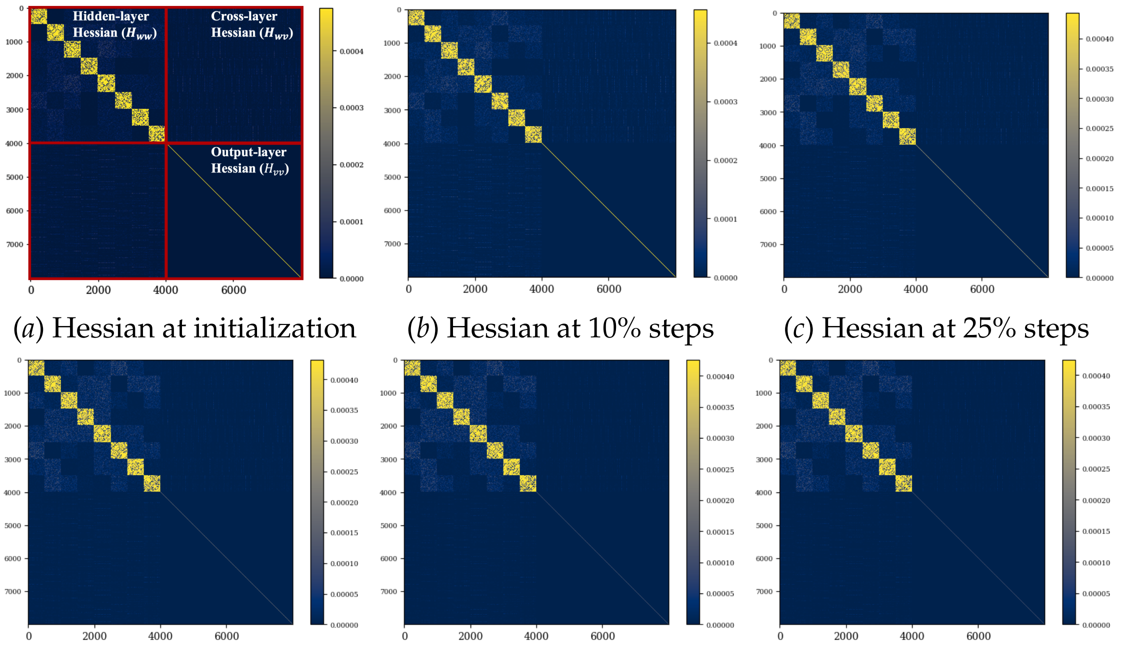
- We highlight some key technical contributions in our proof. The major challenge lies in characterizing the limit distribution of *non-independent* random matrix products, which in general is a difficult problem in random matrix theory. For the Hessian of NNs, we find that such dependency arises from ReLU activation and CE loss, and the dependency diminishes as $d \to \infty$. Subsequently, we propose a systematic procedure to address this type of "diminishing dependencies" caused by ReLU activation and CE loss. Our approach implements *the Lindeberg interpolation principle*, which is originally proposed to prove the Central Limit Theorem (CLT).

2. Empirical Observations: Two Forces Shaping the Hessian Structure

Now we conduct more fine-grained experiments on Hessian structures of NNs. In particular, we explore the 1-hidden-layer network on a Gaussian synthetic dataset. We consider both Mean-Square (MSE) and Cross-Entropy (CE) loss. The detailed experimental setups are presented in Appendix G.3.

We emphasize that these experiments can reveal more Hessian properties not shown in the CIFAR-100 experiments in Figure 1. We highlight the new changes as follows.

[1] $C = 32k$ in Llama 2 [71] and $C = 128k$ in many recent models such as DeepSeek-V3 [41].



(*a*) Hessian at initialization (*b*) Hessian at 10% steps (*c*) Hessian at 25% steps

(*d*) Hessian at 50% steps (*e*) Hessian at 75% steps (*f*) Hessian at 100% steps

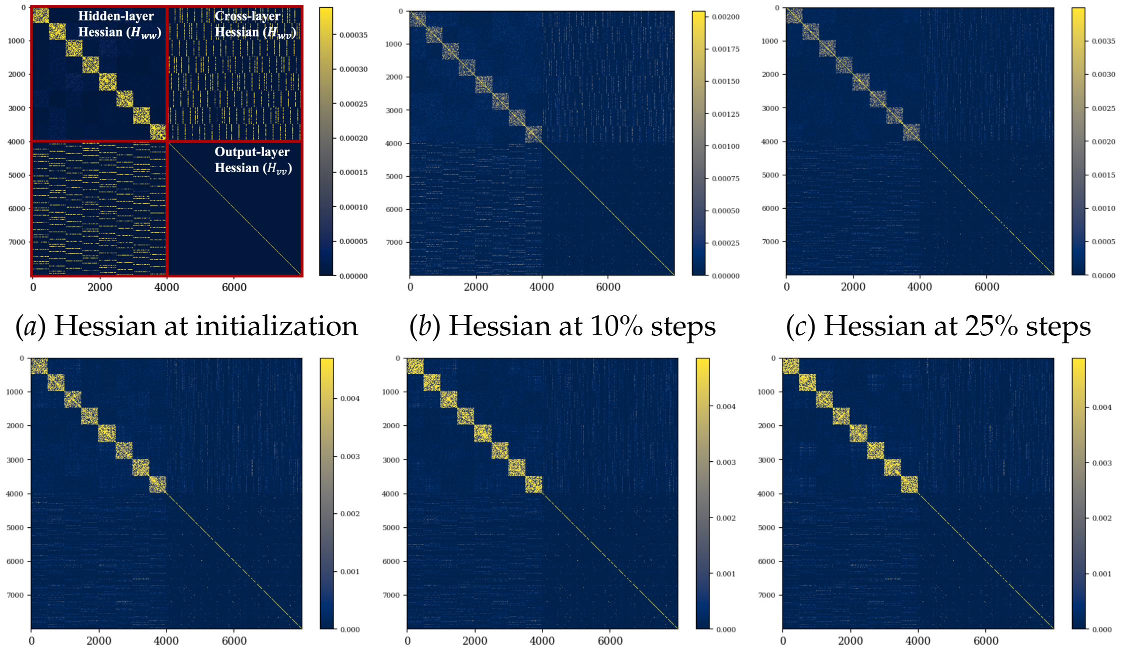
Figure 2: **(a-f):** The Hessian of a 1-hidden-layer network on Gaussian synthetic data under MSE loss. We notice the near-block-diagonal patterns in H_{ww} and H_{vv} with $m + C = 508$ blocks in total, and they maintain along training.

- We change the input dimension d and # classes C to amplify the effect of cross-layer Hessian components H_{wv}. In the CIFAR-100 example in Figure 1, the proportions of the Hessian of hidden layer and output layer, which we abbreviate as H_{ww} and H_{vv}, are largely imbalanced. Here, we change $(d, C) = (3072, 100)$ to $(d, C) = (500, 500)$ so that H_{ww}, H_{vv}, and H_{wv} are proportionally balanced within the Hessian.
- We change the dataset from CIFAR-100 to Gaussian synthetic dataset. Such change suggests that the Hessian structure might be inherently general and is not overfitted to one specific dataset like CIFAR-100.
- In addition to the Hessian at random initialization, we present the Hessian along training until convergence. We present the loss curves in Appendix G.3.

The results are shown in Figure 2 and 3. We summarize two findings.

- **Finding 1:** For both MSE loss and CE loss, we observe near-block-diagonal structures in H_{ww} and H_{vv} and such structures maintains along training.
- **Finding 2:** For CE loss, we observe new special structures in H_{wv} at random initialization: H_{wv} exhibits a "block-circulant" pattern with periodic stripes. When using CE loss, the full Hessian matrix appears to be a combination of "block-circulant" matrix (for H_{wv}) and block-diagonal matrix (for H_{ww} and H_{vv}). We refer to it as "*block-circulant-block-diagonal* matrix". We observe that the "block-circulant" pattern in H_{wv} vanishes as training goes on, while the near-block-diagonal patterns in H_{ww} and H_{vv} remain obvious.

Main Takeaways and insights from the experiments. Based on the findings from Figure 2 and 3, we find that there are at least two forces shaping the Hessian structure.



(a) Hessian at initialization (b) Hessian at 10% steps (c) Hessian at 25% steps

(d) Hessian at 50% steps (e) Hessian at 75% steps (f) Hessian at 100% steps

Figure 3: **(a-f):** The Hessian of a 1-hidden-layer network on Gaussian synthetic data under CE loss. At initialization, we observe the "block-circulant" pattern in H_{wv}, and the near-block-diagonal structure in H_{ww} and H_{vv} (with $m + C = 508$ blocks in total). We refer to it as *"block-circulant-block-diagonal matrix"*. We notice that the "block-circulant" pattern in H_{wv} vanishes along training, while the near-block-diagonal patterns in H_{ww} and H_{vv} are preserved.

- **(1) A "static force" rooted in the architecture design.** For both MSE and CE loss, this force compresses the cross-neuron components in H_{ww} and H_{vv}. This force is effective in both initialization stage and the training stage.
- **(2) A "dynamic force" arisen from training.** When using CE loss, this force gradually erases the initial "block-circulant" pattern in the cross-layer component H_{wv} along training.

In the sequel, we study how both forces shape the Hessian structure. We will primarily focus on the effect of "static force" at random initialization, particularly, how the architecture shapes the Hessian structure for H_{ww} and H_{vv} for both CE loss and MSE loss. As for "how the 'dynamic force' eliminates the block-circulant pattern in H_{wv} along training", we find that it can be explained directly from Hessian expressions. We provide an initial analysis in Appendix A and leave more fine-grained analysis as future direction.

3. Main Results

We now present our rigorous statements. Since the Frobenius norm of Hessian blocks involves the 2nd-order moments of eigenvalues of random matrices, we will resort to the tools from random matrix theory. We first state some standard assumptions.

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, $W_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{d})$, $i \in [m], j \in [d]$, $V_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{m})$, $i \in [C], j \in [m]$.*

Note that Assumption 2 is widely adopted in NNs [67]. Assumption 1 on data distribution is standard in random matrix theory [57]. It is possible to extend the Gaussian distribution to, e.g., Gaussian orthogonal ensembles and more general i.i.d. distribution. However, such generalization is non-trivial and each case may require an independent paper (e.g. Pastur [56], Pastur and Slavin [58]). We now state our results for linear models and 1-hidden-layer networks under Assumptions 1 and 2.

Theorem 1 **(Linear models.)** *Consider the Hessian expressions in (17) and assume Assumptions 1 and 2 hold. Suppose $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$, then for fixed $C \geq 2$, it holds almost surely that*

$$\lim_{d,N\to\infty} \frac{1}{d} \left\| \frac{\partial^2 \ell_{CE}(V)}{\partial v_i \partial v_i^\top} \right\|_{\mathrm{F}} = g_{ii}(\gamma, C), \quad \forall i \in [C], \tag{1}$$

$$\lim_{d,N\to\infty} \frac{1}{d} \left\| \frac{\partial^2 \ell_{CE}(V)}{\partial v_i \partial v_j^\top} \right\|_{\mathrm{F}} = g_{ij}(\gamma, C), \quad \forall i, j \in [C], i \neq j, \tag{2}$$

where functions g_{ii}, g_{ij} are given in Appendix E.1. Furthermore,

$$\lim_{C\to\infty} C^2 g_{ii}(\gamma, C) = \gamma e + 1, \tag{3}$$

$$\lim_{C\to\infty} C^4 g_{ij}(\gamma, C) = \gamma e^2 + 1. \tag{4}$$

Theorem 1 implies that we have the following relation between the diagonal and off-diagonal blocks:

$$\lim_{d,N\to\infty} \frac{\left\| \frac{\partial^2 \ell_{CE}(V)}{\partial v_i \partial v_j^\top} \right\|_{\mathrm{F}}^2}{\left\| \frac{\partial^2 \ell_{CE}(V)}{\partial v_i \partial v_i^\top} \right\|_{\mathrm{F}}^2} = \frac{g_{ij}(\gamma, C)}{g_{ii}(\gamma, C)}, \quad \lim_{C\to\infty} \frac{C^2 g_{ij}(\gamma, C)}{g_{ii}(\gamma, C)} = \frac{\gamma e^2 + 1}{\gamma e + 1}. \tag{5}$$

When $C \to \infty$, the ratio vanishes at the rate $\mathcal{O}(1/C^2)$, and the block-diagonal structure emerges.

The next theorem presents a similar result for 1-hidden-layer networks.

Theorem 2 **(1-hidden-layer networks.)** *Consider the Hessian expressions in (20) to (27), 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*

$$\lim_{d,N\to\infty} \frac{1}{d} \mathbf{E} \left[\left\| \frac{\partial^2 \ell_{CE}(W, V)}{\partial w_i \partial w_i^\top} \right\|_{\mathrm{F}}^2 \right] = h_{ii}(\gamma, C), \quad \lim_{d,N\to\infty} \frac{1}{d} \mathbf{E} \left[\left\| \frac{\partial^2 \ell_{CE}(W, V)}{\partial w_i \partial w_j^\top} \right\|_{\mathrm{F}}^2 \right] = h_{ij}(\gamma, C), \tag{6}$$

$$\lim_{d,N\to\infty} \frac{1}{d} \mathbf{E} \left[\left\| \frac{\partial^2 \ell_{MSE}(W, V)}{\partial w_i \partial w_i^\top} \right\|_{\mathrm{F}}^2 \right] = u_{ii}(\gamma, C), \quad \lim_{d,N\to\infty} \frac{1}{d} \mathbf{E} \left[\left\| \frac{\partial^2 \ell_{MSE}(W, V)}{\partial w_i \partial w_j^\top} \right\|_{\mathrm{F}}^2 \right] = u_{ij}(\gamma, C), \tag{7}$$

$$\lim_{d,N\to\infty} \mathbf{E}\left[\left\|\frac{\partial^2 \ell_{CE}(W,V)}{\partial v_i \partial v_i^\top}\right\|_{\mathrm{F}}^2\right] = q_{ii}(\gamma, C), \quad \lim_{d,N\to\infty} \mathbf{E}\left[\left\|\frac{\partial^2 \ell_{CE}(W,V)}{\partial v_i \partial v_j^\top}\right\|_{\mathrm{F}}^2\right] = q_{ij}(\gamma, C), \tag{8}$$

where functions $h_{ii}, h_{ij}, u_{ii}, u_{ij}, q_{ii}, q_{ij}$ given in Appendix E.2. Furthermore, we have

$$\lim_{C\to\infty} h_{ii}(\gamma, C) = \frac{1+2\gamma}{4m^2}, \quad \lim_{C\to\infty} Ch_{ij}(\gamma, C) = \frac{\gamma(m-1)^2}{2^m(m-2)^3 m}\left(\sqrt{\frac{m}{m-2}}+1\right)^{m-2}, \tag{9}$$

$$\lim_{C\to\infty} \frac{u_{ii}(\gamma, C)}{C^2} = \frac{1+2\gamma}{4m^2}, \quad \lim_{C\to\infty} \frac{u_{ij}(\gamma, C)}{C} = \frac{1+4\gamma}{16m^2}, \tag{10}$$

$$\lim_{C\to\infty} C^2 q_{ii}(\gamma, C) = m a_{12} b_1^{m-1} + m(m-1) a_{11}^2 b_1^{m-2} \tag{11}$$

$$\lim_{C\to\infty} C^4 q_{ij}(\gamma, C) = m a_{22} b_2^{m-1} + m(m-1) a_{21}^2 b_2^{m-1}, \tag{12}$$

where the constant terms $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$ are presented in (67) in Appendix E.2.

Similar to the implication of Theorem 1 in (5), Theorem 2 implies that the ratios

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\mathrm{F}}^2\right]}{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\mathrm{F}}^2\right]}, \quad \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\mathrm{F}}^2\right]}{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\mathrm{F}}^2\right]}, \quad \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_j^\top}\right\|_{\mathrm{F}}^2\right]}{\mathbf{E}\left[\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top}\right\|_{\mathrm{F}}^2\right]} \tag{13}$$

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

The above results rigorously quantify the block-diagonal structure in H_{ww} and H_{wv}. As for the "block-circulant" structure in the cross-layer component H_{wv} and how it vanishes along training (particularly for CE loss), we find that it can be explained directly from Hessian expressions. We provide an initial analysis in Appendix A and leave more rigorous analysis as a future direction.

4. Intuitive Understanding and Key Proof Ideas

The complete proof of Theorem 1 and 2 is rather long. To help readers boost understanding and grasp the key ideas in the proof, we provide the following short sections.

- In Appendix A, we provide preliminaries and intuitive understandings on how the special "block-circulant-block-diagonal" or "block-diagonal" Hessian structure arises. This part only involve elementary calculus and simple probability.
- In Appendix D, we will explain the major technical challenges in our proofs: analyzing the spectrum of *non-independent* random matrix product. We will introduce our new systematic approach to address the difficulties. Our approach implements the *Lindeberg principle*, which is originally proposed to prove the Central Limit Theorem (CLT).
- In Appendix F, we provide more experiments to support our theory: the block-diagonal structure in H_{ww} and H_{vv} emerges as C increases and the rate matches our theoretical prediction.

The complete proof of Theorem 1 and 2 can be seen in Appendix E.1 and E.2.

5. Conclusions

In this work, we reveal two forces that shape the near-block-diagonal Hessian structure of NNs: a "static force" rooted in the architecture design, and a "dynamic force" arisen from training. We then provide a rigorous theoretical analysis of "static force" of linear and 1-hidden-layer NNs at random initialization. It is intriguing to extend our study beyond initialization and simple models. We provide more discussions in Appendix H.

Acknowledgement

Z. Dong would like to thank Prof. Zhenyu Liao, Prof. Shurong Zheng and Prof. Jianfeng Yao for organizing a reading group on RMT and Machine Learning in 2023-2024, where Pastur's technique is introduced and discussed. Y. Zhang would like to thank Weronika Ormaniec, Sidak Pal Singh, Jeremy Cohen, Prof. Yinyu Ye, and Prof. Madeleine Udell for the valuable discussion. J. Yao's research is partially supported by the NSFC grant RFIS 12350710179. R. Sun's research is partially supported by NSFC (No. 12326608); Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

References

[1] Guillaume Alain, Nicolas Le Roux, and Pierre-Antoine Manzagol. Negative eigenvalues of the hessian in deep neural networks. *arXiv preprint arXiv:1902.02366*, 2019.

[2] Kang An, Yuxing Liu, Rui Pan, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

[3] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[4] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer New York, 2010.

[5] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18(2):425–442, 2008.

[6] Sourav Chatterjee. A generalization of the lindeberg principle. *Annals of Probability*, 34 (6):2061 – 2076, 2006.

[7] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[8] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[9] Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

[10] Benoît Collins and Tomohiro Hayase. Asymptotic freeness of layerwise jacobians caused by invariance of multilayer perceptron: The haar orthogonal case. *Communications in Mathematical Physics*, 397(1):85 – 109, 2023.

[11] Ronan Collobert. Large scale machine learning. Technical report, Université de Paris VI, 2004.

[12] Felix Dangel, Stefan Harmeling, and Philipp Hennig. Modular block-diagonal curvature approximations for feedforward architectures. In *International Conference on Artificial Intelligence and Statistics*, pages 799–808. PMLR, 2020.

[13] Rudrajit Das, Naman Agarwal, Sujay Sanghavi, and Inderjit S Dhillon. Towards quantifying the preconditioning effect of adam. *arXiv preprint arXiv:2402.07114*, 2024.

[14] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

[15] Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, et al. Natural neural networks. *Advances in neural information processing systems*, 28, 2015.

[16] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

[17] Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

[18] Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in neural information processing systems*, 31, 2018.

[19] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[20] Donald Goldfarb, Yi Ren, and Achraf Bahamou. Practical quasi-newton methods for training deep neural networks. *Advances in Neural Information Processing Systems*, 33: 2386–2396, 2020.

[21] Friedrich Götze, Holger Kösters, and Alexander Tikhomirov. Asymptotic spectra of matrix-valued functions of independent random matrices and free probability. *Random Matrices: Theory and Applications*, 4(02):1550005, 2015.

[22] Diego Granziol, Timur Garipov, Dmitry Vetrov, Stefan Zohren, Stephen Roberts, and Andrew Gordon Wilson. Towards understanding the true loss surface of deep neural networks using random matrix theory and iterative spectral methods. 2019.

[23] Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *Journal of Machine Learning Research*, 23(173):1–65, 2022.

[24] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[25] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[26] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[27] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019.

[28] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

[29] Stanislaw Jastrzkebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

[30] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[31] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL `https://kellerjordan.github.io/posts/muon/`.

[32] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. *arXiv preprint arXiv:2402.19449*, 2024.

[35] Giovanni Lavezzi, Kidus Guye, and Marco Ciarcià. Nonlinear programming solvers for unconstrained and constrained optimization problems: a benchmark analysis. *arXiv preprint arXiv:2204.05297*, 2022.

[36] Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 2002.

[37] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020.

[38] Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2020.

[39] Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

[40] J.W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211 – 225, 1922.

[41] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[42] Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

[43] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

[44] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

[45] Lucas Maes, Tianyue H Zhang, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions. *arXiv preprint arXiv:2410.19964*, 2024.

[46] Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik, and Dan Alistarh. Pushing the limits of large language model quantization via the linearity theorem. *arXiv preprint arXiv:2411.17525*, 2024.

[47] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[48] V A Marčenko and L A Pastur. Distribution of eigenvalues for some set of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[49] J. Mingo and R. Speicher. *Free Probability and Random Matrices*. Springer, 2017.

[50] Moonshot AI. Kimi k2: Open agentic intelligence, July 2025. URL https://moonshotai.github.io/Kimi-K2/.

[51] Weronika Ormaniec, Felix Dangel, and Sidak Pal Singh. What does it mean to be a transformer? insights from a theoretical hessian analysis. *arXiv preprint arXiv:2410.10986*, 2024.

[52] Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

[53] Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. *arXiv preprint arXiv:1901.08244*, 2019.

[54] Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *The Journal of Machine Learning Research*, 21(1):10197–10260, 2020.

[55] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[56] L. Pastur. Eigenvalue distribution of large random matrices arising in deep neural networks: Orthogonal case. *Journal of Mathematical Physics*, 63(6), 2022.

[57] Leonid Pastur. On random matrices arising in deep neural networks: Gaussian case. *Pure and Applied Functional Analysis*, 5(6):1395 – 1424, 2020.

[58] Leonid Pastur and Victor Slavin. On random matrices arising in deep neural networks: General i.i.d. case. *Random Matrices: Theory and Application*, 12(1), 2023.

[59] Barak A Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1): 147–160, 1994.

[60] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International conference on machine learning*, pages 2798–2806. PMLR, 2017.

[61] Zhaonan Qu, Wenzhi Gao, Oliver Hinder, Yinyu Ye, and Zhengyuan Zhou. Optimal diagonal preconditioning: Theory and practice. *arXiv preprint arXiv:2209.00809*, 2022.

[62] Nicolas Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. *Advances in neural information processing systems*, 20, 2007.

[63] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[64] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[65] Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9481–9488, 2021.

[66] Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. *Advances in Neural Information Processing Systems*, 34:23914–23927, 2021.

[67] Ruoyu Sun. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

[68] Ruoyu Sun and Yinyu Ye. Worst-case complexity of cyclic coordinate descent: O (nˆ 2) o (n 2) gap with randomized version. *Mathematical Programming*, 185:487–520, 2021.

[69] M. Talagrand. *Spin Glasses: A Challenge for Mathematicians. Cavity and Mean Field Models*. Springer, 2003.

[70] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[71] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[72] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

[73] Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Lei Wu, et al. The sharpness disparity principle in transformers for accelerating language model pre-training. *arXiv preprint arXiv:2502.19002*, 2025.

[74] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35:9983–9994, 2022.

[75] Mingwei Wei and David J Schwab. How noise affects the hessian spectrum in overparameterized neural networks. *arXiv preprint arXiv:1910.00195*, 2019.

[76] Eugene P Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67(2):325–327, 1958.

[77] Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

[78] Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

[79] Jianfeng Yao, Shurong Zheng, and Zhidong Bai. *Large sample covariance matrices and high-dimensional data analysis*. 2015.

[80] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[81] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

[82] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

[83] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32, 2019.

[84] Huishuai Zhang, Caiming Xiong, James Bradbury, and Richard Socher. Block-diagonal hessian-free optimization for training neural networks. *arXiv preprint arXiv:1712.07296*, 2017.

[85] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024.

[86] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

Table of Contents for the Appendix

Appendix A. Preliminaries and Intuitive Understanding

Notations. For a matrix $X \in \mathbb{R}^{m\times n}$, we denote $X^\top$ as the transpose of X. We use $I_{n\times n}$ and $0_{n\times n}$ to denote the identity matrix and the zero matrix of size $n \times n$. We denote $\|X\|_{\mathrm{F}}$ as the Frobenius norm of X. We denote $[n]$ as the index set $\{1, \cdots, n\}$. We say $x \overset{d}{=} y$ if the random variable (r.v.) x and y share the same distribution. We denote the Dirac measure at x by δ_x. We denote the support of measure μ by $\mathrm{supp}(\mu)$ and the expectation of x by $\mathbf{E}[x]$. We use $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{X}^2(\mu, \sigma^2)$ to denote Gaussian and chi-square distribution with mean μ and variance σ^2. We use $\Im(z)$ to denote the image part of a complex number $z \in \mathbb{C}$. We denote $\mathbb{C}^+ = \{z \in \mathbb{C} | \Im(z) > 0\}$. In this paper, we will intermittently employ the notations H_{ww}, H_{vv}, and H_{wv} to denote the hidden-layer, output-layer, and cross-layer Hessian, respectively, of a 1-hidden-layer network.

Our settings. We consider a general setting of multi-class classification problems. Given a classification dataset with n samples $\{(x_n, y_n)\}_{n=1}^N$, where $x_n \in \mathbb{R}^d$ is the input data, $y_n \in \{1, \cdots, C\}$ is the label, and C is the number of classes. This setting is quite general: it covers simple logistic regressions, as well as the most advanced LLMs. we consider the following four cases.

Case 1: linear models with MSE loss. Consider the linear model $f(V; x) = Vx \in \mathbb{R}^C$, where $V = (v_1^\top; \cdots; v_C^\top) \in \mathbb{R}^{C\times d}$ is the weight matrix, and $v_i \in \mathbb{R}^d$ is the weight associated with the i-th class (or output neuron). Consider minimizing the MSE loss as follows:

$$\min_V \ell_{\mathrm{MSE}}(V) := \frac{1}{N} \sum_{n=1}^{N} \|Vx_n - \mathcal{Y}_n\|_2^2, \tag{14}$$

where $\mathcal{Y}_n \in \{0, 1\}^C$ is a C-dimensional one-hot vector with 1 at the index for the class of y_n and 0 elsewhere. The Hessian matrix is, for $i, j \in [C]$:

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{MSE}}(V)}{\partial v_i \partial v_i^\top} = \frac{1}{N} \sum_{n=1}^N x_n x_n^\top, \\ \frac{\partial^2 \ell_{\mathrm{MSE}}(V)}{\partial v_i \partial v_j^\top} = 0_{d\times d}. \end{cases} \tag{15}$$

Here, the Hessian is always block-diagonal with C blocks. Note that the expression in (15) holds for general real-valued vector $\mathcal{Y}_n \in \mathbb{R}^C$, so the same Hessian structure also arises in the regression tasks.

Case 2: linear models with CE loss. We now change the loss function in **Case 1** to the CE loss.

$$\min_V \ell_{\mathrm{CE}}(V) := -\frac{1}{N} \sum_{n=1}^{N} \log \left(\frac{\exp(v_{y_n}^\top x_n)}{\sum_{c=1}^C \exp(v_c^\top x_n)} \right). \tag{16}$$

Define $p_{n,i} := \exp(v_i^\top x_n) / \left(\sum_{c=1}^C \exp(v_c^\top x_n)\right)$. The Hessian matrix is, for $i, j \in [C]$.

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top} = \frac{1}{N} \sum_{n=1}^N p_{n,i}(1 - p_{n,i}) x_n x_n^\top, \\ \frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_j^\top} = -\frac{1}{N} \sum_{n=1}^N p_{n,i} p_{n,j} x_n x_n^\top. \end{cases} \tag{17}$$

Intuitive understanding: at random initialization, suppose each entry in 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:

$$\frac{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_j^\top}\right\|_{\mathrm{F}}}{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top}\right\|_{\mathrm{F}}} \approx \frac{\sum_{n=1}^N p_{n,i} p_{n,j}}{\sum_{n=1}^N p_{n,i}(1-p_{n,i})} \approx \frac{\frac{1}{C^2}}{\frac{1}{C}\left(1-\frac{1}{C}\right)} = \frac{1}{C-1}, \tag{18}$$

which pushes the Hessian to become block-diagonal as $C \to \infty$.

Case 3: 1-hidden-layer networks with MSE loss. We now consider the 1-hidden-layer network with m hidden neurons: $f(W, V; x) = V\sigma(Wx) \in \mathbb{R}^C$, where $W \in (w_1^\top; \cdots, w_m^\top) \in \mathbb{R}^{m \times d}$; $\sigma(z) = \max\{0, z\}$ is the ReLU activation and is applied elementwise to Wx; $V = (v_1^\top; \cdots; v_C^\top) \in \mathbb{R}^{C \times m}$. Consider the MSE loss as follows.

$$\min_{W,V} \ell_{\mathrm{MSE}}(W, V) := \frac{1}{N} \sum_{n=1}^N \|V\sigma(Wx_n) - \mathcal{Y}_n\|_2^2. \tag{19}$$

The hidden-layer Hessian H_{ww} is: for $i, j \in [m]$,

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_i^\top} = \frac{1}{N}\left(\sum_{c=1}^C v_{c,i}^2\right)\left(\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0) x_n x_n^\top\right), \\ \frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_j^\top} = \frac{1}{N}\left(\sum_{c=1}^C v_{c,i} v_{c,j}\right)\left(\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0)\mathbf{1}(w_j^\top x_n > 0) x_n x_n^\top\right). \end{cases} \tag{20}$$

The output-layer Hessian H_{vv} is: for $i, j \in [C]$,

$$\begin{cases} \frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial v_i \partial v_i^\top} = \frac{1}{N}\sum_{n=1}^N \sigma(Wx_n)\sigma(Wx_n)^\top, \\ \frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial v_i \partial v_j^\top} = 0_{d \times d}, \end{cases} \tag{21}$$

The output-layer Hessian is block-diagonal. We now discuss the hidden-layer Hessian.

Intuitive understanding: at random initialization, suppose entries in $v_i \in \mathbb{R}^d$ follow an i.i.d. zero-mean Gaussian distribution, then

$$\frac{\left\|\frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\mathrm{F}}}{\left\|\frac{\partial^2 \ell_{\mathrm{MSE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\mathrm{F}}} \approx \frac{\left(\sum_{c=1}^C v_{c,i} v_{c,j}\right)}{\left(\sum_{c=1}^C v_{c,i}^2\right)} \overset{C\to\infty}{=} \frac{\mathrm{Cov}(v_{i,i}, v_{i,j})}{\mathrm{Var}(v_{i,i})}. \tag{22}$$

As $v_{i,i}, v_{i,j}$ are independent, $\mathrm{Cov}(v_{i,i}, v_{i,j}) = 0$ and thus the block-diagonal structure in H_{ww} occurs as $C \to \infty$.

We now discuss the cross-layer component H_{wv} under MSE loss:

$$\frac{\partial^2 \ell_{\mathrm{MSE}}}{\partial w_i \partial v_j^\top} = \frac{2}{N}\sum_{n=1}^N \left[\left(\sigma(Wx_n)^\top v_j - \mathcal{Y}_{n,j}\right)\mathbf{1}(w_i^\top x_n > 0) x_n e_i^\top + v_{j,i}\mathbf{1}(w_i^\top x_n > 0) x_n \sigma(Wx_n)^\top\right], \tag{23}$$

where $e_i \in \mathbb{R}^m$ is an one-hot vector with the i-th component equals to 1. Note that the 2nd term has expectation 0 when $v_{j,i}$ is initialized as a zero-mean Gaussian distribution. As for the 1st term, it is a matrix of the form

$$\begin{bmatrix} 0 & \cdots & a_{1,i} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & a_{d,i} & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{d\times m}, \tag{24}$$

which is a matrix with one non-zero column at position i with

$$a_{d',i} = \frac{1}{N}\sum_{n=1}^{N}\left(\sigma(Wx_n)^\top v_j - \mathcal{Y}_{n,j}\right)\mathbf{1}(w_i^\top x_n > 0)x_{n,d'}, \quad d' \in [d].$$

Note that v_j is initialized as a zero-mean Gaussian distribution, so the inner product $\sigma(Wx_n)^\top v_j$ has expectation 0. Further, as the training goes, $(\sigma(Wx_n)^\top v_j - \mathcal{Y}_{n,j}) \to 0$ and thus the 1st term in H_{wv} shall approach 0 along training. Numerically, we observe that H_{wv} under MSE loss is indeed negligible compared to H_{ww} and H_{vv}. This is observed throughout the training, including at the initialization (see Figure 2).

Case 4: 1-hidden-layer networks with CE loss. We now consider 1-hidden-layer networks with CE loss.

$$\min_{W,V} \ell_{\text{CE}}(W,V) := -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{\exp(v_{y_n}^\top\sigma(Wx_n))}{\sum_{c=1}^{C}\exp(v_c^\top\sigma(Wx_n))}\right). \tag{25}$$

The hidden-layer Hessian H_{ww} is: for $i,j \in [m]$,

$$\begin{cases} \frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial w_i \partial w_i^\top} = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{c=1}^{C} p_{n,c}v_{c,i}^2 - \left(\sum_{c=1}^{C} p_{n,c}v_{c,i}\right)^2\right)\mathbf{1}(w_i^\top x_n > 0)x_n x_n^\top, \\ \frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial w_i \partial w_j^\top} = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{c=1}^{C} p_{n,c}v_{c,i}v_{c,j} - \left(\sum_{c=1}^{C} p_{n,c}v_{c,i}\right)\left(\sum_{c=1}^{C} p_{n,c}v_{c,j}\right)\right)\mathbf{1}(w_i^\top x_n > 0)\mathbf{1}(w_j^\top x_n > 0)x_n x_n^\top. \end{cases} \tag{26}$$

The output-layer Hessian H_{vv} is: for $i,j \in [C]$,

$$\begin{cases} \frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial v_i \partial v_i^\top} = \frac{1}{N}\sum_{n=1}^{N} p_{n,i}(1-p_{n,i})\sigma(Wx_n)\sigma(Wx_n)^\top, \\ \frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial v_i \partial v_j^\top} = -\frac{1}{N}\sum_{n=1}^{N} p_{n,i}p_{n,j}\sigma(Wx_n)\sigma(Wx_n)^\top. \end{cases} \tag{27}$$

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:

$$\frac{\left\|\frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial w_i \partial w_j^\top}\right\|_{\text{F}}}{\left\|\frac{\partial^2\ell_{\text{CE}}(W,V)}{\partial w_i \partial w_i^\top}\right\|_{\text{F}}} \approx \frac{\left(\sum_{c=1}^{C} v_{c,i}v_{c,j} - \left(\sum_{c=1}^{C} v_{c,i}\right)\left(\sum_{c=1}^{C} v_{c,j}\right)\right)/C}{\left(\sum_{c=1}^{C} v_{c,i}^2 - \left(\sum_{c=1}^{C} v_{c,i}\right)^2\right)/C} \overset{C\to\infty}{=} \frac{\text{Cov}(v_{i,i}, v_{i,j})}{\text{Var}(v_{i,i})}. \tag{28}$$

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

$$\frac{\left\| \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_j^\top} \right\|_{\mathrm{F}}}{\left\| \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top} \right\|_{\mathrm{F}}} \approx \frac{\sum_{n=1}^{N} p_{n,i} p_{n,j}}{\sum_{n=1}^{N} p_{n,i}(1 - p_{n,i})} \approx \frac{\frac{1}{C^2}}{\frac{1}{C}\left(1 - \frac{1}{C}\right)} = \frac{1}{C-1}, \tag{29}$$

and thus the block-diagonal structure in H_{vv} arises as $C \to \infty$.

We now discuss the cross-layer component H_{wv} under CE loss. We will explain the block-circulant structure at initialization (i.e., Figure 3 (a)) and why it vanishes along training (i.e., Figure 3 (b-f)). We find that this phenomenon can be seen by a direct Hessian calculation. The cross-layer Hessian is

$$\frac{\partial^2 \ell_{\mathrm{CE}}}{\partial w_i \partial v_j^\top} = \frac{1}{N} \sum_{n=1}^{N} \left[(p_{n,j} - \delta_{y_n,j}) \mathbf{1}(w_i^\top x_n > 0) x_n e_i^\top + \sum_{c=1}^{C} (\delta_{j,c} - p_{n,c}) p_{n,j} v_{c,i} \mathbf{1}(w_i^\top x_n > 0) x_n \sigma(W x_n)^\top \right]. \tag{30}$$

When C is large, by the law of large number and approximating $p_{n,j} \approx 1/C$, for the 2nd-term we have

$$\begin{aligned} &\sum_{c=1}^{C} (\delta_{j,c} - p_{n,c}) p_{n,j} v_{c,i} \mathbf{1}(w_i^\top x_n > 0) x_n \sigma(W x_n)^\top \\ \approx & p_{n,j} v_{j,i} \mathbf{1}(w_i^\top x_n > 0) x_n \sigma(W x_n)^\top - p_{n,j} x_n \sigma(W x_n)^\top \left(\frac{1}{C} \sum_{c=1}^{C} v_{c,i} \mathbf{1}(w_i^\top x > 0) \right) \\ \approx & \frac{1}{C} \cdot v_{j,i} \mathbf{1}(w_i^\top x_n > 0) x_n \sigma(W x_n)^\top. \end{aligned} \tag{31}$$

Thus

$$\frac{\partial^2 \ell_{\mathrm{CE}}}{\partial w_i \partial v_j^\top} \approx \frac{1}{N} \sum_{n=1}^{N} (p_{n,j} - \delta_{y_n,j}) \mathbf{1}(w_i^\top x_n > 0) x_n e_i^\top + \mathcal{O}\left(\frac{1}{C}\right), \tag{32}$$

As such, the leading term of H_{wv} under CE loss is a matrix of the form

$$\begin{bmatrix} 0 & \cdots & a_{1,i} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & a_{d,i} & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{d \times m}, \tag{33}$$

which is a matrix with one non-zero column at position i with

$$a_{d',i} = \frac{1}{N} \sum_{n=1}^{N} \left(p_{n,c} - \delta_{y_n,c}\right) \mathbf{1}\left(w_i^\top x_n \geq 0\right) x_{n,d'}, \quad d' \in [d].$$

This explains the initial block-circulant structure in H_{wv} under CE loss. As the training goes, $\left(p_{n,c} - \delta_{y_n,c}\right) \to 0$ and the block-circulant structure disappears. This shows the "dynamic force" arisen from training.

Appendix B. Related works

Hessian spectrum analysis. Most studies on Hessian of NNs focus on Hessian eigenvalue distribution, a.k.a., the spectrum. Chaudhari et al. [7], Dauphin et al. [14], Ghorbani et al. [19], Granziol et al. [22], LeCun et al. [36], Sagun et al. [63, 64], Yao et al. [81] reported that the Hessian spectra of NNs consist of a “bulk" together with a few “outliers". Fort and Ganguli [17], Liao and Mahoney [39], Papyan [54], Pennington and Bahri [60], Singh et al. [66], Wu et al. [78] studied the shape of the Hessian spectrum and Hessian rank in theory. Papyan [52, 53], Sankar et al. [65] numerically studied the relation between the spectrum of Hessian and that of Gauss-Newton matrix. Granziol et al. [23], Keskar et al. [32], Yao et al. [80], Zhang et al. [83] studied the connection between the Hessian spectrum of NNs and some training phenomena such as the effect of batch sizes. Ghorbani et al. [19], Yao et al. [81] explained the effectiveness of training techniques such as BatchNorm via the shape of Hessian spectrum. Zhang et al. [85] numerically studied the blockwise Hessian spectrum of CNNs and Transformers. They further connect the blockwise spectra to the effectiveness of Adam. Another line of works studied the interplay between Hessian extreme eigenvalues and the trajectories of gradient methods (e.g., [1, 3, 7–9, 16, 25, 27–30, 38, 44, 55, 74, 75, 77]).

Different from all these works, we study the macroscopic structure of the Hessian rather than its spectrum. Note that these two topics are rather orthogonal: it is possible to change the matrix structure without changing its eigenvalues, and vice versa. Specifically, we focus on the ratio between diagonal Hessian blocks and off-diagonal ones, which is not covered in the spectrum analysis.

Hessian structure analysis. Collobert [11] empirically observed the following phenomenon: when using a neural network to solve a binary classification problem under CE loss, the Hessian is near-block-diagonal. They also reported that the near-block-diagonal structure disappears when changing to Mean-Square (MSE) loss. Collobert [11] thereby conjectured that the near-block-diagonal Hessian stems from CE loss, and they provided an one-line informal explanation (re-stated later in Section C). The near-block-diagonal structure was also reported recently under CE loss for various models including 1-hidden-layer network [85], 1-hidden-layer Transformers [86], and linear models [34]. Similar Hessian structure is later numerically reported on more practical models including GPT-2 [45], and OPT-125M [46]. These results show that the near-block-diagonal structure appeared in a wide range of architectures. We point out that these works primarily focus on empirical observations, and the rigorous theoretical analysis is still missing.

Very recently, Ormaniec et al. [51] employed matrix calculus to derive the Hessian expression of a 1-hidden-layer Transformer to understand the difficulties in training Transformers. It is valuable and non-trivial to derive the Hessian expression of Transformers due to their complicated design. However, the subsequent analysis of Hessian structure is relatively simplified: e.g., they view the weights and data as constant matrices and did not incorporate their random distributions. Consequently, the exact distribution of each Hessian block has not been characterized yet, and the origin of the near-block-diagonal structure remains unexplored.

Different from the aforementioned work, we establish the first rigorous theory on the Hessian structure of linear and 1-hidden-layer network via random matrix theory. Our theory reveals that the number of classes C is one major cause of the near-block-diagonal or block-circulant-block-diagonal structure.

Algorithm design. Multiple algorithm designs are proposed by approximating Hessian (or other curvature matrices) by block-diagonal matrices (e.g., [2, 12, 15, 18, 20, 24, 31, 47, 62, 72, 84]). Our theory can explain why these methods work. The special Hessian structure also has strong connections to diagonal preconditioned methods (e.g., [33, 42]).

Appendix C. Existing Wisdom

Here, we revisit the results in [11], which has remained for two-decades the dominating understanding of the near-block-diagonal Hessian structure. The author attributes the near-block-diagonal structure to CE loss. We will point out that this perspective might not be accurate.

Collobert [11] considered the binary classification problem: minimizing $\ell_{\mathrm{CE}}(f(\theta;x),y)$ where $\ell_{\mathrm{CE}}(\cdot,\cdot)$ is CE loss, $f(\theta;x)=\sum_{i=1}^{n} v_i\sigma(w_i^\top x)\in\mathbb{R}$ is an single-output-1-hidden-layer neural network with input $x\in\mathbb{R}^d$, weight $w_i\in\mathbb{R}^d$, $v_i\in\mathbb{R}$, and label $y\in\{0,1\}$. The author focused on the hidden-layer Hessian H_{ww}, and they point out that off-diagonal-blocks in H_{ww} would contain

$$\frac{\partial^2\ell_{\mathrm{CE}}(f(\theta;x),y)}{\partial w_i\partial w_j^\top}=p\,(1-p)v_iv_j\sigma'\left(w_i^\top x\right)\sigma'\left(w_j^\top x\right)xx^\top \quad \text{for } i\neq j, \tag{34}$$

where $p=1/(1+\exp(-yf(\theta,x)))$ denotes the probability of correct prediction, and $\sigma'(\cdot)$ is the derivative of $\sigma(\cdot)$. Collobert [11] argued that CE loss is the key factor for the near-block-diagonal structure. The author provided a one-line intuitive explanation: since the training objective is to maximize p, the term $p\,(1-p)$ will decay to zero, which pushes the off-diagonal blocks to zero. Numerically, Collobert [11] reported that CE loss brings the near-block-diagonal structure in H_{ww}, while MSE loss does not (their Figure 7.3 & 7.5, also restated in Figure 10 in Appendix G.1). The author argued that this is because MSE loss does not produce the term $p\,(1-p)$.

We point out that the arguments in [11] might not be accurate. In particular:

1. For binary classification, the term $p(1-p)$ occurs for both diagonal and off-diagonal blocks. This can be easily inferred in our latter analysis in **Case 1** with $C=2$. Therefore, it cannot serve as a distinguishing factor between the diagonal and off-diagonal blocks. For CE loss, the observed block-diagonal structure in H_{ww} might be due to other properties.

2. Our numerical results in Figure 1 show that the near-block-diagonal structure in H_{ww} occurs not only during the training, but also at initialization. As such, the special structure is not the result of "maximizing p" or "minimizing $(1-p)$".

In our analysis, we show that the number of classes C, instead of the CE loss, is one key factor. Specifically, the near-block-diagonal structure in H_{ww} arises as $C\to\infty$ for *both* the MSE and the CE loss. [11] did not observe the special structure under the MSE loss because binary classification with $C=2$ was considered.

We emphasize that we do not claim "large C" as the *only* cause for the near-block-diagonal structure in H_{ww}, but just that it is a sufficient condition. It is also possible that the special structure arises with small C (Figure 7.3 in [11]) due to other reasons, which we have not explored yet.

Appendix D. Proof Sketches and Technical Challenges

Now we explain the major technical challenges and the main ideas in our proofs. We primarily introduce the proof procedure for Theorem 1, i.e., linear models with CE loss (**Case 2**). Despite the simple form of linear models, we find that it is rather non-trivial to characterize its Hessian structure, and the classical random matrix approaches *cannot* be directly applied. After introducing the proof for Theorem 1, we will discuss how to extend our analysis to Theorem 2 (**Case 3** and **4**). **Since the complete proof of Theorem 1 and 2 are rather long and technical, we present them in Appendix E.1 and E.2, respectively.**

Challenges for proving Theorem 1. We first rewrite the Hessian expression in **Case 2** as follows, and then we discuss why the classical random matrix approaches cannot be directly applied here. Due to the limited space, we only discuss the diagonal blocks $\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top}$. The same challenges and solutions also apply to the off-diagonal blocks $\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_j^\top}$, which we omit here.

$$\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top} \overset{(17)}{=} \frac{1}{N} \sum_{n=1}^{N} p_{n,i}(1-p_{n,i}) x_n x_n^\top := \frac{1}{N} X_N \Lambda_N X_N^\top \in \mathbb{R}^{d \times d}, \tag{35}$$

where $X_N = (x_1, \cdots, x_N) \in \mathbb{R}^{d \times N}$, $\Lambda_N = \mathrm{diag}(p_{1,i}(1-p_{1,i}), \cdots, p_{N,i}(1-p_{N,i})) \in \mathbb{R}^{N \times N}$, and $p_{n,i} := \exp(v_i^\top x_n) / \left(\sum_{c=1}^{C} \exp(v_c^\top x_n)\right)$. Note that both X_N and Λ_N are random matrices. How to characterize $\| \frac{1}{N} X_N \Lambda_N X_N^\top \|_{\mathrm{F}}$? We first recall some classical results in random matrix theory.

Classical results from random matrix theory and why they cannot be directly applied. We first introduce some basic concepts in random matrix theory.

- Eigenvalue distribution and weak convergence of probability measures. For a symmetric matrix $A \in \mathbb{R}^{d \times d}$, we define $\mu_A = \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i(A)}$ as the empirical eigenvalue distribution of A. μ_A is a probability measure on $\mathbb{R}$ that assigns equal probability $\frac{1}{d}$ to each eigenvalue. Note that when A is a random matrix, μ_A is a random measure. For a sequence of random matrices $(A_n)_{n=1}^{\infty}$, we will consider the weak convergence of its eigenvalue distribution $(\mu_{A_n})_{n=1}^{\infty}$.
- Stieltjes transform. For a probability measure ν on $\mathbb{R}$, The Stieltjes transform of ν is defined as
$$s_\nu(z) = \int_{\mathbb{R}} \frac{1}{x-z} d\nu(x), \quad z \in \mathbb{C}^+ \backslash \operatorname{supp}(\nu).$$
A probability measure is uniquely characterized by its Stieltjes transform. For a symmetric matrix A, we write $s_{\mu_A}(z)$ as $s_A(z)$ for short. A sequence of eigenvalue distributions $(\mu_{A_n})_{n=1}^{\infty}$ converges weakly to a probability measure μ if and only if $s_{A_n}(z) \to s_\mu(z)$, $\forall z \in \mathbb{C}^+$.

Now we notice that $\|A\|_{\mathrm{F}}^2$ is nothing but the 2nd-order moment of μ_A. Moreover, we can retrieve the moments of μ_A from $s_A(z)$ by

$$s_A(z) = -\frac{1}{z} - \frac{m_1}{z^2} - \frac{m_2}{z^3} - \cdots, \quad z \to \infty, \tag{36}$$

where $m_k = \int_{\mathbb{R}} t^k d\mu_A(t)$ denotes the k-th order moment. Therefore, the calculation of $\|A\|_F^2$ can be achieved by finding the limiting eigenvalue distribution of A as the matrix size $d \to \infty$, which is a classical topic in random matrix theory. Typically, the random matrices of the type $X_N \Lambda_N X_N^\top$ are closely related to the sample covariance matrices. The case that Λ_N is deterministic or independent of X_N has already been deeply studied (e.g., Bai and Silverstein [4], Bai and Zhou [5], Marčenko and Pastur [48], Yao et al. [79]). For the limiting eigenvalue distribution, we have the following classical result, the generalized Marcenko-Pastur theorem.

Proposition 1 *[48] Consider random matrices $X_N \in \mathbb{R}^{d\times N}$ with entries i.i.d. with mean 0 and variance 1; and $\Lambda_N \in \mathbb{R}^{N\times N}$ which is either deterministic or independent of X_N. Suppose that the eigenvalue distribution of Λ_N converges weakly almost surely to a deterministic probability measure ν. Let $A_N = \frac{1}{d} X_N \Lambda_N X_N^\top$, then as $N, d \to \infty, d/N \to \gamma \in (0, +\infty)$, μ_{A_N} converges weakly almost surely to a deterministic probability measure μ. Here μ is uniquely specified by a functional equation of its Stietjes transform $s(z)$:*

$$s_\mu(z) = \frac{1}{\frac{1}{\gamma}\int_{\mathbb{R}} \frac{t d\nu(t)}{1+ts_\mu(z)} - z}, \quad \forall z \in \mathbb{C}^+. \tag{37}$$

Unfortunately, Proposition 1 can *not* be directly applied to our case. This is because Proposition 1 requires Λ_N and X_N to be independent, while Λ_N and X_N in (35) are clearly dependent.

Key observation: asymptotic independence. For our matrix of interests (35), although $\Lambda_N \in \mathbb{R}^{N\times N}$ and $X_N \in \mathbb{R}^{d\times N}$ are not independent for any fixed d, we observe that they are *asymptotically independent* as $d \to \infty$. This is because:

- Recall $v_i \sim \mathcal{N}(0, \frac{1}{d})$ and denote $z = v_i^\top x_n$, then $z|x \sim \mathcal{N}(0, \frac{\|x\|_2^2}{d})$. Further, since $x \sim \mathcal{N}(0,1)$, we have $\frac{\|x\|_2^2}{d} \sim \mathcal{X}^2(1, \frac{2}{d})$, which concentrates to 1 as $d \to \infty$.
- As such, z asymptotically follows $\mathcal{N}(0,1)$ and thus is independent of x. Therefore, Λ_N and X_N are asymptotically independent.

Therefore, as $d, N \to \infty$, it seems possible to obtain the same limiting eigenvalue distribution as in Proposition 1 for our matrix (35). The remaining question is how to prove it rigorously. One possible path is from free probability theory [49], proving the asymptotic freeness between Λ_N and $X_N X_N^\top$ [10]. We will instead take another path in this work.

Our solutions. In this work, we use a rather classical decoupling technique, motivated by *the Lindeberg interpolation principle*. The Lindeberg principle is originally an elegant proof for the Central Limit Theorem (CLT) [40], by replacing the random variables with Gaussian ones incrementally and proving the impact is negligible under certain conditions. The Lindeberg principle is also applicable for random matrices [6, 21, 57]. We find that such methods are useful for handling asymptotic independence in our case.

We now illustrate our proof strategy. The idea is to first decouple and then apply Proposition 1.

- **Step 1.** For our matrix (35) (denoted as H_{ii}^{CE}), we introduce the decoupled matrix

$$\widetilde{H}_{ii}^{\text{CE}} = \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}(1-\widetilde{p}_{n,i})x_n x_n^\top, \quad \widetilde{p}_{n,i} := \frac{\exp(v_i^\top \widetilde{x}_n)}{\sum_{c=1}^{C}\exp(v_c^\top \widetilde{x}_n)}, \tag{38}$$

  where $\widetilde{X}_N = (\widetilde{x}_1, \cdots, \widetilde{x}_N) \in \mathbb{R}^{d\times n}$ is an independent copy of X_N. The goal is to prove that

$$\lim_{N\to\infty}\left(s_{H_{ii}^{\text{CE}}}(z) - s_{\widetilde{H}_{ii}^{\text{CE}}}(z)\right) = 0, \quad \text{a.s.} \quad \forall z \in \mathbb{C}^+.$$

  From standard measure concentration results, $s_{H_{ii}^{\text{CE}}}(z), s_{\widetilde{H}_{ii}^{\text{CE}}}(z)$ concentrates around their means as $N \to \infty$. Therefore, it suffices to prove that $\lim_{N\to\infty}\left(\mathbf{E}[s_{H_{ii}^{\text{CE}}}(z)] - \mathbf{E}[s_{\widetilde{H}_{ii}^{\text{CE}}}(z)]\right) = 0$.

- **Step 2.** Following the Lindeberg principle, we define the matrix interpolation process

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

  Note that $X_N(0) = X_N$ and $X_N(1) = \widetilde{X}_N$. We then define

$$H_{ii}^{\text{CE}}(t) = \frac{1}{N}\sum_{n=1}^{N} p_{n,i}(t)(1-p_{n,i}(t))x_n x_n^\top, \quad \text{where } p_{n,i}(t) := \frac{\exp(v_i^\top x_n(t))}{\sum_{c=1}^{C}\exp(v_c^\top x_n(t))}.$$

  Using basic matrix calculus, we arrive at the following equation:

$$\mathbf{E}[s_{H_{ii}^{\text{CE}}}(z)] - \mathbf{E}[s_{\widetilde{H}_{ii}^{\text{CE}}}(z)] = \int_0^1 \mathbf{E}\left[\frac{d}{dt}s_{H_{ii}^{\text{CE}}(t)}\right]dt. \tag{39}$$

- **Step 3.** We notice that the r.h.s. of (39) can be bounded in the form of $\mathbf{E}[Zf(Z)]$. We then bound it using Stein's Lemma: for $Z \sim \mathcal{N}(0,1)$ and differentiable function $f : \mathbb{R} \to \mathbb{C}$ with sub-exponential decay at infinity, we have:

$$\mathbf{E}[Zf(Z)] = \mathbf{E}[f'(Z)]. \tag{40}$$

  We then prove that the r.h.s. of (40) decays to zero at rate $\mathcal{O}(1/\sqrt{N})$, and thus H_{ii}^{CE} shares the same limit distribution as $\widetilde{H}_{ii}^{\text{CE}}$. Note that Stein's Lemma requires Gaussian data X_N in Assumption 1. We refer to Appendix A.6 of [69] for the proof of Stein's Lemma.

- **Step 4.** Apply Proposition 1. The decoupled matrix $\widetilde{H}_{ii}^{\text{CE}}$ has the type $X_N\widetilde{\Lambda}_N X_N^\top$ where $\widetilde{\Lambda}_N$ is independent of X_N. Then Proposition 1 is applicable to obtain the limiting eigenvalue distribution of $\widetilde{H}_{ii}^{\text{CE}}$. Then we apply the expansion (36) in the functional equation (37) to get the limiting second moment of $\mu_{\widetilde{H}_{ii}^{\text{CE}}}$, which is also the limit of $\|H_{ii}^{\text{CE}}\|_{\text{F}}^2$. This concludes the proof.

Challenges for the hidden-layer Hessian in 1-hidden-layer networks. Now we extend our analysis on linear models to 1-hidden-layer networks. Similar as before, We primarily discuss the diagonal blocks of the Hessian. The same challenges and solutions also apply to the off-diagonal blocks, which we omit here. We first discuss the hidden-layer Hessian.

$$\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_i \partial w_i^\top} = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\left(\sum_{c=1}^{C} p_{n,c} v_{c,i}^2 - \left(\sum_{c=1}^{C} p_{n,c} v_{c,i}\right)^2\right)}_{(a)} \underbrace{\mathbf{1}(w_i^\top x_n > 0)}_{(b)} x_n x_n^\top := \frac{1}{N} X_N \Lambda_N X_N^\top, \tag{41}$$

Before applying the same decoupling technique as in linear models, we first show that X_N and Λ_N are also asymptotic independent as $d \to \infty$. We present the following reasons:

- We start with (b) first. Recall $w_i \sim \mathcal{N}(0, \frac{1}{d})$ and denote $z_{n,i} = w_i^\top x_n$. Following the same argument as in the linear model case, we have $z_{n,i} | x_n \sim \mathcal{N}(0,1)$ as $d \to \infty$. So $z_{n,i}$ is asymptotically independent of x_n, and thus $\mathbf{1}(z_{n,i} > 0)$ is also asymptotically independent of x_n, $\forall n \in [N]$. Therefore, (b) and X_N are asymptotically independent.
- Now we discuss (a). Denote $y_{n,c} = (z_{n,1}, \cdots, z_{n,m})^\top v_c$. Since $(z_{n,1}, \cdots, z_{n,m})^\top$ is independent of x_n as $d \to \infty$, so do $y_{n,c}$ and $p_{n,c}$. As such, (a) and X_N are asymptotically independent.

As such, (a) and (b) are asymptotically independent of X_N, thus it is reasonable that we can apply our previous decoupling technique here. A similar procedure also applies to the hidden weights with MSE loss. We present the detailed proof in Appendix E.2.1 and E.2.2, respectively.

Challenges for the output-layer Hessian in 1-hidden-layer networks. Now we discuss the new challenges for the output-layer Hessian. We rewrite $\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top}$ as follows.

$$\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top} = \frac{1}{N}\sum_{n=1}^{N} p_{n,i}(1-p_{n,i})\sigma(Wx_n)\sigma(Wx_n)^\top := \frac{1}{N} F_N \Lambda_N F_N^\top, \tag{42}$$

where $F_N = (\sigma(Wx_1), \cdots, \sigma(Wx_N)) \in \mathbb{R}^{m \times N}$. We highlight two major difference with the $X_N \Lambda_N X_N^\top$ in linear models (35) and the hidden weights in 1-hidden-layer networks in (41).

- **First,** in the previous cases (35) and (41), the matrices have growing dimension. Now the matrix in (42) has fixed dimension m, which is away from the standard setting of random matrix theory.
- **Second,** the matrices F_N and Λ_N have more complicated dependence structure. The dependence between Z_N and Λ_N is caused by both W and X_N, which means that we need to decouple Λ_N with $\{W, X_N\}$ at the same time. Meanwhile, in previous cases (35) and (41), we only need to decouple Λ_N with X_N.

To tackle the above challenges, we choose to handle (42) using a largely different approach from above. We consider fixed m and conduct the following steps. Since the complete proof is rather long and technical, we relegate it to Appendix E.2.3.

- **Step 1.** Replace WX_N with Z_N, where $Z_N \in \mathbb{R}^{m \times N}$ has i.i.d. $N(0,1)$ entries. This can be done by letting $d \to \infty$ and applying the Lindeberg principle. This step decouples Λ_N with H_N.

- **Step 2.** Calculate the expectation of the entry-wise second moment of the Hessian matrices. Note that the decoupling in Step 1 is essential, otherwise, the calculation in Step 2 would be complicated.

Appendix E. Proofs of the main theorems

E.1. Proof of Theorem 1

Before delving into the proof, we first present the functions g_{ii}, g_{ij} in Theorem 1. Let $\mathbf{z}_C = (Z_1, \cdots, Z_C) \sim \mathcal{N}_C(0, I_C)$, define

$$h_1(\mathbf{z}_C) = \frac{e^{Z_1}}{\sum_{l=1}^{C} e^{Z_l}} \left(1 - \frac{e^{Z_1}}{\sum_{l=1}^{C} e^{Z_l}}\right), \quad h_2(\mathbf{z}_C) = \frac{e^{Z_1+Z_2}}{\left(\sum_{l=1}^{C} e^{Z_l}\right)^2},$$

then

$$g_{ii}(\gamma, C) := \gamma \mathbf{E}[h_1(\mathbf{z}_C)^2] + (\mathbf{E}[h_1(\mathbf{z}_C)])^2, \quad g_{ij}(\gamma, C) := \gamma \mathbf{E}[h_2(\mathbf{z}_C)^2] + (\mathbf{E}[h_2(\mathbf{z}_C)])^2.$$

We now introduce some notations that will be used in the proof. Let $\widetilde{X}_N = (\widetilde{x}_1, \cdots, \widetilde{x}_N) \in \mathbb{R}^{d\times N}$ be an independent copy of $X_N \in \mathbb{R}^{d\times N}$. Denote

$$p_i(x) := \frac{\exp(v_i^\top x)}{\sum_{c=1}^{C} \exp(v_c^\top x)}, \quad x \in \mathbb{R}^d,$$

$$\alpha_n = p_i(x_n)(1 - p_i(x_n)), \quad \Lambda_n = \mathrm{diag}(\alpha_1, \cdots, \alpha_N),$$
$$\widetilde{\alpha}_n = p_i(\widetilde{x}_n)(1 - p_i(\widetilde{x}_n)), \quad \widetilde{\Lambda}_n = \mathrm{diag}(\widetilde{\alpha}_1, \cdots, \widetilde{\alpha}_N),$$
$$\beta_n = -p_i(x_n)p_j(x_n), \quad \Gamma_n = \mathrm{diag}(\beta_1, \cdots, \beta_N),$$
$$\widetilde{\beta}_n = -p_i(\widetilde{x}_n)p_j(\widetilde{x}_n), \quad \widetilde{\Gamma}_n = \mathrm{diag}(\widetilde{\beta}_1, \cdots, \widetilde{\beta}_N).$$

To ease notations, we write $H_{ii} = \frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_i^\top} \in \mathbb{R}^{d\times d}$, $H_{ij} = \frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_i \partial v_j^\top} \in \mathbb{R}^{d\times d}$, then

$$H_{ii} = \frac{1}{N} X_N \Lambda_N X_N^\top,\ H_{ij} = \frac{1}{N} X_N \Gamma_N X_N^\top.$$

Similarly, we define

$$\widetilde{H}_{ii} = \frac{1}{N} X_N \widetilde{\Lambda}_N X_N^\top, \quad \widetilde{H}_{ij} = \frac{1}{N} X_N \widetilde{\Gamma}_N X_N^\top.$$

Now we prove Theorem 1. The proof consists of two steps. First, we "decouple" X_N, Λ_N and Γ_N. That is, we prove that H_{ii} and $\widetilde{H}_{ii}$ share the same Stieltjes transform as N and d grow proportionally to infinity. Similarly for H_{ij} and $\widetilde{H}_{ij}$. Second, with the help of Proposition 1, we find the second moments of limit eigenvalue distribution of $\widetilde{H}_{ii}$ and $\widetilde{H}_{ij}$. Recall that the second moment is the Frobenius norm, so the proof is concluded.

We now"decouple" X_N, Λ_N and Γ_N using the following Lemma 1.

Lemma 1 *For any $z \in \mathbb{C}^+$, as $d, N \to \infty$, $d/N \to \gamma \in (0, +\infty)$, it holds almost surely that*

$$s_{H_{ii}}(z) - s_{\widetilde{H}_{ii}}(z) = O\left(N^{-\frac{1}{2}}\right), \tag{43}$$

$$s_{H_{ij}}(z) - s_{\widetilde{H}_{ij}}(z) = O\left(N^{-\frac{1}{2}}\right). \tag{44}$$

Proof Here, we only present the proof for $s_{H_{ii}}$. The proof for $s_{H_{ij}}$ is done following the same procedure.

For $t \in [0, 1]$, let

$$X_N(t) = \sqrt{t} X_N + \sqrt{1-t} \widetilde{X}_N.$$

Then $X_N(t) = (x_1(t), \cdots, x_N(t)) \in \mathbb{R}^{d \times N}$, where

$$x_n(t) = \sqrt{t} x_n + \sqrt{1-t} \widetilde{x}_n, \quad n \in [N].$$

Denote

$$\begin{gathered}
\alpha_n(t) = p_i(x_n(t)) \left[1 - p_i(x_n(t))\right], \\
\Lambda_N(t) = \operatorname{diag}(\alpha_1(t), \cdots, \alpha_N(t)), \\
H_{ii}(t) = \frac{1}{N} X_N \Lambda_N(t) X_N^\top, \\
\mathcal{G}_N(z, t) = (H_{ii}(t) - z I_{d \times d})^{-1} \in \mathbb{R}^{d \times d},
\end{gathered}$$

By the definition of $s_{H_{ii}}$, it is easy to see that $s_{H_{ii}} = \frac{1}{d} \operatorname{tr}\left(\mathcal{G}_N(z, t)\right)$.

We first prove that $s_{H_{ii}}(z), s_{\widetilde{H}_{ii}}(z)$ concentrate around their mean. By treating $s_{H_{ii}}(z), s_{\widetilde{H}_{ii}}(z)$ as Lipchitz functions of the Gaussian vectors $v_1, \cdots, v_C, x_1, \cdots, x_N$, we have the following results from Talagrand's inequality,

$$\mathbf{P}\left(|s_{H_{ii}}(z) - \mathbf{E}[s_{H_{ii}}(z)]| \geq t\right) \leq c_1 e^{-pc_2 t^2},$$

$$\mathbf{P}\left(|s_{\widetilde{H}_{ii}}(z) - \mathbf{E}[s_{\widetilde{H}_{ii}}(z)]| \geq t\right) \leq \widetilde{c}_1 e^{-d\widetilde{c}_2 t^2},$$

where $t > 0$ and constants $c_1, c_2, \widetilde{c}_1, \widetilde{c}_2 > 0$. Then from the Borel-Cantelli Lemma,

$$\begin{aligned}
s_{H_{ii}}(z) - \mathbf{E}[s_{H_{ii}}(z)] &\overset{a.s.}{\to} 0, \\
s_{\widetilde{H}_{ii}}(z) - \mathbf{E}[s_{\widetilde{H}_{ii}}(z)] &\overset{a.s.}{\to} 0.
\end{aligned} \tag{45}$$

Now we prove that

$$\delta_N(z) = \mathbf{E}[s_{H_{ii}}(z)] - \mathbf{E}[s_{\widetilde{H}_{ii}}(z)] = O\left(N^{-\frac{1}{2}}\right). \tag{46}$$

Recall for any function $A(t)$ valued in invertible matrices, we have

$$\frac{d}{dt} A^{-1}(t) = -A^{-1}(t) \frac{d}{dt} A(t) A^{-1}(t). \tag{47}$$

Then we have

$$\begin{aligned}\delta_N(z) &= \frac{1}{d}\mathbf{E}\left[\operatorname{tr}\left(\mathcal{G}_N(z,1)-\mathcal{G}_N(z,0)\right)\right]\\ &= \frac{1}{d}\int_0^1 \frac{d}{dt}\mathbf{E}\left[\operatorname{tr}\left(\mathcal{G}_N(z,t)\right)\right]dt\\ &= \frac{1}{d}\int_0^1 \frac{d}{dt}\mathbf{E}\left[\operatorname{tr}\left(\left(H_{ii}(t)-zI_{d\times d}\right)^{-1}\right)\right]dt \qquad (48)\\ &\overset{(47)}{=} -\frac{1}{d}\int_0^1 \mathbf{E}\left[\operatorname{tr}\left(\mathcal{G}_N(z,t)^2\frac{d}{dt}H_{ii}(t)\right)\right]dt\\ &= -\frac{1}{dN}\int_0^1 \mathbf{E}\left[\operatorname{tr}\left(X_N^\top\mathcal{G}_N(z,t)^2X_N\frac{d}{dt}\Lambda_N(t)\right)\right]dt,\end{aligned}$$

Define

$$\Delta_N(z) = -\frac{1}{dN}\sum_{n=1}^{N}\int_0^1 \mathbf{E}\left[\left(X_N^\top\mathcal{G}_N(z,t)X_N\right)_{nn}\frac{d}{dt}\alpha_n(t)\right]dt,$$

then we have

$$\delta_N(z) \overset{(47)}{=} \frac{d}{dz}\Delta_N(z).$$

We now bound $\delta_N(z)$ by bounding $\Delta_N(z)$. We first define $D_\zeta = \{z|z\in\mathbb{C}^+, \Im z \geq \zeta > 0\}$, where $\Im z$ denotes the image part of z. Since all eigenvalues of $H_{ii}(t)\in\mathbb{R}$, $\Delta_N(z)$ is an analytic function in D_ζ. Based on Cauchy's integral formula, for any $z\in D_\zeta$ and arbitrary circle $\gamma\in D_\zeta$ containing z. we have

$$\delta_N(z) = \frac{d}{dz}\Delta_N(z) = \frac{1}{2\pi i}\oint_\gamma \frac{\Delta_N(s)}{(s-z)^2}ds.$$

Then we have:

$$\delta_N(z) \leq \frac{\text{length}(\gamma)}{2\pi}\max_{s\in\gamma}\frac{1}{(s-z)^2}\max_{s\in\gamma}|\Delta_N(s)| \leq \text{Const.}\max_{z\in D_\zeta}|\Delta_N(z)|,$$

where Const. is some positive constant. Now we aim to prove the following equation:

$$\max_{z\in D_\zeta}|\Delta_N(z)| = O\left(N^{-\frac{1}{2}}\right) \qquad (49)$$

If this is true, then $|\delta_N(z)| = O(N^{-\frac{1}{2}})$ for all $z\in D_\zeta$. Let $\zeta\to 0$, we will get $|\delta_N(z)| = O(N^{-\frac{1}{2}})$, which converges to 0 (pointwise) as $N\to\infty$. In the following analysis, we aim to prove (49). We first rewrite $\Delta_N(z)$ as follows.

$$\Delta_N(z) = -\frac{1}{2dN}\sum_{n=1}^{N}\sum_{l=1}^{C}\sum_{s=1}^{d}\int_0^1 \mathbf{E}\left[x_n^\top\mathcal{G}_N(z,t)x_nB_{ln}(t)V_{ls}\left(\frac{X_{sn}}{\sqrt{t}}-\frac{\widetilde{X}_{sn}}{\sqrt{1-t}}\right)\right]dt, \qquad (50)$$

where

$$B_{ln}(t) = \Big(1-2p_i(x_n(t))\Big)p_i(x_n(t))\Big(\delta_{il}-p_l(x_n(t))\Big),$$

and X_{sn} is the abbreviation for the (s,n)-th entry in X_N. Similar abbreviation also applies to $\widetilde{X}_{sn}$. We define $\delta_{il} = 1$ if $i = l$ and $\delta_{il} = 0$ if otherwise. Note that trivial upper bound of (50) does not vanish with N. To better evaluate the expectation in (50), we will use Stein's Lemma

$$\mathbf{E}[Zf(Z)] = \mathbf{E}[f'(Z)] \tag{51}$$

for $Z \sim \mathcal{N}(0,1)$ and differentiable function $f : \mathbb{R} \to \mathbb{C}$ with sub-exponential decay at infinity. Then we have

$$\begin{gathered}\Delta_N(z) = -\frac{1}{2dN}\sum_{n=1}^{N}\sum_{l=1}^{C}\sum_{s=1}^{d}\int_0^1 \left(\frac{1}{\sqrt{t}}\mathbf{E}\left[\frac{\partial F_N(n,l,z,t)}{\partial X_{sn}}V_{ls}\right] - \frac{1}{\sqrt{1-t}}\mathbf{E}\left[\frac{\partial F_N(n,l,z,t)}{\partial \widetilde{X}_{sn}}V_{ls}\right]\right) dt, \\ F_N(n,l,z,t) = x_n^\top \mathcal{G}_N(z,t) x_n B_{ln}(t),\end{gathered} \tag{52}$$

Write $\mathcal{G}_N = \mathcal{G}_N(z,t)$ for short, then

$$\begin{aligned}&\frac{1}{\sqrt{1-t}}\mathbf{E}\left[\frac{\partial F_N(n,l,z,t)}{\partial \widetilde{X}_{sn}}V_{ls}\right] - \frac{1}{\sqrt{t}}\mathbf{E}\left[\frac{\partial F_N(n,l,z,t)}{\partial X_{sn}}V_{ls}\right] \\ =&\frac{2}{\sqrt{t}}\mathbf{E}\left[\frac{\delta_{ns}}{N}B_{ln}(t)V_{ls}\alpha_n(t)(\mathcal{G}_N x_n)_n(\mathcal{G}_N x_n)^\top x_n - B_{ln}(t)V_{ls}(\mathcal{G}_N x_n)_s\right].\end{aligned} \tag{53}$$

Hence $\Delta_N(z) = \Delta_{N,1}(z) - \Delta_{N,2}(z)$, where

$$\begin{aligned}\Delta_{N,1}(z) &= \frac{1}{dN^2}\sum_{l=1}^{C}\sum_{n=1}^{\min(d,N)}\int_0^1 \mathbf{E}\left[B_{ln}(t)V_{ln}\alpha_n(t)(\mathcal{G}_N x_n)_n(\mathcal{G}_N x_n)^\top x_n\right]\frac{dt}{\sqrt{t}}, \\ \Delta_{N,2}(z) &= \frac{1}{dN}\sum_{n=1}^{N}\sum_{l=1}^{C}\sum_{s=1}^{d}\int_0^1 \mathbf{E}\left[B_{ln}(t)V_{ls}(\mathcal{G}_N x_n)_s\right]\frac{dt}{\sqrt{t}}.\end{aligned} \tag{54}$$

From Hölder's inequality,

$$\mathbf{E}[|V_{ln}(\mathcal{G}_N x_n)_n(\mathcal{G}_N x_n)^\top x_n|] \leq \left(\mathbf{E}[V_{ln}^4]\right)^{\frac{1}{4}}\left(\mathbf{E}[(\mathcal{G}_N x_n)_n^2]\right)^{\frac{1}{2}}\left(\mathbf{E}[((\mathcal{G}_N x_n)^\top x_n)^4]\right)^{\frac{1}{4}},$$

$$\mathbf{E}[|V_{ls}(\mathcal{G}_N x_n)_s|] \leq \left(\mathbf{E}[V_{ls}^2]\right)^{\frac{1}{2}}\left(\mathbf{E}[(\mathcal{G}_N x_n)_s^2]\right)^{\frac{1}{2}}.$$

Recall that

$$|B_{ln}(t)| \leq \frac{1}{9},\ |\alpha_n(t)| \leq \frac{1}{4},\ V_{ls} \sim \mathcal{N}(0,\frac{1}{d}),$$

$$\|\mathcal{G}_N\| \leq \min_{z\in D_\zeta}\left|\frac{1}{\lambda_{H_{ii}} - z}\right| \leq \min_{z\in D_\zeta}\frac{1}{\Im(z)} \leq \frac{1}{\zeta},$$

$$\mathbf{E}[(\mathcal{G}_N x_n)_s^2] = \frac{1}{d}\mathbf{E}[\|\mathcal{G}_N x_n\|^2] \leq \mathbf{E}[\|\mathcal{G}_N\|^2] \leq \frac{1}{\zeta},$$

we have

$$|\Delta_{N,1}(z)| \leq \frac{3^{1/4}C\left(\mathbf{E}[\|x_n\|^8]\right)^{1/4}}{18N^2 d^{1/2}\zeta^2}, \quad |\Delta_{N,2}(z)| \leq \frac{2C}{9d^{1/2}\zeta}.$$

As $d, N \to \infty$, $d/N \to \gamma > 0$, we have the following equations for any $z \in D_\zeta$

$$|\Delta_{N,1}(z)| = O\left(N^{-3/2}\right), \quad |\Delta_{N,2}(z)| = O\left(N^{-1/2}\right).$$

Set $\zeta \to 0$, then we have (49) and hence (46), together with (45) implying

$$s_{H_{ii}}(z) - s_{\widetilde{H}_{ii}}(z) = O\left(N^{-1/2}\right) \quad a.s., \forall z \in \mathbb{C}^+.$$

The proof for $s_{H_{ij}}$ is done following the same procedure.

■

Now we can apply Proposition 1 to characterize the limiting eigenvalue distribution of $\widetilde{H}_{ii}, \widetilde{H}_{ij}$, which are identical to the distributions of H_{ii}, H_{ij}.

Proposition 2 *Fix $C \geq 2$, as $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$, we have*

1. *$\mu_{H_{ii}}$ converges almost surely to a deterministic measure μ_{11}^H, and its Stieltjes transform $s_{\mu_{11}^H}(z)$ is uniquely specified by the functional equation*

$$s(z) = \frac{1}{\int_{\mathbb{R}^C} \frac{h_1(\mathbf{t})\varphi_C(\mathbf{t})}{1+\gamma s(z) h_1(\mathbf{t})} dt_1 \cdots dt_C - z}, \quad \forall z \in \mathbb{C}^+. \tag{55}$$

   Here $\mathbf{t} = (t_1, \cdots, t_C)$, and

$$h_1(\mathbf{t}) = \frac{e^{t_1}}{\sum_{l=1}^C e^{t_l}} \left(1 - \frac{e^{t_1}}{\sum_{l=1}^C e^{t_l}}\right),$$

$$\varphi_C(\mathbf{t}) = (2\pi)^{-\frac{C}{2}} e^{-\frac{1}{2}\sum_{l=1}^C t_l^2}.$$

2. *For $i \neq j$, $\mu_{H_{ij}}$ converges weakly almost surely to a deterministic measure μ_{12}, and its Stieltjes transform $s_{\mu_{12}}(z)$ is uniquely specified by the functional equation*

$$s(z) = \frac{1}{\int_{\mathbb{R}^C} \frac{h_2(\mathbf{t})\varphi_C(\mathbf{t})}{1+\gamma s(z) h_2(\mathbf{t})} dt_1 \cdots dt_C - z}, \quad \forall z \in \mathbb{C}^+. \tag{56}$$

   Here

$$h_2(\mathbf{t}) = \frac{e^{t_1+t_2}}{\left(\sum_{l=1}^C e^{t_l}\right)^2}.$$

Proof From Lemma 1, it suffices to prove the convergence of $\mu_{\widetilde{H}_{ii}}, \mu_{\widetilde{H}_{ij}}$ to the limiting measure specified by (55), (56) respectively. For the case of $\widetilde{H}_{ii}$, let

$$Y_{n,i}^{(N)} = p_i(\widetilde{x}_n) = \frac{\exp(v_i^\top \widetilde{x}_n)}{\sum_{l=1}^C \exp(v_l^\top \widetilde{x}_n)},$$

then $\widetilde{\alpha}_n = Y_{n,i}^{(N)}(1 - Y_{n,i}^{(N)})$. Recall that $\widetilde{H}_{ii} = \frac{1}{N} X_N \widetilde{\Lambda}_N X_N^\top$, $\widetilde{\Lambda}_N = \mathrm{diag}(\widetilde{\alpha}_1, \cdots, \widetilde{\alpha}_N)$. Then $\widetilde{\Lambda}_N$ is independent of X_N, and the eigenvalue distribution of $\widetilde{\Lambda}_N$ is the counting measure $\nu_N = \frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{\alpha}_n}$. Note that $(\widetilde{\alpha}_n)_{n=1}^N$ are identically distributed but dependent, therefore we need to prove that ν_N converges almost surely to a deterministic measure before applying Proposition 1.

From the strong law of large number, it holds almost surely that

$$\begin{aligned} &\lim_{d\to\infty} \|w_c\|^2 = 1 \quad \forall c \in [C], \\ &\lim_{d\to\infty} w_c^\top w_{c'} = 0 \quad \forall c, c' \in [C],\ c \neq c'. \end{aligned} \tag{57}$$

Now we restrict the probability space to a subspace that (57) holds. Note that this restriction does not change validity of the proof since (57) holds almost surely. Let $\mathcal{F}_V$ be the σ-algebra generated by $\{V_{ij}\}_{i\in[C], j\in\mathbb{N}^+}$, then from CLT, the conditional distribution of

$$(v_1^\top \widetilde{x}_1, \cdots, v_C^\top \widetilde{x}_1)$$

given $\mathcal{F}_V$ converges weakly to $\mathcal{N}_C(0, I_{C\times C})$. Let ν be the deterministic probability measure such that for any interval $\Delta \subset \mathbb{R}$,

$$\nu(\Delta) = \mathbf{P}\left(\frac{e^{Z_1}}{\sum_{l=1}^C e^{Z_l}}\left(1 - \frac{e^{Z_1}}{\sum_{l=1}^C e^{Z_l}}\right) \in \Delta\right), \tag{58}$$

where $(Z_l)_{l=1}^C$ are i.i.d. $\mathcal{N}(0,1)$. Let $f(x)$ be a bounded continuous function on $\mathbb{R}$, from the conditional independence of $(\widetilde{\alpha}_n)_{n=1}^N$ given $\mathcal{F}_V$, as $N \to \infty$,

$$\begin{aligned} &\int_{\mathbb{R}} f(x)\nu_N(dx) - \int_{\mathbb{R}} f(x)\nu(dx) \\ =&\left(\frac{1}{N}\sum_{n=1}^N f(\widetilde{\alpha}_n) - \mathbf{E}\left[f(\widetilde{\alpha}_1)|\mathcal{F}_V\right]\right) + \left(\mathbf{E}\left[f(\widetilde{\alpha}_1)|\mathcal{F}_V\right] - \int_{\mathbb{R}} f(x)\nu(dx)\right) \\ \to& 0, \quad a.s.. \end{aligned} \tag{59}$$

Here the first term converges a.s to 0 because of the strong law of large number. And the second term converges to 0 from Portmanteau theorem. Then ν_N converges weakly almost surely to ν.

Now let $\widetilde{T}_N = \frac{N}{d}\widetilde{H}_{ii}$. Then $\widetilde{T}_N = \frac{1}{d} X_N \widetilde{\Lambda}_N X_N^\top$, and the eigenvalue distribution of $\widetilde{\Lambda}_N$ converges weakly almost surely to ν. From Proposition 1, $s_{\widetilde{T}_N}$ converges weakly almost surely to the unique solution of

$$s(z) = \frac{1}{\frac{1}{\gamma}\int_{\mathbb{R}} \frac{t\nu(dt)}{1+ts(z)} - z}, \quad \forall z \in \mathbb{C}^+. \tag{60}$$

From (58),

$$\int_{\mathbb{R}} \frac{t\nu(dt)}{1+ts(z)} = \int_{\mathbb{R}^C} \frac{h_1(\mathbf{t})\varphi_C(\mathbf{t})}{1+s(z)h_1(\mathbf{t})} dt_1 \cdots dt_C.$$

Note that $s_{\widetilde{T}_N}(z) = \frac{d}{N} s_{\widetilde{H}_{ii}}\left(\frac{d}{N} z\right)$ and $d/N \to \gamma$. Then with a change of variable $z' = \gamma z$ in (60), it implies that $s_{\widetilde{H}_{ii}}$ converges weakly almost surely to the unique solution of (55). Then we finish the proof for the H_{ii} case.

The proof for H_{ij} case is in the same procedure. The variables $(\widetilde{\beta}_n)_{n=1}^N$ are identically distributed, and $\widetilde{\beta}_n = -Y_{n,i}^{(N)} Y_{n,j}^{(N)}$. Define the counting measure $\eta_N = \frac{1}{N}\sum_{n=1}^N \delta_{\widetilde{\beta}_n}$, then η_N converges weakly almost surely to a deterministic probability measure ν, where for any interval $\Delta \subset \mathbb{R}$,

$$\eta(\Delta) = \mathbf{P}\left(\frac{e^{Z_1}}{\sum_{l=1}^C e^{Z_l}} \frac{e^{Z_2}}{\sum_{l=1}^C e^{Z_l}} \in \Delta\right).$$

Let $\widetilde{S}_N = \frac{N}{d}\widetilde{H}_{ij}$, then from Proposition 1, $s_{\widetilde{S}_N}$ converges weakly almost surely to the unique solution of

$$s(z) = \frac{1}{\frac{1}{\gamma}\int_{\mathbb{R}} \frac{t\eta(dt)}{1+ts(z)} - z}, \quad \forall z \in \mathbb{C}^+.$$

We have

$$\int_{\mathbb{R}} \frac{t\eta(dt)}{1+ts(z)} = \int_{\mathbb{R}^C} \frac{h_2(\mathbf{t})\varphi_C(\mathbf{t})}{1+s(z)h_2(\mathbf{t})} dt_1 \cdots dt_C. \tag{61}$$

Then from $s_{\widetilde{S}_N}(z) = \frac{d}{N} s_{\widetilde{H}_{ij}}\left(\frac{d}{N} z\right)$, $d/N \to \gamma$, and a change of variable $z' = \gamma z$ in (61), it implies that $s_{\widetilde{H}_{ij}}$ converges weakly almost surely to the unique solution of (56). This concludes the whole proof. ■

The next proposition is to extract the second moment of the limiting eigenvalue distribution from the implicit equations (55) and (56). This leads to Theorem 1.

Proposition 3 *Suppose that as $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$. Then for $i \neq j$, it holds almost surely that*

$$\lim_{d,N\to\infty} \frac{\|H_{ii}\|_{\mathrm{F}}^2}{d} = \gamma \int_{\mathbb{R}^C} h_1(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C + \left(\int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C\right)^2, \tag{62}$$

$$\lim_{d,N\to\infty} \frac{\|H_{ij}\|_{\mathrm{F}}^2}{d} = \gamma \int_{\mathbb{R}^C} h_2(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C + \left(\int_{\mathbb{R}^C} h_2(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C\right)^2, \tag{63}$$

$$\lim_{C\to\infty} \lim_{d,N\to\infty} \frac{C^2\|H_{ii}\|_{\mathrm{F}}^2}{d} = \gamma e + 1, \tag{64}$$

$$\lim_{C\to\infty} \lim_{d,N\to\infty} \frac{C^4\|H_{ij}\|_{\mathrm{F}}^2}{d} = \gamma e^2 + 1. \tag{65}$$

Proof Recall that as $z \to \infty$ in $\mathbb{C}^+$,

$$s_\mu(z) = -\frac{1}{z} - \frac{1}{z^2}\int_{\mathbb{R}} x\mu(dx) - \frac{1}{z^3}\int_{\mathbb{R}} x^2\mu(dx) + O\left(\frac{1}{z^4}\right).$$

Then in (55) as $z \to \infty$,

$$\begin{aligned}
s_{\mu_{11}^H}(z) &= \frac{1}{\int_{\mathbb{R}^C} \frac{h_1(\mathbf{t})\varphi_C(\mathbf{t})}{1+\gamma s_{\mu_{11}^H}(z) h_1(\mathbf{t})} dt_1 \cdots dt_C - z} \\
&= \frac{1}{\int_{\mathbb{R}^C} \frac{h_1(\mathbf{t})\varphi_C(\mathbf{t})}{1-\gamma h_1(\mathbf{t}) z^{-1} + O(z^{-2})} dt_1 \cdots dt_C - z} \\
&= \frac{1}{\int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) \left(1 + \gamma h_1(\mathbf{t}) z^{-1} + O(z^{-2})\right) dt_1 \cdots dt_C - z} \\
&= -\frac{1}{z} - \frac{1}{z^2} \int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C \\
&\quad - \frac{1}{z^3} \left[\gamma \int_{\mathbb{R}^C} h_1(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C + \left(\int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C \right)^2 \right] + O\left(\frac{1}{z^4}\right).
\end{aligned} \tag{66}$$

Hence

$$\int_{\mathbb{R}} x^2 \mu_{11}^H(dx) = \gamma \int_{\mathbb{R}^C} h_1(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C + \left(\int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C \right)^2.$$

Then we obtain (1) since from Proposition 2,

$$\lim_{d,N\to\infty} \frac{\|H_{ii}\|_{\mathrm{F}}^2}{d} = \lim_{d,N\to\infty} \int_{\mathbb{R}} x^2 \mu_{H_{ii}}(dx) = \int_{\mathbb{R}} x^2 \mu_{11}^H(dx) \quad a.s..$$

The proof of (2) follows the same procedure, i.e., expanding $s(z)$ as $z \to \infty$ in (56). Now let $\mathbf{q}_C = (q_1, \cdots, q_C) \sim \mathcal{N}_C(0, I_{C\times C})$. From the strong law of large number, as $C \to \infty$,

$$\frac{e^{q_1} + \cdots + e^{q_C}}{C} \xrightarrow{a.s.} \mathbf{E}[e^{q_1}] = \sqrt{e}.$$

Then from Slutsky's theorem,

$$C\sqrt{e} \cdot h_1(\mathbf{q}_C) = \frac{C\sqrt{e} \cdot e^{q_1}}{e^{q_1} + \cdots + e^{q_C}} \left(1 - \frac{e^{q_1}}{e^{q_1} + \cdots + e^{q_C}}\right) \xrightarrow{d} \text{Lognormal}(0,1),$$

$$C^2 e \cdot h_2(\mathbf{q}_C) = \frac{C\sqrt{e} \cdot e^{q_1}}{e^{q_1} + \cdots + e^{q_C}} \cdot \frac{C\sqrt{e} \cdot e^{q_2}}{e^{q_1} + \cdots + e^{q_C}} \xrightarrow{d} \text{Lognormal}(0,1) \otimes \text{Lognormal}(0,1).$$

Here $\otimes$ denotes the multiplicative convolution. That is, $\text{Lognormal}(0,1) \otimes \text{Lognormal}(0,1)$ is the distribution of $\xi_1\xi_2$ where ξ_1, ξ_2 are iid $\text{Lognormal}(0,1)$. Then from $\mathbf{E}[\xi] = \sqrt{e}$, $\mathbf{E}[\xi^2] = e^2$, we have

$$\begin{aligned}
\lim_{C\to\infty} C \int_{\mathbb{R}^C} h_1(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C &= \lim_{C\to\infty} \mathbf{E}\left[C h_1(\mathbf{q}_C)\right] = 1, \\
\lim_{C\to\infty} C^2 \int_{\mathbb{R}^C} h_1(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C &= \lim_{C\to\infty} \mathbf{E}\left[(C h_1(\mathbf{q}_C))^2\right] = e, \\
\lim_{C\to\infty} C^2 \int_{\mathbb{R}^C} h_2(\mathbf{t})\varphi_C(\mathbf{t}) dt_1 \cdots dt_C &= \lim_{C\to\infty} \mathbf{E}\left[C^2 h_2(\mathbf{q}_C)\right] = 1, \\
\lim_{C\to\infty} C^4 \int_{\mathbb{R}^C} h_2(\mathbf{t})^2 \varphi_C(\mathbf{t}) dt_1 \cdots dt_C &= \lim_{C\to\infty} \mathbf{E}\left[(C^2 h_2(\mathbf{q}_C))^2\right] = e^2.
\end{aligned}$$

Then, (3) and (4) follows from (1) and (2) by taking $C \to \infty$, respectively. ■

E.2. Proof of Theorem 2

Before delving into the proof, we first present the constant terms in Theorem 2. We remark that these constants are bounded as long as $m \geq 2$.

$$\begin{aligned}
a_{11} &= \frac{1}{2\pi}\int_0^{+\infty}\int_0^{+\infty} xy\exp\left(\frac{xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy,\\
a_{12} &= \frac{1}{2\pi}\int_0^{+\infty}\int_0^{+\infty} x^2y^2\exp\left(\frac{xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy,\\
a_{21} &= \frac{1}{2\pi}\int_0^{+\infty}\int_0^{+\infty} xy\exp\left(\frac{2xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy,\\
a_{22} &= \frac{1}{2\pi}\int_0^{+\infty}\int_0^{+\infty} x^2y^2\exp\left(\frac{2xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy,\\
b_1 &= \frac{3}{4}+\frac{1}{8\pi}\int_0^{+\infty}\int_0^{+\infty} \exp\left(\frac{xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy,\\
b_2 &= \frac{3}{4}+\frac{1}{8\pi}\int_0^{+\infty}\int_0^{+\infty} \exp\left(\frac{2xy}{m}-\frac{x^2}{2}-\frac{y^2}{2}\right)dxdy.
\end{aligned} \tag{67}$$

The functions h_{ii}, h_{ij}, u_{ii}, u_{ij}, q_{ii}, q_{ij} are given by the right hand side of (91), (92), (100), (101), (103), (104), repectively.

E.2.1. Proof for the Hidden-layer Hessian with CE Loss

Our proof uses the following strategy. Firstly we consider the case that $V \in \mathbb{R}^{C\times m}$ is deterministic. In this case, the techniques in the proof of Theorem 1 is available. Then for the target case that entries of V are i.i.d. Gaussian, we use

$$\mathbf{E}\left[\left\|\frac{\partial^2\ell_{\mathrm{CE}}(W,V)}{\partial w_i\partial w_j^\top}\right\|_{\mathrm{F}}^2\right] = \mathbf{E}\left[\mathbf{E}\left[\left\|\frac{\partial^2\ell_{\mathrm{CE}}(W,V)}{\partial w_i\partial w_j^\top}\right\|_{\mathrm{F}}^2 \,\middle|\, V\right]\right]$$

and apply the deterministic case results in the conditional expectation.

Now for short of notations we write $G_{ii} = \frac{\partial^2\ell_{\mathrm{CE}}(W,V)}{\partial w_i\partial w_i^\top}$, $G_{ij} = \frac{\partial^2\ell_{\mathrm{CE}}(W,V)}{\partial w_i\partial w_j^\top}$. Then

$$\begin{aligned}
G_{ii} &= \frac{1}{N}\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0)\left[\sum_{k=1}^C q_k(x_n)V_{ki}^2 - \left(\sum_{k=1}^C q_k(x_n)V_{ki}\right)^2\right]x_nx_n^\top,\\
G_{ij} &= \frac{1}{N}\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0)\mathbf{1}(w_j^\top x_n > 0)\left[\sum_{k=1}^C q_k(x_n)V_{ki}V_{kj} - \left(\sum_{k=1}^C q_k(x_n)V_{ki}\right)\left(\sum_{k=1}^C q_k(x_n)V_{kj}\right)\right]x_nx_n^\top,
\end{aligned} \tag{68}$$

where

$$q_k(x_n) = \frac{\exp(\sigma(Wx_n)^\top v_k)}{\sum_{l=1}^C \exp(\sigma(Wx_n)^\top v_l)}.$$

Let $\widetilde{X}_N = (\widetilde{x}_1, \cdots, \widetilde{x}_N) \in \mathbb{R}^{d\times N}$ be an independent copy of X_N. Define

$$
\begin{aligned}
\widetilde{G}_{ii} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}(w_i^\top \widetilde{x}_n > 0)\left[\sum_{k=1}^{C} q_k(\widetilde{x}_n)V_{ki}^2 - \left(\sum_{k=1}^{C} q_k(\widetilde{x}_n)V_{ki}\right)^2\right] x_n x_n^\top, \\
\widetilde{G}_{ij} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}(w_i^\top \widetilde{x}_n > 0)\mathbf{1}(w_j^\top \widetilde{x}_n > 0)\left[\sum_{k=1}^{C} q_k(\widetilde{x}_n)V_{ki}V_{kj} - \left(\sum_{k=1}^{C} q_k(\widetilde{x}_n)V_{ki}\right)\left(\sum_{k=1}^{C} q_k(\widetilde{x}_n)V_{kj}\right)\right] x_n x_n^\top.
\end{aligned} \tag{69}
$$

Lemma 2 *Suppose that $C \geq 2,\ m \geq 3$ are fixed, $V \in \mathbb{R}^{C\times m}$ is deterministic. For any $z \in \mathbb{C}^+$, as $d, N \to \infty,\ d/N \to \gamma \in (0, +\infty)$, it holds almost surely that*

$$
s_{G_{ii}}(z) - s_{\widetilde{G}_{ii}}(z) = O\left(N^{-\frac{1}{2}}\right), \tag{70}
$$

$$
s_{G_{ij}}(z) - s_{\widetilde{G}_{ij}}(z) = O\left(N^{-\frac{1}{2}}\right). \tag{71}
$$

Proof Here, we only present the proof for $s_{G_{ii}}$. The proof for $s_{G_{ij}}$ is done following the same procedure.

For $t \in [0, 1]$, let

$$
X_N(t) = \sqrt{t}X_N + \sqrt{1-t}\widetilde{X}_N.
$$

Then $X_N(t) = (x_1(t), \cdots, x_N(t)) \in \mathbb{R}^{d\times N}$, where

$$
x_n(t) = \sqrt{t}x_n + \sqrt{1-t}\widetilde{x}_n, \quad n \in [N].
$$

Denote

$$
\theta_n(t) = \mathbf{1}(w_i^\top x_n(t) > 0)\left[\sum_{k=1}^{C} q_k(x_n(t))V_{ki}^2 - \left(\sum_{k=1}^{C} q_k(x_n(t))V_{ki}\right)^2\right],
$$

$$
\Theta_N(t) = \mathrm{diag}(\theta_1(t), \cdots, \theta_N(t)),
$$

$$
G_{ii}(t) = \frac{1}{N}X_N\Theta_N(t)X_N^\top,
$$

$$
\mathcal{G}_N(z, t) = (G_{ii}(t) - z)^{-1}.
$$

By treating $s_{H_{ii}}(z), s_{\widetilde{H}_{ii}}(z)$ as Lipchitz functions of the Gaussian vectors $w_1, \cdots, w_m, x_1, \cdots, x_N$, from Talagrand's inequality,

$$
\mathbf{P}\left(|s_{G_{ii}}(z) - \mathbf{E}[s_{G_{ii}}(z)]| \geq t\right) \leq c_1 e^{-pc_2t^2},
$$

$$
\mathbf{P}\left(|s_{\widetilde{G}_{ii}}(z) - \mathbf{E}[s_{\widetilde{G}_{ii}}(z)]| \geq t\right) \leq \widetilde{c}_1 e^{-d\widetilde{c}_2t^2},
$$

for $t > 0$ and constants $c_1, c_2, \widetilde{c}_1, \widetilde{c}_2 > 0$. Then from the Borel-Cantelli Lemma,

$$
\begin{aligned}
&s_{G_{ii}}(z) - \mathbf{E}[s_{G_{ii}}(z)] \overset{a.s.}{\to} 0, \\
&s_{\widetilde{G}_{ii}}(z) - \mathbf{E}[s_{\widetilde{G}_{ii}}(z)] \overset{a.s.}{\to} 0.
\end{aligned} \tag{72}
$$

Now we prove that

$$\delta_N(z) = \mathbf{E}[s_{H_{ii}}(z)] - \mathbf{E}[s_{\widetilde{H}_{ii}}(z)] = O\left(N^{-\frac{1}{2}}\right). \tag{73}$$

Following (48), we have

$$\Delta_N(z) = -\frac{1}{dN}\int_0^1 \mathbf{E}\left[\mathrm{tr}\left(X_N^\top \mathcal{G}_N(z,t)^2 X_N \frac{d}{dt}\Theta_N(t)\right)\right] dt. \tag{74}$$

Then $\delta_N(z) = \frac{d}{dz}\Delta_N(z)$, where

$$\Delta_N(z) = -\frac{1}{dN}\sum_{n=1}^{N}\int_0^1 \mathbf{E}\left[\left(X_N^\top \mathcal{G}_N(z,t)X_N\right)_{nn}\frac{d}{dt}\theta_n(t)\right] dt. \tag{75}$$

Similarly to the proof of Theorem 1, it suffices to prove that for any open set $O \subset \mathbb{C}^+$ such that $\zeta = \inf_{z\in O}|\Im z| > 0$,

$$\max_{z\in O}|\Delta_N^{(1)}(z)| = O\left(N^{-\frac{1}{2}}\right). \tag{76}$$

We have

$$\frac{d}{dt}q_k(x_n(t)) = \frac{1}{2}\sum_{l=1}^{C}\sum_{h=1}^{m} q_k(x_n(t))[\delta_{kl} - q_l(x_n(t))]\mathbf{1}(w_h^\top x_n(t) > 0)V_{lh}w_h^\top\left(\frac{x_n}{\sqrt{t}} - \frac{\widetilde{x}_n}{\sqrt{1-t}}\right),$$

$$\frac{d}{dt}\theta_n(t) = \mathbf{1}(w_i^\top x_n(t) > 0)\sum_{k=1}^{C}\left(V_{ki}^2 - 2V_{ki}\sum_{k'=1}^{C} q_{k'}(x_n(t))V_{k'i}\right)\frac{d}{dt}q_k(x_n(t)).$$

Therefore

$$\Delta_N(z) = -\frac{1}{2dN}\sum_{n=1}^{N}\sum_{k=1}^{C}\sum_{l=1}^{C}\sum_{h=1}^{m}\sum_{s=1}^{d}\int_0^1 \mathbf{E}\left[x_n^\top \mathcal{G}_N(z,t)x_n Q_{klhn}(t)W_{hs}\left(\frac{X_{sn}}{\sqrt{t}} - \frac{\widetilde{X}_{sn}}{\sqrt{1-t}}\right)\right] dt, \tag{77}$$

where

$$Q_{klhn}(t) = \left(V_{ki}^2 - 2V_{ki}\sum_{k'=1}^{C} q_{k'}(x_n(t))V_{k'i}\right) q_k(x_n(t))[\delta_{kl} - q_l(x_n(t))]\mathbf{1}(w_h^\top x_n(t) > 0)V_{lh}.$$

From Stein's Lemma (51),

$$\Delta_N(z) = -\frac{1}{2dN}\sum_{n=1}^{N}\sum_{k=1}^{C}\sum_{l=1}^{C}\sum_{h=1}^{m}\sum_{s=1}^{d}\int_0^1 \left(\frac{1}{\sqrt{t}}\mathbf{E}\left[\frac{\partial A_N^{(k,l,h,n)}(z,t)}{\partial X_{sn}}W_{hs}\right] - \frac{1}{\sqrt{1-t}}\mathbf{E}\left[\frac{\partial A_N^{(k,l,h,n)}(z,t)}{\partial \widetilde{X}_{sn}}W_{hs}\right]\right) dt, \tag{78}$$

where

$$A_N^{(k,l,h,n)}(z,t) = x_n^\top \mathcal{G}_N(z,t)x_n Q_{klhn}(t).$$

Then

$$\begin{aligned}&\frac{1}{\sqrt{t}}\mathbf{E}\left[\frac{\partial A_N^{(k,l,h,n)}(z,t)}{\partial X_{sn}}W_{hs}\right] - \frac{1}{\sqrt{1-t}}\mathbf{E}\left[\frac{\partial A_N^{(k,l,h,n)}(z,t)}{\partial \widetilde{X}_{sn}}W_{hs}\right]\\ =&\frac{2}{\sqrt{t}}\mathbf{E}\left[Q_{klhn}(t)W_{hs}\left((\mathcal{G}_N x_n)_s - \frac{\delta_{ns}}{N}\theta_n(t)(\mathcal{G}_N x_n)_n(\mathcal{G}_N x_n)^\top x_n\right)\right].\end{aligned} \tag{79}$$

Hence $\Delta_N(z) = \Delta_{N,1}(z) - \Delta_{N,2}(z)$, where

$$\begin{aligned}\Delta_{N,1}(z) &= \frac{1}{dN^2} \sum_{k=1}^{C} \sum_{l=1}^{C} \sum_{h=1}^{m} \sum_{n=1}^{\min(d,N)} \int_0^1 \mathbf{E}\left[Q_{klhn}(t)\theta_n(t) W_{hn} (\mathcal{G}_N x_n)_n (\mathcal{G}_N x_n)^\top x_n \right] \frac{dt}{\sqrt{t}}, \\ \Delta_{N,2}(z) &= \frac{1}{dN} \sum_{n=1}^{N} \sum_{k=1}^{C} \sum_{l=1}^{C} \sum_{h=1}^{m} \sum_{s=1}^{d} \int_0^1 \mathbf{E}\left[Q_{klhn}(t) W_{hs} (\mathcal{G}_N x_n)_s \right] \frac{dt}{\sqrt{t}}.\end{aligned} \tag{80}$$

Let $M_V = \max_{k\in[C],i\in[m]} |V_{ki}|$. From Hölder's inequality,

$$\mathbf{E}[|W_{hn}(\mathcal{G}_N x_n)_n (\mathcal{G}_N x_n)^\top x_n|] \le \left(\mathbf{E}[W_{hn}^4]\right)^{\frac{1}{4}} \left(\mathbf{E}[(\mathcal{G}_N x_n)_n^2]\right)^{\frac{1}{2}} \left(\mathbf{E}[((\mathcal{G}_N x_n)^\top x_n)^4]\right)^{\frac{1}{4}},$$

$$\mathbf{E}[|W_{hs}(\mathcal{G}_N x_n)_s|] \le \left(\mathbf{E}[W_{hs}^2]\right)^{\frac{1}{2}} \left(\mathbf{E}[(\mathcal{G}_N x_n)_s^2]\right)^{\frac{1}{2}},$$

together with

$$|Q_{klhn}(t)| \le 3M_v^4,\ |\theta_n(t)| \le 2M_v^2,\ \|\mathcal{G}_N\| \le \frac{1}{\zeta},\ W_{hs} \sim N(0, \frac{1}{d}),\ \mathbf{E}[(\mathcal{G}_N x_n)_s^2] = \frac{1}{d}\mathbf{E}[\|\mathcal{G}_N x_n\|^2] \le \frac{1}{\zeta},$$

we have

$$|\Delta_{n,1}(z)| \le \frac{3^{\frac{1}{4}} 12 C^2 m M_V^5 \left(\mathbf{E}[\|x_n\|^8]\right)^{\frac{1}{4}}}{\zeta^2 d^{\frac{3}{2}} N}, \quad |\Delta_{n,2}(z)| \le \frac{6mC^2 M_V^3}{\zeta d^{\frac{1}{2}}}.$$

Then as $d, N \to \infty,\ d/N \to \gamma \in (0, +\infty)$,

$$|\Delta_{N,1}(z)| = O\left(N^{-\frac{3}{2}}\right), \quad |\Delta_{N,1}(z)| = O\left(N^{-\frac{1}{2}}\right).$$

Then we have (76) and then (70). The proof in the case of G_{ij} is similar and is omitted. ■

Proposition 4 *Suppose that $C \ge 2,\ m \ge 3$ are fixed, $V \in \mathbb{R}^{C\times m}$ is deterministic. As $d, N \to \infty, \frac{d}{N} \to \gamma \in (0, +\infty)$, we have*

1. *$\mu_{G_{ii}}$ converges weakly almost surely to a deterministic measure μ_{11}^G, where $s_{\mu_{11}^G}(z)$ is uniquely specified by the functional equation*

$$s(z) = \frac{1}{\int_{\mathbb{R}} \frac{t\nu_1(dt)}{1+\gamma s(z) t} dt - z}, \quad \forall z \in \mathbb{C}^+. \tag{81}$$

Here ν_1 is defined as follows. Let $\mathbf{z} = (z_1, \cdots, z_m) \sim \mathcal{N}_m(0, I_m)$, define random variables

$$r_k = \frac{\exp(\sigma(\mathbf{z})^\top v_k)}{\sum_{l=1}^{C} \exp(\sigma(\mathbf{z})^\top v_l)}, \quad k \in [C],$$

$$\xi_C(V) = \mathbf{1}(z_1 > 0) \left[\sum_{k=1}^{C} r_k V_{k1}^2 - \left(\sum_{k=1}^{C} r_k V_{k1} \right)^2 \right].$$

Then ν_1 is given by that for all intervals $\Delta \subset \mathbb{R}$, $\nu_1(\Delta) = \mathbf{P}(\xi_C(V) \in \Delta)$.

2. *For $i \neq j$, $\mu_{G_{ij}}$ converges weakly almost surely to a deterministic measure μ_{12}^G, where $s_{\mu_{12}^G}(z)$ is uniquely specified by the functional equation*

$$s(z) = \frac{1}{\int_{\mathbb{R}} \frac{t\nu_2(dt)}{1+\gamma s(z) t} dt - z}, \quad \forall z \in \mathbb{C}^+. \tag{82}$$

Here ν_2 is defined as follows. Let $\mathbf{z} = (z_1, \cdots, z_m) \sim \mathcal{N}_m(0, I_m)$, define random variables

$$\eta_C(V) = \mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0) \left[\sum_{k=1}^{C} r_k V_{k1} V_{k2} - \left(\sum_{k=1}^{C} r_k V_{k1} \right) \left(\sum_{k=1}^{C} r_k V_{k2} \right) \right].$$

Then ν_2 is given by that for all intervals $\Delta \subset \mathbb{R}$, $\nu_2(\Delta) = \mathbf{P}(\eta_C(V) \in \Delta)$.

Proof We give a proof for the G_{ii} case, the G_{ij} case is in the same procedure. From Lemma 2, it suffices to prove the convergence of $\mu_{\widetilde{H}_{ii}}$ to the limiting measure specified by (4). Let

$$\widetilde{\theta}_n = \mathbf{1}(w_i^\top \widetilde{x}_n > 0) \left[\sum_{k=1}^{C} q_k(\widetilde{x}_n) V_{ki}^2 - \left(\sum_{k=1}^{C} q_k(\widetilde{x}_n) V_{ki} \right)^2 \right],$$

$$\widetilde{\Theta}_N = \operatorname{diag}\{\theta_1, \cdots, \theta_N\},$$

then $\widetilde{H}_{ii} = \frac{1}{N} X_N \widetilde{\Theta}_N X_N^\top$, and $\widetilde{\Theta}_N$ is independent of X_N. The eigenvalue distribution of $\widetilde{\Theta}_N$ is the counting measure $\tau_N = \frac{1}{N} \sum_{n=1}^{N} \delta_{\widetilde{\theta}_n}$. Since m is fixed, from the strong law of large number, it holds almost surely that

$$\begin{aligned} &\lim_{d\to\infty} \|w_h\|^2 = 1 \quad \forall h \in [m], \\ &\lim_{d\to\infty} w_h^\top w_{h'} = 0 \quad \forall h, h' \in [m],\ h \neq h'. \end{aligned} \tag{83}$$

Without loss of generality we can restrict the probability space to a subspace that (83) holds. Let $\mathcal{F}_W$ be the σ-algebra generated by $\{W_{hs}\}_{h\in[m], s\in\mathbb{N}^+}$, then from CLT, the conditional distribution of

$$(w_1^\top \widetilde{x}_1, \cdots, w_m^\top \widetilde{x}_1)$$

given $\mathcal{F}_W$ converges weakly to $\mathcal{N}_m(0, I_m)$. Since V is deterministic, $(\widetilde{\theta}_n)_{n=1}^N$ are independent conditioning on $\mathcal{F}_W$. Then for any $f \in C_b(\mathbb{R})$, as $N \to \infty$,

$$\begin{aligned} &\int_{\mathbb{R}} f(x)\tau_N(dx) - \int_{\mathbb{R}} f(x)\nu_1(dx) \\ =& \left(\frac{1}{N} \sum_{n=1}^{N} f(\widetilde{\theta}_n) - \mathbf{E}\left[f(\widetilde{\theta}_1) | \mathcal{F}_W \right] \right) + \left(\mathbf{E}\left[f(\widetilde{\theta}_1) | \mathcal{F}_W \right] - \int_{\mathbb{R}} f(x)\nu_1(dx) \right) \\ \to& 0, \quad a.s. \end{aligned} \tag{84}$$

Therefore τ_N converges weakly almost surely to ν_1.

Now let $\widetilde{T}_N = \frac{N}{d}\widetilde{G}_{ii}$. Then $\widetilde{T}_N = \frac{1}{d}X_N\widetilde{\Theta}_N X_N^\top$, and the eigenvalue distribution of $\widetilde{\Theta}_N$ converges weakly almost surely to ν_1. From Proposition 1, $s_{\widetilde{T}_N}$ converges weakly almost surely to the unique solution of

$$s(z) = \frac{1}{\frac{1}{\gamma}\int_{\mathbb{R}} \frac{t\nu_1(dt)}{1+ts(z)} - z}, \quad \forall z \in \mathbb{C}^+. \tag{85}$$

Then with a change of variable $z' = \gamma z$ in (85), it implies that $s_{\widetilde{G}_{ii}}$ converges weakly almost surely to the unique solution of (81). ■

The next proposition is exactly (6) in Theorem 2.

Proposition 5 *Suppose that $m \geq 3$ is fixed, and $V \in \mathbb{R}^{C\times m}$ has i.i.d. $\mathcal{N}(0, \frac{1}{m})$ entries. If as $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$, then for $i \neq j$,*

$$\begin{aligned} \lim_{C\to\infty}\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|G_{ii}\|_{\mathrm{F}}^2\right]}{d} &= \frac{2\gamma+1}{4m^2}, \\ \lim_{C\to\infty}\lim_{d,N\to\infty} \frac{C\mathbf{E}\left[\|G_{ij}\|_{\mathrm{F}}^2\right]}{d} &= \frac{\gamma(m-1)^2}{2^m(m-2)^3 m}\left(\sqrt{\frac{m}{m-2}}+1\right)^{m-2}. \end{aligned} \tag{86}$$

Proof Let $\bar{V}$ be a deterministic realization of V. By expanding (81) and (82) at $z = \infty$, we have almost surely

$$\lim_{d,N\to\infty} \frac{\|G_{ii}\|_{\mathrm{F}}^2}{d} = \int_{\mathbb{R}} x^2 \mu_{11}^G(dx) = \gamma\mathbf{E}[\xi_C(\bar{V})^2] + (\mathbf{E}[\xi_C(\bar{V})])^2, \tag{87}$$

$$\lim_{d,N\to\infty} \frac{\|G_{ij}\|_{\mathrm{F}}^2}{d} = \int_{\mathbb{R}} x^2 \mu_{12}^G(dx) = \gamma\mathbf{E}[\eta_C(\bar{V})^2] + (\mathbf{E}[\eta_C(\bar{V})])^2. \tag{88}$$

Then we have

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|G_{ii}\|_{\mathrm{F}}^2 \mid V\right]}{d} = \gamma\mathbf{E}[\xi_C(V)^2|V] + (\mathbf{E}[\xi_C(V)|V])^2, \tag{89}$$

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|G_{ij}\|_{\mathrm{F}}^2 \mid V\right]}{d} = \gamma\mathbf{E}[\eta_C(V)^2|V] + (\mathbf{E}[\eta_C(V)|V])^2. \tag{90}$$

Taking expectation both sides we have

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|G_{ii}\|_{\mathrm{F}}^2\right]}{d} = \gamma\mathbf{E}[\xi_C(V)^2] + (\mathbf{E}[\xi_C(V)])^2, \tag{91}$$

$$\lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|G_{ij}\|_{\mathrm{F}}^2\right]}{d} = \gamma\mathbf{E}[\eta_C(V)^2] + (\mathbf{E}[\eta_C(V)])^2. \tag{92}$$

Write for short that $\xi_C = \xi_C(V)$, $\eta_C = \eta_C(V)$. By the strong law of large number, as $C \to \infty$,

$$\mathbf{E}\left[\sum_{k=1}^C r_k V_{k1}^2 \,\middle|\, \mathbf{z}\right] = \mathbf{E}\left[\frac{\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)V_{k1}^2}{\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)} \,\middle|\, \mathbf{z}\right] \to \frac{\mathbf{E}[\exp(\sigma(\mathbf{z})^\top v_k)V_{k1}^2 \mid \mathbf{z}]}{\mathbf{E}[\exp(\sigma(\mathbf{z})^\top v_k) \mid \mathbf{z}]} = \frac{\sigma(z_1)^2}{m^2} + \frac{1}{m},$$

$$\mathbf{E}\left[\left(\sum_{k=1}^C r_k V_{k1}\right)^2 \,\middle|\, \mathbf{z}\right] = \mathbf{E}\left[\left(\frac{\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k) V_{k1}}{\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)}\right)^2 \,\middle|\, \mathbf{z}\right] \to \left(\frac{\mathbf{E}[\exp(\sigma(\mathbf{z})^\top v_k)V_{k1} \mid \mathbf{z}]}{\mathbf{E}[\exp(\sigma(\mathbf{z})^\top v_k) \mid \mathbf{z}]}\right)^2 = \frac{\sigma(z_1)^2}{m^2}.$$

Therefore

$$\mathbf{E}[\xi_C] = \mathbf{E}[\mathbf{E}[\xi_C|\mathbf{z}]] = \mathbf{E}\left[\frac{\mathbf{1}(z_1 > 0)}{m}\right] = \frac{1}{2m}.$$

Similarly,

$$\mathbf{E}[\xi_C^2] = \mathbf{E}[\mathbf{E}[\xi_C^2|\mathbf{z}]] = \mathbf{E}\left[\frac{\mathbf{1}(z_1 > 0)}{m^2}\right] = \frac{1}{2m^2}.$$

Therefore

$$\lim_{C\to\infty}\lim_{d,N\to\infty}\frac{\|H_{ii}\|_{\mathrm{F}}^2}{d} = \frac{2\gamma+1}{4m^2}.$$

For the case of G_{ij}, by repeating all arguments above, we have

$$\begin{aligned}
\eta_C =& \mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\left[\sum_{k=1}^C r_k V_{k1} V_{k2} - \left(\sum_{k=1}^C r_k V_{k1}\right)\left(\sum_{k=1}^C r_k V_{k2}\right)\right] \\
=& \mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\sum_{k=1}^C\sum_{l=1}^C r_k r_l V_{k1}(V_{k2} - V_{l2}) \\
=& \frac{1}{2}\cdot\mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\sum_{k=1}^C\sum_{l=1}^C r_k r_l (V_{k1} - V_{l1})(V_{k2} - V_{l2}) \\
=& \frac{1}{2}\cdot\mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\frac{\sum_{k=1}^C\sum_{l=1}^C \exp(\sigma(\mathbf{z})^\top v_k)\exp(\sigma(\mathbf{z})^\top v_l)(V_{k1} - V_{l1})(V_{k2} - V_{l2})}{\left[\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)\right]^2}.
\end{aligned} \tag{93}$$

Let

$$h(k,l) = \exp(\sigma(\mathbf{z})^\top v_k)\exp(\sigma(\mathbf{z})^\top v_l)(V_{k1} - V_{l1})(V_{k2} - V_{l2}).$$

Then for $k \neq k' \neq l \neq l'$,

$$\mathbf{E}[h(k,l) \mid \mathbf{z}] = 0, \quad \mathbf{E}[h(k,l)^2 \mid \mathbf{z}] < \infty, \quad \mathbf{E}[h(k,l)h(k',l') \mid \mathbf{z}] = 0,$$

$$\mathbf{E}[h(k,l)h(k,l') \mid \mathbf{z}] = \left(\frac{\sigma(z_1)^2\sigma(z_2)^2}{m^4} + \frac{\sigma(z_1)^2 + \sigma(z_2)^2}{m^3} + \frac{1}{m^2}\right)\exp\left(\frac{3}{m}\|\sigma(\mathbf{z})\|^2\right).$$

Then as $C \to \infty$,

$$\begin{aligned}
\mathbf{E}[\sqrt{C}\eta_C] &= \mathbf{E}[\mathbf{E}\sqrt{C}\eta_C|\mathbf{z}]] \\
&= \mathbf{E}\left[\frac{1}{2}\mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\mathbf{E}\left[\frac{\frac{1}{C^{3/2}}\sum_{k=1}^C\sum_{l=1}^C h(k,l)}{\left[\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)\right]^2}\middle|\mathbf{z}\right]\right] \\
&= \mathbf{E}\left[\frac{1}{2}\mathbf{1}(z_1 > 0)\mathbf{1}(z_2 > 0)\frac{\frac{1}{C^{3/2}}\sum_{k=1}^C\sum_{l=1}^C \mathbf{E}\left[h(k,l)\middle|\mathbf{z}\right]}{\mathbf{E}\left[\exp(\sigma(\mathbf{z})^\top v_k)\middle|\mathbf{z}\right]^2}\right] + o(1) \\
&= o(1),
\end{aligned} \tag{94}$$

$$\begin{aligned}
\mathbf{E}[C\eta_C^2] &= \mathbf{E}[\mathbf{E}[C\eta_C^2|\mathbf{z}]] \\
&= \mathbf{E}\left[\frac{1}{4}\mathbf{1}(z_1>0)\mathbf{1}(z_2>0)\mathbf{E}\left[\frac{\frac{1}{C^3}\left[\sum_{k=1}^C\sum_{l=1}^C h(k,l)\right]^2}{\left[\frac{1}{C}\sum_{k=1}^C \exp(\sigma(\mathbf{z})^\top v_k)\right]^4}\middle|\mathbf{z}\right]\right] \\
&\to \mathbf{E}\left[\frac{1}{4}\mathbf{1}(z_1>0)\mathbf{1}(z_2>0)\frac{4\mathbf{E}\left[h(k,l)h(k,l')\middle|\mathbf{z}\right]}{\mathbf{E}\left[\exp(\sigma(\mathbf{z})^\top v_k)\middle|\mathbf{z}\right]^4}\right] \\
&= \mathbf{E}\left[\mathbf{1}(z_1>0)\mathbf{1}(z_2>0)\left(\frac{\sigma(z_1)^2\sigma(z_2)^2}{m^4}+\frac{\sigma(z_1)^2+\sigma(z_2)^2}{m^3}+\frac{1}{m^2}\right)\exp\left(\frac{1}{m}\|\sigma(\mathbf{z})\|^2\right)\right] \\
&= \left(\mathbf{E}\left[\left(\frac{\sigma(z_1)^2}{m^2}+\frac{1}{m}\right)\exp\left(\frac{\sigma(z_1)^2}{m}\right)\mathbf{1}(z_1>0)\right]\right)^2\left(\mathbf{E}\left[\exp\left(\frac{\sigma(z_1)^2}{m}\right)\right]\right)^{m-2} \\
&= \frac{(m-1)^2}{2^m(m-2)^3m}\left(\sqrt{\frac{m}{m-2}}+1\right)^{m-2}.
\end{aligned} \tag{95}$$

Then from (92) we finish the proof. ■

E.2.2. Proof for the Hidden-layer Hessian with MSE Loss

For short of notations we write $K_{ii} = \frac{\partial^2 \ell_{\text{MSE}}(W,V)}{\partial w_i \partial w_i^\top}$, $K_{ij} = \frac{\partial^2 \ell_{\text{MSE}}(W,V)}{\partial w_i \partial w_j^\top}$. Then

$$K_{ii} = \left(\sum_{k=1}^C V_{ki}^2\right) L_{ii}, \quad K_{ij} = \left(\sum_{k=1}^C V_{ki}V_{kj}\right) L_{ij},$$

where

$$L_{ii} = \frac{1}{N}\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0)x_n x_n^\top, \quad L_{ij} = \frac{1}{N}\sum_{n=1}^N \mathbf{1}(w_i^\top x_n > 0)\mathbf{1}(w_j^\top x_n > 0)x_n x_n^\top.$$

The following decoupling lemma is motivated by [26].

Lemma 3 *Under the assumptions in Theorem 2, we have:*

1. *There exists random matrices $\hat{X}_N, \hat{\Lambda}_N$ in the same probability space, such that $\hat{X}_N \overset{d}{=} X_N$, $\hat{\Lambda}_N$ is a N-dimensional diagonal matrix with entries i.i.d.* ber$(\frac{1}{2})$ *random variables independent of $\hat{X}_N$, and $L_{ii} \overset{d}{=} \frac{1}{N}\hat{X}_N\hat{\Lambda}_N\hat{X}_N'$.*
2. *For $i \neq j$, there exists random matrices $\hat{X}_N, \hat{\Lambda}_N$ in the same probability space, such that $\hat{X}_N \overset{d}{=} X_N$, $\hat{\Lambda}_N$ is a N-dimensional diagonal matrix with entries i.i.d.* ber$(\frac{1}{4})$ *random variables independent of $\hat{X}_N$, and $L_{ij} \overset{d}{=} \frac{1}{N}\hat{X}_N\hat{\Lambda}_N\hat{X}_N'$.*

Proof Let $\xi_1, \cdots, \xi_p, \eta_1, \cdots, \eta_N$ be i.i.d. Radamacher random variables. Let

$$\tilde{X}_N = \operatorname{diag}(\xi) X_N \operatorname{diag}(\eta), \quad \tilde{\Lambda}_N = \operatorname{diag}(1(w_i' \tilde{x}_1 > 0), \cdots, 1(w_i' \tilde{x}_N > 0)).$$

Here $\widetilde{x}_1, \cdots, \widetilde{x}_N$ are column vectors of $\widetilde{X}_N$. Since entries of X_N are i.i.d. centered, $\{\tilde{X}_N, \tilde{\Lambda}_N\} \overset{d}{=} \{X_N, \Lambda_N\}$. Hence

$$\begin{aligned} G_{ii} \overset{d}{=} \frac{1}{N} \tilde{X}_N \tilde{\Lambda}_N \tilde{X}_N &= \frac{1}{N} \operatorname{diag}(\xi) X_N \operatorname{diag}(\eta) \tilde{\Lambda}_N \operatorname{diag}(\eta) X_N \operatorname{diag}(\xi) \\ &= \frac{1}{N} \operatorname{diag}(\xi) X_N \tilde{\Lambda}_N X_N \operatorname{diag}(\xi). \end{aligned} \tag{96}$$

Clearly $\operatorname{diag}(\xi) X_N \overset{d}{=} X_N$, then it suffices to show that $\tilde{\Lambda}_n$ is diagonal with i.i.d. $\text{ber}(\frac{1}{2})$ entries independent of $\{\xi, X_N\}$. To see this, we have

$$\tilde{\Lambda}_N(n,n) = 1(w_i' \operatorname{diag}(\xi) x_n \eta_n > 0), \quad 1 \le n \le N.$$

Then the required property follows from that $(\eta_n)_n$ are i.i.d. valuing in $\{1, -1\}$ with equal probability.

For $G_{ij} (i \neq j)$, the proof is the same, except that

$$\tilde{\Lambda}_N(n,n) = 1(w_i' \operatorname{diag}(\xi) x_n \eta_n > 0) \cdot 1(w_j' \operatorname{diag}(\xi) x_n \eta_n > 0), \quad 1 \le n \le N$$

are i.i.d. $\text{ber}(\frac{1}{4})$. ■

A probability measure μ is said to have the Marchenko-Pastur distribution with parameter $y > 0$ and $\sigma^2 > 0$, denoted as $\text{MP}(y, \sigma^2)$, if

$$\mu(dx) = (1 - y^{-1}) 1(y > 1) \delta_0(dx) + \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\sigma^2 y x} 1_{(\lambda_-, \lambda_+)}(x) dx, \ \lambda_+ = \sigma^2(1 + \sqrt{y})^2, \ \lambda_- = \sigma^2(1 - \sqrt{y})^2.$$

The Stieltjes transform of $\text{MP}(y, \sigma^2)$ is given by

$$s(z) = \frac{\sigma^2(1-y) - z + \sqrt{(z - \sigma^2 - y\sigma^2)^2 - 4y\sigma^4}}{2yz\sigma^2},$$

or equivalently by the equation (writing $s = s_\mu(z)$ for short)

$$yz\sigma^2 s^2 + (z - \sigma^2(1-y))s + 1 = 0. \tag{97}$$

Proposition 6 *As $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$, we have*

1. *$\mu_{L_{ii}}$ converges weakly almost surely to $MP(2\gamma, \frac{1}{2})$.*
2. *For $i \neq j$, $\mu_{L_{ij}}$ converges weakly almost surely to $MP(4\gamma, \frac{1}{4})$.*

Proof From Lemma 3, we have $L_{ii} \overset{d}{=} \frac{d}{N} T_N$, where $T_N = \frac{1}{d}\hat{X}_N \hat{\Lambda}_N \hat{X}'_N$. From Proposition 1, μ_{T_N} converges weakly almost surely to a deterministic measure μ_T. And $s_{\mu_T}(z)$ is specified by the equation

$$s(z) = \frac{1}{\frac{1}{\gamma}\int_{\mathbb{R}} \frac{t\nu_1(dt)}{1+ts(z)} - z}, \quad \forall z \in \mathbb{C}^+, \tag{98}$$

where $\nu_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. The equation can be simplified as

$$zs^2 + \left(z - \frac{1}{2\gamma} + 1\right) s + 1 = 0.$$

This is exactly (97) with $y = 2\gamma$, $\sigma^2 = \frac{1}{2\gamma}$. Then $\mu_{T_N} \to MP(2\gamma, \frac{1}{2\gamma})$ and hence $\mu_{L_{ii}} \to MP(2\gamma, \frac{1}{2})$. The L_{ij} case in the same procedure, replacing ν_1 in (98) with $\nu_2 = \frac{3}{4}\delta_0 + \frac{1}{4}\delta_1$. ■

Proposition 7 *Suppose that m is fixed and as $d, N \to \infty$, $\frac{d}{N} \to \gamma \in (0, +\infty)$. Then for $i \neq j$, we have*

$$\begin{aligned} \lim_{C\to\infty} \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|H_{ii}\|_{\mathrm{F}}^2\right]}{C^2 d} &= \frac{1+2\gamma}{4m^2}, \\ \lim_{C\to\infty} \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|H_{ij}\|_{\mathrm{F}}^2\right]}{Cd} &= \frac{1+4\gamma}{16m^2}. \end{aligned} \tag{99}$$

Proof From Proposition 6 we have

$$\frac{\|L_{ii}\|_{\mathrm{F}}^2}{d} = \int_{\mathbb{R}} x^2 \mu_{L_{ii}}(dx) \to \int_{\mathbb{R}} x^2 \mu_{MP,2\gamma,\frac{1}{2}}(dx) = \frac{1+2\gamma}{4} \quad a.s. \tag{100}$$

For $i \neq j$,

$$\frac{\|L_{ij}\|_{\mathrm{F}}^2}{d} = \int_{\mathbb{R}} x^2 \mu_{L_{ij}}(dx) \to \int_{\mathbb{R}} x^2 \mu_{MP,4\gamma,\frac{1}{4}}(dx) = \frac{1+4\gamma}{16} \quad a.s. \tag{101}$$

Since entries of V are i.i.d. $\mathcal{N}(0, \frac{1}{m})$ independent of $\{L_{ii}, L_{ij}\}$,

$$\lim_{C\to\infty} \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|H_{ii}\|_{\mathrm{F}}^2\right]}{C^2 d} = \lim_{C\to\infty} \mathbf{E}\left[\left(\frac{\sum_{k=1}^C V_{ki}^2}{C}\right)^2\right] \lim_{d,N\to\infty} \mathbf{E}\left[\frac{\|L_{ii}\|_{\mathrm{F}}^2}{d}\right] = \frac{1+2\gamma}{4m^2},$$

$$\lim_{C\to\infty} \lim_{d,N\to\infty} \frac{\mathbf{E}\left[\|H_{ij}\|_{\mathrm{F}}^2\right]}{Cd} = \lim_{C\to\infty} \mathbf{E}\left[\left(\frac{\sum_{k=1}^C V_{ki}V_{kj}}{\sqrt{C}}\right)^2\right] \lim_{d,N\to\infty} \mathbf{E}\left[\frac{\|L_{ii}\|_{\mathrm{F}}^2}{d}\right] = \frac{1+4\gamma}{16m^2}.$$

■

E.2.3. Proof for the Output-layer Hessian with CE Loss

Denote that

$$G_{ii} := \frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial v_i \partial v_i^\top} = \frac{1}{N}\sum_{n=1}^{N} p_{n,i}(1-p_{n,i})\sigma(Wx_n)\sigma(Wx_n)^\top,$$

and for $i \neq j$,

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

Let Z be a $d \times N$ random matrix with entries i.i.d. $N(0,1)$ independent of V, and $z_1, \cdots, z_N$ are column vectors of Z. Define

$$\widetilde{p}_{i,n} = \frac{\exp(\sigma(z_n)^\top v_i)}{\sum_{c=1}^{C}\exp(\sigma(z_n)^\top v_c)}, \quad n \in [N], i \in [C],$$

$$\widetilde{G}_{ii} = \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}(1-\widetilde{p}_{n,i})\sigma(z_n)\sigma(z_n)^\top,$$

$$\widetilde{G}_{ij} = \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}\widetilde{p}_{n,j}\sigma(z_n)\sigma(z_n)^\top.$$

Following the proof of Lemma 12 with the Lindeberg principle, one can show that for $k,l \in [m]$, $k \neq l$,

$$\lim_{d,N\to\infty}\left(\mathbf{E}\left[G_{ii}(k,k)^2\right] - \mathbf{E}\left[\widetilde{G}_{ii}(k,k)^2\right]\right) = 0,$$

$$\lim_{d,N\to\infty}\left(\mathbf{E}\left[G_{ii}(k,l)^2\right] - \mathbf{E}\left[\widetilde{G}_{ii}(k,l)^2\right]\right) = 0,$$

$$\lim_{d,N\to\infty}\left(\mathbf{E}\left[G_{ij}(k,k)^2\right] - \mathbf{E}\left[\widetilde{G}_{ij}(k,k)^2\right]\right) = 0,$$

$$\lim_{d,N\to\infty}\left(\mathbf{E}\left[G_{ij}(k,l)^2\right] - \mathbf{E}\left[\widetilde{G}_{ij}(k,l)^2\right]\right) = 0.$$

Then since m is fixed and

$$\begin{aligned}\mathbf{E}\left[\|G_{ii}\|_{\mathrm{F}}^2\right] &= m\mathbf{E}\left[G_{ii}(k,k)^2\right] + m(m-1)\mathbf{E}\left[G_{ii}(k,l)^2\right], \\ \mathbf{E}\left[\|G_{ij}\|_{\mathrm{F}}^2\right] &= m\mathbf{E}\left[G_{ij}(k,k)^2\right] + m(m-1)\mathbf{E}\left[G_{ij}(k,l)^2\right],\end{aligned} \tag{102}$$

$$\mathbf{E}\left[\|\widetilde{G}_{ii}\|_{\mathrm{F}}^2\right] = m\mathbf{E}\left[\widetilde{G}_{ii}(k,k)^2\right] + m(m-1)\mathbf{E}\left[\widetilde{G}_{ii}(k,l)^2\right], \tag{103}$$

$$\mathbf{E}\left[\|\widetilde{G}_{ij}\|_{\mathrm{F}}^2\right] = m\mathbf{E}\left[\widetilde{G}_{ij}(k,k)^2\right] + m(m-1)\mathbf{E}\left[\widetilde{G}_{ij}(k,l)^2\right], \tag{104}$$

we have

$$\begin{aligned}\lim_{d,N\to\infty}\left(\mathbf{E}\left[\|G_{ii}\|_{\mathrm{F}}^2\right] - \mathbf{E}\left[\|\widetilde{G}_{ii}\|_{\mathrm{F}}^2\right]\right) &= 0, \\ \lim_{d,N\to\infty}\left(\mathbf{E}\left[\|G_{ij}\|_{\mathrm{F}}^2\right] - \mathbf{E}\left[\|\widetilde{G}_{ij}\|_{\mathrm{F}}^2\right]\right) &= 0.\end{aligned} \tag{105}$$

From

$$\begin{aligned}
\widetilde{G}_{ii}(k,k) &= \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}(1-\widetilde{p}_{n,i})\sigma(z_{n,k})^2,\\
\widetilde{G}_{ii}(k,l) &= \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}(1-\widetilde{p}_{n,i})\sigma(z_{n,k})\sigma(z_{n,l}),\\
\widetilde{G}_{ij}(k,k) &= \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}\widetilde{d}_{n,j}\sigma(z_{n,k})^2,\\
\widetilde{G}_{ij}(k,l) &= \frac{1}{N}\sum_{n=1}^{N} \widetilde{p}_{n,i}\widetilde{p}_{n,j}\sigma(z_{n,k})\sigma(z_{n,l}),
\end{aligned} \tag{106}$$

we have

$$\begin{aligned}
\lim_{p,N\to\infty} \mathbf{E}\left[\widetilde{G}_{ii}(k,k)^2\right] &= \mathbf{E}\left[\widetilde{p}_{1,i}\widetilde{p}_{2,i}(1-\widetilde{p}_{1,i})(1-\widetilde{p}_{2,i})\sigma(z_{1,k})^2\sigma(z_{2,k})^2\right],\\
\lim_{p,N\to\infty} \mathbf{E}\left[\widetilde{G}_{ii}(k,l)^2\right] &= \mathbf{E}\left[\widetilde{p}_{1,i}\widetilde{p}_{2,i}(1-\widetilde{p}_{1,i})(1-\widetilde{p}_{2,i})\sigma(z_{1,k})\sigma(z_{1,l})\sigma(z_{2,k})\sigma(z_{2,l})\right],\\
\lim_{p,N\to\infty} \mathbf{E}\left[\widetilde{G}_{ij}(k,k)^2\right] &= \mathbf{E}\left[\widetilde{p}_{1,i}\widetilde{p}_{2,i}\widetilde{p}_{1,j}\widetilde{p}_{2,j}\sigma(z_{1,k})^2\sigma(z_{2,k})^2\right],\\
\lim_{p,N\to\infty} \mathbf{E}\left[\widetilde{G}_{ij}(k,l)^2\right] &= \mathbf{E}\left[\widetilde{p}_{1,i}\widetilde{p}_{2,i}\widetilde{p}_{1,j}\widetilde{p}_{2,j}\sigma(z_{1,k})\sigma(z_{1,l})\sigma(z_{2,k})\sigma(z_{2,l})\right].
\end{aligned} \tag{107}$$

Therefore

$$\begin{aligned}
&\lim_{C\to\infty}\lim_{p,N\to\infty} C^2\mathbf{E}\left[\widetilde{G}_{ii}(k,k)^2\right]\\
=&\lim_{C\to\infty}\mathbf{E}\left[\mathbf{E}\left[\frac{\exp\left((\sigma(z_1)+\sigma(z_2))^\top v_i\right)}{\left(\frac{1}{C}\sum_{c=1}^{C}\exp(\sigma(z_1)^\top v_c)\right)\left(\frac{1}{C}\sum_{c=1}^{C}\exp(\sigma(z_2)^\top v_c)\right)}(1-\widetilde{p}_{1,i})(1-\widetilde{p}_{2,i})\sigma(z_{1,k})^2\sigma(z_{2,k})^2\,\middle|\, z_1,z_2\right]\right]\\
=&\mathbf{E}\left[\frac{\mathbf{E}\left[\exp\left((\sigma(z_1)+\sigma(z_2))^\top v_i\right)\sigma(z_{1,k})^2\sigma(z_{2,k})^2\,\middle|\, z_1,z_2\right]}{\left(\mathbf{E}\left[\exp\left(\sigma(z_1)^\top v_i\right)\,\middle|\, z_1\right]\right)\left(\mathbf{E}\left[\exp\left(\sigma(z_2)^\top v_i\right)\,\middle|\, z_2\right]\right)}\right]\\
=&\mathbf{E}\left[\frac{\exp\left(\frac{1}{2m}|\sigma(z_1)+\sigma(z_2)|^2\right)\sigma(z_{1,k})^2\sigma(z_{2,k})^2}{\exp\left(\frac{1}{2m}(|\sigma(z_1)|^2+|\sigma(z_2)|^2)\right)}\right]\\
=&\mathbf{E}\left[\exp\left(\frac{1}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\sigma(z_{1,1})^2\sigma(z_{1,2})^2\right]\left(\mathbf{E}\left[\exp\left(\frac{1}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\right]\right)^{m-1}\\
=&a_{12}b_1^{m-1},
\end{aligned} \tag{108}$$

$$
\begin{aligned}
&\lim_{C\to\infty}\lim_{p,N\to\infty} C^2\mathbf{E}\left[\widetilde{G}_{ii}(k,l)^2\right]\\
=&\lim_{C\to\infty}\mathbf{E}\left[\mathbf{E}\left[\frac{\exp\left((\sigma(z_1)+\sigma(z_2))^\top v_i\right)(1-\widetilde{p}_{1,i})(1-\widetilde{p}_{2,i})}{\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_1)^\top v_c)\right)\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_2)^\top v_c)\right)}\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})\,\middle|\, z_1,z_2\right]\right]\\
=&\mathbf{E}\left[\frac{\mathbf{E}\left[\exp\left((\sigma(z_1)+\sigma(z_2))^\top v_i\right)\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})\middle|\, z_1,z_2\right]}{\left(\mathbf{E}\left[\exp\left(\sigma(z_1)^\top v_i\right)\middle|\, z_1\right]\right)\left(\mathbf{E}\left[\exp\left(\sigma(z_2)^\top v_i\right)\middle|\, z_2\right]\right)}\right]\\
=&\mathbf{E}\left[\frac{\exp\left(\frac{1}{2m}|\sigma(z_1)+\sigma(z_2)|^2\right)\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})}{\exp\left(\frac{1}{2m}(|\sigma(z_1)|^2+|\sigma(z_2)|^2)\right)}\right]\\
=&\mathbf{E}\left[\exp\left(\frac{1}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\sigma(z_{1,1})\sigma(z_{1,2})\right]^2\left(\mathbf{E}\left[\exp\left(\frac{1}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\right]\right)^{m-2}\\
=&a_{11}^2 b_1^{m-2},
\end{aligned}
\tag{109}
$$

$$
\begin{aligned}
&\lim_{C\to\infty}\lim_{p,N\to\infty} C^4\mathbf{E}\left[\widetilde{G}_{ij}(k,k)^2\right]\\
=&\lim_{C\to\infty}\mathbf{E}\left[\mathbf{E}\left[\frac{\exp\left((\sigma(z_1)+\sigma(z_2))^\top (v_i+v_j)\right)}{\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_1)^\top v_c)\right)^2\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_2)^\top v_c)\right)^2}\sigma(z_{1,k})^2\sigma(z_{2,k})^2\,\middle|\, z_1,z_2\right]\right]\\
=&\mathbf{E}\left[\frac{\mathbf{E}\left[\exp\left((\sigma(z_1)+\sigma(z_2))^\top (v_i+v_j)\right)\sigma(z_{1,k})^2\sigma(z_{2,k})^2\middle|\, z_1,z_2\right]}{\left(\mathbf{E}\left[\exp\left(\sigma(z_1)^\top v_i\right)\middle|\, z_1\right]\right)^2\left(\mathbf{E}\left[\exp\left(\sigma(z_2)^\top v_i\right)\middle|\, z_2\right]\right)^2}\right]\\
=&\mathbf{E}\left[\frac{\exp\left(\frac{1}{m}|\sigma(z_1)+\sigma(z_2)|^2\right)\sigma(z_{1,k})^2\sigma(z_{2,k})^2}{\exp\left(\frac{1}{m}(|\sigma(z_1)|^2+|\sigma(z_2)|^2)\right)}\right]\\
=&\mathbf{E}\left[\exp\left(\frac{2}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\sigma(z_{1,1})^2\sigma(z_{1,2})^2\right]\left(\mathbf{E}\left[\exp\left(\frac{2}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\right]\right)^{m-1}\\
=&a_{22} b_2^{m-1},
\end{aligned}
\tag{110}
$$

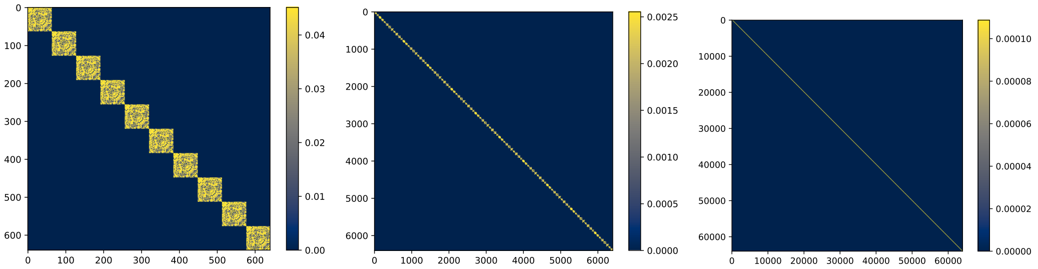
$$
\begin{aligned}
&\lim_{C\to\infty}\lim_{p,N\to\infty} C^4\mathbf{E}\left[\widetilde{G}_{ij}(k,k)^2\right]\\
=&\lim_{C\to\infty}\mathbf{E}\left[\mathbf{E}\left[\frac{\exp\left((\sigma(z_1)+\sigma(z_2))^\top(v_i+v_j)\right)}{\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_1)^\top v_c)\right)^2\left(\frac{1}{C}\sum_{c=1}^C\exp(\sigma(z_2)^\top v_c)\right)^2}\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})\,\middle|\,z_1,z_2\right]\right]\\
=&\mathbf{E}\left[\frac{\mathbf{E}\left[\exp\left((\sigma(z_1)+\sigma(z_2))^\top(v_i+v_j)\right)\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})\,\middle|\,z_1,z_2\right]}{\left(\mathbf{E}\left[\exp\left(\sigma(z_1)^\top v_i\right)\,\middle|\,z_1\right]\right)^2\left(\mathbf{E}\left[\exp\left(\sigma(z_2)^\top v_i\right)\,\middle|\,z_2\right]\right)^2}\right]\\
=&\mathbf{E}\left[\frac{\exp\left(\frac{1}{m}|\sigma(z_1)+\sigma(z_2)|^2\right)\sigma(z_{1,k})\sigma(z_{2,k})\sigma(z_{1,l})\sigma(z_{2,l})}{\exp\left(\frac{1}{m}(|\sigma(z_1)|^2+|\sigma(z_2)|^2)\right)}\right]\\
=&\mathbf{E}\left[\exp\left(\frac{2}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\sigma(z_{1,1})\sigma(z_{1,2})\right]^2\left(\mathbf{E}\left[\exp\left(\frac{2}{m}\sigma(z_{1,1})\sigma(z_{1,2})\right)\right]\right)^{m-2}\\
=&a_{21}^2b_2^{m-2}.
\end{aligned}
\tag{111}
$$

Then from (103)-(105) and (108)-(111) we obtain (8). The whole proof is then completed.

Appendix F. More Numerical Results

We now provide some more numerical evidence to support our theory. We use the the same Gaussian synthetic dataset as in Appendix 2 (which follows Assumption 1) and LeCun initialization (which follows Assumption 2), and try different C. More details can be seen in Appendix G.3.

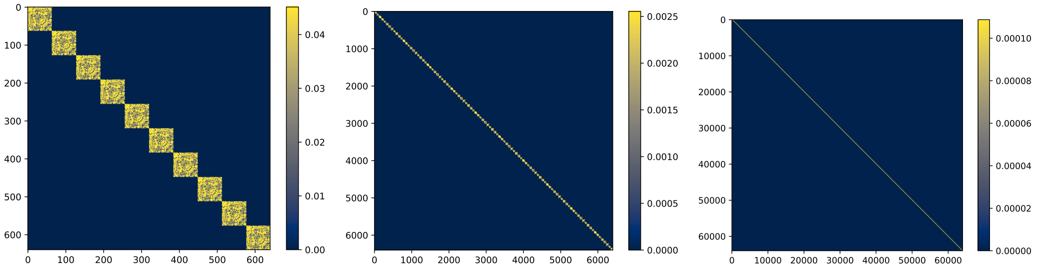
Case 1: linear models with MSE loss. In Figure 4, we present the Hessian of linear models under MSE loss. By the calculation of (15) in Appendix A, the Hessian is strictly block-diagonal. The numerical results match the calculation.



(a) $C = 10$ (b) $C = 100$ (c) $C = 1000$

Figure 4: **(a-c):** The Hessian of **Case 1: linear models with MSE loss**. We observe that the block-diagonal Hessian structure arises for all C. This is because the off-diagonal blocks are strictly zero in this case (see Eq. (15)).

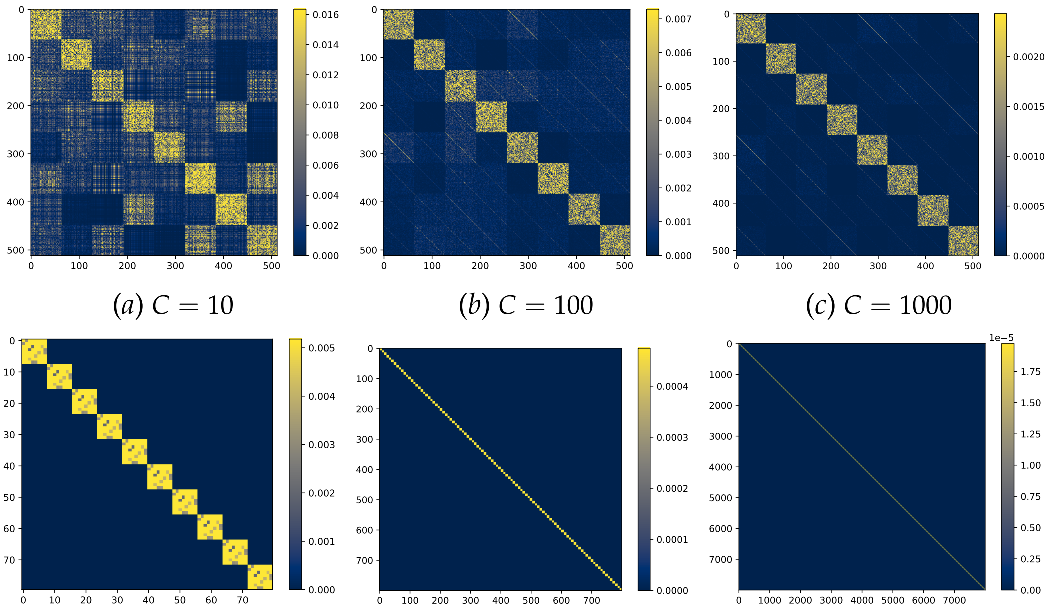
Case 2: linear models with CE loss. In Figure 5, we present the Hessian of linear models under CE loss. The block-diagonal Hessian structure becomes clear when C increases, which matches our theoretical prediction in Theorem 1.



(a) $C = 10$ (b) $C = 100$ (c) $C = 1000$

Figure 5: **(a-c):** The Hessian of **Case 2: linear models with CE loss**.We observe that the block-diagonal Hessian structure becomes clear when C increases.

Case 3 and 4: 1-hidden-layer networks with MSE and CE loss. In Figure 6 and 7, we consider 1-hidden-layer network at random initialization. We present the hidden-layer Hessian H_{ww} and output weights H_{vv} to see if they match our theoretical prediction. It can be seen that the block-diagonal structure becomes clearer as the number of classes C



(*a*) $C = 10$ (*b*) $C = 100$ (*c*) $C = 1000$

(*d*) $C = 10$ (*e*) $C = 100$ (*f*) $C = 1000$

Figure 6: The Hessian in **Case 3: 1-hidden-layer network with MSE loss**. The network has 8 hidden neurons. **(a, b, c):** The hidden-layer Hessian H_{ww}. **(e, f, g):** The output-layer Hessian H_{vv}. We observe that the block-diagonal Hessian structure in H_{ww} becomes clearer as C increases. H_{vv} is always strictly block-diagonal, as expected by Eq. (21).

increases, which matches the theoretical prediction. These results hold for both MSE and CE loss.

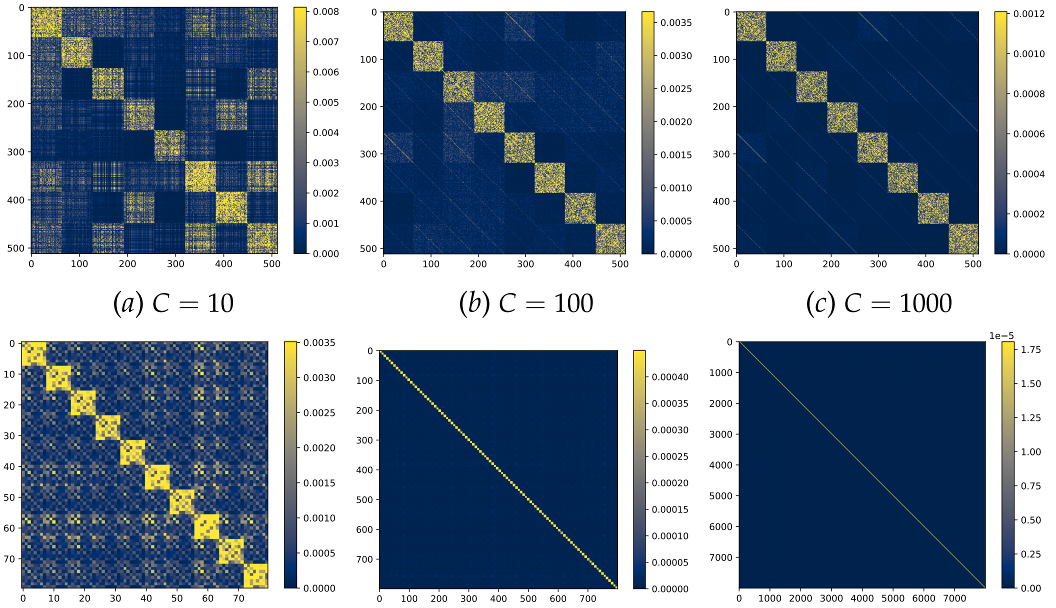
On the Frobenius Norm of Hessian Blocks for Case 2. We now investigate the following quantities, which appeared in Theorem 1:

$$H_{11}^{\mathrm{CE}} := \frac{C^2}{d}\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_1 \partial v_1^\top}\right\|_{\mathrm{F}}^2, \quad H_{12}^{\mathrm{CE}} := \frac{C^4}{d}\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_1 \partial v_2^\top}\right\|_{\mathrm{F}}^2, \quad r := \left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_1 \partial v_2^\top}\right\|_{\mathrm{F}}^2 \Big/ \left\|\frac{\partial^2 \ell_{\mathrm{CE}}(V)}{\partial v_1 \partial v_1^\top}\right\|_{\mathrm{F}}^2.$$

For each C, we simulate 1000 H_{11}^{CE} and H_{12}^{CE} with LeCun initialization (Assumption 2), and track their changes with C. The results are shown in Figure 8. We make the following observations. These observations match our theoretical prediction.

- **First**, for each C, the realizations of H_{11}^{CE} and H_{12}^{CE} concentrate around the red curves, which are their theoretical means in (1) and (2) in Proposition 3 (shown later in Appendix E.1).
- **Second,** as $C \to \infty$, we find that the H_{11} and H_{12} approach the green lines, which are their theoretical limits in Theorem 1. These results justify the results in Theorem 1.
- **Third**, as $C \to \infty$, we have $r \to 0$, and the decay rate matches our theoretical prediction. This means that off-diagonal blocks become relatively negligible as C increases.

On the Frobenius Norm of Hessian Blocks for Case 4. We now consider 1-hidden-layer networks with CE loss. We introduce the quantities as follows.



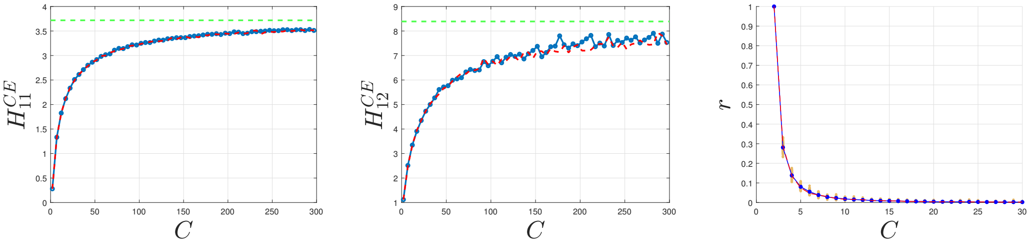
(*a*) $C = 10$ (*b*) $C = 100$ (*c*) $C = 1000$

(*d*) $C = 10$ (*e*) $C = 100$ (*f*) $C = 1000$

Figure 7: The Hessian in **Case 4: 1-hidden-layer network with CE loss**. The network has 8 hidden neurons. **(a, b, c):** The hidden-layer Hessian H_{ww}. **(e, f, g):** The output-layer Hessian H_{vv}. For both H_{ww} and H_{vv}, we observe that the block-diagonal Hessian structure becomes clearer as C increases.

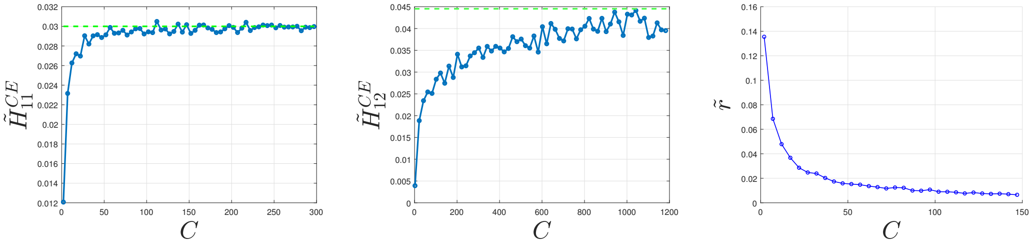
$$\tilde{H}_{11}^{\mathrm{CE}} := \frac{1}{d}\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_1 \partial w_1^\top}\right\|_{\mathrm{F}}^2, \quad \tilde{H}_{12}^{\mathrm{CE}} = \frac{C}{d}\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_1 \partial w_2^\top}\right\|_{\mathrm{F}}^2, \quad \tilde{r} = \frac{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_1 \partial w_2^\top}\right\|_{\mathrm{F}}^2}{\left\|\frac{\partial^2 \ell_{\mathrm{CE}}(W,V)}{\partial w_2 \partial w_2^\top}\right\|_{\mathrm{F}}^2}$$

The results are shown in Figure 9. Similarly as in Figure 8, we find that the $\tilde{H}_{11}^{\mathrm{CE}}$ and $\tilde{H}_{12}^{\mathrm{CE}}$ approach the green lines, which are their theoretical limits in Theorem 2. Further, $\tilde{r}$ decays to 0 with the same rate as we predicted in Theorem 2. These results support the results in Theorem 2.



(a) C v.s. H_{11}^{CE} (b) C v.s. H_{12}^{CE} (c) C v.s. r

Figure 8: **(a, b, c):** The evolution of H_{11}^{CE}, H_{12}^{CE}, and r as C increases. For each C, the realizations of H_{11}^{CE} and H_{12}^{CE} concentrate around the red curves, which are their theoretical means in (1) and (2) in Proposition 3 (shown later in Appendix E.1). As $C \to \infty$, H_{11}^{CE} and H_{12}^{CE} approach the green lines, which are their theoretical limits in Theorem 1. Further, r vanishes to 0 as C increases, and the decay rate matches Theorem 1. This means that off-diagonal blocks become relatively negligible as C increases.



(a) C v.s. $\tilde{H}_{11}^{\mathrm{CE}}$ (b) C v.s. $\tilde{H}_{12}^{\mathrm{CE}}$ (c) C v.s. $\tilde{r}$

Figure 9: **(a,b):** The evolution of $\tilde{H}_{11}^{\mathrm{CE}}$ and $\tilde{H}_{12}^{\mathrm{CE}}$ as C increases. We find that the $\tilde{H}_{11}^{\mathrm{CE}}$ and $\tilde{H}_{12}^{\mathrm{CE}}$ approach the green lines, which are their theoretical limits in Theorem 2. **(c):** $\tilde{r}$ decays to 0 with the same rate as we predicted in Theorem 2.

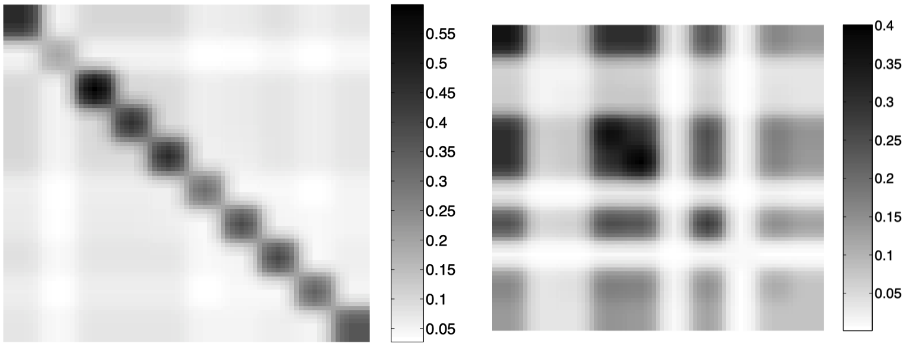
Appendix G. More Numerical Results and Experimental Details

Now we present the numerical results. We first re-state the experiments in [11] and then present some more of our numerical results. All experimental details of our results are explained in Appendix G.3.

G.1. Results from [11]

In Figure 10, we restate Figure 7.3 and 7.5 from [11]. The authors reported block-diagonal Hessian structure for a 1-hidden-layer network with CE loss on a binary-classification dataset. Such structure disappeared when changing to MSE loss.

We make two comments here. First, for MSE loss in Figure 10 (b), it is not surprising to see non-block-diagonal structure since our theory states that such structure arises when $C \to \infty$, while C only equals to 2 here. Second, for CE loss in Figure 10 (a), the near-block-diagonal structure arises despite the small C. This is not covered in our theory since we focus on large C. Nevertheless, it does not contradict our theory, either. We leave the exploration of binary classification with CE loss as a future direction.



(*a*) CE loss

(*b*) MSE loss

Figure 10: **(a,b):** Figure 7.3 and 7.5 from [11]. . The Hessian matrices of a 1-hidden-layer network with 10 hidden neurons on the Forest binary-classification dataset. For CE loss, the Hessian is computed after 1 iteration. For MSE loss, the Hessian is computed after 10 iterations.

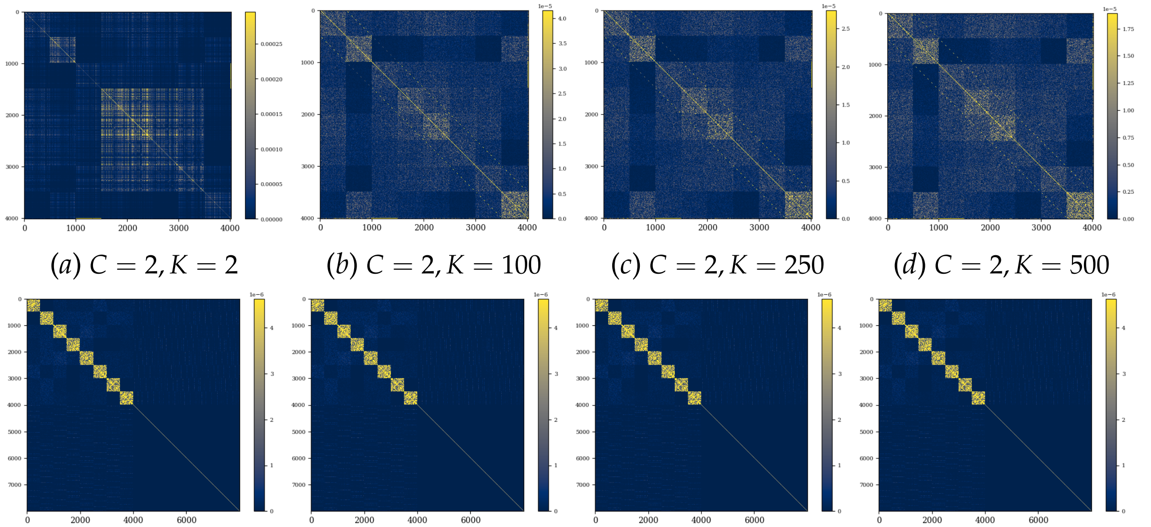
G.2. More Ablation Studies

We now conduct some more ablation studies on other Factors contributing to the Hessian structure. Li et al. [37] argue that K-feature-clustered dataset will bring K-ranked Jacobian. We now investigate how this relates to the block-diagonal Hessian structure.

We construct the K-clustered dataset following the descriptions in [37]: *"assume that the input $x_n \in \mathbb{R}^d$ come from K clusters which are located on the unit Euclidean ball; assume our dataset consists of $C \leq K$ classes where each class can be composed of multiple clusters."*. We attach the code for data generation below.

We present the results in Figure 11 and 12. We report two findings here: (1) When # classes $C = 2$ is small, the Hessian has no block-diagonal structure, regardless of # clusters

K. (2) When $C = 500$ is large, the block-diagonal pattern appears regardless of # clusters K. This suggests that large C plays a more critical role than K in the Hessian structure.



(*a*) $C = 2, K = 2$ (*b*) $C = 2, K = 100$ (*c*) $C = 2, K = 250$ (*d*) $C = 2, K = 500$

(*e*) $C = 500, K = 500$ (*f*) $C = 500, K = 750$ (*g*) $C = 500, K = 1000$ (*h*) $C = 500, K = 1250$

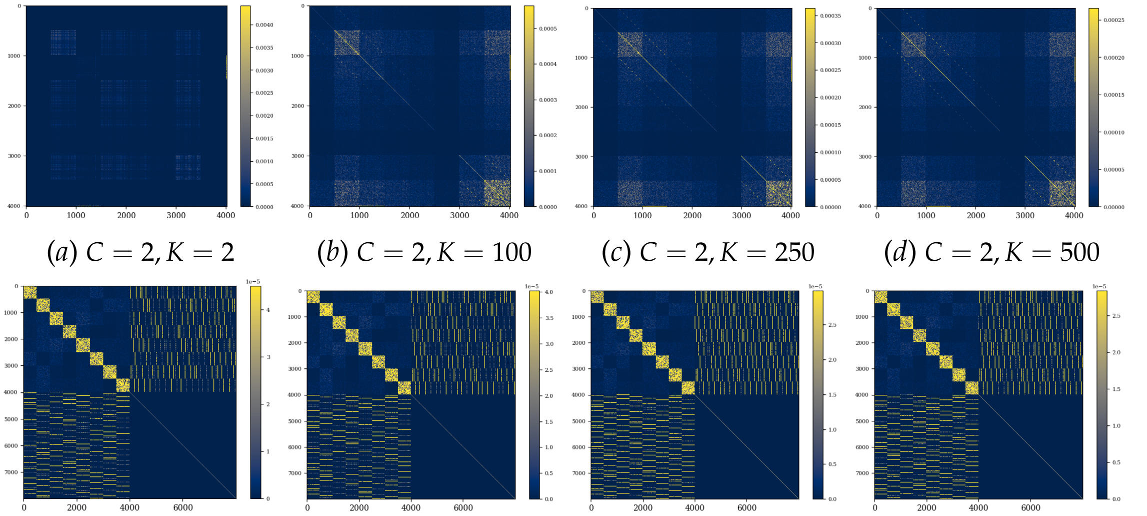
Figure 11: Ablation studies for the effect of # cluster K on the Hessian structure. We construct K-clustered dataset following the setup in [Li et al. 19]: "assume that the input $x_n \in \mathbb{R}^d$ come from K clusters which are located on the unit Euclidean ball; assume our dataset consists of $C \leq K$ classes where each class can be composed of multiple clusters." We use MSE loss and random intialization. We find that: **(a-d):** When $C = 2$ is small, the Hessian has no clear structure, regardless of # clusters K. **(e-h):** When $C = 500$ is large, the block-diagonal pattern appears regardless of # clusters K. This suggests that large C plays a more critical role than K in the Hessian structure.

```python

def generate_cluster_data(n_total, n_classes, n_clusters, input_dim):
    # Generate clustered synthetic data for specified dimensions
    # used for ablation study
    # n_total is the total number of samples
    # n_classes is the number of classes (smaller than n_clusters)
    # input_dim is the dimension of the data
    # raise error if n_cluster is larger than n_classes
    assert n_classes<= n_clusters, f"n_cluster = {n_classes} is not smaller than n_classes = {n_classes}"

    # n_samples_per_class is the number of samples per class
    X = []
    y = []
    n_cluster_per_class = n_clusters // n_classes
    n_samples_per_cluster = n_total // n_clusters
    cluster_idx = 0
    for class_idx in range(n_classes):

        for _ in range(n_cluster_per_class):
            cluster_idx += 1
            if input_dim == 2:
```



(*a*) $C = 2, K = 2$ (*b*) $C = 2, K = 100$ (*c*) $C = 2, K = 250$ (*d*) $C = 2, K = 500$

(*e*) $C = 500, K = 500$ (*f*) $C = 500, K = 750$ (*g*) $C = 500, K = 1000$ (*h*) $C = 500, K = 1250$

Figure 12: Same ablation studies as in Figure 11 except that we change to CE loss. We find that: **(a-d):** When $C = 2$ is small, the Hessian has no clear structure, regardless of # clusters K. **(e-h):** When $C = 500$ is large, the block-diagonal pattern appears in H_{ww} and H_{vv} regardless of # clusters K. This suggests that large C plays a more critical role than K in the Hessian structure.

```
                center = np.array([np.cos(2 * np.pi * cluster_idx / n_clusters), np.sin(2 * np.pi * (cluster_idx) / n_clusters)]) * 5  # Class centers on a circle
            else:
                #extend the 2D case to higher dimension
                # Generate random points in higher dimensions and project onto hypersphere
                center = np.random.randn(input_dim)
                # Normalize to create a unit vector (point on unit hypersphere)
                center = center / np.linalg.norm(center)

            cluster_samples = np.random.randn(n_samples_per_cluster, input_dim) * 0.05 + center  # Add some noise
            X.append(cluster_samples)
            # assign label
            y.extend([class_idx]*n_samples_per_cluster)

    X = np.vstack(X)  # Combine all class samples
    y = np.array(y)    # Convert labels to a NumPy array
    return X, y
```

G.3. Experimental Details

Now we present the experimental details. All experiments are conducted on one NVIDIA V100 GPU.

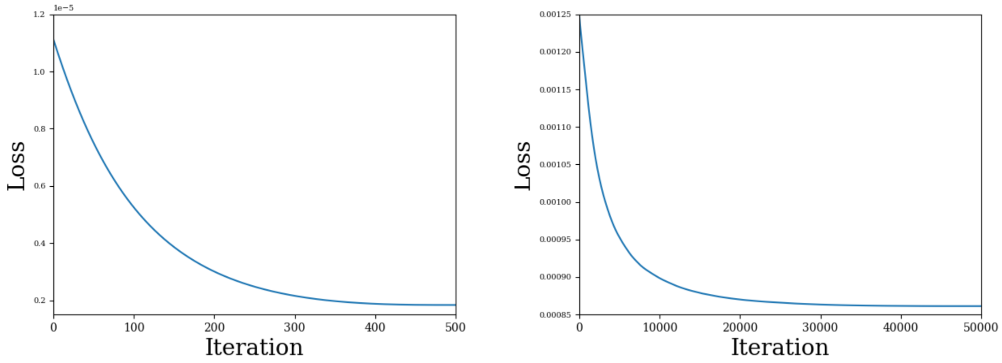
Implementation details for Figure 1. We calculate Hessian on a randomly selected 128 images from CIFAR-100. We calculate Hessian via two backpropagation passes [59], our code is modified based on open-source Hessian-vector-product implementation [2]. We consider a 1-hidden-layer network with ReLU activation, 8 hidden neurons, and 100 output neurons at random initialization. For all Hessian matrices reported in the paper, we report the absolute value of each Hessian entry.

Implementation details for Appendix 2 . We first introduce the implementation details for the synthetic dataset used in both Appendix 2 and Appendix G. We build the dataset following Assumption 1 and we assign the label randomly. We attach the code for data generation here. In Section 2, we use `input_dim` = 500, `n_classes` = 500, `n_samples_per_class` = 10.

```python
def generate_gaussian_data(n_samples_per_class, n_classes, input_dim):
    X = []
    y = []
    for i in range(n_classes):
        class_samples = np.random.randn(n_samples_per_class, input_dim)
        X.append(class_samples)
        y.extend([i] * n_samples_per_class) # not used

    X = np.vstack(X)
    y = np.array(y)
    return X, y
```

Now we describe the model configurations in Appendix 2. We use 1-hidden-layer networks with 8 hidden neurons and ReLU activation. All the models are trained using Adam with `lr` = 1e-4 with cosine annealing schedule with `lr_min` = 0. We train the models until convergence. The loss curves are reported in Figure 13.



(*a*) Loss curve under MSE loss

(*b*) Loss curve under CE loss

Figure 13: The loss curve of the models trained in Appendix 2. We train the models until convergence.

Implementation details for Appendix F and Appendix G. For these experiments, we use the same setups as in Appendix 2. We use `input_dim` = 64, `n_samples_per_class` = 1, and the total number of samples $N = C$.

[2] https://github.com/zyushun/hessian-spectrum

Implementation details for Figure 8 and Figure 9. For Figure 8, we consider $N = 1000, d = 1000$. For each C, we randomly sample 1000 realization of H_{11}, H_{12} and report their Frobenius norms. We use 10^6 repetitions in Monte-Carlo integrals. For Figure 9, we use $N = 300$, $d = 300$, and 200 repetitions for each C.

Appendix H. More Discussions

More discussions on our theory. For completeness, we provide discuss more discussions on our theory, including some clarifications and future directions.

- **First,** our theory requires N and d proportionally grows to infinity, which is also known as "the proportional asymptotic regime". We believe this regime is meaningful. First, the proportional asymptotic regime in standard in random matrix theory (e.g., see [70, 76]). Second, in the proportional asymptotic regime, we obtained new insights into the Hessian structure (e.g., the effect of large C) and our insights matched a wide range of finite-dimensional experiments. We will try to extend our analysis to other regimes such as non-asymptotic or over-parameterized regime in the future.

- **Second,** our theory focuses on random initialization, and it does not cover the whole training process. Interestingly, we numerically find that the block-diagonal Hessian structure remains throughout the training process (see Figure 2 and 3). This suggests that block-diagonal structure continuously influences the behavior of optimizers, *not just at initialization*. It is possible to extend our theory to the whole training process, but it requires substantially new mathematical tools and we leave as a future direction.

- **Third,** our theory focuses on linear and 1-hidden-layer networks, and it currently does not cover deeper models. We believe our theory on linear and 1-hidden-layer networks is meaningful since it already provides new insights, e.g., the effect of large C. It is possible to extend the results to deeper models by recursively applying our decoupling methods, but substantial effort is needed. For deeper models, we conjecture the block-diagonal structure will be primarily driven by the number of output neurons in each layer (a.k.a., the "fan-out dimension"). It is also intriguing to explore other potential factors that will reshape the Hessian structure of deep models. We leave it as a future direction.

- **Forth,** our theory focuses on block-wise Frobenius norm instead of block-wise spectrum. We focus on on Frobenius norm is it is more relevant to our current goal: justifying the Hessian structure. One future direction is to theoretically characterize the block-wise spectrum and provide guidance for optimizer design. Wang et al. [73], Zhang et al. [85, 86] did some initial attempts in this direction, but these works focused on numerical exploration and did not establish rigorous theory on characterizing block-wise spectrum. Based on our theory so far, it is possible to theoretically analyze the block-wise spectrum by more fine-grained analysis of the Steiltjes transform of the limit eigenvalue distribution, which we leave as a future direction.