# How the Hessian-Spectrum of Neural Networks Depends on Data

**Jasraj Singh**[*] JASRAJ.SINGH00150@GMAIL.COM

**Enea Monzio Compagnoni** ENEAMONZIOCOMPAGNONI@GMAIL.COM
*Universität Basel, Basel, Switzerland*

**Antonio Orvieto** ANTONIO@TUE.ELLIS.EU
*ELLIS Institute, Tübingen, Germany*
*MPI-IS, Tübingen, Germany*

## Abstract

The Hessian matrix is an important quantity of interest when it comes to studying the loss landscape and optimization dynamics in deep learning, as well as designing measures of generalization, second-order learning algorithms, etc. Prior works have focused on empirical results or pursued a theoretical treatment under overly simplified settings. In this work, we derive the eigenvalues of the Hessian of linear networks with arbitrary widths and depths, and datasets with an arbitrary number of samples, features, and labels. Importantly, for classification tasks with MSE loss, we identify that the sharpness of the solution is directly related to the maximum proportion of samples belonging to any class. We empirically validate our predictions and systematically analyze the effects of shedding the impractical assumptions one at a time, as well as incorporating nonlinearities. We observe that our predictions are considerably robust in most cases, allowing us to extend our conclusions to more practical learning setups.

## 1. Introduction

Neural networks induce highly nonconvex loss landscapes that are complicated to study. Despite this complexity, the Hessian matrix of the loss, encoding only its second-order information, has played a central role in designing efficient optimization algorithms [9, 23], deriving measures of generalization [2, 15] – which, in turn, inspired the design of well-generalizing optimizers [6, 20] – as well as practical algorithms for pruning [10, 21], quantization [7], continual learning [24, 25], etc.

Our motivation resembles that of [31, 32], in that we aim to precisely characterize the full spectrum of the Hessian. However, while these works prioritize minimal assumptions at the cost of model complexity, we leverage standard assumptions in deep learning theory in exchange for networks with *arbitrary widths* and *depths*, and datasets with *arbitrary number of samples*, *features* and *labels*. Our methodology falls closest to [8], which studies the spectrum of the Hessian and training dynamics at the edge-of-stability (EoS) [3] for matrix factorization problems. In contrast, we consider the problem of supervised learning, with the overarching goal of understanding the influence of data-geometry on the sharpness of the learnt solution.

[*] Work done during an internship at the ELLIS Institute, Tübingen, and MPI-IS.

© J. Singh, E.M. Compagnoni & A. Orvieto.

## 2. Setup

Consider the inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^{1\times d_0}$ collected in $\mathbf{X} \in \mathbb{R}^{n\times d_0}$, and outputs $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^{1\times d_L}$ collected in $\mathbf{Y} \in \mathbb{R}^{n\times d_L}$. Our model is an $L$-layer linear neural network, defined as

$$\mathbf{f}\left(\mathbf{x}; \mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L\right) = \mathbf{x}\mathbf{W}_1^\intercal \mathbf{W}_2^\intercal \ldots \mathbf{W}_L^\intercal \tag{1}$$

with $\mathbf{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ is the learnable weight matrix in the $\ell^{\text{th}}$ layer, $\forall \ell \in [L] \coloneqq \{1, 2, \ldots, L\}$.[1] In what follows, we denote the model parameters by $\mathbf{W}_{1:L} \coloneqq (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L)$. The loss function is the Mean-Squared Error (MSE):

$$\mathcal{L}\left(\mathbf{W}_{1:L}; \mathbf{X}, \mathbf{Y}\right) \coloneqq \frac{1}{2n} \left\|\mathbf{f}\left(\mathbf{X}; \mathbf{W}_{1:L}\right) - \mathbf{Y}\right\|_F^2 \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of the argument.

### 2.1. Approximating the Hessian-spectrum

We use the generalized Gauss-Newton (GGN) approximation of the Hessian, which gets more and more precise as loss reduces with training:

$$\frac{\partial^2}{\partial \mathbf{W}_{1:L}^2} \mathcal{L}\left(\mathbf{W}_{1:L}; \mathbf{X}, \mathbf{Y}\right) \approx \frac{1}{n} \left(\frac{\partial \mathbf{f}\left(\mathbf{X}\right)}{\partial \mathbf{W}_{1:L}}\right)^\intercal \left(\frac{\partial \mathbf{f}\left(\mathbf{X}\right)}{\partial \mathbf{W}_{1:L}}\right) = \frac{1}{n} \mathbf{J}^\intercal \mathbf{J} \in \mathbb{R}^{p\times p} \tag{3}$$

with the total number of parameters given by $p = \sum_{\ell=1}^{L} d_\ell d_{\ell-1}$, and the Jacobian w.r.t. predictions given by $\mathbf{J} \coloneqq \partial \mathbf{f}\left(\mathbf{X}\right) / \partial \mathbf{W}_{1:L} = \begin{bmatrix}\mathbf{J}_1 & \mathbf{J}_2 & \ldots & \mathbf{J}_L\end{bmatrix} \in \mathbb{R}^{d_L n \times p}$, where

$$\mathbf{J}_\ell = \left(\underbrace{\mathbf{W}_{\ell+1}^\intercal \mathbf{W}_{\ell+2}^\intercal \ldots \mathbf{W}_L^\intercal}_{\mathbf{S}_\ell \in \mathbb{R}^{d_\ell \times d_L}}\right)^\intercal \otimes \left(\mathbf{X} \underbrace{\mathbf{W}_1^\intercal \mathbf{W}_2^\intercal \ldots \mathbf{W}_{\ell-1}^\intercal}_{\mathbf{P}_\ell \in \mathbb{R}^{d_0 \times d_{\ell-1}}}\right) \tag{4}$$

where $\otimes$ is the Kronecker outer-product. The GGN matrix shares its non-zero eigenvalues with the Neural Tangent Kernel (NTK) matrix [14], which is simpler to study:

$$\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal = \sum_{\ell=1}^{L} \left(\mathbf{S}_\ell^\intercal \mathbf{S}_\ell\right) \otimes \left(\frac{1}{n}\mathbf{X}\mathbf{P}_\ell \mathbf{P}_\ell^\intercal \mathbf{X}^\intercal\right) \in \mathbb{R}^{d_L n \times d_L n} \tag{5}$$

**Assumption 1** *We have an overdetermined system, i.e. $n \geq d_0$, and the feature matrix is full-rank, i.e. rank $(\mathbf{X}) = d_0$.*

Going forward, we will make use of the Singular Value Decomposition (SVD) of the feature matrix, $\mathbf{X} = \mathbf{U}_\mathbf{X} \mathbf{\Sigma} \mathbf{V}_\mathbf{X}^\intercal$, where $\mathbf{U}_\mathbf{X} \in \mathbb{R}^{n\times d_0}$ is a semi-orthogonal matrix, $\mathbf{\Sigma} \in \mathbb{R}^{d_0 \times d_0}$ is a diagonal matrix with $\mathbf{\Sigma}_{11} \geq \mathbf{\Sigma}_{22} \geq \ldots \geq \mathbf{\Sigma}_{d_0 d_0} > 0$, and $\mathbf{V}_\mathbf{X} \in \mathbb{R}^{d_0 \times d_0}$ is an orthogonal matrix. At times, we will consider isotropic features, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma}_{11} \mathbf{I}_{d_0}$, to draw insights from the bounds on the eigenvalues of the NTK.

1. The parameterization follows the implementation of a linear layer in the PyTorch framework [27].

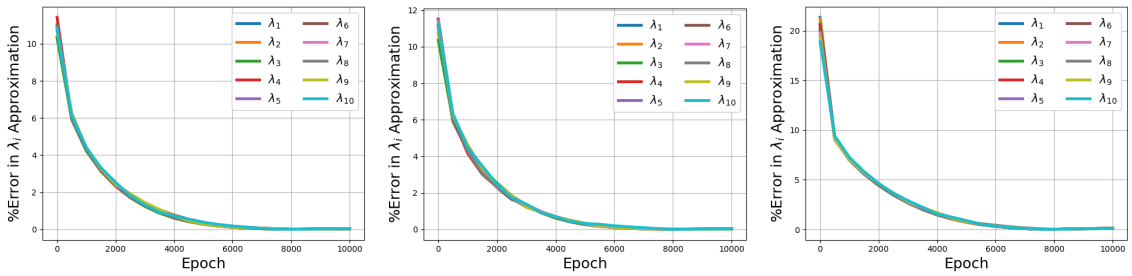


**Figure 1:** Percentage error in approximating top-10 eigenvalues of the loss-Hessian (using Theorem 1) for a 2-layer MLP on (a) MNIST, (b) FashionMNIST, and (c) CIFAR10.

### 2.2. Shallow Networks

**Theorem 1** *Under Assumption 1, the non-zero eigenvalues of shallow linear networks ($L = 2$) are characterized by the pairwise-sums of squared singular values of the two layers:*

$$\lambda_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^{\intercal}\right) \leq \frac{\|\mathbf{X}\|_2^2}{n}\left(\sigma_i^2\left(\mathbf{W}_1\right) + \sigma_j^2\left(\mathbf{W}_2\right)\right) \tag{6}$$

*where $i \in [\min\{d_0, d_1\}]$ and $j \in [\min\{d_1, d_2\}]$, and $\sigma_i(\mathbf{W})$ denotes the $i^{th}$ largest singular value of $\mathbf{W}$. Furthermore, the bound is attained if the features are isotropic, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$.*

We visualize the case of isotropic features in Figure 1, where we note that the approximation becomes increasingly precise as training progresses. Theorem 1 implies that, of the $d_0 d_1 + d_1 d_2$ eigenvalues of the Hessian, at most $\min\{d_0, d_1\} \times \min\{d_1, d_2\}$ may be non-zero; we comment on this at the end of Section 3.

Consider the case where $\mathbf{\Sigma} = \sqrt{n}\mathbf{I}_{d_0}$. The spectral norm (largest eigenvalue for symmetric matrices) of the NTK is given by $\lambda_{\max}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^{\intercal}\right) = \|\mathbf{W}_1\|_2^2 + \|\mathbf{W}_2\|_2^2$. This contradicts the conclusion drawn in Singh et al. [32, Appendix D.3.3], suggesting that the maximum eigenvalue is given by $\max\left\{\|\mathbf{W}_1\|_2^2, \|\mathbf{W}_2\|_2^2\right\}$; we believe that the discrepancy is caused by the block-diagonal structure of the Hessian assumed in their analysis.

As for the sum of eigenvalues, $\mathrm{Tr}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^{\intercal}\right) = d_2\|\mathbf{W}_1\|_F^2 + d_0\|\mathbf{W}_2\|_F^2$. This relation provides insight into the progressive flattening and subsequent sharpening observed in early- and intermediate-training [16, 17], respectively, before training reaches EoS [3, 4, 13] – sharpness decreases since the parameter norm decreases in early-training, as the model output evolves to align with the labels, and both start increasing after a point [30, Theorem 2]. In Figure 2, we demonstrate this relation for a shallow network as well as for a deep network, showing a high amount of agreement in the dynamics of $\mathrm{Tr}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^{\intercal}\right)$ and the squared parameter-norm.

## 3. Strongly Balanced Deep Networks

In this section, we derive analogous results to Theorem 1 but for networks of arbitrary depth, while relying on an additional assumption on the balance between layers.

**Assumption 2** *Assume that consecutive layers are strongly balanced, i.e. $\mathbf{W}_\ell\mathbf{W}_\ell^{\intercal} = \mathbf{W}_{\ell+1}^{\intercal}\mathbf{W}_{\ell+1}$, $\forall \ell \in [L-1]$.*

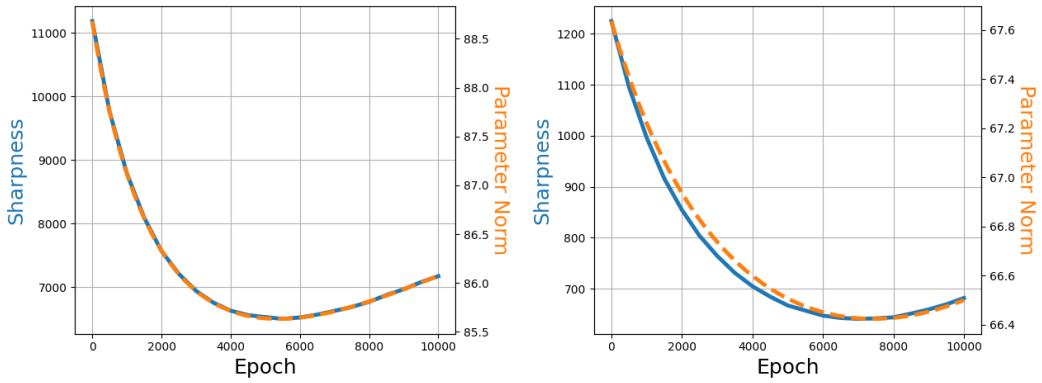


**Figure 2:** Sum of eigenvalues of the loss-Hessian, computed numerically (Hutchinson's method [12]), and squared parameter-norm of (a) a 2-layer MLP and (b) a 4-layer MLP trained on CIFAR10.

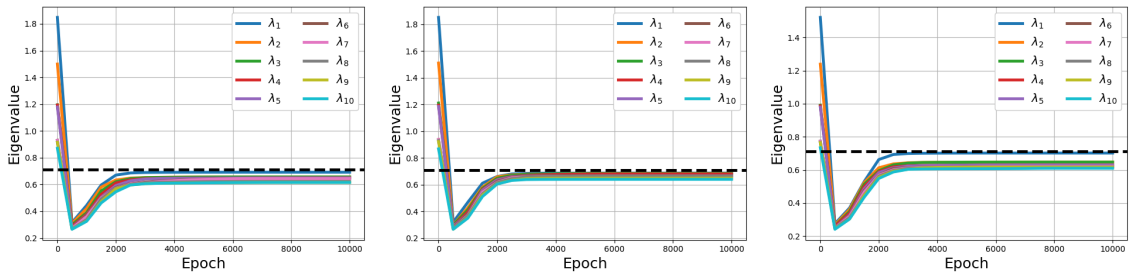


**Figure 3:** Top-10 eigenvalues of the loss-Hessian and the predicted values (in black, using Theorem 2) for a 4-layer MLP on (a) MNIST, (b) FashionMNIST, and (c) CIFAR10.

In essence, this implies that the left singular vectors of a layer are aligned with the right singular vectors of the next layer. Moreover, all layers share their non-zero singular values; we denote them by $\sigma_i(\mathbf{W})$. While arbitrary points in the parameter space need not correspond to balanced layers, initially balanced layers remain balanced under gradient flow training, provided their is no nonlinearity between them [1, 5]. Even if initially unbalanced, Singh et al. [30, Theorem 1] showed that the layers become increasingly balanced under weight-decay training, while Ghosh et al. [8, Proposition 2] showed that gradient descent training at EoS [3] balances the singular values of the weight matrices.Therefore, the results hereon can be seen as characterizing the local geometry of the minimizers.

**Theorem 2** *Under Assumptions 1 and 2, the non-zero eigenvalues of the NTK are given by*

$$\lambda_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^{\intercal}\right) \leq \frac{\|\mathbf{X}\|_2^2}{n} \cdot \begin{cases} \frac{\sigma_i^{2L}(\mathbf{W})-\sigma_j^{2L}(\mathbf{W})}{\sigma_i^{2}(\mathbf{W})-\sigma_j^{2}(\mathbf{W})}\ , & \sigma_i(\mathbf{W}) \neq \sigma_j(\mathbf{W}) \\ L \cdot \sigma^{2(L-1)}(\mathbf{W})\ , & \sigma(\mathbf{W}) := \sigma_i(\mathbf{W}) = \sigma_j(\mathbf{W}) \end{cases} \tag{7a}$$

*with equality if* $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$.

Hence, of the $p := \sum_{\ell=1}^{L} d_\ell d_{\ell-1}$ eigenvalues of the Hessian, at most $(\min\{d_0, d_1, \ldots, d_L\})^2$ are non-zero. This corresponds to the empirical observations suggesting that the bulk of eigenvalues of the Hessian are near-zero [26, 28, 29]. We visualize the case of isotropic features in Figure 3, where we note that the approximation becomes increasingly precise as training progresses.

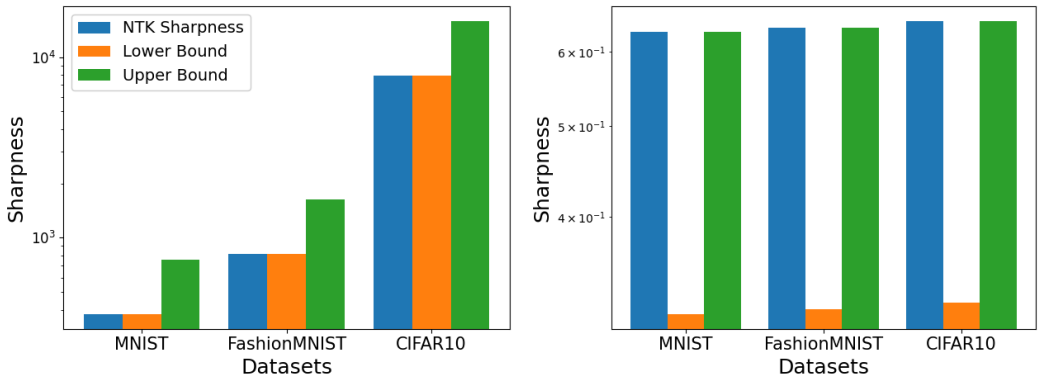


**Figure 4:** Sharpness of the NTK of the ordinary-least squares (OLS) solution, using (a) original datasets and (b) whitened features, along with corresponding lower and upper bounds.

**Corollary 1** *Under Assumptions 1 and 2, the spectral norm of the NTK can be bounded as*

$$\frac{1}{n}\left\|\mathbf{X}\right\|_2^2\left\|\mathbf{W}\right\|_2^{2(L-1)} \leq \lambda_{\max}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{L}{n}\left\|\mathbf{X}\right\|_2^2\left\|\mathbf{W}\right\|_2^{2(L-1)} \tag{8}$$

We distinguish two cases: 1. when the features are highly anisotropic (typical case for real-world datasets), and 2. when the features are isotropic (more idealistic). We observe that in the former case, the lower bound is tight, while in the latter case the upper bound is tight; see Figure 4. This suggests that the true sharpness floats between the two bounds as the degree of isotropy in the features changes.

## 4. Dependence on Data

**Theorem 3** *Say the network implements the ordinary least-squares (OLS) solution, i.e.* $\mathbf{W}_1^\intercal\mathbf{W}_2^\intercal\ldots\mathbf{W}_L^\intercal = \mathbf{X}^\dagger\mathbf{Y}$*, where* $\cdot^\dagger$ *denotes the Moore-Penrose inverse. If* $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$*, then under Assumption 2,*

$$\sigma_i\left(\mathbf{W}\right) = \mathbf{\Sigma}_{11}^{-1/L}\cdot\sigma_i^{1/L}\left(\mathbf{U}_\mathbf{X}^\intercal\mathbf{Y}\right) \tag{9}$$

*In particular, the spectral norm is given by*

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) = \frac{L}{n}\mathbf{\Sigma}_{11}^{2/L}\cdot\left\|\mathbf{U}_\mathbf{X}^\intercal\mathbf{Y}\right\|_2^{2(1-1/L)} \leq L\left(\frac{\left\|\mathbf{X}\right\|_2^2}{n}\right)^{1/L}\left(\frac{\left\|\mathbf{Y}\right\|_2^2}{n}\right)^{1-1/L} \tag{10}$$

*This bound is attained when* $\mathcal{L}\left(\mathbf{W}_{1:L};\mathbf{X},\mathbf{Y}\right) = 0$.

We visualize this correspondence in the left panel in Figure 5, where we note that the theory predicts the non-zero eigenvalues exactly. For a centered feature matrix $\mathbf{X}$, i.e. the mean of each feature is zero, $\{\sigma_i^2\left(\mathbf{X}\right)/n\}_{i=1}^{d_0}$ represents the amount of variance explained by the principle components. For one-hot encoded labels, $\{\sigma_i^2\left(\mathbf{Y}\right)/n\}_{i=1}^{d_L}$ is the empirical distribution of labels, i.e. proportion of labels from each class in the training set; in what follows, we will use $\alpha_i := \sigma_i^2\left(\mathbf{Y}\right)/n$ to denote these proportions. Therefore, the sharpness of balanced solutions is independent of dataset size $n$ (see Figure 7), in disagreement with Cohen et al. [3, Figure 18], which suggests that sharpness of

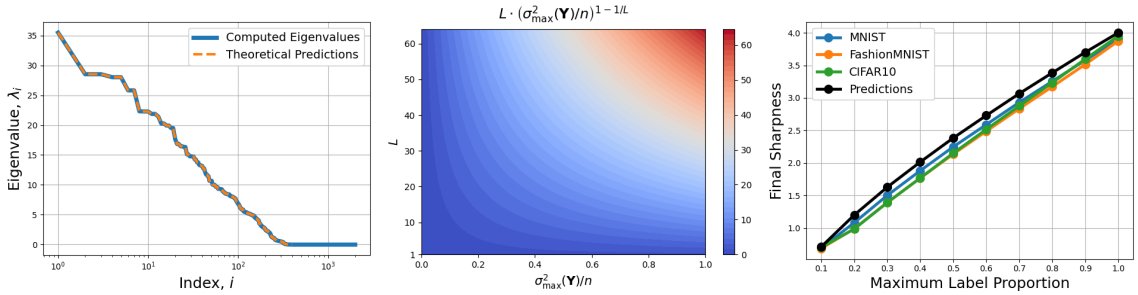


**Figure 5:** Computing eigenvalues of the Hessian of strongly balanced solutions using the labels, $\mathbf{Y}$. (a) Computed numerically and theoretically (Theorems 2 and 3) with random data. (b) Contour plot on the effect of depth and spectral norm. (c) Spectral norm of trained models and theoretical predictions (Theorem 3).

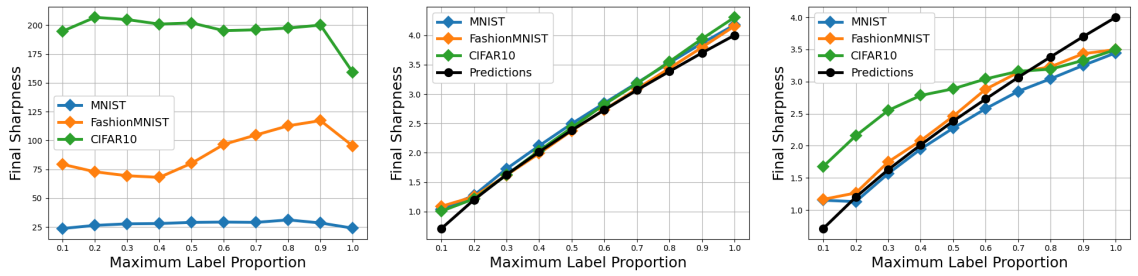


**Figure 6:** Ablating model assumptions: (a) raw inputs (shedding Assumption 1), (b) Kaiming uniform initialization (shedding Assumption 2), and (c) unbalanced initialization and Tanh (nonlinear) activation.

solutions increases with $n$. Sharpness increases with depth mainly due to the leading factor of $L$ in eigenvalue computation – we visualize this in the middle panel in Figure 5.

The discussion above for classification tasks implies that the model learns sharper solutions when there is one class with disproportionately large number of labels; we illustrate this in the right panel in Figure 5. Such datasets are, intuitively, simpler to learn, e.g. as an extreme case, a dataset with *all inputs belonging to one class* should be easier to learn than a dataset with uniform label distribution. This intuition is in disagreement with Cohen et al. [3, Caveat 2], which suggests that the model achieves lower peak-sharpness on simpler datasets.

**Ablations.** We repeat the experiment in Figure 5, shedding each of Assumptions 1 and 2 one-at-a-time. In the left panel in Figure 6, we present the results shedding the assumption on whitened inputs. In this case, our theory breaks down immediately – the sharpness of the learnt solution turns out to have inconsistent variations with maximum label proportion in the dataset, suggesting that peak-sharpness is strongly influenced by the input-geometry. In the middle panel, we present the results shedding the assumption on balanced weights, and use the standard Kaiming uniform initialization [11] implemented in the PyTorch framework. Here, we note the theoretical predictions made in previous section stay intact, as the final sharpness increasing with label imbalance. Finally, we inspect if incorporating nonlinearity breaks our theory. In the right panel, we present results for a Tanh-MLP, noting that while the results are slightly different from the theoretical predictions, they are qualitatively similar, allowing us to extend our conclusions on the effects of label distribution to such nonlinear networks.

## References

[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 244–253. PMLR, 10–15 Jul 2018.

[2] 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. In *International Conference on Learning Representations*, 2017.

[3] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[4] Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability, 2024.

[5] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[7] Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, pages 4475–4488. Curran Associates, Inc., 2022.

[8] Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *International Conference on Learning Representations*, volume 2025, pages 17753–17800, 2025.

[9] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 10–15 Jul 2018.

[10] B. Hassibi, D.G. Stork, and G.J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299 vol.1, 1993.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[12] M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communication in Statistics- Simulation and Computation*, 18:1059–1076, 01 1989.

[13] Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert Gower. Non-euclidean gradient descent operates at the edge of stability. In *Proceedings of the 43rd International Conference on Machine Learning*, ICML, 2026.

[14] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd, 2018.

[16] Dayal Singh Kalra and Maissam Barkeshli. Phase diagram of early training dynamics in deep neural networks: effect of the learning rate, depth, and width. In *Advances in Neural Information Processing Systems*, volume 36, pages 51621–51662. Curran Associates, Inc., 2023.

[17] Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. In *International Conference on Learning Representations*, volume 2025, pages 55966–56000, 2025.

[18] Andrew V. Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2): 517–541, 2001.

[19] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[20] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 18–24 Jul 2021.

[21] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989.

[22] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*, 2, 2010.

[23] James Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, ICML, page 735–742, Madison, WI, USA, 2010. Omnipress.

[24] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 7308–7320. Curran Associates, Inc., 2020.

[25] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*, 2021.

[26] Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size, 2019.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] Levent Sagun, Léon Bottou, and Yann LeCun. Singularity of the hessian in deep learning. *CoRR*, abs/1611.07476, 2016.

[29] Levent Sagun, Utku Evci, V. Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks, 2018.

[30] Jasraj Singh, Enea Monzio Compagnoni, and Antonio Orvieto. Unified perspectives on balancedness and parameter-norm evolution in neural nets. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026.

[31] Sidak Pal Singh and Thomas Hofmann. Closed form of the hessian spectrum for some neural networks. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

[32] Sidak Pal Singh, Weronika Ormaniec, and Thomas Hofmann. Cracking the hessian: Closed-form hessian spectra for fundamental neural networks, 2026.

[33] Andreas Stathopoulos and Kesheng Wu. A block orthogonalization procedure with constant synchronization requirements. *SIAM Journal on Scientific Computing*, 23(6):2165–2182, 2002.

[34] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

## Appendix A. Omitted Proofs

**Proof** [Theorem 1]

For $L = 2$, following Equation 5,

$$\mathbf{S}_1^\intercal \mathbf{S}_1 = \mathbf{W}_2 \mathbf{W}_2^\intercal \implies \lambda_i \left(\mathbf{S}_1^\intercal \mathbf{S}_1\right) = \sigma_i^2 \left(\mathbf{W}_2\right) \tag{11a}$$

$$\mathbf{X}\mathbf{P}_1\mathbf{P}_1^\intercal \mathbf{X}^\intercal = \mathbf{X}\mathbf{X}^\intercal \implies \lambda_i \left(\mathbf{X}\mathbf{P}_1\mathbf{P}_1^\intercal \mathbf{X}^\intercal\right) = \sigma_i^2 \left(\mathbf{X}\right) \leq \mathbf{\Sigma}_{11}^2 \tag{11b}$$

$$\mathbf{S}_2^\intercal \mathbf{S}_2 = \mathbf{I}_{d_2} \implies \lambda_i \left(\mathbf{S}_2^\intercal \mathbf{S}_2\right) = 1 \tag{11c}$$

$$\mathbf{X}\mathbf{P}_2\mathbf{P}_2^\intercal \mathbf{X}^\intercal = \mathbf{X}\mathbf{W}_1^\intercal \mathbf{W}_1 \mathbf{X}^\intercal \implies \lambda_i \left(\mathbf{X}\mathbf{P}_2\mathbf{P}_2^\intercal \mathbf{X}^\intercal\right) \leq \mathbf{\Sigma}_{11}^2 \sigma_i^2 \left(\mathbf{W}_1\right) \tag{11d}$$

where both upper bounds are tight if $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$. Therefore,

$$\lambda_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{\|\mathbf{X}\|_2^2}{n} \cdot \left(\sigma_i^2\left(\mathbf{W}_1\right) + \sigma_j^2\left(\mathbf{W}_2\right)\right) \tag{12}$$

Furthermore, if $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$, the two summands share their (unnormalized) eigenvectors:

$$\mathbf{q}_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) = \mathbf{u}_i \otimes \left(\mathbf{X}\mathbf{v}_j\right) \tag{13}$$

where $\mathbf{u}_i$ is the $i^{\text{th}}$ left singular vector of $\mathbf{W}_2$, and $\mathbf{v}_j$ is the $j^{\text{th}}$ right singular vector of $\mathbf{W}_1$. Therefore, *all* non-zero eigenvalues of the NTK are given by pairwise-sums of squared singular values of the two layers:

$$\lambda_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) = \frac{\|\mathbf{X}\|_2^2}{n}\left(\sigma_i^2\left(\mathbf{W}_1\right) + \sigma_j^2\left(\mathbf{W}_2\right)\right) \tag{14}$$

where $i \in [\min\{d_0, d_1\}]$ and $j \in [\min\{d_1, d_2\}]$.

■

**Proof** [Theorem 2]

Under Assumption 2, we have $\mathbf{S}_\ell^\intercal \mathbf{S}_\ell = \left(\mathbf{W}_L\mathbf{W}_L^\intercal\right)^{L-\ell}$ and $\mathbf{P}_\ell\mathbf{P}_\ell^\intercal = \left(\mathbf{W}_1^\intercal\mathbf{W}_1\right)^{\ell-1}$. Therefore, following Equation 5,

$$\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal = \frac{1}{n}\sum_{\ell=1}^{L}\left(\mathbf{W}_L\mathbf{W}_L^\intercal\right)^{L-\ell} \otimes \left(\mathbf{X}\left(\mathbf{W}_1^\intercal\mathbf{W}_1\right)^{\ell-1}\mathbf{X}^\intercal\right) \tag{15}$$

Using triangle inequality, the eigenvalues of the NTK are upper bounded as

$$\lambda_{ij}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{1}{n}\sum_{\ell=1}^{L}\sigma_i^{2(L-\ell)}\left(\mathbf{W}\right)\lambda_j\left(\mathbf{X}\left(\mathbf{W}_1^\intercal\mathbf{W}_1\right)^{\ell-1}\mathbf{X}^\intercal\right) \tag{16a}$$

$$\leq \frac{\|\mathbf{X}\|_2^2}{n} \cdot \sum_{\ell=1}^{L}\sigma_j^{2(L-\ell)}\left(\mathbf{W}\right)\sigma_i^{2(\ell-1)}\left(\mathbf{W}\right) \tag{16b}$$

$$= \frac{\|\mathbf{X}\|_2^2}{n} \cdot \begin{cases} \frac{\sigma_i^{2L}(\mathbf{W})-\sigma_j^{2L}(\mathbf{W})}{\sigma_i^2(\mathbf{W})-\sigma_j^2(\mathbf{W})}\ , & \sigma_i\left(\mathbf{W}\right) \neq \sigma_j\left(\mathbf{W}\right) \\ L \cdot \sigma^{2(L-1)}\left(\mathbf{W}\right)\ , & \sigma\left(\mathbf{W}\right) \coloneqq \sigma_i\left(\mathbf{W}\right) = \sigma_j\left(\mathbf{W}\right) \end{cases} \tag{16c}$$

If $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$, then we further have that the (unnormalized) eigenvectors of each summand are given by the Kronecker products $\mathbf{q}_{ij} = \mathbf{u}_i \otimes (\mathbf{X}\mathbf{v}_j)$, where $\mathbf{u}_i$ is the $i^{\text{th}}$ left singular vector of $\mathbf{W}_L$, and $\mathbf{v}_j$ is the $j^{\text{th}}$ right singular vector of $\mathbf{W}_1$. Hence, the eigenvalues add up over the summands, and the upper bound derived above becomes tight.

■

**Proof** [Theorem 3]

The OLS solution to the problem is given by

$$\mathbf{W}_1^\intercal \mathbf{W}_2^\intercal \dots \mathbf{W}_L^\intercal = \mathbf{X}^\dagger \mathbf{Y} = \mathbf{V}_\mathbf{X} \mathbf{\Sigma}^{-1} \mathbf{U}_\mathbf{X}^\intercal \mathbf{Y} \tag{17}$$

Under Assumption 2, the shared singular values simplify to

$$\sigma_i^L(\mathbf{W}) = \sigma_i\left(\mathbf{V}_\mathbf{X} \mathbf{\Sigma}^{-1} \mathbf{U}_\mathbf{X}^\intercal \mathbf{Y}\right) = \sigma_i\left(\mathbf{\Sigma}^{-1} \mathbf{U}_\mathbf{X}^\intercal \mathbf{Y}\right) \tag{18}$$

Assuming isotropic features,

$$\sigma_i^L(\mathbf{W}) = \mathbf{\Sigma}_{11}^{-1} \cdot \sigma_i\left(\mathbf{U}_\mathbf{X}^\intercal \mathbf{Y}\right) \implies \sigma_i(\mathbf{W}) = \mathbf{\Sigma}_{11}^{-1/L} \cdot \sigma_i^{1/L}\left(\mathbf{U}_\mathbf{X}^\intercal \mathbf{Y}\right) \tag{19}$$

In particular, if $\mathbf{X}\mathbf{W}_1^\intercal \mathbf{W}_2^\intercal \dots \mathbf{W}_L^\intercal = \mathbf{Y}$, then

$$\mathbf{X}\left(\mathbf{W}_1^\intercal \mathbf{W}_1\right)^L \mathbf{X}^\intercal = \mathbf{Y}\mathbf{Y}^\intercal \implies \mathbf{\Sigma}_{11}^2 \sigma_i^{2L}(\mathbf{W}) = \sigma_i^2(\mathbf{Y}) \tag{20a}$$

$$\implies \sigma_i(\mathbf{W}) = \mathbf{\Sigma}_{11}^{-1/L} \cdot \sigma_i^{1/L}(\mathbf{Y}) \tag{20b}$$

All that is left is to substitute the shared singular values in Theorem 2.

■

## Appendix B. Other Results

### B.1. Summary of Eigenvalues

Several works have used Hessian-based measures like the maximum eigenvalue (spectral norm for symmetric matrices) or the sum of eigenvalues (trace) to measure generalization. Intuitively, at a local minimum, the spectral norm measures the worst-case change in loss caused by a small perturbation in any direction, while the trace measures the expected change in loss caused by a random isotropic perturbation with small variance.

**Lemma 1** *Under Assumption 1, the spectral norm and trace of the NTK are given by*

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{\|\mathbf{X}\|_2^2}{n} \cdot \sum_{\ell=1}^{L} \|\mathbf{S}_\ell\|_2^2 \|\mathbf{P}_\ell\|_2^2 \tag{21a}$$

$$Tr\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{\|\mathbf{X}\|_2^2}{n} \cdot \sum_{\ell=1}^{L} \|\mathbf{S}_\ell\|_F^2 \|\mathbf{P}_\ell\|_F^2 \tag{21b}$$

*with equality holding for trace if* $\mathbf{\Sigma} = \mathbf{\Sigma}_{11}\mathbf{I}_{d_0}$.

**Proof**

Following Equation 5,

$$\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal = \frac{1}{n}\sum_{\ell=1}^{L}\left(\mathbf{S}_\ell^\intercal\mathbf{S}_\ell\right)\otimes\left(\mathbf{X}\mathbf{P}_\ell\mathbf{P}_\ell^\intercal\mathbf{X}^\intercal\right) \tag{22}$$

Using triangle inequality,

$$\lambda_{\max}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) \leq \frac{1}{n}\sum_{\ell=1}^{L}\lambda_{\max}\left(\left(\mathbf{S}_\ell^\intercal\mathbf{S}_\ell\right)\otimes\left(\mathbf{X}\mathbf{P}_\ell\mathbf{P}_\ell^\intercal\mathbf{X}^\intercal\right)\right) \tag{23a}$$

$$= \frac{1}{n}\sum_{\ell=1}^{L}\|\mathbf{S}_\ell\|_2^2\,\lambda_{\max}\left(\mathbf{X}\mathbf{P}_\ell\mathbf{P}_\ell^\intercal\mathbf{X}^\intercal\right) \tag{23b}$$

$$= \frac{\|\mathbf{X}\|_2^2}{n}\cdot\sum_{\ell=1}^{L}\|\mathbf{S}_\ell\|_2^2\,\|\mathbf{P}_\ell\|_2^2 \tag{23c}$$

As for the sum of eigenvalues,

$$\mathrm{Tr}\left(\frac{1}{n}\mathbf{J}\mathbf{J}^\intercal\right) = \frac{1}{n}\sum_{\ell=1}^{L}\mathrm{Tr}\left(\left(\mathbf{S}_\ell^\intercal\mathbf{S}_\ell\right)\otimes\left(\mathbf{X}\mathbf{P}_\ell\mathbf{P}_\ell^\intercal\mathbf{X}^\intercal\right)\right) \tag{24a}$$

$$= \frac{1}{n}\sum_{\ell=1}^{L}\|\mathbf{S}_\ell\|_F^2\,\mathrm{Tr}\left(\mathbf{X}\mathbf{P}_\ell\mathbf{P}_\ell^\intercal\mathbf{X}^\intercal\right) \tag{24b}$$

$$\leq \frac{\|\mathbf{X}\|_2^2}{n}\cdot\sum_{\ell=1}^{L}\|\mathbf{S}_\ell\|_F^2\,\|\mathbf{P}_\ell\|_F^2 \tag{24c}$$

■

## B.2. Figures

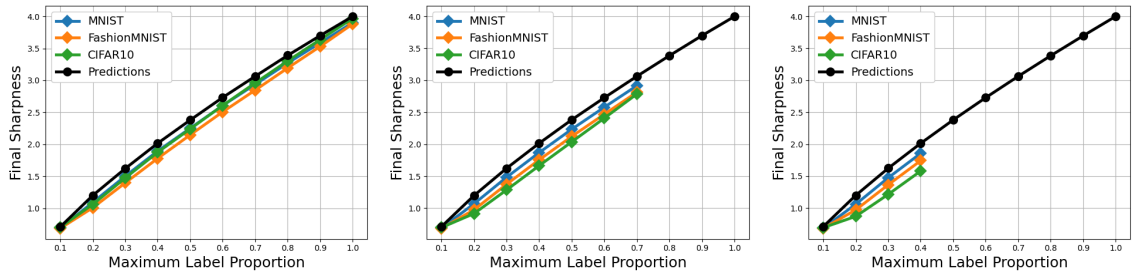


**Figure 7:** Effect of dataset size on spectral norm of trained networks: (a) $n = 4$k, (b) 7k and (c) 10k. Theoretical predictions correspond to Theorems 2 and 3. Some high $\alpha_{\max}$ are not presented for $n = 7$k and 10k since they requires more samples with label 0 than available – 5k in the datasets considered.

## Appendix C. Experimental Details

**Data Whitening.** Given the SVD of a feature matrix $\mathbb{R}^{n\times d_0} \ni \mathbf{X}' = \mathbf{U}_\mathbf{X}\mathbf{\Sigma}\mathbf{V}_\mathbf{X}^\intercal$, we redefine the features to be proportional to the polar factor, i.e. $\mathbf{X} = \sqrt{n}\sigma_\mathbf{x}\mathbf{U}_\mathbf{X}\mathbf{V}_\mathbf{X}^\intercal$.

**Sampling Balanced Weights.** Assume we are given a sequence of $k = \min\{d_0, d_1, \ldots, d_L\}$ singular values; we arrange them to create a diagonal matrix, $\mathbf{S} \in \mathbb{R}^{k\times k}$. To construct a sequence of balanced weight matrices, we start by sampling the right singular vectors of $\mathbf{W}_1$ by orthogonalizing a random $d_0 \times k$ matrix with $\mathcal{N}(0,1)$ entries; denote it by $\mathbf{V}_1 \in \mathbb{R}^{d_0\times k}$. Then, for each $\ell \in [L]$, do

1. Sample the left singular vectors of $\mathbf{W}_\ell$ by orthogonalizing a random $d_\ell \times k$ matrix with $\mathcal{N}(0,1)$ entries; denote it by $\mathbf{U}_\ell \in \mathbb{R}^{d_1\times k}$.
2. Define the weight matrix as $\mathbf{W}_\ell \coloneqq \mathbf{U}_\ell\mathbf{S}\mathbf{V}_\ell^\intercal \in \mathbb{R}^{d_\ell\times d_{\ell-1}}$.
3. Set the right singular vectors of the next layer as the left singular vectors of the previous layers, $\mathbf{V}_{\ell+1} \coloneqq \mathbf{U}_\ell$.

**Subsampling.** With real-world datasets MNIST [22], FashionMNIST [34] and CIFAR10 [19], we use subsets of $n = 5,000$ samples. If the maximum label proportion $\alpha_{\max}$ is set, we take $\lfloor(1-\alpha_{\max})n\rfloor$ samples from each of the classes $1-9$, where $\lfloor\cdot\rfloor$ denotes the floor function, and the rest of the samples from class $0$.

**Training.** All models are trained using full-batch gradient descent for $10,000$ epochs, with an initial learning rate $\eta_{\max}$ which is successively halved anytime $< 0.1\%$ improvement in loss is observed over $100$ epochs.

**Figure 1:** Top-10 eigenvalues of the Hessian, computed numerically using LOBPCG and analytically using Theorem 1. Datasets are orthogonalized with $\sigma_\mathbf{x} = 1$, and our model is a 2-layer Linear-MLP with width 256, trained with $\eta_{\max} = 1e-3$.

**Figure 2:** Sum of eigenvalues of the Hessian is computed numerically using Hutchinson's method [12]. Datasets are orthogonalized with $\sigma_\mathbf{x} = 1$, and our models are a 2-layer Linear-MLP with width 256 and a 4-layer Linear-MLP with width 64, trained with $\eta_{\max} = 5e-4$.

**Figure 3:** Top-10 eigenvalues of the Hessian, computed numerically using LOBPCG and analytically using Theorem 2. Datasets are orthogonalized with $\sigma_\mathbf{x} = 1$, and our model is a 4-layer Linear-MLP with width 64, trained with $\eta_{\max} = 1e-2$.

**Figure 4:** The OLS solution is computed as $\mathbf{W}^\intercal \coloneqq \mathbf{W}_1^\intercal\mathbf{W}_2^\intercal\ldots\mathbf{W}_L^\intercal = \mathbf{X}^\dagger\mathbf{Y}$, and a network with balanced layers is constructed to simulate this solution by setting $\mathbf{V}_1$ and $\mathbf{U}_L$ as the right and left singular vectors of $\mathbf{W}$. Our model is a 2-layer Linear-MLP with width 128.

**Figure 5:**

- **Left:** We sample $\mathbf{X}' \in \mathbb{R}^{20\times 100}$ with standard normal entries, and then orthogonalize it with $\sigma_\mathbf{x} = 3$. We sample $d = 20$ singular values from $\mathcal{U}[0,1]$, and use it sample a sequence of $L = 5$ balanced weight matrices, with $\mathbf{W}_\ell \in \mathbb{R}^{d\times d}$. Finally, set $\mathbf{Y} = \mathbf{W}_5\mathbf{W}_4\ldots\mathbf{W}_1\mathbf{X}$ so that the GGN approximation is exact.
- **Right:** Spectral norm of the Hessian is computed numerically using LOPPCG [18, 33] and analytically using Theorems 2 and 3. Datasets are orthogonalized with $\sigma_\mathbf{x} = 1$, and our model

is a 4-layer Linear-MLP with width 64, initialized with balanced weights, and trained with $\eta_{\max} = 1e - 2$.

**Figure 6:** Spectral norm of the Hessian is computed numerically using LOBPCG and analytically using Theorems 2 and 3.

- **Left:** Datasets are *not* orthogonalized, and our model is a 4-layer Linear-MLP with width 64, initialized with balanced weights.
- **Middle:** Datasets are orthogonalized with $\sigma_{\mathbf{x}} = 1$, and our model is a 4-layer Linear-MLP with width 64, initialized with Kaiming uniform initialization [11] (standard in PyTorch).
- **Right:** Datasets are orthogonalized with $\sigma_{\mathbf{x}} = 1$, and our model is a 4-layer Tanh-MLP with width 64, initialized with balanced weights.

All models are trained with $\eta_{\max} = 1e - 2$.

**Figure 7:** Spectral norm of the Hessian, computed numerically using LOPPCG and analytically using Theorems 2 and 3. Datasets are orthogonalized with $\sigma_{\mathbf{x}} = 1$, and our model is a 4-layer Linear-MLP with width 64, trained with $\eta_{\max} = 1e - 2$.