A simple connection from loss flatness to compressed neural representations

Shirui Chen SC256@UW.EDU
University of Washington, Seattle

Stefano Recanatesi STEFANO.RECANATESI@GMAIL.COM
Allen Institute

Eric Shea-Brown ETSB@UW.EDU
University of Washington, Seattle

Abstract

Sharpness, a geometric measure in the parameter space that reflects the flatness of the loss landscape, has long been studied for its potential connections to neural network behavior. While sharpness is often associated with generalization, recent work highlights inconsistencies in this relationship, leaving its true significance unclear. In this paper, we build on the original approach of Ma and Ying to investigate how sharpness influences the local geometric features of neural representations in feature space, offering a new perspective on its role. We introduce this problem and study the Maximum Local Sensitivity (MLS) and Network MLS (NMLS), measuring robustness of the network output with respect to internal neural representations. We show that MLS/NMLS correlate with the flatness of the loss around the local minima, and that this correlation is predicted by a relatively simple mathematical relationship: a flatter loss corresponds to a lower upper bound on the compression metrics of neural representations. Our work builds upon the linear stability trick by Ma and Ying, deriving inequalities between various compression metrics and quantities involving sharpness. Our inequalities readily extend to reparametrization-invariant sharpness as well. Through empirical experiments on various feedforward, convolutional, and transformer architectures, we find that our inequalities predict a consistently positive correlation between local representation compression and sharpness.

1. Introduction

There has been a long-lasting interest in sharpness, a geometric metric in the *parameter* space that measures the flatness of the loss landscape at local minima. Flat minima refer to regions in the loss landscape where the loss function has a relatively large basin, and the loss does not change much in different directions around the minimum. Empirical studies and theoretical analyses have shown that training deep neural networks using stochastic gradient descent (SGD) with a small batch size and a high learning rate often converges to flat and wide minima [5, 13, 16, 23, 25, 41, 43, 46]. Many works conjecture that flat minima lead to a simpler model (shorter description length), and thus are less likely to overfit and more likely to generalize well [15, 19, 41, 44]. Based on this rationale, sharpness-aware minimization (SAM) has been a popular method for improving a model's generalization ability. However, recent work has shown that SAM does not *only* minimize sharpness to achieve superior generalization performance [1, 36]. More confusingly, it remains unclear whether flatness correlates positively with the generalization capacity of the network [2, 8, 45], and even when it does, the correlation is not perfect [18, 28]. In particular, Dinh et al. [8] argues that one can construct very sharp networks that generalize well through reparametrization; while [2] shows that even reparametrization-invariant sharpness cannot capture the relationship between sharpness and generalization.

As an alternative to this contentious relationship between sharpness and generalization, we show that there exists a different, more consistent perspective by investigating how sharpness near interpolation solutions in the *parameter* space influences local geometric features of neural representations in the *feature* space. By building a relationship between sharpness and the local compression of neural representations, we argue that sharpness, in its essence, measures the compression of neural representations. Specifically, we show that as sharpness decreases and the minimum flattens, certain compression measures set a lower bound on sharpness-related quantities, meaning that the neural representation must also undergo some degree of compression.

More specifically, our work makes the following novel contributions:

1. We identify two feature space quantities that quantify compression and are bounded by sharpness – maximum local sensitivity (MLS) and Network MLS (NMLS) – and give new explicit formulas for these bounds.
2. We improve the bound on MLS in [25] and propose Network MLS (NMLS), ensuring that the bound consistently predicts a positive correlation between both sides of the inequality in various experimental settings.
3. We conducted empirical experiments with VGG-11, LeNet, MLP, and ViT networks and found that MLS/NMLS are indeed strongly correlated with their sharpness-related bound.

Our paper proceeds as follows. First, we review arguments of Ma and Ying [25] that flatter minima can constrain the gradient of network output with respect to network input and extend the formulation to the multidimensional input case (Section 2). Next, we prove that lower sharpness implies a lower upper bound on metrics of the compression of the representation manifold in feature space: the maximum local sensitivity (MLS) and network MLS (NMLS) (Section 2.1). Finally, we empirically verify our theory by calculating various compression metrics, their theoretical bounds, and sharpness for models during training as well as pretrained ones (Section 3).

2. Background and setup

Consider a feedforward neural network f with input data $\mathbf{x} \in \mathbb{R}^M$ and parameters $\boldsymbol{\theta}$. The output of the network is:

$$\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta}) \, , \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ $(N < M)$. We consider a quadratic loss $L(\mathbf{y}, \mathbf{y}_{\text{true}}) = \frac{1}{2}||\mathbf{y} - \mathbf{y}_{\text{true}}||^2$, a function of the outputs and ground truth $\mathbf{y}_{\text{true}}$. In the following, we will simply write $L(\mathbf{y})$, $L(f(\mathbf{x}, \boldsymbol{\theta}))$ or simply $L(\boldsymbol{\theta})$ to highlight the dependence of the loss on the output, the network, or its parameters.

Sharpness Ma and Ying [25] measures how much the loss gradient changes when the network parameters are perturbed, and is defined by the sum of the eigenvalues of the Hessian:

$$S(\boldsymbol{\theta}) = \text{Tr}(H) \, , \tag{2}$$

with $H = \nabla^2 L(\boldsymbol{\theta})$ being the Hessian. The trace of the Hessian, $\text{Tr}\left(\nabla^2 L(\theta)\right)$, is not the only definition of sharpness, but many sharpness minimization methods have been theoretically shown to reduce this quantity in interpolating models. Specifically, assuming the training loss minimizers lie on a smooth manifold [6, 10], methods like Sharpness-Aware Minimization (SAM) [11] when used with batch size 1 and sufficiently small learning rate and perturbation radius [3, 37], or Label Noise SGD with a small enough learning rate [4, 7, 22], tend to favor interpolating solutions with a low Hessian trace. Therefore, we focus our analysis on the trace of the Hessian.

In particular, for small MSE training loss, Ma and Ying [25] show that

$$S(\boldsymbol{\theta}^*) \approx \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^2, \tag{3}$$

where $\|\cdot\|_F$ is the Frobenius norm. We state a proof of this equality, which appears in Ma and Ying [25] and Wen et al. [36], in Appendix C. In practice, Equation (3) is a good enough approximation of sharpness since the training loss drops quickly after the onset of training (see error bounds in Theorem 2).

We note that for the cross-entropy loss function, the Hessian vanishes as the cross-entropy (CE) loss approaches 0 [14, 40]. Therefore, the sharpness of CE loss cannot differentiate between local minima with different traces of the Hessian. As a result, Granziol [14] showed that SGD may find a flatter minimum with lower loss overfitted to the training data, leading to worse generalization performance. However, our result readily extends to logistic loss with label smoothing (ref. Lemma A.13 in Wen et al. [36]).

To see why minimizing the sharpness of the solution leads to more compressed representations, we need to move from the parameter space to the input space. To do so we clear up the proof of Equation (4) in Ma and Ying [25] that relates adversarial robustness to sharpness in the following. The improvements we made are summarized in Appendix E. The proof is given in Appendix A. As a result, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_2^k \leq \frac{\|\mathbf{W}\|_2^k}{\min_i \|\mathbf{x}_i\|_2^k}\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^k. \tag{4}$$

Equation (4) holds for any positive k, and reveals the impact of flatness on the input sensitivity when $k = 2$. Note that Equation (4) corresponds to Equation (4) in Ma and Ying [25] with multivariable output. In the next section, we will improve this bound to relate sharpness to various metrics measuring compression of representations (for a list of improvements, see Appendix E). Moreover, we show that the underlying theory readily extends to networks with residual connections in Appendix B.

2.1. Sharpness bounds over compression metrics

We introduce Maximum Local Sensitivity (MLS) and Network MLS (NMLS) in this section to understand how sharpness affects the robustness of network output against perturbation of input or internal neural representations. *We collectively term MLS, NMLS, and those introduced in Appendix H.2 as "compression metrics", because these quantities measure how compressed/concentrated a set of noise-perturbed input/internal neural representations is after going through the network.*

In the following, we denote the input to the l-th linear layer as $\mathbf{x}_i^l$ for $l = 1, 2, \cdots, L$. In particular, $\mathbf{x}_i^1 = \mathbf{x}_i$ is the input of the entire network. Similarly, $\mathbf{W}_l$ is the weight matrix of l-th linear/convolutional layer, and $\mathbf{W}_1 = \mathbf{W}$ is the first layer weights. Finally, we use f_l to denote the mapping from the input of the l-th layer to the final output.

Definition 2.1 *The **Maximum Local Sensitivity (MLS)** of network f is defined to be* $\mathrm{MLS}_f = \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}_i)\|_2$*, which is the sample mean of the largest singular value of* $\nabla_{\mathbf{x}} f$.

Intuitively, MLS is the largest possible average local change of $f(\mathbf{x})$ when the norm of the perturbation to $\mathbf{x}$ is regularized. MLS is also referred to as adversarial robustness or Lipschitz constant of the model function in Ma and Ying [25]. Given this definition, we can obtain a bound on MLS below.

Proposition 2.2 *The maximum local sensitivity is upper bounded by a sharpness-related quantity:*

$$\mathrm{MLS}_f = \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_2 \leq \|\mathbf{W}\|_2 \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{\|\mathbf{x}_i\|_2^2} S(\boldsymbol{\theta}^*)^{1/2}} \,. \tag{5}$$

The derivation of the above bound is included in Appendix E. We include more analysis of the tightness of this bound in Appendix H and discuss its connection to other works therein.

A straightforward extension of MLS is the Network MLS (NMLS), which we define as the average of MLS w.r.t. input to each linear layer, i.e. all internal neural representations.

Definition 2.3 *The **Network Maximum Local Sensitivity (NMLS)** of network f is defined as the sum of* MLS_{f_l} *for all l, i.e.* $\sum_{l=1}^{L} \mathrm{MLS}_{f_l}$.

Recall that $\mathbf{x}_i^l$ is the input to the l-th linear/convoluational layer for sample $\mathbf{x}_i$ and f_l is the mapping from the input of l-th layer to the final output. Again we have the following inequality:

Proposition 2.4 *The network maximum local sensitivity is upper bounded by a sharpness related quantity:*

$$\mathrm{NMLS} = \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}^l} f^l(\mathbf{x}_i^l, \boldsymbol{\theta}^*)\|_2 \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L} \frac{\|\mathbf{W}_l\|_2^2}{\|\mathbf{x}_i^l\|^2}} \cdot S(\boldsymbol{\theta}^*)^{1/2}. \tag{6}$$

The derivation is in Appendix E. The advantage of NMLS is that instead of only considering the robustness of the final output w.r.t. the input, NMLS considers the robustness of the output w.r.t. all hidden-layer representations. This allows us to derive a bound that not only considers the weights in the first linear layer but also all other linear weights.

2.2. Connection to neural collapse

The neural collapse phenomenon [29, 47] indicates that the within-class variance of the features in the penultimate layer vanishes at the terminal phase of training. To apply our method to study the penultimate-layer features, we can adapt the linear stability trick in Equation (9) to establish a relationship between their robustness and sharpness. More concretely, we can show that

$$\|\nabla_{\mathbf{x}} g(\mathbf{W}\mathbf{x}; \bar{\boldsymbol{\theta}})\|_F \leq \frac{\|\mathbf{W}\|_2}{\sigma_{\min}(\mathbf{W}_L)\|\mathbf{x}\|_2} \|\nabla_{\mathbf{W}} f(\mathbf{W}\mathbf{x}; \bar{\boldsymbol{\theta}})\|_F \,. \tag{7}$$

Here again, $\mathbf{W}$ is the first-layer weights, $\mathbf{W}_L$ is the last-layer linear classifier weights, and $g(x)$ is the penultimate-layer feature. $\sigma_{\min}(\mathbf{W}_L)$ is defined as the square root of the smallest eigvalue of $\mathbf{W}_L^T\mathbf{W}_L$. The proof is given in Appendix A.1.

Let the feature dimension be d and the number of classes be K. Then, $\mathbf{W}_L \in \mathbb{R}^{K\times d}$, and $\sigma_{\min}(\mathbf{W}_L) = 0$ if $d > K$; otherwise, it is the smallest singular value of $\mathbf{W}_L$. It is interesting to observe that an effective bound on the robustness of penultimate-layer features is not obtained unless $d \leq K$, i.e. when the number of classes is larger than the feature dimension. This indicates a less-than-straightforward relationship between neural collapse and adversarial robustness [33]. On the other hand, our theory then broadly applies to cases where the number of classes is much larger than the feature dimension, such as language modeling, retrieval systems, and face recognition applications, where the generalized neural collapse can occur [17].

3. Experiments

All networks are trained with MSE loss except for the pretrained ViTs in Appendix H.4.

3.1. Sharpness and compression metrics during training: verifying the theory

The theoretical results derived above show that when the training loss is low, measures of compression of the network's representation are upper-bounded by a function of the sharpness of the loss function in parameter space. This links sharpness and representation compression: the flatter the loss landscape, the lower the upper bound on the representation's compression metrics.

To empirically verify whether these bounds are sufficiently tight to show a clear relationship between sharpness and representation compression, we trained a VGG-11 network [32] to classify images from the CIFAR-10 dataset [20] and calculated the (approximate) sharpness (Equation (3)), MLS (Definition 2.1) and NMLS (Definition 2.3) during the training phase (Fig 1 and 2).

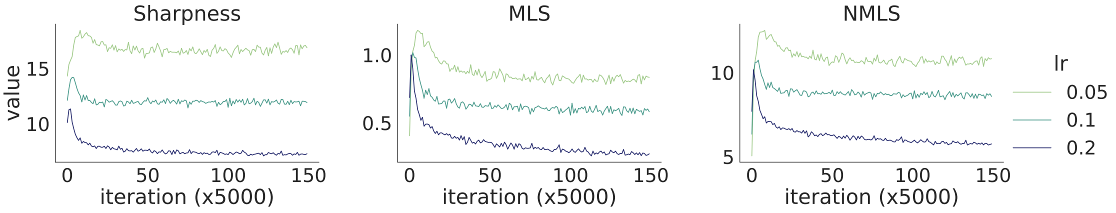

Figure 1: Trends in key variables across SGD training of the VGG-11 network with fixed batch size (equal to 20) and varying learning rates (0.05, 0.1 and 0.2). Higher learning rates lead to lower sharpness, MLS, and NMLS during training. From left to right: sharpness (square root of Equation (3)),MLS, NMLS.

We trained the network using SGD on CIFAR-10 images and explored the influence of two specific parameters that previous work has shown to affect the network's sharpness: learning rate and batch size [16]. For each combination of learning rate and batch size parameters, we computed all quantities across 100 input samples and averaged across five different random initializations for network weights.

In Figure 1, we study the link between sharpness and representation compression with a fixed batch size (of 20). We observe that the sharpness correlates with compression metrics, MLS and NMLS during training. The trend is consistent across multiple learning rates for a fixed batch size.

In Figure 2, we repeated the experiments while keeping the learning rate fixed at 0.1 and varying the batch size. The same broadly consistent trends emerged, linking sharpness to compression in the neural representation.

We repeated the experiments with an MLP trained on the FashionMNIST dataset [42] (Figure I.8 and Figure I.9). The sharpness again follows the same trend as MLS and NMLS, consistent with our bound. We also show that sharpness strongly correlates with MLS/NMLS for LeNet and ViTs (Appendix H.3 and Appendix H.4).

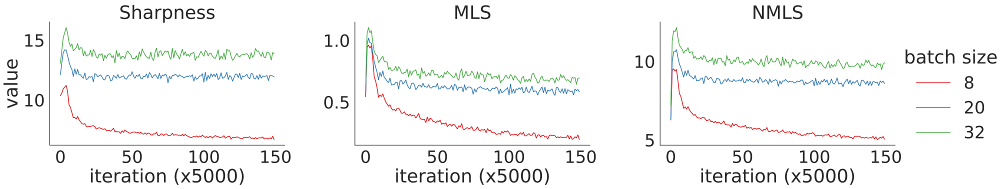

Figure 2: Trends in key variables across SGD training of the VGG-11 network with fixed learning rate size (equal to 0.1) and varying batch size (8, 20, and 32). Smaller batch sizes lead to lower sharpness, MLS, and NMLS during training. From left to right: sharpness (square root of Equation (3)),MLS, NMLS.

4. Discussion and Conclusion

We introduced a dual perspective linking parameter-space sharpness with feature-space compression via maximum local sensitivity (MLS). Our bounds, validated on feedforward, convolutional, and attention-based models, show that MLS closely tracks sharpness.

Although generalization was not our primary focus, we find that sharpness aligns more with compression than with generalization gaps (see Appendix H.3). This suggests that sharpness principally governs representation robustness, and only with additional biases does it improve generalization. For example, higher learning rates reduce sharpness and compression (see Section 3) but can harm out-of-distribution performance [39]. In tasks like information retrieval in long texts, output compression may be undesirable.

Our dual-perspective framework clarifies when sharpness-driven compression benefits or hinders performance, guiding future work on training strategies that balance robustness, compression, and generalization.

References

[1] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pages 639–668. PMLR, 2022.

[2] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

[3] Peter L Bartlett, Shahar Mendelson, and Joe Neeman. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022.

[4] Guy Blanc, Tengyu Ma, and Andrej Risteski. Implicit regularization of stochastic gradient descent for mean-field neural networks. In *Conference on Learning Theory (COLT)*, 2019.

[5] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process, 2020. URL http://arxiv.org/abs/1904.09080.

[6] Charles Cooper. Generalization bounds for deep learning. *arXiv preprint arXiv:1808.09540*, 2018.

[7] Valentin Damian, Vikrant Thakur, Yasaman Bahri, and Benjamin Recht. Label noise sgd induces implicit bias to minima of lower complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[8] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[9] Yuguang Fang, K.A. Loparo, and Xiangbo Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994. doi: 10.1109/9.362841.

[10] Benjamin Fehrman, Benjamin Gess, and Arnulf Jentzen. Generalization error bounds for training neural networks with gradient descent. *arXiv preprint arXiv:2007.07169*, 2020.

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

[12] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, page 214262, 2017.

[13] Mario Geiger, Leonardo Petrini, and Matthieu Wyart. Landscape and training regimes in deep learning. *Physics Reports*, 924:1–18, 2021. ISSN 0370-1573. doi: 10.1016/j.physrep.2021.04.001. URL https://www.sciencedirect.com/science/article/pii/S0370157321001290.

[14] Diego Granziol. Flatness is a false friend, 2020. URL https://arxiv.org/abs/2006.09091.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[16] Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three Factors Influencing Minima in SGD, September 2018. URL http://arxiv.org/abs/1711.04623. arXiv:1711.04623 [cs, stat].

[17] Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

[18] 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.

[19] 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.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[21] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021.

[22] Zhiyuan Li, Tengyu Liang, and Andrej Risteski. On the implicit bias of gradient descent for mean-field neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

[23] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? –a mathematical framework, 2022. URL http://arxiv.org/abs/2110.06914.

[24] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and LF Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

[25] Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks, 2021. URL http://arxiv.org/abs/2105.13462.

[26] Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Stimuli Reduce the Dimensionality of Cortical Activity. *Frontiers in Systems Neuroscience*, 10, February 2016. ISSN 1662-5137. doi: 10.3389/fnsys.2016.00011. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4756130/.

[27] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022. URL https://arxiv.org/abs/2110.02178.

[28] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[29] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[30] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34: 18420–18432, 2021.

[31] Stefano Recanatesi, Serena Bradde, Vijay Balasubramanian, Nicholas A Steinmetz, and Eric Shea-Brown. A scale-dependent measure of system dimensionality. *Patterns*, 3(8), 2022.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[33] Jingtong Su, Ya Shi Zhang, Nikolaos Tsilivis, and Julia Kempe. On the robustness of neural collapse and the neural collapse of robustness. *arXiv preprint arXiv:2311.07444*, 2023.

[34] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis, 2019.

[35] Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. Fastvit: A fast hybrid vision transformer using structural reparameterization, 2023. URL https://arxiv.org/abs/2303.14189.

[36] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization, 2023.

[37] Zhe Wen, Hongyang Zhang, and Yisen Yang. On the interplay between sharpness-aware minimization and adversarial robustness. *arXiv preprint arXiv:2206.01235*, 2022.

[38] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971, June 2022.

[40] Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.

[41] Lei Wu, Chao Ma, and Weinan E. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.,

2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/6651526b6fb8f29a00507de6a49ce30f-Abstract.html.

[42] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL http://arxiv.org/abs/1708.07747.

[43] Zeke Xie, Issei Sato, and Masashi Sugiyama. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima, January 2021. URL http://arxiv.org/abs/2002.03495. arXiv:2002.03495 [cs, stat].

[44] Ning Yang, Chao Tang, and Yuhai Tu. Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions. *Physical Review Letters*, 130(23): 237101, 2023. doi: 10.1103/PhysRevLett.130.237101. URL https://link.aps.org/doi/10.1103/PhysRevLett.130.237101. Publisher: American Physical Society.

[45] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733, 2021.

[46] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects, June 2019. URL http://arxiv.org/abs/1803.00195. arXiv:1803.00195 [cs, stat].

[47] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features, 2021.

Appendix A. Linear stability trick and proof of Equation 4 in Ma & Ying, 2021

Let $\mathbf{W}$ be the input weights (the parameters of the first linear layer) of the network, and $\bar{\boldsymbol{\theta}}$ be the rest of the parameters. Following [25], as the weights $\mathbf{W}$ multiply the inputs $\mathbf{x}$, we have the following identities:

$$\begin{aligned} \|\nabla_{\mathbf{W}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F &= \sqrt{\sum_{i,j,k} J_{jk}^2 x_i^2} \\ &= \|J\|_F \|\mathbf{x}\|_2 \geq \|J\|_2 \|\mathbf{x}\|_2 \ , \\ \nabla_{\mathbf{x}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}}) &= J\mathbf{W} \ , \end{aligned} \tag{8}$$

where $J = \frac{\partial f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})}{\partial(\mathbf{W}\mathbf{x})}$ is a complex expression computed with backpropagation. From Equation (8) and the sub-multiplicative property of the Frobenius norm and the matrix 2-norm [1], we have:

$$\begin{aligned} \|\nabla_{\mathbf{x}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_2 &\leq \|\nabla_{\mathbf{x}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F \\ &\leq \frac{\|\mathbf{W}\|_2}{\|\mathbf{x}\|_2} \|\nabla_{\mathbf{W}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F \ . \end{aligned} \tag{9}$$

We call Equation (9) the linear stability trick. As a result, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_2^k &\leq \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^k \\ &\leq \frac{\|\mathbf{W}\|_2^k}{\min_i \|\mathbf{x}_i\|_2^k} \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{W}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^k \\ &\leq \frac{\|\mathbf{W}\|_2^k}{\min_i \|\mathbf{x}_i\|_2^k} \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^k . \end{aligned} \tag{10}$$

A.1. Linear stability trick on penultimate-layer features

we define $g(x)$ to be the penultimate-layer features such that $f(x) = \mathbf{W}_L g(x) + b$. With slight abuse of notation, we define $J = \frac{\partial g(\mathbf{W}\mathbf{x})}{\partial \mathbf{W}\mathbf{x}}$. Similar to Equation (8), we have

$$\begin{aligned} \|\nabla_{\mathbf{W}} f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F &= \|\mathbf{W}_L J\|_F \|\mathbf{x}\|_2 \\ \nabla_{\mathbf{x}} g(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}}) &= J\mathbf{W} \ , \end{aligned} \tag{11}$$

Lemma 1 $\|AB\|_F \geq \lambda_{\min}(A^T A)\|B\|_F$, *where* $\lambda_{\min}$ *is the smallest eigenvalue.*

Proof By the definition of Frobenius norm,

$$\|AB\|_F^2 = \mathrm{Tr}(ABB^T A^T) = \mathrm{Tr}(A^T ABB^T). \tag{12}$$

From Fang et al. [9], we have that for positive semidefinite matrices P and Q,

$$\lambda_{\min}(P)\,\mathrm{Tr}(Q) \leq \mathrm{Tr}(PQ) \tag{13}$$

[1] $\|AB\|_F \leq \|A\|_F \|B\|_2$, $\|AB\|_2 \leq \|A\|_2 \|B\|_2$

Therefore,

$$\operatorname{Tr}(A^TABB^T) \geq \lambda_{\min}(A^TA)\operatorname{Tr}(BB^T) = \lambda_{\min}(A^TA)\|B\|_F^2 \tag{14}$$

■

As a result, $\|\mathbf{W}_LJ\|_F \geq \sqrt{\lambda_{\min}(\mathbf{W}_L^T\mathbf{W}_L)}\|J\|_F$. Let d be the feature dimension, and K be the number of classes, and $\mathbf{W}_L \in \mathbb{R}^{d\times K}$. Then, $\lambda_{\min}(\mathbf{W}_L^T\mathbf{W}_L)$ vanishes when $K > d$, otherwise $\sqrt{\lambda_{\min}(\mathbf{W}_L^T\mathbf{W}_L)} = \sigma_{\min}(\mathbf{W}_L)$, the smallest singular value of $\mathbf{W}_L$. Therefore,

$$\begin{aligned}\|\nabla_{\mathbf{x}}g(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F &= \|J\mathbf{W}\|_F \\ &\leq \|J\|_F\|\mathbf{W}\|_2 \\ &\leq \frac{\|\mathbf{W}_LJ\|_F}{\sqrt{\lambda_{\min}(\mathbf{W}_L^T\mathbf{W}_L)}}\|\mathbf{W}\|_2 \\ &= \frac{\|\nabla_{\mathbf{W}}f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})\|_F}{\|\mathbf{x}\|_2\sqrt{\lambda_{\min}(\mathbf{W}_L^T\mathbf{W}_L)}}\|\mathbf{W}\|_2\end{aligned} \tag{15}$$

Appendix B. Adaptation of Inequality 4 to Residual Layers

We need to slightly adapt the proof in Eq. 4 and 5. Consider a network whose first layer has a residual connection: $y = g(x + f(Wx))$, where f is the nonlinearity with bias (e.g. $f(x) = \tanh(x + b)$), and g is the rest of the mappings in the network. Then we have

$$\begin{aligned}\|\nabla_W g(x + f(Wx))\|_F = \|JK\|_F\,\|x\|_2 \\ \nabla_x g(x + f(Wx)) = J + JKW\end{aligned} \tag{16}$$

where $J = \frac{\partial g(x+f(Wx))}{\partial(x+f(Wx))}$ and $K = \frac{\partial f(Wx)}{\partial(Wx)}$.

Therefore, $||\nabla_x g(x+f(Wx))||_2 \leq ||J||_2+||JK||_2||W||_2 \leq ||J||_2+\frac{||\nabla_W g(x+f(Wx))||_F}{||x||_2}||W||_2$. Now, we get the bound for the *difference* between MLS of input and the MLS of input to the next layer:

$$\|\nabla_x g(x + f(Wx))\|_2 - \|J\|_2 \leq \frac{\|\nabla_W g(x + f(Wx))\|_F}{\|x\|_2}\,\|W\|_2 \tag{17}$$

Notice that if we apply this inequality to every residual layer in the network, and sum the left-hand side, we will get a telescoping sum on the left-hand side. Assuming the last layer is linear with weights W_L, we get $||\nabla_x g(x + f(W_1x))||_2 - ||W_L||_2 \leq \sum_{l=1}^{L-1}\frac{||W_l||_2}{||x_l||_2}||\nabla_W g_l(x_l + f(W_lx_l))||_F$. The right-hand side is bounded by sharpness due to Cauchy, see also Equation (33).

Appendix C. Proof of Equation (3)

Lemma 2 *If $\boldsymbol{\theta}$ is an approximate interpolation solution, i.e. $\|f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i\| < \varepsilon$ for $i \in \{1, 2, \cdots, n\}$, and second derivatives of the network function $\|\nabla^2_{\theta_j} f(\mathbf{x}_i, \boldsymbol{\theta})\| < M$ is bounded, then*

$$S(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\boldsymbol{\theta}}f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^2 + O(\varepsilon) \tag{18}$$

Proof Using basic calculus we get

$$
\begin{aligned}
S(\boldsymbol{\theta}) &= \mathrm{Tr}(\nabla^2 L(\boldsymbol{\theta})) \\
&= \frac{1}{2n}\sum_{i=1}^{n} \mathrm{Tr}(\nabla^2_{\boldsymbol{\theta}} \left\|f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i\right\|^2) \\
&= \frac{1}{2n}\sum_{i=1}^{n} \mathrm{Tr}\, \nabla_{\boldsymbol{\theta}}(2(f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i,\boldsymbol{\theta})) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}_j}((f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i,\boldsymbol{\theta}))_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\partial}{\partial \boldsymbol{\theta}_j}(f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i,\boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \left\|\nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i,\boldsymbol{\theta})\right\|_2^2 + (f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i,\boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i,\boldsymbol{\theta})\|_F^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} (f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i,\boldsymbol{\theta}).
\end{aligned}
$$

Therefore

$$
\left| S(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i,\boldsymbol{\theta})\|_F^2 \right| < \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |(f(\mathbf{x}_i,\boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i,\boldsymbol{\theta})| < mM\varepsilon = O(\varepsilon). \tag{19}
$$

■

In other words, when the network reaches zero training error and enters the interpolation phase (i.e. it classifies all training data correctly), Equation (3) will be a good enough approximation of the sharpness because the quadratic training loss is sufficiently small.

Appendix D. Proof of Proposition H.3 and Proposition H.5

For notation simplicity, we write $f_i := f(\mathbf{x}_i, \boldsymbol{\theta}^*)$ in what follows. Because of Equation (9), we have the following inequality due to Cauchy-Swartz inequality,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f_i\|_F^k \leq \|\mathbf{W}\|_2^k \frac{1}{n}\sum_{i=1}^{n} \frac{\|\nabla_{\mathbf{W}} f_i\|_F^k}{\|\mathbf{x}_i\|_2^k} \\
\leq \frac{1}{n}\|\mathbf{W}\|_2^k \sqrt{\sum_{i=1}^{n} \frac{1}{\|\mathbf{x}_i\|_2^{2k}}} \cdot \sqrt{\sum_{i=1}^{n} \|\nabla_{\mathbf{W}} f_i\|_F^{2k}}.
\end{aligned}
\tag{20}
$$

Since the input weights $\mathbf{W}$ is just a part of all the weights ($\boldsymbol{\theta}$) of the network, we have $\|\nabla_{\mathbf{W}} f_i\|_F^k \leq \|\nabla_{\boldsymbol{\theta}} f_i\|_F^k$.

We next show the correctness of Proposition H.3 with a standard lemma.

Lemma 3 *For vector* $\mathbf{x}$, $\|\mathbf{x}\|_p \geq \|\mathbf{x}\|_q$ *for* $1 \leq p \leq q \leq \infty$.

Proof First we show that for $0 < k < 1$, we have $(|a| + |b|)^k \leq |a|^k + |b|^k$. It's trivial when either a or b is 0. So W.L.O.G, we can assume that $|a| < |b|$, and divide both sides by $|b|^k$. Therefore it suffices to show that for $0 < t < 1$, $(1+t)^k < t^k + 1$. Let $f(t) = (1+t)^k - t^k - 1$, then $f(0) = 0$, and $f'(t) = k(1+t)^{k-1} - kt^{k-1}$. Because $k - 1 < 0$, $1 + t > 1$ and $t < 1$, $t^{k-1} > 1 > (1+t)^{k-1}$. Therefore $f'(t) < 0$ and $f(t) < 0$ for $0 < t < 1$. Combining all cases, we have $(|a| + |b|)^k \leq |a|^k + |b|^k$ for $0 < k < 1$. By induction, we have $(\sum_n |a_n|)^k \leq \sum_n |a_n|^k$.

Now we can prove the lemma using the conclusion above,

$$\left(\sum_n |x_n|^q\right)^{1/q} = \left(\sum_n |x_n|^q\right)^{p/q \cdot 1/p} \leq \left(\sum_n (|x_n|^q)^{p/q}\right)^{1/p} = \left(\sum_n |x_n|^p\right)^{1/p}$$

■

Now we can prove Proposition H.3

Proposition *The local volumetric ratio is upper bounded by a sharpness related quantity:*

$$dV_{f(\boldsymbol{\theta}^*)} \leq \frac{N^{-N/2}}{n} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta}^*)\|_F^N \leq \frac{1}{n} \sqrt{\sum_{i=1}^{n} \frac{\|\mathbf{W}\|_2^{2N}}{\|\mathbf{x}_i\|_2^{2N}}} \left(\frac{nS(\boldsymbol{\theta}^*)}{N}\right)^{N/2} \tag{21}$$

for all $N \geq 1$.

Proof Take the x_i in Lemma 3 to be $\|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F^2$ and let $p = 1, q = k$, then we get

$$\left(\sum_{i=1}^{n} (\|\nabla_{\boldsymbol{\theta}} f_i\|_F^2)^k\right)^{1/k} \leq \sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f_i\|_F^2. \tag{22}$$

Therefore,

$$\begin{aligned}\frac{1}{n}\|\mathbf{W}\|_2^k \sqrt{\sum_{i=1}^{n} \frac{1}{\|\mathbf{x}_i\|_2^{2k}}} \cdot \sqrt{\sum_{i=1}^{n} \|\nabla_{\mathbf{W}} f_i\|_F^{2k}} &\leq n^{k/2-1} \|\mathbf{W}\|_2^k \sqrt{\sum_{i=1}^{n} \frac{1}{\|\mathbf{x}_i\|_2^{2k}}} \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f_i\|_F^2\right)^{k/2} \\ &= n^{k/2-1} \|\mathbf{W}\|_2^k \sqrt{\sum_{i=1}^{n} \frac{1}{\|\mathbf{x}_i\|_2^{2k}}} S(\boldsymbol{\theta}^*)^{k/2}\end{aligned} \tag{23}$$

■

Next, we show that the first inequality in Equation (23) can be tightened by considering all linear layer weights.

Proposition *The network volumetric ratio is upper bounded by a sharpness related quantity:*

$$\sum_{l=1}^{L} dV_{f_l} \leq \frac{N^{-N/2}}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}^l} f_i^l\|_F^N \leq \frac{1}{n} \sqrt{\sum_{l=1}^{L} \sum_{i=1}^{n} \frac{\|\mathbf{W}_l\|_2^{2N}}{\|\mathbf{x}_i^l\|_2^{2N}}} \cdot \left(\frac{nS(\boldsymbol{\theta}^*)}{N}\right)^{N/2}. \tag{24}$$

Proof Recall that the input to l-th linear layer as x_i^l for $l = 1, 2, \cdots, L$. In particular, x_i^1 is the input of the entire network. Similarly, $\mathbf{W}_l$ is the weight matrix of l-th linear/convolutional layer. With a slight abuse of notation, we use f^l to denote the mapping from the activity of l-th layer to the final output, and $f_i^l := f^l(\mathbf{x}_i, \boldsymbol{\theta}^*)$. We can apply Cauchy-Swartz inequality again to get

$$\begin{aligned}\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}^l} f_i^l\|_F^k &\leq \frac{1}{n}\sum_{l=1}^{L}\sqrt{\sum_{i=1}^{n}\frac{\|\mathbf{W}_l\|_2^{2k}}{\|\mathbf{x}_i^l\|_2^{2k}}}\cdot\sqrt{\sum_{i=1}^{n}\|\nabla_{\mathbf{W}_l} f_i^l\|_F^{2k}} \\ &\leq \sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\frac{\|\mathbf{W}_l\|_2^{2k}}{\|\mathbf{x}_i^l\|_2^{2k}}}\cdot\sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\nabla_{\mathbf{W}_l} f_i^l\|_F^{2k}}.\end{aligned} \tag{25}$$

Using Lemma 3 again we have

$$\begin{aligned}\left(\sum_{l=1}^{L}(\|\nabla_{\boldsymbol{W}_l} f_i^l\|_F^2)^k\right)^{1/k} &\leq \sum_{l=1}^{L}\|\nabla_{\boldsymbol{W}_l} f_i^l\|_F^2 = \|\nabla_{\boldsymbol{\theta}} f_i\|_F^2, \\ \left(\sum_{i=1}^{n}(\|\nabla_{\boldsymbol{\theta}} f_i\|_F^2)^k\right)^{1/k} &\leq \sum_{i=1}^{n}\|\nabla_{\boldsymbol{\theta}} f_i\|_F^2 = nS(\boldsymbol{\theta}^*),\end{aligned} \tag{26}$$

The second equality holds because both sides represent the same gradients in the computation graph. Therefore from Equation (25), we have

$$\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}^l} f_i^l\|_F^k \leq \sqrt{\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\frac{\|\mathbf{W}_l\|_2^{2k}}{\|\mathbf{x}_i^l\|_2^{2k}}}\cdot\sqrt{n^{k-1}S(\boldsymbol{\theta}^*)^k} \tag{27}$$

■

Appendix E. Proof of Proposition 2.2 and Proposition 2.4

Compared to Equation (4) of Ma and Ying [25], we make the following improvements:

1. We replace the reciprocal of the minimum with the quadratic mean to achieve a more stable bound (Proposition 2.2). This term remains relevant as common practice in deep learning does *not* normalize the input by its 2-norm, as this would erase information about the modulus of the input.

2. While Ma and Ying [25] only considers scalar output, we extend the result to consider networks with multivariable output throughout the paper.

3. We introduce Network MLS (Definition 2.3) and their sharpness-related bounds (Proposition 2.4), which have two advantages compared to prior results (cf. the right-hand side of Equation (4)):

   (a) our metrics consider all linear weights so that bounds remain stable to weight changes during training.

(b) they avoid the gap between derivative w.r.t. the first layer weights and the derivative w.r.t. all weights, i.e. the second inequality in Eq. 6, thus tightening the bound.

Below we give the proof of Proposition 2.2.

Proposition *The maximum local sensitivity is upper bounded by a sharpness related quantity:*

$$\mathrm{MLS} = \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_2 \leq \|\mathbf{W}\|_2 \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}} S(\boldsymbol{\theta}^*)^{1/2}\ . \tag{28}$$

Proof From Equation (9), we get

$$\mathrm{MLS} = \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} f_i\|_2 \leq \|\mathbf{W}\|_2 \frac{1}{n}\sum_{i=1}^{n}\frac{\|\nabla_{\mathbf{W}} f_i\|_F}{\|\mathbf{x}_i\|_2}. \tag{29}$$

Now the Cauchy-Schwarz inequality tells us that

$$\left(\sum_{i=1}^{n}\frac{\|\nabla_{\mathbf{W}} f_i\|}{\|\mathbf{x}_i\|_2}\right)^2 \leq \left(\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}\right)\cdot\left(\sum_{i=1}^{n}\|\nabla_{\mathbf{W}} f_i\|_F^2\right). \tag{30}$$

Therefore

$$\begin{aligned}\mathrm{MLS} &\leq \|\mathbf{W}\|_2 \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{W}} f_i\|_F^2} \\ &\leq \|\mathbf{W}\|_2 \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}} \cdot S(\boldsymbol{\theta}^*)^{1/2}.\end{aligned} \tag{31}$$

■

Now we can prove Proposition 2.4.

Proposition *The network maximum local sensitivity is upper bounded by a sharpness related quantity:*

$$\mathrm{NMLS} = \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}^l} f^l(\mathbf{x}_i^l, \boldsymbol{\theta}^*)\|_2 \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\frac{\|\mathbf{W}_l\|_2^2}{\|\mathbf{x}_i^l\|^2}} \cdot S(\boldsymbol{\theta}^*)^{1/2}. \tag{32}$$

Proof We can apply Equation (31) to every linear layer and again apply the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\mathrm{NMLS} &= \frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} f_l(\mathbf{x}_i^l, \boldsymbol{\theta}^*)\|_2 \\
&\leq \sum_{l=1}^{L}\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{\|\mathbf{W}_l\|_2^2}{\|\mathbf{x}_i^l\|_2^2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{W}_l} f_i^l\|_F^2}\right) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\frac{\|\mathbf{W}_l\|_2^2}{\|\mathbf{x}_i^l\|_2^2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\|\nabla_{\mathbf{W}_l} f_i^l\|_F^2} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{L}\frac{\|\mathbf{W}_l\|_2^2}{\|\mathbf{x}_i^l\|_2^2}}\cdot S(\boldsymbol{\theta}^*)^{1/2}.
\end{aligned} \tag{33}$$

Note that the gap in the last inequality is significantly smaller than that of Equation (31) since now we consider all linear weights. ■

Appendix F. Reparametrization-invariant sharpness and input-invariant MLS

Dinh et al. [8] argues that a robust sharpness metric should have the reparametrization-invariant property, meaning that scaling the neighboring linear layer weights should not change the metric. While the bounds in Proposition 2.2 and Proposition 2.4 are not strictly reparametrization-invariant, metric that redesign the sharpness [34] to achieve invariance can be proved to tighten our bounds (see Appendix F.1). Another more aggressive reparametrization-invariant sharpness is proposed in Andriushchenko et al. [2], Kwon et al. [21], and we again show that it upper-bounds input-invariant MLS in Appendix F.2. We also empirically evaluate the relative flatness [30], which is also reparametrization-invariant in Appendix H.3, but no significant correlation is observed. Overall, we provide a novel perspective: reparametrization-invariant sharpness is characterized by the robustness of outputs to internal neural representations.

F.1. Reparametrization-invariant sharpness in Tsuzuku et al. [34]

In this appendix, we show that the reparametrization-invariant sharpness metrics introduced in Tsuzuku et al. [34] can seen as an effort to tighten the bound that we derived above. For matrix-normalized sharpness (cf. Equation 13), the connection is immediately seen from Equation (31). Let

$$\overline{\mathbf{x}} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}\right)^{-\frac{1}{2}}. \tag{34}$$

Then from Equation (31) we have

$$\sum_{l=1}^{L}\overline{\mathbf{x}}^l\cdot \mathrm{MLS}^l \leq \sum_{l=1}^{L}\|\mathbf{W}_l\|_2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{W}_l} f_i^l\|_F^2} \approx \sum_{l=1}^{L}\|\mathbf{W}_l\|_2\sqrt{S(\mathbf{W}_l)}, \tag{35}$$

where $S(\mathbf{W}_l)$ is the trace of Hessian of the loss w.r.t. the weights of the l-th layer. The right-hand side of Equation (35) is exactly what Tsuzuku et al. [34] refer to as the matrix-normalized sharpness. Note that a similar inequality holds if we use Frobenius norm instead of 2-norm of the weights.

Tsuzuku et al. [34] also pose an interesting optimization problem (cf. Equation 17) to define the normalized sharpness:

$$\min_{\boldsymbol{\sigma},\boldsymbol{\sigma'}} \sum_{i,j}\left(\frac{\partial^2 L}{\partial W_{i,j}\partial W_{i,j}}(\sigma_i\sigma'_j)^2 + \frac{W_{i,j}^2}{4\lambda^2(\sigma_i\sigma'_j)^2}\right). \tag{36}$$

Note that by Lemma 2, $\frac{\partial^2 L}{\partial W_{i,j}\partial W_{i,j}} \approx \|\nabla_{\mathbf{W}_{i,j}} f\|_2^2$. Moreover, we have

$$\begin{aligned}\sum_{i,j}\left(\|\nabla_{\mathbf{W}_{i,j}} f\|^2(\sigma_i\sigma'_j)^2 + \frac{W_{i,j}^2}{4\lambda^2(\sigma_i\sigma'_j)^2}\right) &\geq \frac{1}{\lambda}\sqrt{\sum_{i,j}(\nabla_{\mathbf{W}_{i,j}} f)^2(\sigma_i\sigma'_j)^2}\cdot\sqrt{\sum_{i,j}\frac{W_{i,j}^2}{(\sigma_i\sigma'_j)^2}}\\ &\geq \frac{1}{\lambda}\|\mathrm{diag}(\boldsymbol{\sigma})J\|_F\|\mathrm{diag}(\boldsymbol{\sigma'})\mathbf{x}\|_2\|\mathrm{diag}(\boldsymbol{\sigma}^{-1})\mathbf{W}\,\mathrm{diag}(\boldsymbol{\sigma'}^{-1})\|_F\\ &\geq \frac{1}{\lambda}\|\mathrm{diag}(\boldsymbol{\sigma'}^{-1})W^T J\|_F\|\mathrm{diag}(\boldsymbol{\sigma'})\mathbf{x}\|_2\\ &= \frac{1}{\lambda}\|\mathrm{diag}(\boldsymbol{\sigma'}^{-1})\nabla_{\mathbf{x}} f\|_F\|\mathrm{diag}(\boldsymbol{\sigma'})\mathbf{x}\|_2,\end{aligned} \tag{37}$$

where $J = \frac{\partial f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})}{\partial(\mathbf{W}\mathbf{x})}$ (see some of the calculations in Equation (8)). Therefore, the optimization problem Equation (36) is equivalent to choosing $\boldsymbol{\sigma},\boldsymbol{\sigma'}$ to minimize the upper bound on a scale-invariant MLS-like quantity (the quantity is invariant under the transformation of the first layer from $\mathbf{W}\mathbf{x}$ to $\mathbf{W}\,\mathrm{diag}(\boldsymbol{\sigma}^{-1})(\mathrm{diag}(\boldsymbol{\sigma})\mathbf{x})$, where $\mathrm{diag}(\boldsymbol{\sigma})\mathbf{x}$ becomes the new input). For simplicity, we do not scale the original dataset in our work and only compare MLS within the same dataset. As a result, we can characterize those reparametrization-invariant sharpness metrics by the robustness of output to the input. If we consider all linear weights in the network, then those metrics indicate the robustness of output to internal network representations.

F.2. Reparametrization-invariant sharpness upper-bounds input-invariant MLS

In this appendix, we consider the adaptive average-case n-sharpness considered in Andriushchenko et al. [2], Kwon et al. [21]:

$$S^{\rho}_{\text{avg}}(\mathbf{w},|\mathbf{w}|) \triangleq \frac{2}{\rho^2}\mathbb{E}_{S\sim P_n,\delta\sim\mathcal{N}(0,\rho^2\text{diag}(|\mathbf{w}|^2))}\left[L_S(\mathbf{w}+\delta) - L_S(\mathbf{w})\right], \tag{38}$$

which is shown to be *elementwise* adaptive sharpness in Andriushchenko et al. [2]. They also show that for a thrice differentiable loss, $L(w)$, the average-case elementwise adaptive sharpness can be written as

$$S^{\rho}_{\text{avg}}(\mathbf{w},|\mathbf{w}|) = \mathbb{E}_{S\sim P_n}\left[\mathrm{Tr}\left(\nabla^2 L_S(\mathbf{w})\odot|\mathbf{w}||\mathbf{w}|^{\top}\right)\right] + O(\rho). \tag{39}$$

Definition F.1 *We define the* ***Elementwise-Adaptive Sharpness*** $S_{\textit{adaptive}}$ *to be*

$$S_{\textit{adaptive}}(\mathbf{w}) \triangleq \lim_{\rho\to 0} S^{\rho}_{\textit{avg}}(\mathbf{w},|\mathbf{w}|) = \mathbb{E}_{S\sim P_n}\left[\mathrm{Tr}\left(\nabla^2 L_S(\mathbf{w})\odot|\mathbf{w}||\mathbf{w}|^{\top}\right)\right] \tag{40}$$

In this appendix, we focus on the property of S_{adaptive} instead of the approximation Equation (39). Adapting the proof of $Theorem$ 2, we have the following lemma.

Lemma 4 *If $\boldsymbol{\theta}$ is an approximate interpolation solution, i.e. $\|f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i\| < \varepsilon$ for $i \in \{1, 2, \cdots, n\}$, $|\boldsymbol{\theta}_j|^2 \|\nabla^2_{\theta_j} f(\mathbf{x}_i, \boldsymbol{\theta})\| < M$ for all j, and L is MSE loss, then*

$$S_{adaptive}(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 \left\|\nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta})\right\|_2^2 + O(\varepsilon), \tag{41}$$

where m is the number of parameters.

Proof Using basic calculus we get

$$\begin{aligned}
S_{\text{adaptive}}(\boldsymbol{\theta}) &= \frac{1}{2n}\sum_{i=1}^{n} \operatorname{Tr}(\nabla^2_{\boldsymbol{\theta}} \|f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i\|^2 \odot |\boldsymbol{\theta}||\boldsymbol{\theta}|^\top) \\
&= \frac{1}{2n}\sum_{i=1}^{n} \operatorname{Tr} \nabla_{\boldsymbol{\theta}}(2(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta})) \odot |\boldsymbol{\theta}||\boldsymbol{\theta}|^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 \frac{\partial}{\partial \boldsymbol{\theta}_j}((f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta}))_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 \frac{\partial}{\partial \boldsymbol{\theta}_j}(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 \left\|\nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta})\right\|_2^2 + |\boldsymbol{\theta}_j|^2 (f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 \left\|\nabla_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta})\right\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |\boldsymbol{\theta}_j|^2 (f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta})
\end{aligned}$$

Therefore

$$\left| S(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \|\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i, \boldsymbol{\theta})\|_F^2 \right| < \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} |(f(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i)^T |\boldsymbol{\theta}_j|^2 \nabla^2_{\boldsymbol{\theta}_j} f(\mathbf{x}_i, \boldsymbol{\theta})| < mM\varepsilon = O(\varepsilon). \tag{42}$$

■

Definition F.2 *We define the* ***Input-invariant MLS*** *of a network $f : \mathbb{R}^N \to \mathbb{R}^M$ to be*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{p=1}^{N} \left\|\nabla_{x_p^{(i)}} f\right\|_2^2 (x_p^{(i)})^2 \,, \tag{43}$$

where $x_p^{(i)}$ is the p-th entry of i-th training sample.

It turns out that again the adaptive sharpness upper bounds the input-invariant MLS.

Proposition F.3 *Assuming that the condition of Theorem 4 holds, then elementwise-adaptive sharpness upper-bounds input-invariant MLS:*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}|\boldsymbol{\theta}_j|^2\left\|\nabla_{\boldsymbol{\theta}_j}f(\mathbf{x}_i,\boldsymbol{\theta})\right\|_2^2 \geq \frac{1}{nd}\sum_{i=1}^{n}\sum_{p=1}^{N}\left\|\nabla_{x_p^{(i)}}f\right\|_2^2(x_p^{(i)})^2 \tag{44}$$

Proof Now we adapt the linear stability trick. For $\boldsymbol{\theta} = \mathbf{W}$ the first layer weight, we have

$$\begin{aligned}\sum_{j=1}^{m}|\boldsymbol{\theta}_j|^2\left\|\nabla_{\boldsymbol{\theta}_j}f(\mathbf{x},\boldsymbol{\theta})\right\|_2^2 &= \sum_{i,j,k}J_{jk}^2\mathbf{W}_{ki}^2x_p^2\\ &= \sum_{i,j}\left(\sum_{k=1}^{d}J_{jk}^2\mathbf{W}_{ki}^2\right)x_p^2\\ &\geq \frac{1}{d}\sum_{i}\left\|\nabla_{x_p}f\right\|_2^2x_p^2\end{aligned} \tag{45}$$

where same as in Equation (8), $J = \frac{\partial f(\mathbf{Wx};\bar{\boldsymbol{\theta}})}{\partial(\mathbf{Wx})}$, $\nabla_{\mathbf{x}}f(\mathbf{Wx};\bar{\boldsymbol{\theta}}) = J\mathbf{W}$, and x_p is the p-th entry of $\mathbf{x}$. Taking the sample mean of both sides proves the proposition.

■

Appendix G. Numerical approximation of normalized MLS and elementwise-adaptive sharpness

In this appendix, we detail how we approximate the normalized MLS and adaptive sharpness in Appendix H.4. Note that for all network f the last layer is the sigmoid function, so the output is bounded in $(0, 1)$, and we use MSE loss to be consistent with the rest of the paper.

For the adaptive sharpness, we adopt the definition in Andriushchenko et al. [2] and uses sample mean to approximate the expectation in Equation (38). Therefore, for network $f(\mathbf{w})$,

$$S_{\text{adaptive}}(f) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}L(\mathbf{x}_i;\mathbf{w}+\delta_j) - L(\mathbf{x}_i;\mathbf{w}), \tag{46}$$

where $\delta \sim \mathcal{N}(0, 0.01\,\mathrm{diag}(|w|^2))$.

For normalized MLS, we first reiterate the definition from the main text. We use normalized MLS below as an approximation to the input-invariant MLS (Definition F.2), because the latter is computationally prohibitive for modern large ViTs. On the other hand, there is an efficient way to estimate normalized MLS as detailed below.

Definition G.1 *We define the **normalized MLS** as* $\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2\,\|\nabla_{\mathbf{x}_i}f\|_2^2$

Therefore, to approximate normalized MLS, we need to approximate $\|\nabla_{\mathbf{x}_i}f\|_2$. By definition of matrix 2-norm,

$$\|\nabla_{\mathbf{x}}f\|_2 = \sup_{\delta}\frac{\|\nabla_{\mathbf{x}}f\ \delta\|_2}{\|\delta\|_2} \approx \max_{\delta}\frac{\|f(\mathbf{x}+\delta)-f(\mathbf{x})\|_2}{\|\delta\|_2}. \tag{47}$$

To solve this optimization problem, we start from a randomly sampled vector δ that has the same shape as the network input, and we update δ using gradient descent.

Appendix H. Empirical analysis of the bound

H.1. Tightness of the bound

In this section, we mainly explore the tightness of the bound in Equation (5) for reasons discussed in Section 2.1. First we rewrite Equation (5) as

$$\begin{aligned}
\mathrm{MLS} &= \frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} f(\mathbf{x}_i,\boldsymbol{\theta}^*)\|_2 && := A \\
&\le \frac{\|\mathbf{W}\|_2}{n}\sum_{i=1}^{n}\frac{\|\nabla_{\mathbf{W}} f(\mathbf{x}_i,\boldsymbol{\theta}^*)\|_F}{\|\mathbf{x}_i\|_2} && := B \\
&\le \|\mathbf{W}\|_2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\nabla_{\mathbf{W}} f(\mathbf{x}_i,\boldsymbol{\theta}^*)\|_F^2} && := C \\
&\le \|\mathbf{W}\|_2\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\|\mathbf{x}_i\|_2^2}}\,S(\boldsymbol{\theta}^*)^{1/2} && := D
\end{aligned} \tag{48}$$

Thus Equation (5) consists of 3 different steps of relaxations. We analyze them one by one:

1. ($A \le B$) The equality holds when $\|W^T J\|_2 = \|W\|_2\|J\|_2$ and $\|J\|_F = \|J\|_2$, where $J = \frac{\partial f(\mathbf{W}\mathbf{x};\bar{\boldsymbol{\theta}})}{\partial(\mathbf{W}\mathbf{x})}$. The former equality requires that W and J have the same left singular vectors. The latter requires J to have zero singular values except for the largest singular value. Since J depends on the specific neural network architecture and training process, we test the tightness of this bound empirically (Figure H.3).

2. ($B \le C$) The equality requires $\frac{\|\nabla_{\mathbf{W}} f(\mathbf{x}_i,\boldsymbol{\theta}^*)\|_F}{\|\mathbf{x}_i\|_2}$ to be the same for all i. In other words, the bound is tight when $\frac{\|\nabla_{\mathbf{W}} f(\mathbf{x}_i,\boldsymbol{\theta}^*)\|_F}{\|\mathbf{x}_i\|_2}$ does not vary too much from sample to sample.

3. ($C \le D$) The equality holds if the model is linear, i.e. $\boldsymbol{\theta} = \mathbf{W}$.

We empirically verify the tightness of the above bounds in Figure H.3

H.2. More metrics in the feature space

H.2.1. SHARPNESS BOUNDS LOCAL VOLUMETRIC TRANSFORMATION

Now we quantify how a network compresses its input volumes via the local volumetric ratio, between the volume of a hypercube of side length h at $\mathbf{x}$, $H(\mathbf{x})$, and its image under transformation f, $f(H(\mathbf{x}),\boldsymbol{\theta}^*)$:

$$\begin{aligned}
d\,\mathrm{Vol}\big|_{f(\mathbf{x},\boldsymbol{\theta}^*)} &= \lim_{h\to 0}\frac{\mathrm{Vol}(f(H(\mathbf{x}),\boldsymbol{\theta}^*))}{\mathrm{Vol}(H(\mathbf{x}))} \\
&= \sqrt{\det\left(\nabla_{\mathbf{x}} f^T \nabla_{\mathbf{x}} f\right)}\,,
\end{aligned} \tag{49}$$

which is equal to the square root of the product of all positive eigenvalues of $C_f^{\lim}$ (see Equation (54)).

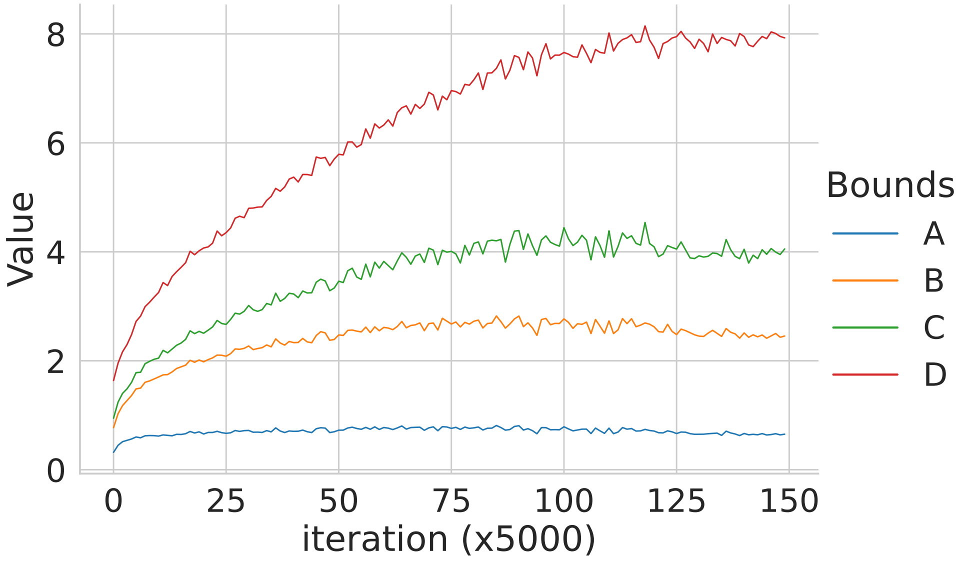

Figure H.3: **Empirical tightness of the bounds.** We empirically verify that the inequalities in Equation (48) hold and test their tightness. The results are shown for a fully connected feedforward network trained on the FashionMNIST dataset. The quantities A, B, C, and D are defined in Equation (48). We see that the gap between C and D is large compared to the gap between A and B or B and C. This indicates that partial sharpness $\|\nabla_{\mathbf{W}} f(\mathbf{x}_i, \boldsymbol{\theta}^*)\|_F$ (sensitivity of the loss w.r.t. only the input weights) is more indicative of the change in the maximum local sensitivity (A). Indeed, correlation analysis shows that bound C is positively correlated with MLS while bound D, perhaps surprisingly, is negatively correlated with MLS (Figure H.5).

Definition H.1 *The **Local Volumetric Ratio at input** $\mathbf{x}$ of a network f with parameters $\boldsymbol{\theta}$ is defined as* $d\,\mathrm{Vol}|_{f(\mathbf{x},\boldsymbol{\theta})} = \sqrt{\det\left(\nabla_{\mathbf{x}} f^T \nabla_{\mathbf{x}} f\right)}$.

Exploiting the bound on the gradients derived earlier in Equation (9), we derive a similar bound for the volumetric ratio:

Lemma 5

$$\begin{aligned} d\,\mathrm{Vol}|_{f(\mathbf{x},\boldsymbol{\theta}^*)} &\leq \left(\frac{\mathrm{Tr}\,\nabla_{\mathbf{x}} f^T \nabla_{\mathbf{x}} f}{N}\right)^{N/2} \\ &= N^{-N/2} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta}^*)\|_F^N \,, \end{aligned} \tag{50}$$

where the first inequality uses the inequality of arithmetic and geometric means, and the second the definition of the Frobenius norm. Next, we introduce a measure of the volumetric ratio averaged across input samples.

Definition H.2 *The **Local Volumetric Ratio (LVR)** of a network f with parameters $\boldsymbol{\theta}$ is defined as the sample mean of Local Volumetric Ratio at different input samples:* $dV_{f(\boldsymbol{\theta})} = \frac{1}{n}\sum_{i=1}^{n} d\,\mathrm{Vol}|_{f(\mathbf{x}_i,\boldsymbol{\theta})}$.

Then we have the following inequality that relates the sharpness to the mean local volumetric ratio:

Proposition H.3 *The local volumetric ratio is upper bounded by a sharpness related quantity:*

$$\begin{aligned} dV_{f(\boldsymbol{\theta}^*)} &\leq \frac{N^{-N/2}}{n} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta}^*)\|_F^N \\ &\leq \frac{1}{n} \sqrt{\sum_{i=1}^{n} \frac{\|\mathbf{W}\|_2^{2N}}{\|\mathbf{x}_i\|_2^{2N}}} \left(\frac{nS(\boldsymbol{\theta}^*)}{N}\right)^{N/2} \end{aligned} \tag{51}$$

for all $N \geq 1$.

The proof of the above inequalities is given in Appendix D.

Next, we give an inequality that is obtained by applying Equation (51) to every intermediate layer. Instead of only considering the input layer, all linear weights (including any convolution layers) are taken into account. Denote the input to the l-th linear layer as $\mathbf{x}_i^l$ for $l = 1, 2, \cdots, L$. In particular, $\mathbf{x}_i^1 = \mathbf{x}_i$ is the input of the entire network. Similarly, $\mathbf{W}_l$ is the weight matrix of l-th linear/convolutional layer. With a slight abuse of notation, we use f_l to denote the mapping from the input of the l-th layer to the final output. Then we define the Network Volumetric Ratio:

Definition H.4 *The* ***Network Volumetric Ratio (NVR)*** *is defined as the sum of the local volumetric ratios* dV_{f_l} *for all* f_l*, that is,* $dV_{net} = \sum_{l=1}^{L} dV_{f_l}$

Then we have the following inequality:

Proposition H.5 *The network volumetric ratio is upper bounded by a sharpness related quantity:*

$$\begin{aligned} \sum_{l=1}^{L} dV_{f_l} &\leq \frac{N^{-N/2}}{n} \sum_{l=1}^{L} \sum_{i=1}^{n} \|\nabla_{\mathbf{x}^l} f_i^l\|_F^N \\ &\leq \frac{1}{n} \sqrt{\sum_{l=1}^{L} \sum_{i=1}^{n} \frac{\|\mathbf{W}_l\|_2^{2N}}{\|\mathbf{x}_i^l\|_2^{2N}}} \cdot \left(\frac{nS(\boldsymbol{\theta}^*)}{N}\right)^{N/2} . \end{aligned} \tag{52}$$

Again, a detailed derivation of the above inequalities is given in Appendix D. Proposition H.3 and Proposition H.5 imply that flatter minima of the loss function in parameter space contribute to local compression of the data's representation manifold.

H.2.2. LOCAL DIMENSIONALITY IS TIED TO, BUT NOT BOUNDED BY, SHARPNESS

Now we introduce a local measure of dimensionality. Consider an input data point $\bar{\mathbf{x}}$ drawn from the training set: $\bar{\mathbf{x}} = \mathbf{x}_i$ for a specific $i \in \{1, \cdots, n\}$. Let the set of all possible perturbations around $\bar{\mathbf{x}}$ in the input space are samples from an isotropic normal distribution, $\mathcal{B}(\bar{\mathbf{x}})_\alpha \sim \mathcal{N}(\bar{\mathbf{x}}, \alpha\mathcal{I})$, where $C_{\mathcal{B}(\bar{\mathbf{x}})} = \alpha\mathcal{I}$, with $\mathcal{I}$ as the identity matrix, is the covariance matrix. We first propagate $\mathcal{B}(\bar{\mathbf{x}})_\alpha$ through the network transforming each point $\mathbf{x}$ into its corresponding image $f(\mathbf{x})$. Following a Taylor expansion for points within $\mathcal{B}(\bar{\mathbf{x}})_\alpha$ as $\alpha \to 0$ with high probability we have:

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + \nabla_{\mathbf{x}} f(\bar{\mathbf{x}}, \boldsymbol{\theta}^*)^T (\mathbf{x} - \bar{\mathbf{x}}) + O(\|\mathbf{x} - \bar{\mathbf{x}}\|^2) . \tag{53}$$

We can express the limit of the covariance matrix $C_{f(\mathcal{B}(\mathbf{x}))}$ of the output $f(\mathbf{x})$ as

$$C_f^{\lim} := \lim_{\alpha\to 0} \frac{C_{f(\mathcal{B}(\mathbf{x})_\alpha)}}{\alpha} = \nabla_{\mathbf{x}} f(\bar{\mathbf{x}}, \boldsymbol{\theta}^*) \nabla_{\mathbf{x}}^T f(\bar{\mathbf{x}}, \boldsymbol{\theta}^*) \,. \tag{54}$$

Our covariance expressions capture the distribution of the samples in $\mathcal{B}(\bar{\mathbf{x}})_\alpha$ as they go through the network $f(\bar{\mathbf{x}}, \boldsymbol{\theta}^*)$. The local Participation Ratio based on this covariance is given by:

$$D_{\text{PR}}(f(\bar{\mathbf{x}})) = \lim_{\alpha\to 0} \frac{\text{Tr}[C_{f(\mathcal{B}(\mathbf{x})_\alpha)}]^2}{\text{Tr}[(C_{f(\mathcal{B}(\mathbf{x})_\alpha)})^2]} = \frac{\text{Tr}[C_f^{\lim}]^2}{\text{Tr}[(C_f^{\lim})^2]} \tag{55}$$

(Recanatesi et al. 31, cf. nonlocal measures in Gao et al. 12, Litwin-Kumar et al. 24, Mazzucato et al. 26).

Definition H.6 *The **Local Dimensionality** of a network f is defined as the sample mean of local participation ratio at different input samples: $D(f) = \frac{1}{n}\sum_{i=1}^n D_{PR}(f(\mathbf{x}_i))$*

This quantity in some sense represents the sparseness of the eigenvalues of $C_f^{\lim}$: If we let $\boldsymbol{\lambda}$ be all the eigenvalues of $C_f^{\lim}$, then the local dimensionality can be written as $D_{\text{PR}} = (\|\boldsymbol{\lambda}\|_1/\|\boldsymbol{\lambda}\|_2)^2$, which attains its maximum value when all eigenvalues are equal to each other, and its minimum when all eigenvalues except for the leading one are zero. Note that the quantity retains the same value when $\boldsymbol{\lambda}$ is arbitrarily scaled. As a consequence, it is hard to find a relationship between the local dimensionality and the fundamental quantity on which our bounds are based: $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \boldsymbol{\theta}^*)\|_F^2$, which is $\|\boldsymbol{\lambda}\|_1$.

H.3. Correlation analysis

We empirically show how different metrics correlate with each other, and how these correlations can be predicted from our bounds. We train 100 VGG-11 networks with different batch sizes, learning rates, and random initialization to classify images from the CIFAR-10 dataset, and plot pairwise scatter plots between different quantities at the end of the training: local dimensionality, sharpness (square root of Equation (3)), log volume (Equation (49)), MLS (Equation (5)), NMLS (Equation (6)), generalization gap (gen gap), D (Equation (48)), bound (right-hand side of Equation (6)) and relative sharpness [30] (see Figure H.4). We only include CIFAR-10 data with 2 labels to ensure that the final training accuracy is close to 100%.

We repeat the analysis on MLPs and LeNets trained on the FashionMNIST dataset and the CIFAR-10 dataset (Figure H.5 and Figure H.6). We find that

1. The bound over NMLS, MLS, and NMLS introduced in Equation (6) and Equation (5) consistently correlates positively with the generalization gap.

2. Although the bound in Equation (51) is loose, log volume correlates well with sharpness and MLS.

3. Sharpness is positively correlated with the generalization gap, indicating that little reparametrization effect [8] is happening during training, i.e. the network weights do not change too much during training. This is consistent with observations in Ma and Ying [25].

4. The bound derived in Equation (6) correlates positively with NMLS in all experiments.

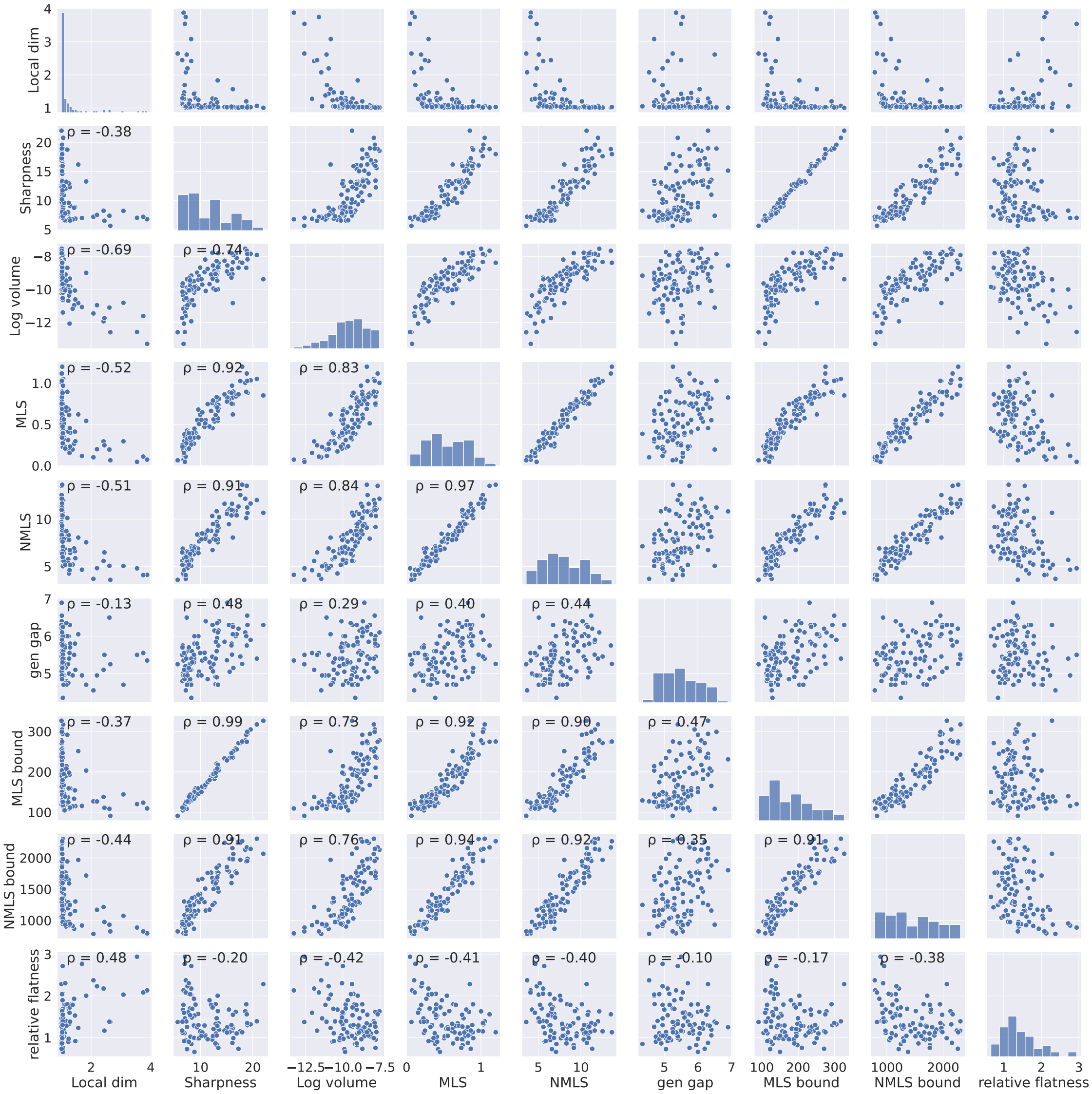

Figure H.4: **Pairwise correlation among different metrics.** We trained 100 different VGG-11 networks on the CIFAR-10 dataset using vanilla SGD with different learning rates, batch sizes, and random initializations and plot pairwise scatter plots between different quantities: local dimensionality, sharpness (square root of Equation (3)), log volume (Equation (49)), MLS (Equation (5)), NMLS (Equation (6)), generalization gap (gen gap), MLS bound (Proposition 2.2), NMLS bound (Proposition 2.4) and relative sharpness ([30]). The Pearson correlation coefficient ρ is shown in the top-left corner for each pair of quantities. See Appendix H.3 for a summary of the findings in this figure.

5. MLS that only consider the first layer weights can sometimes negatively correlate with the bound derived in Equation (5) (Figure H.5).

6. Relative flatness that only consider the last layer weights introduced in [30] shows weak (even negative) correlation with the generalization gap. Note that “relative flatness” is a misnomer that is easier understood as “relative *sharpness*”, and is supposed to be *positively* correlated with the generalization gap.

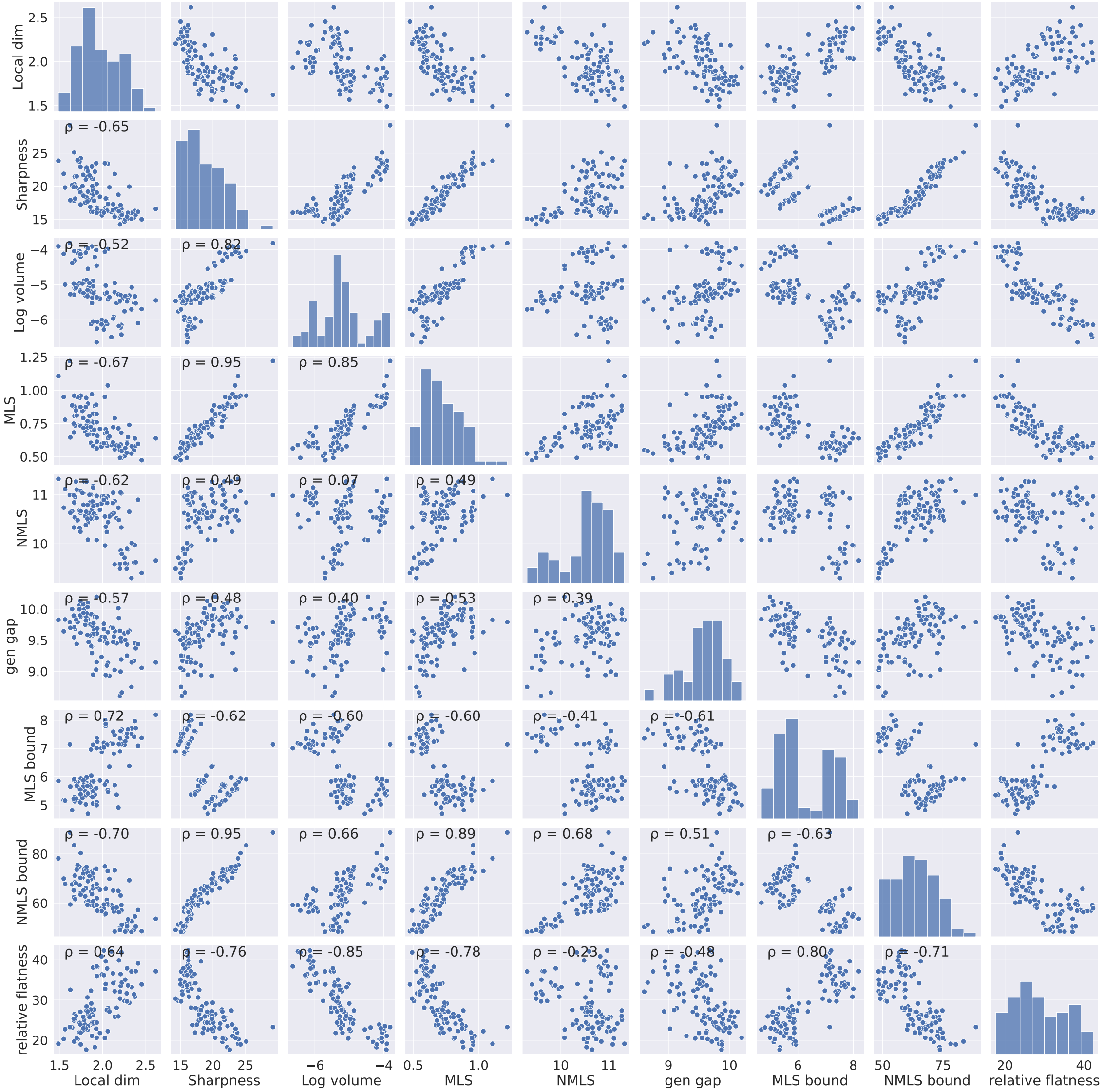

Figure H.5: **Pairwise correlation among different metrics.** We trained 100 different 4-layer MLPs on the FashionMNIST dataset using vanilla SGD with different learning rates, batch size, and random initializations and plot pairwise scatter plots between different quantities: local dimensionality, sharpness (square root of Equation (3)), log volume (Equation (49)), MLS (Equation (5)), NMLS (Equation (6)), generalization gap (gen gap), MLS bound (Proposition 2.2), NMLS bound (Proposition 2.4) and relative sharpness ([30]). The Pearson correlation coefficient ρ is shown in the top-left corner for each pair of quantities. See Appendix H.3 for a summary of the findings in this figure.

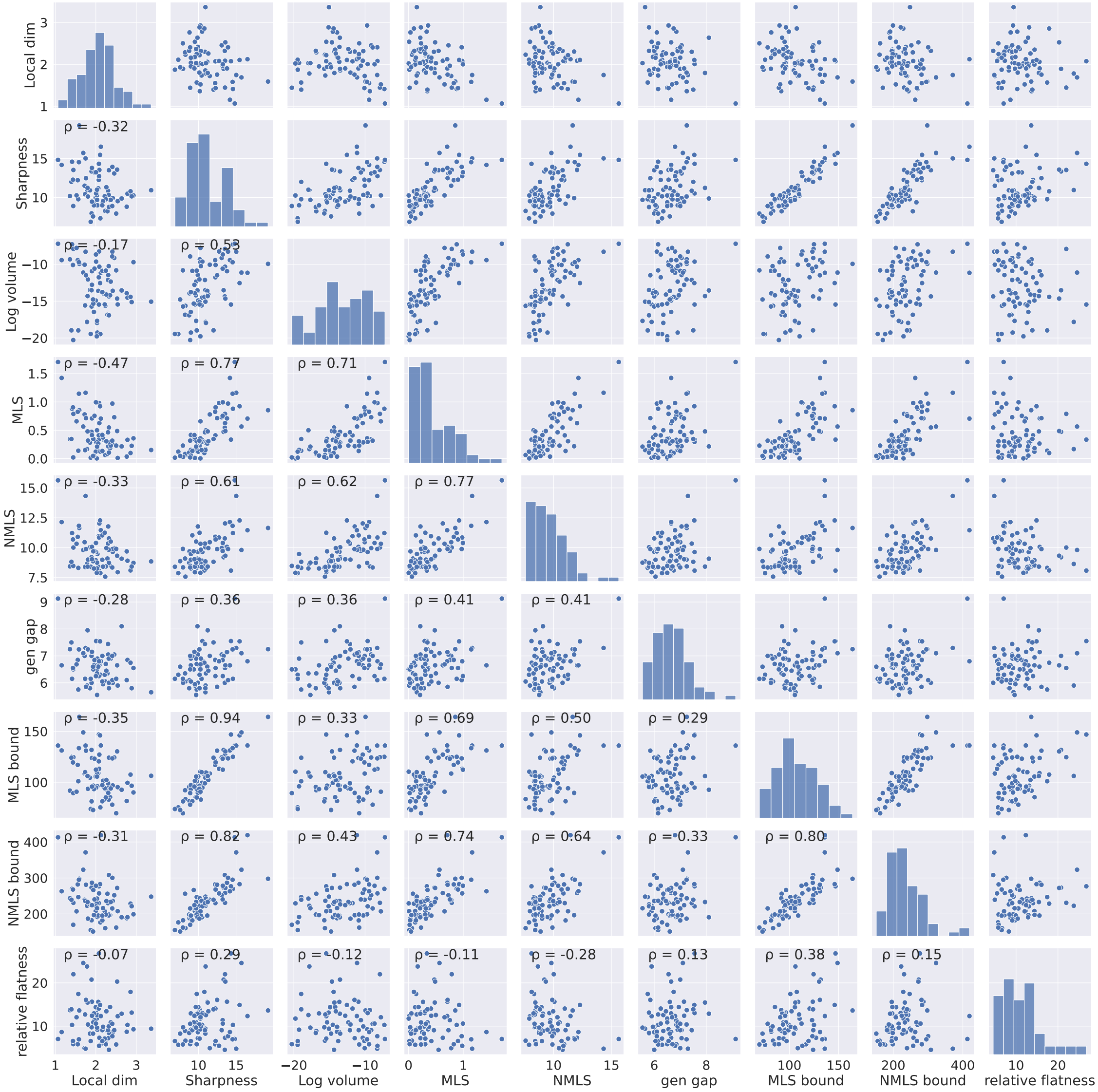

Figure H.6: **Pairwise correlation among different metrics.** We trained 100 different LeNets on the CIFAR-10 dataset using vanilla SGD with different learning rates, batch size, and random initializations and plot pairwise scatter plots between different quantities: local dimensionality, sharpness (square root of Equation (3)), log volume (Equation (49)), MLS (Equation (5)), NMLS (Equation (6)), generalization gap (gen gap), MLS bound (Proposition 2.2), NMLS bound (Proposition 2.4) and relative sharpness ([30]). The Pearson correlation coefficient ρ is shown in the top-left corner for each pair of quantities. See Appendix H.3 for a summary of the findings in this figure.

H.4. Empirical evidence in Vision Transformers (ViTs)

Since our theory applies to linear and convolutional layers as well as residual layers (Appendix B), relationships among sharpness and compression, as demonstrated above for VGG-11 and MLP networks, it should hold more generally in modern architectures such as the Vision Transformer (ViT) and its variants. However, naive ways of evaluating the quantities discussed in previous sections are computationally prohibitive. Instead, we look at the MLS normalized by the norm of the input and the elementwise-adaptive sharpness defined in Andriushchenko et al. [2], Kwon et al. [21]. Both of the metrics can be estimated efficiently for large networks. Specifically, in Figure H.7 we plot the normalized MLS against the elementwise-adaptive sharpness. The analytical relationship between the normalized MLS and the elementwise-adaptive sharpness and the details of the numerical approximation we used are given in Appendix F.2 and Appendix G respectively. For all the models, we attach a sigmoid layer to the output logits and use MSE loss to calculate the adaptive sharpness. Figure H.7 shows the results for 181 pretrained ViT models provided by the `timm` package [38]. We observe that there is a general trend that lower sharpness indeed implies lower MLS. However, there are also outlier clusters that with data corresponding to the same model class; an interesting future direction would be to understand the mechanisms driving this outlier behavior. Interestingly, we did not observe this correlation between unnormalized metrics, indicating that weight scales should be taken into account when comparing between different models.

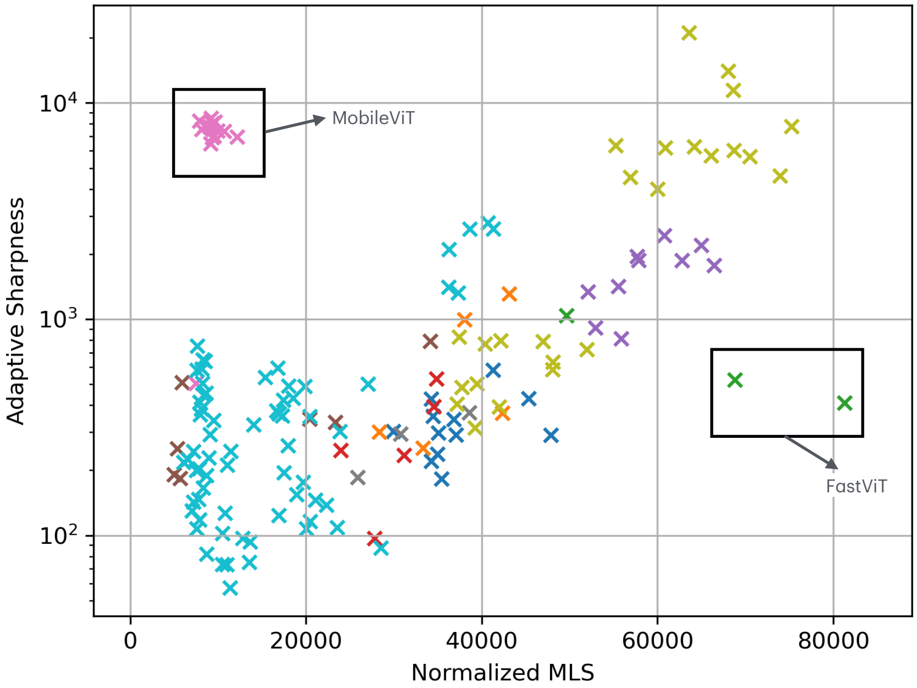

Figure H.7: Adaptive sharpness vs Normalized MLS for 181 ViT models and variants. Different colors represent different model classes. For most models, there is a positive correlation between Sharpness and MLS. However, outlier clusters also exist, for MobileViT [27] models in the upper left corner, and two FastViT [35] models in the lower right corner.

Appendix I. Additional experiments

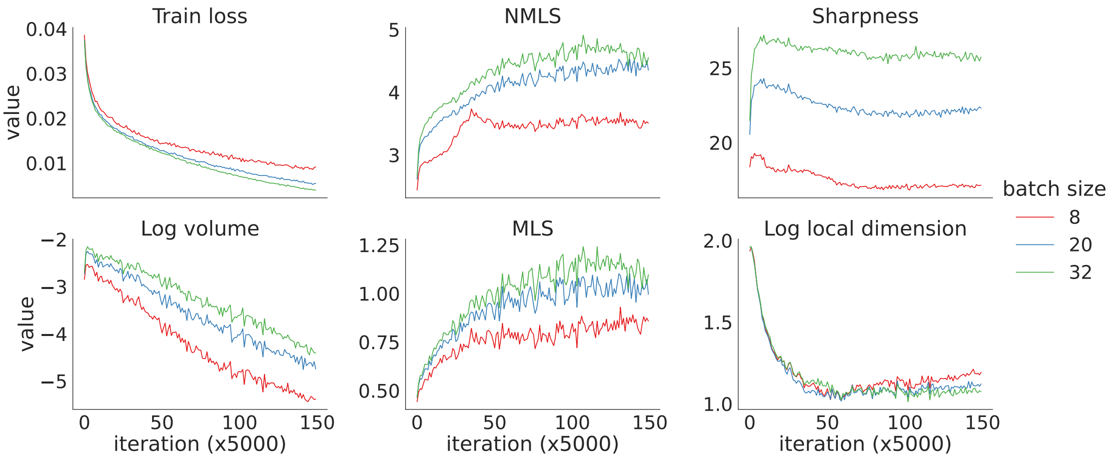

Figure I.8: Trends in key variables across SGD training of a 4-layer MLP with fixed learning rate (equal to 0.1) and varying batch size (8, 20, and 32). MLS/NMLS closely follows the trend of sharpness during the training. From left to right: train loss, NMLS, sharpness (square root of Equation (3)), log volumetric ratio (Equation (49)), MLS (Equation (5)), and local dimensionality of the network output (Equation (55)).

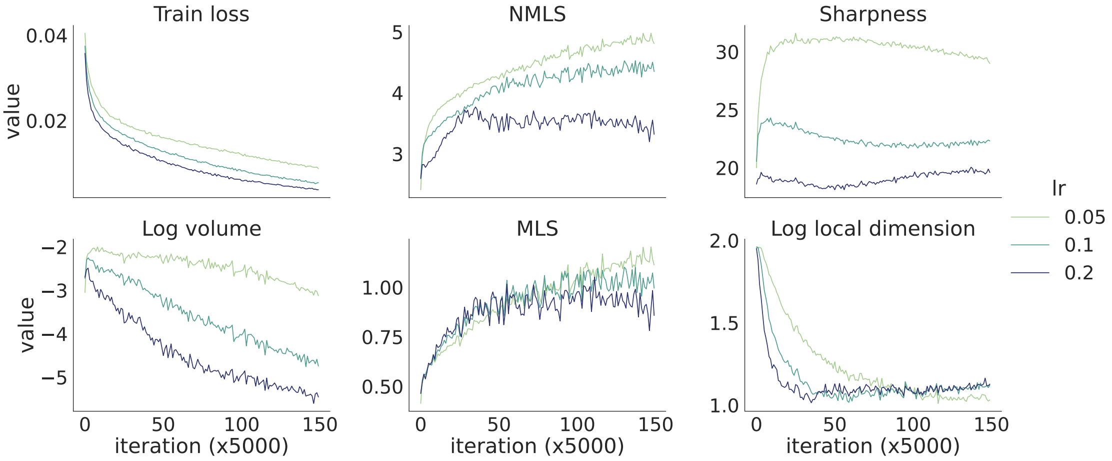

Figure I.9: Trends in key variables across SGD training of a 4-layer MLP with fixed batch size (equal to 20) and varying learning rates (0.05, 0.1 and 0.2). MLS/NMLS closely follows the trend of sharpness during the training. From left to right: train loss, NMLS, sharpness (square root of Equation (3)), log volumetric ratio (Equation (49)), MLS (Equation (5)), and local dimensionality of the network output (Equation (55)).

Appendix J. Sharpness and compression on test set data

Even though Equation (3) is exact for interpolation solutions only (i.e., those with zero loss), we found that the test loss is small enough (Figure J.10) so that it should be a good approximation for test data as well. Therefore we analyzed our simulations to study trends in sharpness and volume for these held-out test data as well (Figure J.10). We discovered that this sharpness increased rather than diminished as a result of training. We hypothesized that sharpness could correlate with the difficulty of classifying testing points. This was supported by the fact that the sharpness of misclassified test data was even greater than that of all test data. Again we see that MLS has the same trend as the sharpness. Despite this increase in sharpness, the volume followed the same pattern as the training set. This suggests that compression in representation space is a robust phenomenon that can be driven by additional phenomena beyond sharpness. Nevertheless, the compression still is weaker for misclassified test samples that have higher sharpness than other test samples. Overall, these results emphasize an interesting distinction between how sharpness evolves for training vs. test data.

Appendix K. Computational resources and code availability

All experiments can be run on one NVIDIA Quadro RTX 6000 GPU. The code will be released after acceptance.

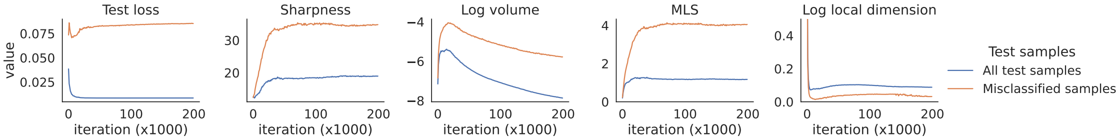

Figure J.10: Trends in key variables across SGD training of the VGG-11 network with fixed learning rate (equal to 0.1) and batch size (equal to 20) for samples of the test set. After the loss is minimized, we compute sharpness and volume on the test set. Moreover, the same quantities are computed separately over the entire test set or only on samples that are misclassified. In order from left to right in row-wise order: test loss, sharpness (Equation (2)), log volumetric ratio (Equation (49)), MLS, and local dimensionality of the network output (Equation (55)).