A Universal Class of Sharpness-Aware Minimization Algorithms

Behrooz Tahmasebi BZT@MIT.EDU
MIT CSAIL

Ashkan Soleymani ASHKANSO@MIT.EDU
MIT LIDS

Dara Bahri DBAHRI@GOOGLE.COM
Google DeepMind

Stefanie Jegelka STEJFE@MIT.EDU
TU Munich and MIT CSAIL

Patrick Jaillet JAILLET@MIT.EDU
MIT LIDS

Abstract

Recently, there has been a surge in interest in developing optimization algorithms for overparameterized models, as achieving generalization is believed to require algorithms with suitable biases. This interest centers on minimizing sharpness of the original loss function; the Sharpness-Aware Minimization (SAM) algorithm has proven effective. However, existing literature focuses on only a few sharpness measures (such as the maximum eigenvalue/trace of the training loss Hessian), which may not necessarily yield meaningful insights for non-convex optimization scenarios (e.g., neural networks). Moreover, many sharpness measures show sensitivity to parameter invariances in neural networks, e.g., they magnify significantly under rescaling parameters. Hence, here we introduce a new class of sharpness measures leading to sharpness-aware objective functions. We prove that these measures are *universally expressive*, allowing any function of the training loss Hessian matrix to be represented by choosing appropriate hyperparameters. Furthermore, we show that the proposed objective functions explicitly bias towards minimizing their corresponding sharpness measures. Finally, as an example of our proposed general framework, we present *Frob-SAM* and *Det-SAM*, which are specifically designed to minimize the Frobenius norm and the determinant of the Hessian of the training loss, respectively. We also demonstrate the advantages of our general framework through an extensive series of experiments.

1. Introduction

Understanding the generalization capabilities of overparameterized networks is a fundamental yet unsolved challenge in deep learning. It is postulated that achieving near-zero training loss alone may be insufficient as there exist many instances where global minima fail to exhibit satisfactory generalization performance. To this end, a dominant observation asserts that the characteristics of the loss landscape play a pivotal role in determining which parameters have low training loss while also exhibiting generalization capabilities. A recently proposed approach to consider the geometric aspects of the loss landscape, with the aim of achieving generalization, entails the avoidance of sharp minima. For example, the celebrated Sharpness-Aware Minimization (SAM) algorithm has shown enhancements in generalization across many practical tasks [18]. While the concept of sharpness lacks a precise definition in a general sense, people often introduce various measures to quantify it in practice [16]. Many sharpness measures in the literature rely on the second-order derivative characteristics of the training loss function, such as the trace or the operator norm of Hessian [13, 29].

Nevertheless, traditional methodologies for quantifying sharpness may not suffice to ensure generalization, given the intricate geometry of the loss landscape, which may necessitate different regularization techniques. Moreover, many existing sharpness measures, e.g., trace of Hessian, fail to encapsulate the genuine essence of sharpness in deep neural networks because the Hessian matrix no longer maintains positive semi-definiteness. Furthermore, neural networks exhibit parameter invariances, wherein different parameterizations can yield identical functions — such as scaling invariances in ReLU networks. Consequently, an effective measure of sharpness should remain invariant in the face of such parameter variations. Unfortunately, conventional approaches for quantifying sharpness frequently fall short in addressing this phenomenon. In Appendix C, we argue with illustrative examples why existing sharpness measures above fail to define a meaningful notion for overparameterized models.

Therefore, a fundamental question is: how can one succinctly represent all measures of sharpness within a parameterized framework that also enables meaningful applications to overparameterized models with parameter invariances? This question holds significance in applications as it allows *learning/designing the regularization* in cases where information about the geometry of the loss landscape or parameter invariances is provided, either empirically or through assumption. To the best of our knowledge, this question has remained fairly unexplored in the deep learning literature. In this paper, we characterize *all* sharpness measures (i.e., functions of the Hessian of the training loss) through an average-based parameterized representation. We prove that by changing the (hyper)parameters, the provided representation spans all the sharpness measures as a function of the Hessian matrix. In other words, it is provably a *universal representation*. We also provide quantitative results on the complexity of the sharpness representation as a function of the data dimension.

Moreover, attached to any representation of sharpness, we provide a new zeroth-order loss function, and we prove that the new loss function is *biased* toward minimizing its corresponding sharpness measure. Since the parameterized representation reduces to SAM (i.e., worst-direction) and average-direction sharpness measures [53] in special cases, it can be considered as a generalized (hyper)parameterized sharpness-aware minimization algorithm. This generalizes the recent study of the explicit bias of a few sharpness-aware minimization algorithms [53] to a comprehensive class of objectives. Furthermore, this allows us to readily design algorithms with *any* bias of interest, while to the best of our knowledge, only algorithms with biases towards minimizing the trace, operator norm of the Hessian matrix, and a few other sharpness measures are known in the literature. As instances of our proposed general algorithm, we present *Frob-SAM* and *Det-SAM*, two new sharpness-aware minimization algorithms that are specifically designed to minimize the Frobenius norm and the determinant of the Hessian of the training loss function, respectively.

In short, in this paper we make the following contributions:

- We propose a new class of sharpness measures, as function of the training loss Hessian. We prove that the new representation is *universally expressive*, meaning that it covers all sharpness measures of the Hessian as its (hyper)parameters change.
- Along with each sharpness measure we provide an optimization objective and prove that the new objective is explicitly *biased* toward minimizing the corresponding sharpness measure.
- We introduce two fundamental illustrative examples of our proposed general representation and the corresponding algorithms: *Frob-SAM* and *Det-SAM*. *Frob-SAM* is geared towards minimizing the Frobenius norm of the Hessian matrix, providing a meaningful and natural solution to the definition problem of sharpness for non-convex optimization problems. Conversely, *Det-SAM* is

focused on minimizing the determinant[1] of Hessian, addressing scale-invariant issues related to parameterization.

2. A New Class of Sharpness Measures and Algorithms

To define a new class of sharpness measures, we take a closer look at the average-based sharpness-aware loss $L_{\text{AVG}}(x)$ that is defined as a function of the training loss $L(x)$ at parameters $\mathbb{R}^d$; using its Taylor expansion [53], we have[2]

$$L_{\text{AVG}}(x) = \mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2})\Big] \approx\ L(x) + \rho\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[\langle\nabla L(x), \frac{v}{\|v\|_2}\rangle\Big]$$
$$+ \rho^2\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[\frac{v^t\nabla^2 L(x)v}{\|v\|_2^2}\Big] = L(x) + \rho^2\frac{\text{tr}(\nabla^2 L(x))}{d}.$$

This intuitively tells us that for a small perturbation parameter ρ, the leading term in the objective function is the training loss $L(x)$, and after we get close to the zero-loss manifold Γ, the leading term becomes $\frac{1}{d}\,\text{tr}(\nabla^2 L(x))$, which is exactly the explicit bias of the average-based sharpness-aware minimization objective [53]. This motivates us to define the following parameterized sharpness measure, as a function of the second-order Taylor expansion of the training loss. For the full definition, see Appendix D.

Definition 1 ((ϕ,ψ,μ)-sharpness measure) *For any continuous functions $\phi,\psi:\mathbb{R}\to\mathbb{R}$ and any (Borel) measure μ on $\mathbb{R}^d$, the (ϕ,ψ,μ)-sharpness measure $S(x;\phi,\psi,\mu)$ is defined as*

$$S(x;\phi,\psi,\mu) := \phi\Big(\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v)\Big). \tag{1}$$

In Appendix D, we extend this definition to having multi-dimensional functions ϕ,ψ (instead of being one-dimensional). We specify several examples of hyperparameters (ϕ,ψ,μ) in Table 1, which shows how (ϕ,ψ,μ)-sharpness measures can represent various notions of sharpness, as a function of the training loss Hessian matrix.

Table 1: Various (ϕ,ψ,μ)-sharpness measures (Appendix D; λ_i, $i\in[d]$, are eigenvalues of Hessian).

Hyperparameters				
$\phi(t)$	$\psi(t)$	m	μ	$S(x;\phi,\psi,\mu)$ (or bias)
t	t	1	Uniform($\mathbb{S}^{d-1}$)	$\frac{1}{2d}\,\text{tr}(\nabla^2 L(x)) = \frac{1}{2d}\sum_{i=1}^d \lambda_i$
$2(t_2-t_1^2)$	(t,t^2)	2	$\mathcal{N}(0,I_d)\otimes\mathcal{N}(0,I_d)$	$\sum_{i=1}^d \lambda_i^2 = \|\nabla^2 L(x)\|_F$
$(2\pi)^d/t^2$	$\exp(-t)$	1	Lebesgue measure on $\mathbb{R}^d$	$\det(\nabla^2 L(x)) = \prod_{i=1}^d \lambda_i$ r
$1/t^2$	$\exp(\sigma t)$	1	$\mathcal{N}(0,I_d)$	$\prod_{i=1}^d(1-\sigma\lambda_i)$

Expressive Power and Universality. By changing the hyperparameters, what functions of Hessian of training loss can be represented? In the paper, we show that our framework is *universal*, meaning that any function of Hessian of training loss is achievable by choosing appropriate ϕ,ψ, and μ.

Theorem 2 (Universality, informal, see Appendix E for more details) *For any continuous function of Hessian of training loss such as $S(\nabla^2 L(x))$, there exists continuous functions ϕ,ψ, and a probability measure μ such that $S(\nabla^2 L(x)) = S(x;\phi,\psi,\mu)$.*

1. To be more precise, the product of non-zero eigenvalues.
2. For more details, please see Appendix B

New Algorithms and Explicit Biases. Given the universality of the framework, the next question is how to build algorithms that (explicitly) minimize $S(x;\phi,\psi,\mu)$ along with the training loss function? Note that minimizing $S(x;\phi,\psi,\mu)$ (and even explicitly computing it) is impossible because it involves computing the Hessian of loss function that is expensive. We need to have an approximation of $S(x;\phi,\psi,\mu)$ that is easy to minimize, meaning that it only depends on the zeroth-order information about the loss function. We achieve this via Taylor series in Appendix H and we derived Algorithm 2. Moreover, we prove that this zeroth-order approximation allows being explicitly biased towards minimizing $S(x;\phi,\psi,\mu)$ along with the training loss.

Theorem 3 (Explicit bias, informal, see Appendix F for more details) *Minimizing the zeroth-order approximation of $S(x;\phi,\psi,\mu)$ given in Definition 8 corresponds to a sharpness-aware minimization algorithm that is explicitly biased towards minimizing the sharpness measure $S(x;\phi,\psi,\mu)$ over the zero-loss manifold.*

Parameter Invariances. We showed that the proposed framework is able to represent any function of Hessian of training loss, and it can also achieve any desired bias via a zeroth-order loss function. The next question is how to achieve sharpness measures that are invariant with respect to some arbitrary group actions (i.e., symmetries) on the parameter space? In other words, if we have $L(g.x) = L(x)$ for all parameters x and all group transformations $g \in G$, then when we have $S(g.x;\phi,\psi,\mu) = S(x;\phi,\psi,\mu)$? It turns out that the structure of the proposed framework allows an easy answer to this question.

Theorem 4 (Parameter invariance, informal, see Appendix G for more details) *If the measure μ is G-invariant, then the sharpness measure $S(x;\phi,\psi,\mu)$ is also G-invariant.*

3. Frobenius-SAM and Determinant-SAM

Now we can replace the specific hyperparameters in Table 1 to obtain sharpness-aware minimization algorithms that are explicitly biased towards minimizing the Frobenius norm or the determinant of the Hessian of the training loss function. In particular, for Frobenius norm, after replacing specific functions in Table 1 in Algorithm 2, we obtain the *Frob-SAM* algorithm (Algorithm 3; see also Appendix N for full details). Achieving *Det-SAM* is also similar, but the only difficulty is that it involves computing an integral with respect to the Lebesgue measure which can be challenging (Table 1). To address this issue, we instead sample a point from the hypercube $[-t,t]^d$ for a hyperparameter $t \in \mathbb{R}$ to approximate the Lebesgue measure.

4. Experiments

The goal of our experiments is twofold. Firstly, we show that minimization of the sharpness-aware loss has the explicit bias of minimizing the sharpness measure. Secondly, we show that our proposed methods are practical and effective on benchmark tasks.

MNIST. We address the first point by training a 6-layer ReLU network (each with 128 units) on MNIST using Adam with constant learning rate 0.001 for 20 epochs. Here, we focus our attention specifically on Frob-SAM, as we can reliably estimate its sharpness measure. Concretely, we estimate the Frobenius norm (squared) as $\frac{1}{n}\sum_{i=1}^{n}\|Hz_i\|_2^2$, where H is the Hessian and $z_i \sim \mathcal{N}(0, I_d)$; we adapt the PyHessian library [55]. Shown in Figure 1, for Frob-SAM, we observe a decrease in the sharpness measure with training and its magnitude is inversely related to strength of the regularization.

CIFAR10, CIFAR100, SVHN – Setup. The full details are deferred to Appendix O; we summarize them here. We train ResNet18 [21] on CIFAR10, CIFAR100, and SVHN using momentum-SGD

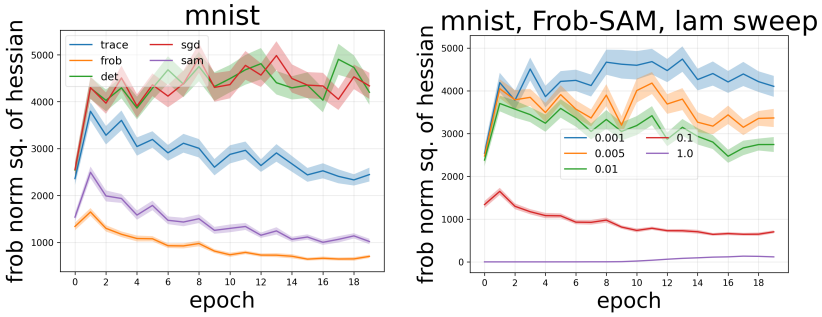

Figure 1: **Left**: Frobenius norm for Frob-SAM is smaller than for others, and it decreases monotonically with training. **Right**: Larger regularization strengths for Frob-SAM encourage lower norms. All settings except 1.0 achieve high (over 90%) test accuracy.

	CIFAR10	CIFAR10-S	CIFAR100	CIFAR100-S	SVHN	SVHN-S
Frob-SAM	93.83	61.82	75.90	29.91	96.35	88.42
Det-SAM	93.94	63.92	75.87	29.60	96.19	88.32
Trace-SAM	93.94	61.67	75.96	28.81	96.35	88.15
SAM	93.74	51.96	75.64	24.97	96.48	62.53
ASAM	94.33	62.00	75.76	23.46	96.22	85.35
SSAM	93.70	50.45	75.65	22.68	96.35	20.28
SGD	93.80	53.33	75.86	29.23	96.17	89.35

Table 2: Final test accuracy for both full and 10% sub-sampled datasets (denoted by -S). Frob-SAM and Det-SAM are competitive, especially when data is scarce. For example, for sub-sampled CIFAR10, Det-SAM has nearly 2% better test accuracy than the others.

and a multi-step learning rate schedule. To understand the efficacy of our methods in the *low data* setting, we additionally sub-sample each of the three datasets to only include the first 10% of training examples. We compare Frob-SAM and Det-SAM to the following baselines: **Trace-SAM** (We use one sample for the estimation of the regularizer and a ρ of 0.01 for all datasets.), **SAM** (We set ρ to 0.05/0.1/0.05 for CIFAR10/CIFAR100/SVHN, following Foret et al. [18]), **Adaptive SAM (ASAM)** (Kwon et al. [32] proposes a modification of SAM that is scale-invariant. We set ρ to 0.5/1/0.5 and η to 0.01/0.1/0.01 for CIFAR10/CIFAR100/SVHN.), **Sparse SAM (SSAM)** (Mi et al. [44] speeds up and improves the performance of SAM by only perturbing important parameters, as determined via Fisher information and sparse dynamic training. We use SSAM-F with ρ set to 0.1/0.2/0.1 for CIFAR10/CIFAR100/SVHN. 50% sparsity is used with 16 samples.).

CIFAR10, CIFAR100, SVHN – Results. Results are reported in Table 2. We find that for the full datasets, Frob-SAM and Det-SAM perform comparable to baselines. However, when only 10% of the training data is available, the regularization conferred by our methods can boost performance. For example, for sub-sampled CIFAR10, Det-SAM achieves nearly 2% better test accuracy than others. Overall, these findings suggest that the explicit biases we propose to optimize can be practically useful, particularly in low-data regimes. We leave it to future work to find even more effective applications.

Acknowledgments

The authors appreciate Joshua Robinson for his insightful comments and valuable suggestions. BT and SJ are supported by the Office of Naval Research award N00014-20-1-2023 (MURI ML-SCOPE), NSF award CCF-2112665 (TILOS AI Institute), NSF award 2134108, and the Alexander von Humboldt Foundation. AS and PJ are supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018).

References

[1] Atish Agarwala and Yann Dauphin. SAM operates far from home: eigenvalue regularization as a dynamical phenomenon. In *Int. Conference on Machine Learning (ICML)*, 2023.

[2] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Int. Conference on Machine Learning (ICML)*, 2022.

[3] Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. *arXiv preprint arXiv:2305.16292*, 2023.

[4] Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *Int. Conference on Machine Learning (ICML)*, 2023.

[5] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Int. Conference on Machine Learning (ICML)*, 2022.

[6] Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *Int. Conference on Learning Representations (ICLR)*, 2019.

[7] Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022.

[8] Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *arXiv preprint arXiv:2210.01513*, 2022.

[9] Kayhan Behdin and Rahul Mazumder. On statistical properties of sharpness-aware minimization: Provable guarantees. *arXiv preprint arXiv:2302.11836*, 2023.

[10] Kayhan Behdin, Qingquan Song, Aman Gupta, David Durfee, Ayan Acharya, Sathiya Keerthi, and Rahul Mazumder. Improved deep neural network generalization using m-sharpness-aware minimization. *arXiv preprint arXiv:2212.04343*, 2022.

[11] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on Learning Theory (COLT)*, 2020.

[12] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

[14] Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, and Aurelien Lucchi. An SDE for modeling sam: Theory and insights. In *Int. Conference on Machine Learning (ICML)*, 2023.

[15] Alex Damian, Tengyu Ma, and Jason D Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[16] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Int. Conference on Machine Learning (ICML)*, 2017.

[17] Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

[19] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Int. Conference on Machine Learning (ICML)*, 2018.

[20] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[23] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[24] Cheongjae Jang, Sungyoon Lee, Frank Park, and Yung-Kyun Noh. A reparametrization-invariant sharpness measure based on information geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[25] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *Int. Conference on Learning Representations (ICLR)*, 2019.

[26] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, 2019.

[27] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *Int. Conference on Learning Representations (ICLR)*, 2020.

[28] Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[29] 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. In *Int. Conference on Learning Representations (ICLR)*, 2017.

[30] Hoki Kim, Jinseong Park, Yujin Choi, Woojin Lee, and Jaewook Lee. Exploring the effect of multi-step ascent in sharpness-aware minimization. *arXiv preprint arXiv:2302.10181*, 2023.

[31] Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher sam: Information geometry and sharpness aware minimisation. In *Int. Conference on Machine Learning (ICML)*, 2022.

[32] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Int. Conference on Machine Learning (ICML)*, 2021.

[33] Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T Kiani. Implicit bias of linear equivariant networks. In *Int. Conference on Machine Learning (ICML)*, 2022.

[34] Thien Le and Stefanie Jegelka. Training invariances and the low-rank phenomenon: beyond linear networks. In *Int. Conference on Learning Representations (ICLR)*, 2022.

[35] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *Int. Conference on Learning Representations (ICLR)*, 2022.

[36] Zhiyuan Li, Tianhao Wang, and Dingli Yu. Fast mixing of stochastic gradient descent with normalization and weight decay. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[37] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *Int. Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[38] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[39] Yong Liu, Siqi Mai, Minhao Cheng, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Random sharpness-aware minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[40] Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *arXiv preprint arXiv:2309.12488*, 2023.

[41] Peng Lu, Ivan Kobyzev, Mehdi Rezagholizadeh, Ahmad Rashid, Ali Ghodsi, and Philippe Langlais. Improving generalization of pre-trained language models via stochastic weight averaging. *arXiv preprint arXiv:2212.05956*, 2022.

[42] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[43] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Int. Conference on Learning Representations (ICLR)*, 2018.

[44] Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[45] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[46] Atsushi Nitanda, Ryuhei Kikuchi, and Shugo Maeda. Parameter averaging for sgd stabilizes the implicit bias towards flat regions. *arXiv preprint arXiv:2302.09376*, 2023.

[47] Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. In *Int. Conference on Machine Learning (ICML)*, 2022.

[48] Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[49] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 2018.

[50] Hao Sun, Li Shen, Qihuang Zhong, Liang Ding, Shixiang Chen, Jingwei Sun, Jing Li, Guangzhong Sun, and Dacheng Tao. AdaSAM: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks. *arXiv preprint arXiv:2303.00565*, 2023.

[51] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *Int. Conference on Machine Learning (ICML)*, pages 9636–9647, 2020.

[52] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[53] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How does sharpness-aware minimization minimize sharpness? In *Int. Conference on Learning Representations (ICLR)*, 2023.

[54] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory (COLT)*, 2020.

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

[56] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Randomized sharpness-aware training for boosting computational efficiency in deep learning. *arXiv preprint arXiv:2203.09962*, 2022.

[57] Qihuang Zhong, Liang Ding, Li Shen, Peng Mi, Juhua Liu, Bo Du, and Dacheng Tao. Improving sharpness-aware minimization with fisher mask for better generalization on language models. *arXiv preprint arXiv:2210.05497*, 2022.

[58] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *Int. Conference on Learning Representations (ICLR)*, 2023.

[59] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *Int. Conference on Learning Representations (ICLR)*, 2022.

Appendix A. Related Work

Foret et al. [18] recently proposed the Sharpness-Aware Minimization (SAM) algorithm to avoid sharp minima. The SAM objective has connections to a similar robust optimization problem that was suggested for the study of adversarial attacks in deep learning [43]. Besides SAM, Nitanda et al. [46] show how parameter averaging for SGD is biased toward flatter minima. Label noise SGD also prefers flat minima [15]. Woodworth et al. [54] studied the role of sharpness in overparametrization from a kernel perspective. See Wang et al. [52] for the applications of flat minima for domain generalization (see also Cha et al. [12]). For applications of SAM in large language models, see Bahri et al. [7] (also Zhong et al. [57], and Qu et al. [47], Shi et al. [48] for federated learning). Besides those applications, Wen et al. [53] prove that current sharpness minimization algorithms sometimes fail to generalize for non-generalizing flattest models.

The (implicit) bias of many optimization algorithms and architectures has been studied, from the Gradient Descent (GD) [26, 49] to mirror descent [6, 19]; see also [20] for linear convolutional networks, and [33] for equivariant networks. Ji and Telgarsky [25] observed that linear neural networks are biased toward weight alignment for different layers (see [34] for non-linear networks). Andriushchenko and Flammarion [2] study implicit bias of SAM for diagonal linear networks, and Wen et al. [53] find the explicit bias of the Gaussian averaging method and other SAM variants.

The role of scale-invariance in generalization in deep learning is emphasized in Neyshabur et al. [45]. Dinh et al. [16] point out that parameter invariances can lead to the different parameterization of the same function, making the definition of flatness challenging; see also [4] for a recent study, and also [51] for a definition using PAC-Bayesian analysis. This motivates the study of sharpness measures that are invariant to such reparametrizations.

There have been a few attempts to address reparametrization problems with sharpness measures recently. Kwon et al. [32] proposed to adaptively calculate the sharpness in a normalized ball around the loss function to achieve scale invariance. However, their method is limited to scaling problems. Kim et al. [31] took a step further and introduced a new SAM algorithm by capturing the neighborhood of the parameters in an ellipsoid induced by the Fisher information. This way, the neighborhood becomes invariant with respect to the parameter invariances in the network. Jang et al. [24] defined an information geometric sharpness measure by investigating the eigenspaces of Fisher Information Matrix (FIM) of distribution parameterized by neural networks. They proved scale-invariance properties for their notion. Even though Kim et al. [31] and Jang et al. [24] enjoy some parameter invariance properties, (1) in practice, their methods are limited to classification tasks because of FIM calculation, (2) the underlying explicit biasing of their algorithms remains a mystery and is not guaranteed.

SAM can provide a strong regularization of the eigenvalues throughout the learning trajectory [1]. Bartlett et al. [8] show that the dynamics of SAM similar to GD on the spectral norm of Hessian. Compagnoni et al. [14] propose an SDE for modeling SAM, while Behdin and Mazumder [9] study the statistical benefits of SAM (see also [35] for a general framework for the dynamics of SGD around the zero-loss manifold). It is shown that SAM can reduce the feature rank (i.e., allowing learning low-rank features) [3]. Blanc et al. [11] proved that SGD is implicitly biased toward minimizing the trace of Hessian.

Kim et al. [30] propose a multi-step ascent approach to improve SAM, while Mi et al. [44] suggested sparsification of SAM. Zhuang et al. [59] improve SAM by changing the directions in the ascent step; their method is called Surrogate Gap Guided Sharpness-Aware Minimization (GSAM)

(see also Behdin et al. [10]). Random smoothing-based SAM (R-SAM) is another SAM variant that is proposed to reduce its computational complexity [39] (see also [17, 38, 50, 56] for more). Adaptive SAM (ASAM) is proposed for applying SAM on scale-invariant neural networks and has shown generalization benefits [32]. Li et al. [36] also prove that scale-invariant loss functions allow faster mixing in function spaces for neural networks. Lyu et al. [42] show how normalization can make GD reduce the sharpness via a continuous sharpness-reduction flow. Liang et al. [37] propose a capacity measure based on information geometry for parameter invariances in overparameterized models (for more on information geometry, see [24, 31]). Jiang et al. [27] empirically compare different complexity measures for overparameterized models. Keskar et al. [29] show how a large batch yields sharp minima but a small batch achieves flat minima.

Stochastic Weight Averaging (SWA) is another way to improve generalization and it relies on finding wider minima by averaging multiple points along the trajectory of SGD [23]. (see e.g., [41] which uses this method for language models). See [28] for the empirical comparison between two popular flat-minima optimization approaches: SWA and SAM. It is also worth mentioning that neural networks trained with large learning rates often generalize better (the edge-of-stability regime); see [5, 40, 58] for the theoretical understanding of this phenomenon.

Appendix B. Settings

Consider a standard learning setup with a labeled dataset $\mathcal{S}$, and a training loss function $L : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, where $L(x)$ denotes the training loss over $\mathcal{S}$ computed for the parameters $x \in \mathbb{R}^d$. The main objective in Empirical Risk Minimization (ERM) is to minimize the training loss $L(x)$ over the feasibility set $\mathcal{X} \subseteq \mathbb{R}^d$. However, achieving parameters satisfying $L(x) \approx 0$ in overparameterized models is often straightforward. This is because in contrast to other models, in overparameterized models, there are *many* global minima, i.e., the set $\Gamma := \{x \in \mathcal{X} : L(x) = 0\}$ is a manifold – it is called the *zero-loss manifold* in the literature. Moreover, in practical scenarios, it is noteworthy that not all global minima exhibit favorable generalization capabilities [18].

Appendix C. Motivation

It is hypothesized that the avoidance of sharp minima can enhance generalization performance [22, 23, 29]. However, it should be noted that the concept of sharpness encompasses a multitude of distinct definitions in practical contexts.

C.1. Background on SAM

The Sharpness-Aware Minimization (SAM) algorithm [18] suggests minimizing the training loss function over a small ball around the parameters:

$$\min_{x \in \mathcal{X}} \Big\{ L_{\text{SAM}}(x) := \max_{\|v\|_2 \leq 1} L(x + \rho v) \Big\},$$

where $\rho \in \mathbb{R}_{\geq 0}$ is the *perturbation parameter*. Note that L_{SAM} can be decomposed into two terms:

$$L_{\text{SAM}}(x) = \underbrace{L(x)}_{\text{empirical loss}} + \underbrace{\max_{\|v\|_2 \leq 1} \big\{ L(x + \rho v) - L(x) \big\}}_{\text{sharpness}}.$$

Foret et al. [18] also suggest alternative average-based sharpness-aware objectives to use PAC bounds on the generalization error of overparameterized models; we follow the definition in [53]:

$$\begin{aligned} L_{\text{AVG}}(x) &:= \mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2})\Big] \\ &= \underbrace{L(x)}_{\text{empirical loss}} + \underbrace{\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2})-L(x)\Big]}_{\text{sharpness}}. \end{aligned}$$

Wen et al. [53] recently proved that minimizing $L_{\text{SAM}}(x)$ will lead to global minima (i.e., $L(x)\approx 0$) with small $\lambda_{\max}(\nabla^2 L(x))$. In other words, SAM is (explicitly) biased towards minimizing $\lambda_{\max}(\nabla^2 L(x))$. Moreover, they show that using $L_{\text{AVG}}(x)$ biases towards minimizing $\frac{1}{d}\operatorname{tr}(\nabla^2 L(x))$. This means that SAM measures the sharpness of a global minimum by $\lambda_{\max}(\nabla^2 L(x))$, while the average-based objective uses $\frac{1}{d}\operatorname{tr}(\nabla^2 L(x))$ to evaluate it.

C.2. Motivating Examples

In the next examples, we argue how both sharpness measures above fail to define a meaningful notion for overparameterized models. In Example 2, a special case of problem with parameter invariances, i.e., under parameter rescalings is discussed.

Example 1 *The sharpness measures $\lambda_{\max}(\nabla^2 L(x))$, $\lambda_{\min}(\nabla^2 L(x))$ and $\operatorname{tr}(\nabla^2 L(x))$ are conceptually meaningful when the objective function $L(x)$ is convex, therefore λ_is are nonnegative. However, the Loss landscape of neural networks is highly nonconvex, and as a result, λ_i can be potentially negative. Consider the toy non-convex example of $L(x_1,x_2)=\frac{1}{2}(x_1^2-x_2^2)$,*

$$\nabla^2 L = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{2}$$

For all $x_1,x_2\in\mathbb{R}$, we know that $\operatorname{tr}(\nabla^2 L)=1+(-1)=0$, which in the Trace measure of sharpness it suggests that all the points $x_1,x_2\in\mathbb{R}$ are equally flat. Are these sharpness notions really capturing the intended concepts? For a better illustration, consider the plot of this function provided in Figure 2. This problem extends to other existing notions.

Example 2 *Consider the loss function $L(x_1,x_2)=x_1^2x_2^2-2x_1x_2+1$ with two parameters $x_1,x_2\in\mathbb{R}$. It is scale-invariant, i.e., $L(kx_1,\frac{x_2}{k})=L(x_1,x_2)$ for all $k\neq 0$. Indeed, the zero-loss manifold $\Gamma=\{(x_1,x_2):x_1x_2=1\}$ contains infinitely many global minima. Straightforward calculation shows $\nabla^2 L(x_1,x_2)=\begin{pmatrix} 2x_2^2 & 4x_1x_2-2 \\ 4x_1x_2-2 & 2x_1^2\end{pmatrix}$. Thus, we have $\frac{1}{2}\operatorname{tr}(\nabla^2 L(x_1,x_2))=x_1^2+x_2^2$. After rescaling, we get $\frac{1}{2}\operatorname{tr}(\nabla^2 L(x_1,x_2))\Big|_{(kx_1,k^{-1}x_2)}=k^2x_1^2+\frac{x_2^2}{k^2}\neq\frac{1}{2}\operatorname{tr}(\nabla^2 L(x_1,x_2))$. Therefore, as a sharpness measure, $\operatorname{tr}(\nabla^2 L(x_1,x_2))$ is not scale-invariant. The problem magnifies in the limit: $\lim_{k\to\infty}\operatorname{tr}(\nabla^2 L(x_1,x_2))\Big|_{(kx_1,k^{-1}x_2)}=\infty$. Similar problems exist for $\lambda_{\max}(\nabla^2 L(x_1,x_2))$. However, $\det(\nabla^2 L(x_1,x_2))$ is scale-invariant; we have $\det(\nabla^2 L(x_1,x_2))\Big|_{(kx_1,k^{-1}x_2)}=\det(\nabla^2 L(x_1,x_2))$ for all $k\neq 0$.*

Note that neural networks are often scale-invariant, e.g., linear networks or ReLU networks after scaling up the parameters of one hidden layer and scaling down the parameters of another hidden layer encode the same functions.

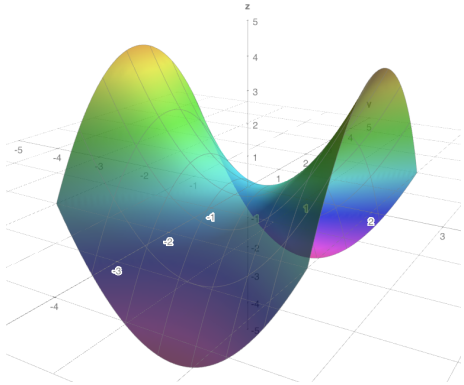

Figure 2: The loss landscape of the non-convex objective function $L(x_1, x_2) = \frac{1}{2}(x_1^2 - x_2^2)$. This examples shows how existing sharpness measures fall short of capturing sharpness meaning in non-convex settings. In particular, for all points $(x_1, x_2) \in \mathbb{R}^2$, $\text{tr}(\nabla^2 L(x_1, x_2)) = 1 + (-1) = 0$.

Appendix D. Definition and Examples of (ϕ, ψ, μ)-Sharpness Measures

Let us first review the definition of the (ϕ, ψ, μ)-Sharpness Measures.

Definition 5 ((ϕ, ψ, μ)-sharpness measure) *For any continuous functions $\phi, \psi : \mathbb{R} \to \mathbb{R}$ and any (Borel) measure μ on $\mathbb{R}^d$, the (ϕ, ψ, μ)-sharpness measure $S(x; \phi, \psi, \mu)$ is defined as*

$$S(x; \phi, \psi, \mu) := \phi\Big(\int \psi\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v) \Big). \tag{3}$$

Similarly, one can consider continuous functions $\psi : \mathbb{R} \to \mathbb{R}^m$ and $\phi : \mathbb{R}^m \to \mathbb{R}$, for some positive integer $m \geq 1$, and (Borel) measures μ_ℓ, $\ell \in [m]$, and define

$$S(x; \phi, \psi, \mu) := \phi\Big(\int \psi_1\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu_1(v), \tag{4}$$

$$\int \psi_2\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu_2(v), \tag{5}$$

$$\dots, \tag{6}$$

$$\int \psi_m\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu_m(v) \Big), \tag{7}$$

where we use $\mu := \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_m$ for the sake of brevity in our notation, and $\psi = (\psi_1, \psi_2, \ldots, \psi_m)^t$.

We specify several examples of hyperparameters (ϕ, ψ, μ) in Table 1, which shows how (ϕ, ψ, μ)-sharpness measures can represent various notions of sharpness, as a function of the training loss Hessian matrix.

In this section, we prove various notions of sharpness can be achieved using the proposed approach in this paper (Table 1). For the last row of Table 1, we refer the reader to the proof of Theorem 6.

- **Trace.** Let $\phi(t) = \psi(t) = t$, and note that

$$S(x;\phi,\psi,\mu) = \int \frac{1}{2} v^t \nabla^2 L(x) v d\mu(v) \tag{8}$$

$$= \frac{1}{2}\mathbb{E}_{v\sim\mu}[v^t \nabla^2 L(x) v], \tag{9}$$

where μ is the uniform distribution over the $(d-1)$-sphere $\mathbb{S}^{(d-1)} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. Denote the entries of $\nabla^2 L(x)$ as $(\nabla^2 L(x))_{i,j}$. Then, by the linearity of expectation

$$\mathbb{E}_{v\sim\mu}[v^t \nabla^2 L(x)] = \sum_{i,j=1}^{d} (\nabla^2 L(x))_{i,j} \mathbb{E}[v_i v_j] = \sum_{i=1}^{d} \frac{1}{d} (\nabla^2 L(x))_{i,i} = \frac{1}{d} \operatorname{tr}(\nabla^2 L(x)), \tag{10}$$

since $\mathbb{E}[v_i v_j] = \frac{1}{d}\delta_{i,j}$, where $\delta_{i,j}$ denotes the Knocker delta function.

- **Determinant.** To achieve the determinant, we choose $\phi(t) = (2\pi)^d/t^2$ and $\psi(t) = \exp(-t)$. Then,

$$S(x;\phi,\psi,\mu) = (2\pi)^d \Big(\int \exp(-\frac{1}{2} v^t \nabla^2 L(x) v) dv \Big)^{-2}, \tag{11}$$

where dv denotes the Lebesgue measure. However, using the multivariate Gaussian integral, we have

$$\int \exp\Big(-\frac{1}{2} v^t \nabla^2 L(x) v \Big) dv = (2\pi)^{d/2} \det(\nabla^2 L(x))^{-1/2}. \tag{12}$$

Replacing this intro the definition of $S(x;\phi,\psi,\mu)$ gives the desired result.

- **Polynomials of eigenvalues.** First assume that $\psi(t) = t^n$ for some $n \geq 0$. Then, for any function $\phi(t)$,

$$S(x;\phi,\psi,\mu) = \phi\Big(\int (\frac{1}{2} v^t \nabla^2 L(x) v)^n d\mu(v) \Big) \tag{13}$$

$$= \phi\Big(\mathbb{E}_{v\sim\mu}\Big[(\frac{1}{2} v^t \nabla^2 L(x) v)^n \Big] \Big), \tag{14}$$

where μ is the uniform distribution over the $(d-1)$-sphere $\mathbb{S}^{(d-1)}$. Since $\nabla^2 L(x)$ is a symmetric matrix, we can find an orthogonal matrix Q such that $\nabla^2 L(x) = Q^t D Q$, where D is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_d$. Now we write $\big(v^t \nabla^2 L(x) v\big)^n = \big(v^t Q^t D Q v\big)^n$. But Qv is distributed uniformly over the $(d-1)$-sphere $\mathbb{S}^{(d-1)}$, similar to v. Thus, we conclude

$$S(x;\phi,\psi,\mu) = \phi\Big(\mathbb{E}_{v\sim\mu}\Big[(\frac{1}{2} v^t \nabla^2 L(x) v)^n \Big] \Big) \tag{15}$$

$$= \phi\Big(\mathbb{E}_{v\sim\mu}\Big[(\frac{1}{2} \sum_{i=1}^{d} \lambda_i v_i^2)^n \Big] \Big). \tag{16}$$

Define

$$q(\lambda_1, \lambda_2, \ldots, \lambda_d) := \mathbb{E}_{v\sim\mu}\Big[\big(\sum_{i=1}^{d} \frac{1}{2}\lambda_i v_i^2\big)^n\Big], \tag{17}$$

which is clearly a polynomial function (by the linearity of expectation).

Note that the above computation is still valid if we replace the uniform distribution on hypersphere with the Gaussian multivariate distribution with identity covariance $\mathcal{N}(0, I_d)$. Indeed, let us compute this polynomial for $n = 2$ with Gaussian distribution. Note that

$$q(\lambda_1, \lambda_2, \ldots, \lambda_d) = \mathbb{E}_{v\sim\mu}\Big[\big(\sum_{i=1}^{d} \frac{1}{2}\lambda_i v_i^2\big)^2\Big] = \frac{1}{4}\sum_{i=1}^{d} \lambda_i^2 \mathbb{E}[Z^4] + \sum_{i\neq j} \frac{1}{4}\lambda_i\lambda_j(\mathbb{E}[Z^2])^2 \tag{18}$$

$$= \frac{3}{4}\sum_{i=1}^{d} \lambda_i^2 + \frac{1}{4}\sum_{i\neq j} \lambda_i\lambda_j, \tag{19}$$

where Z is a zero-mean Gaussian random variable with unit variance, and note that $E[Z^2] = 1$ and $E[Z^4] = 3$.

Now if we take $m = 2$, and $\psi(t) = (t, t^2)$, with $\mu = \mathcal{N}(0, I_d) \otimes \mathcal{N}(0, I_d)$, we have that

$$\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v) = \Big(\frac{1}{2}\sum_{i=1}^{d} \lambda_i, \frac{3}{4}\sum_{i=1}^{d} \lambda_i^2 + \frac{1}{4}\sum_{i\neq j} \lambda_i\lambda_j\Big). \tag{20}$$

Finally, by taking $\phi(t_1, t_2) = 2(t_2 - t_1^2)$, we obtain

$$\phi\Big(\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v)\Big) = \sum_{i=1}^{d} \lambda_i^2. \tag{21}$$

Appendix E. Expressive Power and Universality

In this section, we prove that the proposed class of sharpness measures is *universal*. In other words, for any continuous function $S : \mathbb{R}^d \to \mathbb{R}$, we specify continuous functions ϕ, ψ and a (Borel) probability measure[3] μ on $\mathbb{R}^d$ such that $S(\lambda_1, \lambda_2 \ldots, \lambda_d) = S(x; \phi, \psi, \mu)$, where λ_i, $i \in [d]$, are the eigenvalues of the Hessian matrix $\nabla^2 L(x)$.

Theorem 6 (Universality of the (ϕ, ψ, μ)-sharpness measures for functions of Hessian eigenvalues) *Let $\mathcal{A} \subseteq \mathbb{R}^d$ be a compact set. For any continuous function $S : \mathcal{A} \to \mathbb{R}$, there exist a product (Borel) probability measure μ, a positive integer $m \leq d$, and continuous functions $\phi : \mathbb{R}^m \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^m$, such that $S(\lambda_1, \lambda_2 \ldots, \lambda_d) = S(x; \phi, \psi, \mu)$ for any $x \in \mathcal{A}$, where λ_i, $i \in [d]$, are the eigenvalues of the Hessian matrix $\nabla^2 L(x)$.*

3. We indeed prove that Borel probability measures (as a subset of arbitrary Borel measures) are enough to achieve universality.

We present the proof of Theorem 6 in Appendix I. Note that to achieve universality, we need the functions ϕ, ψ to be of dimension $m = d$. However, as one can see in Table 1, many celebrated sharpness measures can indeed be represented using only small m. We believe that practically small hyperparameter m is enough, as it is motivated from the measures in Table 1.

While we proved the universality of the proposed class of sharpness measures for continuous functions of the Hessian eigenvalues, one may be interested in measuring sharpness with more information about the loss Hessian (e.g., the eigenvectors of the loss Hessian). The following theorem proves the universality for this class of arbitrary functions.

Theorem 7 (Universality of the (ϕ, ψ, μ)-sharpness measures for arbitrary functions of Hessian) *For any continuous function $S : \mathbb{R}^{d\times d} \to \mathbb{R}$, there exist a positive integer $m \leq d(d+1)/2$, (Borel) probability measures μ_ℓ, $\ell \in [m]$, and continuous functions $\phi : \mathbb{R}^m \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^m$, such that $S(\nabla^2 L(x)) = S(x; \phi, \psi, \mu)$ for any $x \in \mathbb{R}^d$, where $\mu := \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_m$ is a product probability measure.*

We present the proof of Theorem 7 in Appendix J. Note that arbitrary functions of the Hessian matrix can be quite hard to compute, e.g., consider the permanent of the Hessian matrix. Moreover, the dimension m must be quite large to allow us to prove the universality in overparameterized models (for d of considerable size), since the generality bound scales as $\mathcal{O}(d^2)$. Nevertheless, in practice, only small m allows to cover many interesting cases.

Appendix F. Explicit Bias

Now that we defined a flexible set of sharpness measures and we proved that it is universally expressive, the following question arises: how can one achieve $S(x; \phi, \psi, \mu)$ as the explicit bias of an objective function that only relies on the zeroth-order information about the training loss, similar to $L_{\text{SAM}}(x)$ and $L_{\text{AVG}}(x)$? To answer this question, we introduce the (ϕ, ψ, μ)-sharpness-aware loss function as follows.

Definition 8 *The (ϕ, ψ, μ)-sharpness-aware loss function*

$$L_{(\phi,\psi,\mu)}(x) := \underbrace{L(x)}_{\text{empirical loss}} + \rho^2 \underbrace{\phi\Big(\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big) d\mu(v)\Big)}_{:= R_\rho(x) \text{ sharpness}} = L(x) + \rho^2 R_\rho(x),$$

where ρ is the perturbation parameter and $R_\rho(x)$ denotes the sharpness regularizer.

Extending this definition to the cases with $m > 1$ is straightforward.

In the above definition, the new regularizer $R_\rho(x)$ is an approximation of the sharpness measure $S(x; \phi, \psi, \mu)$ as $\rho \to 0^+$. As a result, it is expected that minimizing $L_{(\phi,\psi,\mu)}(x)$ lead to minimizing the training loss as well as the sharpness measure $S(x; \phi, \psi, \mu)$. The next theorem formalizes this intuitive observation via characterizing the explicit bias of minimizing the sharpness-aware loss function $L_{(\phi,\psi,\mu)}(x)$.

Theorem 9 (Explicit bias of the (ϕ, ψ, μ)-sharpness-aware loss function) *Given a triplet (ϕ, ψ, μ), $m \geq 1$, and a training loss function $L : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, assume that:*

- *$L(x)$ is third-order continuously differentiable and satisfies the following upper bound*

$$\max_{i,j,k\in\{1,2,3\}} |\partial_i\partial_j\partial_k L(v)| = O(\|v\|^{-1}), \tag{22}$$

  for $v \in \mathbb{R}^d$ as $\|v\|_2 \to \infty$.

- *The two functions ϕ, ψ are continuously differentiable.*

- *For some $C > \max_{x\in\mathcal{X}} \max_{i\in[d]} |\lambda_i(\nabla^2 L(x))|$, we have $\int \|v\|_2^2 \tilde{\psi}_i(v) d\mu(v) < \infty$, $i \in [m]$, where*

$$\tilde{\psi}_i(v) := \max_{|t|\le\|v\|_2} |\psi_i'(Ct^2)|. \tag{23}$$

Then, there exists an open neighborhood $U \supsetneq \Gamma$, where Γ is the zero-loss manifold, for connected U and Γ, such that if for some $u \in U$, one has

$$L(u) + \rho^2 R_\rho(u) - \inf_{x\in U} \Big(L(x) + \rho^2 R_\rho(x) \Big) \le \Delta\rho^2, \tag{24}$$

with some optimally gap $\Delta > 0$, then

$$L(u) \le \underbrace{\inf_{x\in U} L(x)}_{=0} + (\Delta + o_\rho(1))\rho^2, \tag{25}$$

and also

$$S(u;\phi,\psi,\mu) \le \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) + \Delta + o_\rho(1). \tag{26}$$

We present the proof of Theorem 9 in Appendix K. The above theorem shows how using the new objective function $L_{(\phi,\psi,\mu)}(x)$ leads to explicitly biased optimization algorithms towards minimizing the sharpness measure $S(x;\phi,\psi,\mu)$ over the zero-loss manifold Γ. Indeed, it proves that if we are close to the zero-loss manifold (i.e., $u \in U$ for some open neighborhood $U \supsetneq \Gamma$), and also $L_{(\phi,\psi,\mu)}(u)$ is close to its global minimum over U, then (1) the training loss function $L(u)$ is close to zero, and (2) the corresponding sharpness measure $S(u;\phi,\psi,\mu)$ is close to its global minimum over the zero-loss manifold, with respect to an optimality gap Δ.

Appendix G. Invariant Sharpness-Aware Minimization

For which hyperparameters (ϕ,ψ,μ) is the corresponding sharpness measure scale-invariant? The following theorem answers this question.

Theorem 10 (Scale-invariant (ϕ,ψ,μ)-sharpness measures) *Consider a scale-invariant loss function $L(x)$ and let μ be a Borel measure of the form*

$$d\mu(x) = f\Big(\prod_{i=1}^d x_i\Big) \prod_{i=1}^d dx_i, \tag{27}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a measurable function. Then, for any continuous functions ϕ, ψ, the corresponding sharpness measure $S(x;\phi,\psi,\mu)$ is scale-invariant; this means that $S(x;\phi,\psi,\mu) = S(Dx;\phi,\psi,\mu)$ for any diagonal matrix $D \in \mathbb{R}^{d\times d}$ with $\det(D) = 1$.

We present the proof of Theorem 10 in Appendix L.

Example 3 *Note that* $\det(\nabla^2 L(x))$ *is a scale-invariant sharpness measure; for any diagonal matrix* $D \in \mathbb{R}^{d\times d}$ *with* $\det(D) = 1$,

$$\begin{aligned}\det(\nabla^2 L(x))\Big|_{Dx} &= \det(D^{-1}\nabla^2 L(x) D^{-1}) \\ &= \det(D^{-1})^2 \det(\nabla^2 L(x)) = \det(\nabla^2 L(x)).\end{aligned}$$

Note that Theorem 10 also supports the scale-invariance of the determinant; the Lebesgue measure satisfies the condition in Theorem 10 with $f \equiv 1$*, and we have the representation of the determinant in Table 1.*

While in Theorem 10 we only considered scale-invariances, one can generalize it to a general class of parameter invariances in the following theorem.

Theorem 11 (General parameter-invariant (ϕ, ψ, μ)**-sharpness measures)** *Let* G *be a group acting by matrices on* $\mathbb{R}^d$*, and assume that* $L(x)$ *is invariant with respect to the action of* G*. Then, for any* G*-invariant (Borel) measure* μ*, and any continuous functions* ϕ, ψ*, the corresponding sharpness measure* $S(x; \phi, \psi, \mu)$ *is* G*-invariant; this means that* $S(x; \phi, \psi, \mu) = S(A_g x; \phi, \psi, \mu)$ *for any matrix* $A_g \in \mathbb{R}^{d\times d}$ *corresponding to the action of an element* $g \in G$*.*

The proof of this theorem is analogous to Theorem 10 and is deferred to Appendix M. Thus, the strategy to create sharpness measures invariant to any group action G is simply to choose a group action invariant measure μ. Now, for any family of choices of functions ϕ and ψ, we obtain a family of G-invariant sharpness measures. Consequently, there is a family of G-invariant Sharpness-Aware Minimization algorithms, as explained in Appendix H.

Appendix H. (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm

In this section, we present the pseudocode for the (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm (see Algorithm 1). For simplicity, we present the algorithm for the full-batch gradient descent, and first assume that $m = 1$. The idea is to apply (stochastic) gradient decent or other optimization algorithms on the (ϕ, ψ, μ)-sharpness-aware loss function defined in Theorem 8,

$$L_{(\phi,\psi,\mu)} = L(x) + \rho^2 R_\rho(x).$$

However, calculating the sharpness term $R_\rho(x)$ directly is analytically hard to do because of the integration with respect to the probability measure μ. Hence, we propose to estimate the inner integration at each iteration with i.i.d. random variables $\nu_1, \nu_2, \ldots, \nu_n \sim \mu$ as perturbations, i.e.,

$$\tilde{R}_\rho(x) \coloneqq \phi\Big(\frac{1}{n}\sum_{i=1}^{n} \psi\Big(\frac{1}{\rho^2}\big(L(x + \rho v_i) - L(x)\big)\Big)\Big).$$

When ϕ satisfies continuity conditions, for large enough n, the estimator $\tilde{R}_\rho(x)$ will converge to $R_\rho(x)$. Now, we calculate the gradients of $L(x) + \rho^2 \tilde{R}_\rho(x)$. By chain rule,

$$\rho^2 \nabla \tilde{R}_\rho(x) = \phi'\Big(\sum_{i=1}^{n} \frac{1}{n} \psi\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_i) - L(x_t)\big)\Big)\Big)$$

Algorithm 1 (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm (with $m = 1$)

Input: The triplet (ϕ, ψ, μ), Training loss $L(x)$, Step size η, Perturbation parameter ρ, Number of samples n,
Output: Model parameters x_t trained with (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm
Initialization: $x \leftarrow x_0$ and $t \leftarrow 0$
while 1 **do**
Sample $v_1, v_2, \dots, v_n \overset{\text{i.i.d.}}{\sim} \mu$
Compute the following:

$$g_t = \nabla L(x_t) + \phi'\Big(\sum_{i=1}^{n} \frac{1}{n} \psi\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_i) - L(x_t)\big)\Big)\Big) \times \sum_{i=1}^{n} \frac{1}{n}\Big\{\psi'\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_i) - L(x_t)\big)\Big) \times \Big(\nabla L(x_t + \rho v_i) - \nabla L(x_t)\Big)\Big\}.$$

Update the parameters: $x_{t+1} = x_t - \eta g_t$
$t \leftarrow t + 1$
end while

$$\times \sum_{i=1}^{n} \frac{1}{n}\Big\{\psi'\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_i) - L(x_t)\big)\Big) \times \Big(\nabla L(x_t + \rho v_i) - \nabla L(x_t)\Big)\Big\},$$

which leads to Algorithm 1.

Our algorithm needs $n + 1$ gradient evaluations per iteration, which for $n = 1$ matches the SAM algorithm [18]. In practice, small values for n demonstrate the expected results, therefore, the computational overhead of our algorithm is not a barrier.

Note that to recover the original SAM algorithm, one can set the function ϕ, ψ to identity, $m = 1$, and choose μ to be the single-point measure on $\nabla L(x_t)/\|\nabla L(x_t)\|_2$ with $n = 1$ sample for each t.

Moreover, even though to prove universality, we only used probability measures, we proposed a compact representation of determinant with Lebesgue measure with $m = 1$ in Table 1 and Theorem 10. However, integrals with respect to Lebesgue measure cannot be estimated via sampling and we need to truncate the integral to integration over a large hypercube; this allows us to use Algorithm 1 for the scale-invariant sharpness measures. Also, this approximation achieves non-zero sharpness in cases that the Hessian matrix is not full-rank (which happens in overparametrized models), as it gets the product of non-zero eigenvalues. We use this approximation to implement the algorithm.

To propose an algorithm for the general case (i.e., arbitrary m), we compute the gradient of

$$\tilde{R}_\rho(x) := \phi\Big(\frac{1}{n}\sum_{i=1}^{n} \psi_1\Big(\frac{1}{\rho^2}\big(L(x + \rho v_{i,1}) - L(x)\big)\Big), \tag{28}$$

$$\frac{1}{n}\sum_{i=1}^{n} \psi_2\Big(\frac{1}{\rho^2}\big(L(x + \rho v_{i,2}) - L(x)\big)\Big), \tag{29}$$

$$\dots \tag{30}$$

Algorithm 2 (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm (with arbitrary m)

Input: The triplet (ϕ, ψ, μ), Training loss $L(x)$, Step size η, Perturbation parameter ρ, Number of samples n,
Output: Model parameters x_t trained with (ϕ, ψ, μ)-Sharpness-Aware Minimization Algorithm
Initialization: $x \leftarrow x_0$ and $t \leftarrow 0$
while 1 **do**
Sample $v_{i,\ell} \overset{\text{i.i.d.}}{\sim} \mu_\ell$, for any $i \in [n]$ and $\ell \in [m]$
Compute the following:

$$g_t = \nabla L(x_t) + \sum_{\ell=1}^{m} \partial_\ell \phi\Big(\sum_{i=1}^{n} \frac{1}{n} \psi_\ell\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_{i,\ell}) - L(x_t)\big)\Big)\Big) \times \sum_{i=1}^{n} \frac{1}{n}\Big\{ \psi_\ell'\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_{i,\ell}) - L(x_t)\big)\Big) \times \Big(\nabla L(x_t + \rho v_{i,\ell}) - \nabla L(x_t)\Big)\Big\}.$$

Update the parameters: $x_{t+1} = x_t - \eta g_t$
$t \leftarrow t + 1$
end while

$$\frac{1}{n} \sum_{i=1}^{n} \psi_m\Big(\frac{1}{\rho^2}\big(L(x + \rho v_{i,m}) - L(x)\big)\Big), \tag{31}$$

where $v_{i,\ell} \overset{\text{i.i.d.}}{\sim} \mu_\ell$ for each $\ell \in [m]$. Note that $\psi = (\psi_1, \psi_2, \ldots, \psi_m)^t$ for some scalar functions ψ_ℓ, $\ell \in [m]$. Let $\partial_\ell \phi$ denote partial derivatives of the function $\phi : \mathbb{R}^m \to \mathbb{R}$, for any $\ell \in [m]$. Then,

$$\rho^2 \nabla \tilde{R}_\rho(x) = \sum_{\ell=1}^{m} \partial_\ell \phi\Big(\sum_{i=1}^{n} \frac{1}{n} \psi_\ell\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_{i,\ell}) - L(x_t)\big)\Big)\Big) \tag{32}$$

$$\times \sum_{i=1}^{n} \frac{1}{n}\Big\{ \psi_\ell'\Big(\frac{1}{\rho^2}\big(L(x_t + \rho v_{i,\ell}) - L(x_t)\big)\Big) \tag{33}$$

$$\times \Big(\nabla L(x_t + \rho v_{i,\ell}) - \nabla L(x_t)\Big)\Big\}, \tag{34}$$

and this leads to Algorithm 2.

Appendix I. Proof of Theorem 6

Theorem 6 (Universality of the (ϕ, ψ, μ)-sharpness measures for functions of Hessian eigenvalues) *Let $\mathcal{A} \subseteq \mathbb{R}^d$ be a compact set. For any continuous function $S : \mathcal{A} \to \mathbb{R}$, there exist a product (Borel) probability measure μ, a positive integer $m \leq d$, and continuous functions $\phi : \mathbb{R}^m \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^m$, such that $S(\lambda_1, \lambda_2 \ldots, \lambda_d) = S(x; \phi, \psi, \mu)$ for any $x \in \mathcal{A}$, where λ_i, $i \in [d]$, are the eigenvalues of the Hessian matrix $\nabla^2 L(x)$.*

Proof We explicitly construct the (Borel) probability measure μ and the function $\psi : \mathbb{R}^m \to \mathbb{R}$. Let us take $m = d$ to prove the universality theorem, while we believe lower m should be enough for specific practical sharpness measures.

Indeed, let us consider μ to be the multivariate Gaussian probability measure with identity covariance matrix. Also, define a (parameterized) function $\psi_\sigma(t) = \exp(\sigma t)$, for some σ to be set later. We are interested to compute the following quantity:

$$\int \psi_\sigma\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v). \tag{35}$$

Note that μ is the standard Gaussian probability measure on $\mathbb{R}^d$, and specifically, it's invariant under the action of the orthogonal matrices. Indeed, there exists an orthogonal matrix Q such that $\nabla^2 L(x) = Q^t\Lambda Q$, where Λ is a diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_d$. Observe that $u := Qv$ is also distributed according to the Gaussian probability distribution with the identity covariance matrix on $\mathbb{R}^d$, similar to v. Now we write

$$\begin{aligned}
\int \psi_\sigma\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v) &= \int \psi_\sigma\big(\frac{1}{2}v^tQ^t\Lambda Qv\big)d\mu(v) && (36)\\
&= \int \psi_\sigma\big(\frac{1}{2}u^t\Lambda u\big)d\mu(u) && (37)\\
&= \int \psi_\sigma\Big(\frac{1}{2}\sum_{i=1}^d \lambda_i u_i^2\Big)d\mu(u) && (38)\\
&= \int (2\pi)^{-d/2}\psi_\sigma\Big(\frac{1}{2}\sum_{i=1}^d \lambda_i u_i^2\Big)\exp\Big(-\frac{1}{2}\sum_{i=1}^d u_i^2\Big)du && (39)\\
&= \int (2\pi)^{-d/2}\exp\Big(\frac{1}{2}\sum_{i=1}^d \sigma\lambda_i u_i^2\Big)\exp\Big(-\frac{1}{2}\sum_{i=1}^d u_i^2\Big)du && (40)\\
&= \int (2\pi)^{-d/2}\exp\Big(\sum_{i=1}^d \frac{1}{2}(\sigma\lambda_i - 1)u_i^2\Big)du, && (41)
\end{aligned}$$

where du denotes the Lebesgue measure on $\mathbb{R}^d$. Now to compute the integral, note that $du = du_1 \times du_2 \times \ldots \times du_d$ is a product measure and the integrand also takes on a product form; thus,

$$\begin{aligned}
\int \psi_\sigma\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v) &= \int (2\pi)^{-d/2}\exp\Big(\sum_{i=1}^d \frac{1}{2}(\sigma\lambda_i - 1)u_i^2\Big)du && (42)\\
&= \prod_{i=1}^d \int (2\pi)^{-1/2}\exp\Big(\frac{1}{2}(\sigma\lambda_i - 1)u_i^2\Big)du_i && (43)\\
&\overset{(a)}{=} \prod_{i=1}^d \frac{1}{\sqrt{1-\sigma\lambda_i}}\underbrace{\int \sqrt{\frac{1-\sigma\lambda_i}{2\pi}}\exp\Big(\frac{1}{2}(\sigma\lambda_i - 1)u_i^2\Big)du_i}_{=1} && (44)\\
&= \prod_{i=1}^d \frac{1}{\sqrt{1-\sigma\lambda_i}}, && (45)
\end{aligned}$$

where (a) holds by the Gaussian integral identities.

Note that to calculate the integral above, we assumed that $1-\sigma\lambda_i > 0$ for any $i \in [d]$. This is equivalent to having

$$\max_{\lambda_i<0} \lambda_i^{-1} < \sigma < \min_{\lambda_i>0} \lambda_i^{-1}. \tag{46}$$

Now, due to the assumption in the theorem, we study the target sharpness measure $S(\lambda_1, \lambda_2, \dots, \lambda_d)$ only on a compact domain $\mathcal{A} \subseteq \mathbb{R}^d$, and this means that there exists an open interval $I = (-\epsilon, \epsilon)$ with

$$\epsilon := \min_{\left(\lambda_1,\lambda_2,\dots,\lambda_d\right)^t \in \mathcal{A}} \min_i |\lambda_i|^{-1}, \tag{47}$$

such that the above integral is well-defined and finite for all $\sigma \in I$.

Let us define the function $\tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^d$ as follows:

$$\tilde{\phi}(t_1, t_2, \dots, t_d) := (t_1^{-2}, t_2^{-2}, \dots, t_d^{-2}). \tag{48}$$

Finally, consider the following polynomial in one variable with degree d:

$$p(x) := (1-\lambda_1 x) \times (1-\lambda_2 x) \times \dots \times (1-\lambda_d x). \tag{49}$$

Claim 12 *For any $\sigma_i \in I$, $i \in [d]$, we have*

$$\tilde{\phi}\Big(\int \psi_{\sigma_1}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v), \int \psi_{\sigma_2}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v), \dots, \tag{50}$$

$$\int \psi_{\sigma_d}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v)\Big) = \Big(p(\sigma_1), p(\sigma_2), \dots, p(\sigma_d)\Big). \tag{51}$$

The above claim simply follows from the integral we calculated before.

Now we are ready to complete the proof. Choose arbitrary non-zero distinct $\sigma_i \in I, i \in [d]$, and note that having access to $\big(p(\sigma_1), p(\sigma_2), \dots, p(\sigma_d)\big)$ is enough to recover all the eigenvalues. Indeed, assume that $p(x) = p_0 + p_1 x + p_2 x^2 + \ldots + p_d x^d$ and note that $p(0) = 1$ by definition. Let also $V(0, \sigma_1, \sigma_2, \dots, \sigma_d) \in \mathbb{R}^{d\times d}$ denote a Vandermonde matrix of order $d+1$, which is provably invertible by definition, and note that

$$\big(p(\sigma_1), p(\sigma_2), \dots, p(\sigma_d)\big)^t = V(0, \sigma_1, \sigma_2, \dots, \sigma_d) \times \big(p_0, p_1, \dots, p_{d+1}\big)^t \tag{52}$$

$$\implies \big(p_0, p_1, \dots, p_{d+1}\big)^t = V(0, \sigma_1, \sigma_2, \dots, \sigma_d)^{-1} \times \big(p(\sigma_1), p(\sigma_2), \dots, p(\sigma_d)\big)^t. \tag{53}$$

Indeed, this shows that having access to the vector $\big(p(\sigma_1), p(\sigma_2), \dots, p(\sigma_d)\big)^t$ is enough to reconstruct the polynomial $p(x) = p_0 + p_1 x + \ldots + p_d x^d$. Having access to this polynomial is equivalent to having access to its roots, so one can find a continuous function $\phi_1 : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\big(\lambda_1, \lambda_2, \dots, \lambda_d\big)^t = \phi_1 \circ \tilde{\phi}\Big(\int \psi_{\sigma_1}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v), \tag{54}$$

$$\int \psi_{\sigma_2}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v), \dots, \int \psi_{\sigma_d}\big(\frac{1}{2} v^t \nabla^2 L(x) v\big) d\mu(v)\Big). \tag{55}$$

Since the sharpness measure $S(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a continuous function of its coordinates, we conclude that

$$S(\lambda_1,\lambda_2,\ldots,\lambda_d) = S \circ \phi_1 \circ \tilde{\phi}\Big(\int \psi_{\sigma_1}\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v), \tag{56}$$

$$\int \psi_{\sigma_2}\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v), \ldots, \int \psi_{\sigma_d}\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v)\Big). \tag{57}$$

Now to complete the proof, we define a continuous function $\psi : \mathbb{R} \to \mathbb{R}^d$ as $\psi = \big(\psi_{\sigma_1}, \psi_{\sigma_2}, \ldots, \psi_{\sigma_d}\big)^t$, and a continuous function $\phi : \mathbb{R}^d \to \mathbb{R}$ as $\phi = S \circ \phi_1 \circ \tilde{\phi}$, and observe that $S(\lambda_1, \lambda_2 \ldots, \lambda_d) = S(x; \phi, \psi, \mu)$ for any $x \in \mathcal{A}$. This completes the proof. ■

Appendix J. Proof of Theorem 7

Theorem 7 (Universality of the (ϕ, ψ, μ)-sharpness measures for arbitrary functions of Hessian) *For any continuous function $S : \mathbb{R}^{d\times d} \to \mathbb{R}$, there exist a positive integer $m \leq d(d+1)/2$, (Borel) probability measures μ_ℓ, $\ell \in [m]$, and continuous functions $\phi : \mathbb{R}^m \to \mathbb{R}$ and $\psi : \mathbb{R} \to \mathbb{R}^m$, such that $S(\nabla^2 L(x)) = S(x; \phi, \psi, \mu)$ for any $x \in \mathbb{R}^d$, where $\mu := \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_m$ is a product probability measure.*

Proof We explicitly construct a set of functions/probability measures to achieve the desired representation. Indeed, let's take $m = d(d+1)/2$ and consider the following Dirac measures: $\mu_i = \delta_{e_i}, i \in [d]$, and also $\mu_{ij} = \delta_{e_i+e_j}$, for any $i, j \in [d]$ such that $i < j$. Here, e_i denotes the unit vector in the ith coordinate in $\mathbb{R}^d$. Now, note that we have

$$(\nabla^2 L(x))_{i,i} = \int v^t\nabla^2 L(x)v d\mu_i(v) \tag{58}$$

for any $i \in [d]$, and

$$2(\nabla^2 L(x))_{i,j} + (\nabla^2 L(x))_{i,i} + (\nabla^2 L(x))_{j,j} = \int v^t\nabla^2 L(x)v d\mu_{i,j}(v), \tag{59}$$

for any $i, j \in [d]$ such that $i < j$. The above system of linear equations has clearly a unique solution, as the Hessian matrix is symmetric. This means that, similar to the proof of Theorem 6, one can find continuous functions $\psi : \mathbb{R} \to \mathbb{R}^m$ and $\phi : \mathbb{R}^m \to R$, along with m constructed probability measures such that $S(\nabla^2 L(x)) = S(x; \phi, \psi, \mu)$ for any $x \in \mathbb{R}^d$, where $\mu = \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_m$ is a product probability measure. ■

Appendix K. Proof of Theorem 9

Theorem 9 (Explicit bias of the (ϕ, ψ, μ)-sharpness-aware loss function) *Given a triplet (ϕ, ψ, μ), $m \geq 1$, and a training loss function $L : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, assume that:*

- *$L(x)$ is third-order continuously differentiable and satisfies the following upper bound*

$$\max_{i,j,k\in\{1,2,3\}} |\partial_i\partial_j\partial_k L(v)| = O(\|v\|^{-1}), \tag{22}$$

  for $v \in \mathbb{R}^d$ as $\|v\|_2 \to \infty$.

- *The two functions ϕ, ψ are continuously differentiable.*
- *For some $C > \max_{x\in\mathcal{X}} \max_{i\in[d]} |\lambda_i(\nabla^2 L(x))|$, we have $\int \|v\|_2^2 \tilde{\psi}_i(v) d\mu(v) < \infty$, $i \in [m]$, where*

$$\tilde{\psi}_i(v) := \max_{|t|\le\|v\|_2} |\psi_i'(Ct^2)|. \tag{23}$$

Then, there exists an open neighborhood $U \supsetneq \Gamma$, where Γ is the zero-loss manifold, for connected U and Γ, such that if for some $u \in U$, one has

$$L(u) + \rho^2 R_\rho(u) - \inf_{x\in U} \Big(L(x) + \rho^2 R_\rho(x)\Big) \le \Delta\rho^2, \tag{24}$$

with some optimally gap $\Delta > 0$, then

$$L(u) \le \underbrace{\inf_{x\in U} L(x)}_{=0} + (\Delta + o_\rho(1))\rho^2, \tag{25}$$

and also

$$S(u;\phi,\psi,\mu) \le \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) + \Delta + o_\rho(1). \tag{26}$$

Proof For simplicity, we assume that $m = 1$. The general proof for $m > 1$ follows with a similar argument to this special case. Define an open set $U \subseteq \mathbb{R}^d$ as follows:

$$U := \big\{x \in \mathbb{R}^d : \|\nabla L(x)\|_2 < \rho^2\big\}. \tag{60}$$

Note that this set contain the zero-loss manifold, i.e., $\Gamma \subseteq U$. We study the behavior of the loss function on this open set.

Let us denote the sharpness term in the loss function $L_{(\phi,\psi,\mu)}(x)$ by $R_\rho(x)$:

$$\begin{aligned} L_{(\phi,\psi,\mu)}(x) &= L(x) + \rho^2 R_\rho(x) &(61)\\ &= L(x) + \rho^2 \phi\Big(\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big) d\mu(v)\Big). &(62)\end{aligned}$$

We first study the convergence of $R_\rho(x)$ to the corresponding sharpness measure $S(x;\phi,\psi,\mu)$. Fix any point $x \in U$ and note that using Taylor's theorem for the function $L : \mathbb{R}^d \to \mathbb{R}$ and for any $v \in \mathbb{R}^d$, one has

$$\begin{aligned} \frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big) &= \frac{1}{\rho^2}\Big(\rho\langle\nabla L(x), v\rangle + \frac{1}{2}\rho^2 v^t \nabla^2 L(x) v + O_x(\rho^3\|v\|_2^3 \times \|v\|_2^{-1})\Big) &(63)\\ &= \rho^{-1}\langle\nabla L(x), v\rangle + \frac{1}{2} v^t \nabla^2 L(x) v + O_x(\rho\|v\|_2^2 \times \|v\|_2^{-1}), &(64)\end{aligned}$$

where in above we used the fact that $L(x)$ is third-order continuously differentiable and its third-order derivative satisfies the estimate

$$\max_{i,j,k\in\{1,2,3\}} |\partial_i\partial_j\partial_k L(v)| = O(\|v\|^{-1}), \tag{65}$$

for $v \in \mathbb{R}^d$ as $\|v\|_2 \to \infty$.

Note that using the assumption $x \in U$, we have that

$$|\rho^{-1}\langle \nabla L(x), v\rangle| \leq \rho^{-1}\|\nabla L(x)\|_2\|v\|_2 < \rho\|v\|_2. \tag{66}$$

Thus, we have

$$\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big) = \frac{1}{2}v^t\nabla^2 L(x)v + O_x(\rho(\|v\|_2^2 + \|v\|_2)). \tag{67}$$

Note that we study the above approximation only for $x \in U$, and for small enough ρ, we know that U is a precompact set. Therefore, we drop the dependence on x in the error term above.

Now using the above approximation, we have

$$\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big)d\mu(v) = \int \psi\Big(\frac{1}{2}v^t\nabla^2 L(x)v + O(\rho(\|v\|_2^2 + \|v\|_2))\Big)d\mu(v). \tag{68}$$

Let us use Taylor's theorem for the function ψ and write

$$\psi\Big(\frac{1}{2}v^t\nabla^2 L(x)v + O(\rho(\|v\|_2^2 + \|v\|_2))\Big) = \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big) \tag{69}$$

$$+ \rho \times O(\tilde{\psi}(v)(\|v\|_2^2 + \|v\|_2)), \tag{70}$$

where

$$\tilde{\psi}(v) := \max_{|t|\leq\|v\|_2} |\psi'(Ct^2)|, \tag{71}$$

and C is a constant, and it's big enough to absorb the quadratic growth of $\nabla^2 L(x)$; i.e.,

$$C > \max_{x\in\mathcal{X}} \max_{i\in[d]} |\lambda_i(\nabla^2 L(x))|. \tag{72}$$

Therefore, we conclude that

$$\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big)d\mu(v) = \tag{73}$$

$$\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v) + \rho \times O\Big(\int \tilde{\psi}(v)(\|v\|_2^2 + \|v\|_2)d\mu(v)\Big) \tag{74}$$

$$= \int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v) + O(\rho), \tag{75}$$

by the assumption. This allows us to conclude that

$$R_\rho(x) = \phi\Big(\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big)d\mu(v)\Big) \tag{76}$$

$$= \phi\Big(\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v)\Big) + O(\rho) \tag{77}$$

$$= S(x;\phi,\psi,\mu) + O(\rho), \tag{78}$$

again, by assuming that

$$\max_{x\in\mathcal{X}} \phi'\Big(\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)v\big)d\mu(v)\Big) < \infty, \tag{79}$$

which holds by the compactness of $\mathcal{X}$, and also using $U \subseteq \mathcal{X}$.

Now according to the assumption, for some $u \in U$, we have

$$L(u) + \rho^2 R_\rho(u) - \inf_{x\in U}\Big(L(x) + \rho^2 R_\rho(x)\Big) \leq \Delta\rho^2, \tag{80}$$

for some optimally gap Δ. Using the following proven approximation

$$R_\rho(x) = S(x;\phi,\psi,\mu) + O(\rho), \tag{81}$$

we conclude that

$$L(u) + \rho^2 S(u;\phi,\psi,\mu) - \inf_{x\in U}\Big(L(x) + \rho^2 S(x;\phi,\psi,\mu)\Big) \leq (\Delta + O(\rho))\rho^2. \tag{82}$$

Now by proof by contradiction, assume that

$$L(u) \geq \inf_{x\in U} L(x) + (\Delta+\delta)\rho^2, \tag{83}$$

for some strictly positive δ, as $\rho \to 0^+$. Note that $\inf_{x\in U} L(x) = 0$ as $\Gamma \subseteq U$. Thus, we can conclude that

$$\begin{aligned}
\rho^2 S(u;\phi,\psi,\mu) + (\Delta+\delta)\rho^2 &\leq L(u) + \rho^2 S(u;\phi,\psi,\mu) && (84)\\
&\leq \inf_{x\in U}\Big(L(x) + \rho^2 S(x;\phi,\psi,\mu)\Big) + (\Delta + O(\rho))\rho^2 && (85)\\
&\leq \rho^2 \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) + (\Delta + O(\rho))\rho^2, && (86)
\end{aligned}$$

since $\Gamma \subseteq U$. This shows that

$$S(u;\phi,\psi,\mu) \leq \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) - \delta + O(\rho). \tag{87}$$

This must hold for as $\rho \to 0^+$. However, as $\rho \to 0^+$, we have that $U_\rho \to \Gamma$. This means that

$$S(u;\phi,\psi,\mu) \leq \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) - \delta, \tag{88}$$

for some $u \in \Gamma$, which is a contradiction. This shows that

$$L(u) \leq \inf_{x\in U} L(x) + (\Delta + o(1))\rho^2. \tag{89}$$

Also, to prove the next part of the theorem, for any $u \in U$ satisfying the assumptions, similarly we can show

$$\begin{aligned}
\rho^2 S(u;\phi,\psi,\mu) &\leq L(u) + \rho^2 S(u;\phi,\psi,\mu) && (90)\\
&= \inf_{x\in U}\Big(L(x) + \rho^2 S(x;\phi,\psi,\mu)\Big) + (\Delta + O(\rho))\rho^2 && (91)
\end{aligned}$$

$$\leq \rho^2 \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) + (\Delta + O(\rho))\rho^2, \tag{92}$$

which implies that

$$S(u;\phi,\psi,\mu) \leq \inf_{x\in\Gamma} S(x;\phi,\psi,\mu) + (\Delta + O(\rho)). \tag{93}$$

The proof is thus complete.

■

Appendix L. Proof of Theorem 10

Theorem 10 (Scale-invariant (ϕ,ψ,μ)-sharpness measures) *Consider a scale-invariant loss function $L(x)$ and let μ be a Borel measure of the form*

$$d\mu(x) = f\Big(\prod_{i=1}^{d} x_i\Big)\prod_{i=1}^{d} dx_i, \tag{27}$$

where $f:\mathbb{R}\to\mathbb{R}$ is a measurable function. Then, for any continuous functions ϕ,ψ, the corresponding sharpness measure $S(x;\phi,\psi,\mu)$ is scale-invariant; this means that $S(x;\phi,\psi,\mu) = S(Dx;\phi,\psi,\mu)$ for any diagonal matrix $D\in\mathbb{R}^{d\times d}$ with $\det(D)=1$.

Proof Let $D\in\mathbb{R}^{d\times d}$ be an arbitrary diagonal matrix. Then,

$$S(Dx;\phi,\psi,\mu) = \phi\Big(\int \psi\big(\frac{1}{2}v^t\nabla^2 L(x)\Big|_{Dx} v\big)d\mu(v)\Big). \tag{94}$$

But note that by the assumption

$$L(x) = L(Dx) \implies \nabla^2 L(x) = D^t\nabla^2 L(x)\Big|_{Dx} D. \tag{95}$$

Therefore,

$$S(Dx;\phi,\psi,\mu) = \phi\Big(\int \psi\big(\frac{1}{2}v^t D^{-1}\nabla^2 L(x)D^{-1}v\big)d\mu(v)\Big). \tag{96}$$

Now define a new variable $u := D^{-1}v$. Then,

$$d\mu(v) = f(\prod_{i=1}^{d} v_i)\prod_{i=1}^{d} dv_i = f(\prod_{i=1}^{d} D_{i,i}\prod_{i=1}^{d} u_i)\prod_{i=1}^{d} D_{i,i}\prod_{i=1}^{d} du_i \tag{97}$$

$$= f(\det(D)\prod_{i=1}^{d} u_i)\det(D)\prod_{i=1}^{d} du_i \tag{98}$$

$$= f(\prod_{i=1}^{d} u_i)\prod_{i=1}^{d} du_i. \tag{99}$$

Therefore, we conclude that

$$S(Dx;\phi,\psi,\mu) = \phi\Big(\int \psi\big(\frac{1}{2}u^t\nabla^2 L(x)u\big)d\mu(u)\Big) = S(x;\phi,\psi,\mu), \tag{100}$$

and this completes the proof.

■

Appendix M. Proof of Theorem 11

Theorem 11 (General parameter-invariant (ϕ, ψ, μ)-sharpness measures) *Let G be a group acting by matrices on $\mathbb{R}^d$, and assume that $L(x)$ is invariant with respect to the action of G. Then, for any G-invariant (Borel) measure μ, and any continuous functions ϕ, ψ, the corresponding sharpness measure $S(x; \phi, \psi, \mu)$ is G-invariant; this means that $S(x; \phi, \psi, \mu) = S(A_g x; \phi, \psi, \mu)$ for any matrix $A_g \in \mathbb{R}^{d\times d}$ corresponding to the action of an element $g \in G$.*

Proof We start by evaluating $S(A_g x; \phi, \psi, \mu)$.

$$S(Dx; \phi, \psi, \mu) = \phi\Big(\int \psi\big(\frac{1}{2} v^t \nabla^2 L(x)\Big|_{A_g x} v\big) d\mu(v)\Big). \tag{101}$$

But again here, note that by the assumption

$$L(x) = L(A_g x) \implies \nabla^2 L(x) = A_g^t \nabla^2 L(x)\Big|_{A_g x} A_g. \tag{102}$$

Therefore,

$$S(A_g x; \phi, \psi, \mu) = \phi\Big(\int \psi\big(\frac{1}{2} v^t A_g^{-1} \nabla^2 L(x) A_g^{-1} v\big) d\mu(v)\Big). \tag{103}$$

Now define a new variable $u := A_g^{-1} v$. Therefore, we conclude that

$$S(A_g x; \phi, \psi, \mu) = \phi\Big(\int \psi\big(\frac{1}{2} u^t \nabla^2 L(x) u\big) d\mu(u)\Big) = S(x; \phi, \psi, \mu), \tag{104}$$

and this completes the proof.

■

Appendix N. Frobenius-SAM

To be more concrete, we specify our general framework (Algorithm 2) to the case with the Frobenius norm regularization. Note that to achieve this, one needs to specify $\phi(t_1, t_2) = 2(t_2 - t_1^2)$ and $\psi(t) = (t, t^2)$, according to Table 1. Furthermore, since we only need to collect samples from the Gaussian distribution to get the Frobenius norm bias (Table 1), we can use the same samples to estimate both integrals for the functions $\psi_1(t) = t$ and $\psi_2(t) = t^2$. Replacing these assumptions into the general algorithm (Algorithm 2), we get the following update rule:

$$\begin{aligned} g_t = \nabla L(x_t) &+ 4 \sum_{i=1}^{n} \frac{1}{n\rho^2} \Big\{ \big(L(x_t + \rho v_i) - L(x_t)\big) \times \Big(\nabla L(x_t + \rho v_i) - \nabla L(x_t)\Big) \Big\} \\ &- 4 \Big\{ \sum_{i=1}^{n} \frac{1}{n\rho} \big(L(x_t + \rho v_i) - L(x_t)\big) \Big\} \times \Big\{ \sum_{j=1}^{n} \frac{1}{n\rho} \Big(\nabla L(x_t + \rho v_j) - \nabla L(x_t)\Big) \Big\}. \end{aligned}$$

If we take a closer look at this, we observe that

$$g_t = \nabla L(x_t) + \frac{4}{\rho^2} \times \widehat{\text{cov}}\Big(\big(L(x_t + \rho v) - L(x_t)\big), \big(\nabla L(x_t + \rho v) - \nabla L(x_t)\big)\Big), \tag{105}$$

where $\widehat{\text{cov}}$ denotes the (biased) empirical cross-covariance between the scalar random variable $L(x_t+\rho v)-L(x_t)$ and the vector-values random variable $\nabla L(x_t+\rho v)-\nabla L(x_t)$, for $v \sim \mathcal{N}(0, I_d)$. Since the covariance is not sensitive to the means of random variables/vectors, we can further simply the update rule to

$$g_t = \nabla L(x_t) + \frac{4}{\rho^2} \times \widehat{\text{cov}}\Big(L(x_t+\rho v), \nabla L(x_t+\rho v)\Big). \tag{106}$$

We can further replace the unbiased estimator of the cross-covariance instead of $\widehat{\text{cov}}$ which leads to Algorithm 3.

Algorithm 3 *Frob-SAM*

Input: Training loss $L(x)$, Step size η, Perturbation parameter ρ, Number of samples n,
Output: Model parameters x_t trained with Frobenius SAM
Initialization: $x \leftarrow x_0$ and $t \leftarrow 0$
while 1 **do**
Sample $v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ for any $i \in [n]$
Compute the following:

$$g_t = \nabla L(x_t) + 4\sum_{i=1}^{n} \frac{1}{(n-1)\rho^2} L(x_t+\rho v_i)\nabla L(x_t+\rho v_i) - 4\sum_{i=1}^{n} \frac{1}{(n-1)\rho} L(x_t+\rho v_i) \times \sum_{i=1}^{n} \frac{1}{n\rho} \nabla L(x_t+\rho v_i).$$

Update the parameters: $x_{t+1} = x_t - \eta g_t$
$t \leftarrow t+1$
end while

Appendix O. Experimental Details

We now describe details of the experiments that were omitted from the main text. For CIFAR10 and CIFAR100, we apply random crops and random horizontal flips. We use a momentum term of 0.9 for all datasets and a weight decay of 5e-4 for CIFAR10 and SVHN and 1e-3 for CIFAR100. We use batch size 128 and train SGD to 50 epochs. We use a multi-step schedule where the LR is initially 0.1 and decays by a multiplicative factor of 0.25 every 10 epochs. All runs use the same random seed which seeds the initialization of weights. We use 1280 training examples and 100 noise samples to estimate the frobenius norm via Hessian-vector products and shade 1 standard error in the plots.

For Det-SAM, we sample from the hypercube, where each side of the cube has length 0.02, and we set ρ to 1. We sweep λ in {0.01, 0.1, 1.0}. For Frob-SAM, we sweep ρ in {0.005, 0.01} and λ in {0.0001, 0.001, 0.005}, and the number of samples in {2, 10}.