On Scale-Invariant Sharpness Measures

Behrooz Tahmasebi
MIT CSAIL
bzt@mit.edu

Ashkan Soleymani
MIT LIDS
ashkanso@mit.edu

Stefanie Jegelka
TU Munich and MIT CSAIL
stefje@mit.edu

Patrick Jaillet
MIT LIDS
jaillet@mit.edu

Abstract

Recently, there has been a substantial surge of interest in the development of optimization algorithms tailored for overparameterized models. This interest centers around the objective of minimizing a concept of sharpness in conjunction with the original loss function, e.g., the Sharpness-Aware Minimization (SAM) algorithm shown effective in practice. Nevertheless, the majority of sharpness measures exhibit sensitivity to parameter scaling in neural networks, and they may even experience significant magnification when subjected to rescaling operations. Motivated by this issue, in this paper, we introduce a new class of scale-invariant sharpness measures, that gives rise to a new class of scale-invariant sharpness-aware objective functions. Furthermore, we prove that the newly introduced objective functions are explicitly biased towards the minimization of our scale-invariant sharpness measures.

1 Introduction

The success of deep learning [10] is frequently attributed to its overparameterization. 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 is to consider the geometric aspects of the loss landscape, with the aim of achieving generalization; it entails the avoidance of sharp minima. For example, the celebrated Sharpness-Aware Minimization (SAM) algorithm has shown enhancements in generalization across many practical tasks [8]. While the concept of sharpness lacks a precise definition in a general sense, people often introduce various measures to quantify it in practice, while most of them rely on the second-order derivative characteristics of the training loss function, such as the trace or the operator norm of the Hessian matrix.

Nevertheless, given the intricate geometry of the loss landscape, traditional methodologies for quantifying sharpness may not suffice for the study of generalization. Indeed, neural networks exhibit parameter invariances, wherein distinct parameterizations can yield identical functions — such as scale-invariance 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 of addressing this phenomenon.

Therefore, a fundamental question arises: how can one represent measures of sharpness within a compact parameterized framework that also enables meaningful applications to models with parameter invariances? As a step towards answering this question, this paper introduces an average-based parame-

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

terized representation for sharpness measures that is invariant under parameter rescaling. Furthermore, although the provided representation depends on the Hessian of the training loss, we introduce a novel sharpness-aware loss function for any proposed sharpness measures that only relies on the zeroth-order information about the training loss. We also prove that this new loss function is explicitly *biased* towards minimizing the associated sharpness measure. Thus, it can be considered as a generalized parameterized sharpness-aware minimization algorithm. Indeed, this allows us to readily design algorithms with *invariant* biases, 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 not necessarily scale-invariant sharpness measures are known in the literature.

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

- We propose a new parameterized representation for sharpness measures as a function of the training loss's Hessian matrix, and prove that the new representation is *scale-invariant* when applied to neural networks under parameter rescaling.
- Attached to any sharpness representation, an optimization objective, that only depends on the zeroth-order information about the training loss, is provided, and it is proved that the new objective is *biased* towards minimizing the corresponding sharpness measure.

2 Related Work

Foret et al. [8] recently proposed the Sharpness-Aware Minimization (SAM) algorithm to avoid sharp minima. Besides SAM, Nitanda et al. [18] show how parameter averaging for SGD is biased towards flatter minima. Label noise SGD also prefers flat minima [6]. Woodworth et al. [26] study the role of sharpness in overparametrization from a kernel perspective (see [23, 5] for the applications of flat minima for domain generalization). For applications of SAM in large language models, see [4, 27] (and [19, 20] for federated learning). Besides those applications, Wen et al. [25] prove that current sharpness minimization algorithms sometimes fail to generalize for non-generalizing flattest models.

The (implicit) bias of many optimization algorithms and architectures is presently understood, from the Gradient Descent (GD) [14, 21] to the mirror decent [11, 3, 22]; see also [12] for linear convolutional networks, and [15] for equivariant networks. Gatmiry et al. [9] find the bias of flatness regularization for deep matrix factorization. It is also observed that linear neural networks are biased towards weight alignment for different layers [13] (see also [16] for non-linear networks). Andriushchenko and Flammarion [1] study the implicit bias of SAM for diagonal linear networks, and Wen et al. [24] find the explicit bias of the Gaussian averaging method and other SAM variants.

Scale-invariances' role in generalization in deep learning is emphasized in [17]. Dinh et al. [7] point out that parameter invariances can lead to the different parameterization of the same function, making the definition of flatness challenging; see also [2] for a recent study.

3 A Scale-Invariant Sharpness Measure

3.1 Setting

Consider a standard learning setup with a labeled dataset $\mathcal{S}$, and a third-order continuously differentiable 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 [8].

3.2 Background on SAM

It is hypothesized that the avoidance of sharp minima can enhance generalization performance. However, it should be noted that the concept of sharpness encompasses a multitude of distinct definitions in practical contexts. The Sharpness-Aware Minimization (SAM) algorithm suggests minimizing the training loss function over a small ball around the parameters:

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

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

$$L_{\text{SAM}}(x) = \underbrace{L(x)}_{\text{empirical loss}} + \underbrace{\max_{\|v\|_2 \leq 1} \big\{L(x+\rho v) - L(x)\big\}}_{\text{sharpness}}. \tag{2}$$

Foret et al. [8] also suggest alternative average-based sharpness-aware objectives to use PAC bounds on the generalization error of overparameterized models; we follow the definition in Wen et al. [25]:

$$L_{\text{AVG}}(x) := \mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2})\Big] = \underbrace{L(x)}_{\text{empirical loss}} + \underbrace{\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2}) - L(x)\Big]}_{\text{sharpness}}. \tag{3}$$

3.3 Proposed Approach

Wen et al. [25] 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 $L_{\text{AVG}}(x)$ is biased towards minimizing $\frac{1}{d}\text{Tr}(\nabla^2 L(x))$. This means that SAM measures the sharpness of a global minimum by $\lambda_{\max}(\nabla^2 L(x))$, while the average-based objective uses $\frac{1}{d}\text{Tr}(\nabla^2 L(x))$ to evaluate it.

In the next example, we argue how both sharpness measures above fail to define a meaningful notion under parameter rescalings.

Example 1 *Consider the loss function $L(x_1, x_2) = x_1^2 x_2^2 - 2x_1 x_2 + 1$ 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 $\Gamma = \{(x_1, x_2) : x_1 x_2 = 1\}$ 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 $\frac{1}{2}Tr(\nabla^2 L(x_1, x_2)) = x_1^2 + x_2^2$. After rescaling, we get $\frac{1}{2}Tr(\nabla^2 L(x_1, x_2))\Big|_{(kx_1, k^{-1}x_2)} = k^2 x_1^2 + \frac{x_2^2}{k^2} \neq \frac{1}{2}Tr(\nabla^2 L(x_1, x_2))$. Therefore, as a sharpness measure, $Tr(\nabla^2 L(x_1, x_2))$ is not scale-invariant. The problem magnifies drastically in the limit: $\lim_{k\to\infty} Tr(\nabla^2 L(x_1, x_2))\Big|_{(kx_1, k^{-1}x_2)} = \infty$. 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. To define a new scale-invariant sharpness measure, we take a closer look at the average-based sharpness-aware objective $L_{\text{AVG}}(x)$; using its Taylor expansion [25], we have

$$L_{\text{AVG}}(x) = \mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[L(x+\frac{\rho v}{\|v\|_2})\Big] \tag{4}$$

$$\approx\ L(x) + \rho\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[\langle\nabla L(x), \frac{v}{\|v\|_2}\rangle\Big] + \rho^2\mathbb{E}_{v\sim\mathcal{N}(0,I)}\Big[\frac{v^t\nabla^2 L(x)v}{\|v\|_2^2}\Big] \tag{5}$$

$$= L(x) + \rho^2\frac{\text{Tr}(\nabla^2 L(x))}{d}. \tag{6}$$

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. This motivates us to define the following parameterized sharpness measure.

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*

$$S(x; \phi, \psi, \mu) := \phi\Big(\int \psi\big(v^t\nabla^2 L(x)v\big)d\mu(v)\Big). \tag{7}$$

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 Hessian matrix.

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

Table 1: Examples of (ϕ, ψ, μ)-sharpness measures; see Section 4

Hyperparameters			
$\phi(t)$	$\psi(t)$	$d\mu(v)$	$S(x;\phi,\psi,\mu)$
t	t	Uniform$(\mathbb{S}^{d-1})$	$\frac{1}{d}\text{Tr}(\nabla^2 L(x)) = \frac{1}{d}\sum_{i=1}^{d}\lambda_i$
$(2\pi)^d/t^2$	$\exp(-t/2)$	Lebesgue measure on $\mathbb{R}^d$	$\det(\nabla^2 L(x)) = \prod_{i=1}^{d}\lambda_i$
$\phi(t)$	t^n	Uniform$(\mathbb{S}^{d-1})$	$\phi\Big(p(\lambda_1,\lambda_2,\ldots,\lambda_d)\Big)$

*$\phi(t)$ is an arbitrary continuous function and $p(\lambda_1,\lambda_2\ldots,\lambda_d)$ is a specifically constructed polynomial

Theorem 1 *(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*

$$d\mu(x) = f\Big(\prod_{i=1}^{d} x_i\Big)\prod_{i=1}^{d} dx_i, \tag{8}$$

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 $\det(D)=1$.

Example 2 *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 $\det(D)=1$,*

$$\det(\nabla^2 L(x))\Big|_{Dx} = \det(D^{-1}\nabla^2 L(x) D^{-1}) = \det(D^{-1})^2\det(\nabla^2 L(x)) = \det(\nabla^2 L(x)). \tag{9}$$

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

Now that we defined a flexible set of sharpness measures, 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 give answer to this question, we introduce the (ϕ,ψ,μ)-sharpness-aware loss function as follows:

$$L_{(\phi,\psi,\mu)}(x) := \underbrace{L(x)}_{\text{empirical loss}} + \rho^2 \underbrace{\phi\Big(\int \psi\Big(\frac{1}{\rho^2}\big(L(x+\rho v) - L(x)\big)\Big)d\mu(v)\Big)}_{\text{sharpness}}, \tag{10}$$

where ρ is the perturbation parameter (similar to $L_{\text{SAM}}(x)$ and $L_{\text{AVG}}(x)$).

Theorem 2 *(Informal; explicit bias of the (ϕ,ψ,μ)-sharpness-aware loss function). For a large class of triplets (ϕ,ψ,μ) and sequences $\epsilon(\rho)\le\Delta\rho^2$ with some $\Delta>0$, if $L_{(\phi,\psi,\mu)}(x(\rho))\le\epsilon(\rho)$, then $S(x(\rho);\phi,\psi,\mu)\le\Delta+o_\rho(1)$ as $\rho\to 0^+$.*

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.

4 Examples of (ϕ,ψ,μ)-Sharpness Measures

In this section, we prove that various notions of sharpness can be achieved using the proposed approach in this paper (Table 1).

4.1 Trace

Let $\phi(t)=\psi(t)=t$, and note that

$$S(x;\phi,\psi,\mu) = \int v^t\nabla^2 L(x) v\, d\mu(v) \tag{11}$$

$$= \mathbb{E}_{v\sim\mu}[v^t\nabla^2 L(x) v], \tag{12}$$

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

$$\mathbb{E}_{v\sim\mu}[v^t \nabla^2 L(x) v] = \sum_{i,j=1}^{d} (\nabla^2 L(x))_{i,j} \mathbb{E}[v_i v_j] = \sum_{i=1}^{d} \frac{1}{d} (\nabla^2 L(x))_{i,i} = \frac{1}{d} \text{Tr}(\nabla^2 L(x)), \tag{13}$$

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

4.2 Determinant

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

$$S(x;\phi,\psi,\mu) = (2\pi)^d \Big(\int \exp(-\frac{1}{2} v^t \nabla^2 L(x) v) dv \Big)^{-2}, \tag{14}$$

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

$$\int \exp(-\frac{1}{2} v^t \nabla^2 L(x) v) dv = (2\pi)^{d/2} \det(\nabla^2 L(x))^{-1/2}. \tag{15}$$

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

4.3 Polynomials of Eigenvalues

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

$$S(x;\phi,\psi,\mu) = \phi\Big(\int \big(v^t \nabla^2 L(x) v\big)^n d\mu(v) \Big) \tag{16}$$

$$= \phi\Big(\mathbb{E}_{v\sim\mu} \Big[\big(v^t \nabla^2 L(x) v\big)^n \Big] \Big), \tag{17}$$

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 $\big(v^t \nabla^2 L(x) v\big)^n = \big(v^t Q^t \nabla^2 L(x) Q v\big)^n$. But Qv is distributed uniformly over the $(d-1)$-sphere $\mathbb{S}^{(d-1)}$, similar to v. Thus, we conclude

$$S(x;\phi,\psi,\mu) = \phi\Big(\mathbb{E}_{v\sim\mu} \Big[\big(v^t \nabla^2 L(x) v\big)^n \Big] \Big) \tag{18}$$

$$= \phi\Big(\mathbb{E}_{v\sim\mu} \Big[\big(\sum_{i=1}^{d} \lambda_i v_i^2 \big)^n \Big] \Big). \tag{19}$$

Define $p(\lambda_1, \lambda_2, \ldots, \lambda_d) := \mathbb{E}_{v\sim\mu} \Big[\big(\sum_{i=1}^{d} \lambda_i v_i^2 \big)^n \Big]$, which is clearly a polynomial function (by the linearity of expectation).

5 Conclusion

In this paper, we introduced a new family of sharpness measures and we showed how this new parameterized representation can generate many meaningful sharpness notions (Table 1). Moreover, we proved in Theorem 1 how specific Borel measures can lead to scale-invariant sharpness measures (such as the determinant of the Hessian matrix). Furthermore, in Theorem 2, we showed how the corresponding zeroth-order objective function to each sharpness measure is explicitly biased towards minimizing the desired sharpness subject to the zero-loss manifold of the training loss.

Acknowledgments

BT and SJ are supported by the NSF TRIPODS program (award DMS-2022448), the Office of Naval Research grant N00014-20-1-2023 (MURI ML-SCOPE), the NSF award CCF-2112665 (TILOS AI Institute), and the NSF award 2134108. AS and PJ are partially supported by AI Singapore, grant AISG2-RP-2020-018.

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] Khashayar Gatmiry, Zhiyuan Li, Ching-Yao Chuang, Sashank Reddi, Tengyu Ma, and Stefanie Jegelka. The inductive bias of flatness regularization for deep matrix factorization. *arXiv preprint arXiv:2306.13239*, 2023.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[11] 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.

[12] 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.

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

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

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

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

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

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

[19] 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.

[20] 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.

[21] 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.

[22] Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *arXiv preprint arXiv:2306.13853*, 2023.

[23] 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.

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

[25] Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *arXiv preprint arXiv:2307.11007*, 2023.

[26] 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.

[27] 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.