Beyond Sharpness: The Role of Nonuniformity in Generalization

Yingcong Zhou[1], Pingfan Wu[2], Li Wang[3], Zhiguo Fu[1], Fengqin Yang[1]*

[1]School of Information Science and Technology & AI4S Center, Northeast Normal University, China
[2]School of Computing and Artificial Intelligence & Key Laboratory of Interdisciplinary Research of Computation and Economics, Shanghai University of Finance and Economics, China
[3]School of Computer Science and Engineering & School of Software & School of Artificial Intelligence, Guangxi Normal University, China
zhouyc821@nenu.edu.cn, wupingfan994@nenu.edu.cn, wangl028@gxnu.edu.cn, fuzg432@nenu.edu.cn, yangfq147@nenu.edu.cn

Abstract

Sharpness-aware minimization (SAM) is widely recognized for enhancing the generalization performance of deep neural networks. However, recent works have challenged the statement that flatness implies generalization, demonstrating that it is insufficient as the indicator of generalization. In this paper, we reveal an insightful phenomenon: among minima of similar sharpness, stochastic optimization algorithms tend to prefer those with lower nonuniformity. We define nonuniformity by both the magnitude and structure of the gradient noise, and show that it fundamentally differs from sharpness and plays a critical role in generalization. Specifically, we first theoretically prove that the expected generalization gap of models trained via stochastic optimization algorithm is positively correlated with nonuniformity (the magnitude of the gradient noise). Empirically, we show that nonuniformity exhibits a stronger correlation with generalization than sharpness, especially in Transformer models. Furthermore, we demonstrate that the nonuniformity (the structure of the gradient noise) more effectively guides the algorithm towards sparser solutions and exhibits better generalization performance than sharpness-based methods in the high-dimensional sparse regression problem. Finally, extensive experiments on various datasets and models confirm the advantages of nonuniformity for generalization: (1) optimization guided by nonuniformity achieves better generalization compared to those achieved through flatness (including standard training, transfer learning, hyperparameter sensitivity and robustness to label noise); (2) model architecture (such as depth and width) is closely related to nonuniformity.

Code and Extended version —
https://github.com/YingcongZhou/Beyond-Sharpness-The-Role-of-Nonuniformity-in-Generalization

1 Introduction

Modern deep learnings success in achieving ever-better performance on various tasks has relied significantly on ever-heavier overparameterization. Many works have shown that the flatter loss surface often indicates the better generalization (Hochreiter and Schmidhuber 1994; Keskar et al. 2016; Li et al. 2018), and the stochastic optimization methods can implicitly converge to flatter regions of the loss landscape (Zhu et al. 2018; Liu, Ziyin, and Ueda 2020; Wu, Wang, and Su 2022; Wu and Su 2023). Sharpness-aware minimization (SAM) (Foret et al. 2020) and its variants (Zheng, Zhang, and Mao 2020; Du et al. 2021, 2022; Kwon et al. 2021; Liu et al. 2022; Mi et al. 2022; Zhong et al. 2022; Zhuang et al. 2022; Zhang et al. 2023; Li et al. 2024) have successfully leveraged this insight and achieved state-of-the-art results on various tasks (Chen, Hsieh, and Gong 2022; Wang et al. 2024; Chen et al. 2023; Zhang et al. 2022; Yue, Nouiehed, and Kontar 2020; Zhang et al. 2024), emphasizing the crucial role of sharpness in generalization.

By studying the dynamics of optimization with the unbiased noise, (Mori et al. 2022; Xie, Sato, and Sugiyama 2020; Liu, Ziyin, and Ueda 2020; Wojtowytsch 2021; Zhu et al. 2018) have revealed that the anisotropic noise inherent in SGD can effectively help escape from sharp minima. Furthermore, it has been observed that SGD tends to converge to the solution with the small gradient noise variance, which quantifies the consistency among the mini-batch gradients (Smith et al. 2021; HaoChen et al. 2020; Wu, Ma, and Weinan 2018). The gradient noise variance is an important factor beyond sharpness, that influences generalization.

To understand the impact of the flatness and the gradient noise on optimization, we conduct the experiments on the heuristic example suggested by (Wu, Ma, and Weinan 2018) and demonstrate that when minima exhibit the same sharpness but differ in their gradient noise variances, optimizers (SGD, MSGD, NSGD and SAM) tend to converge to the solution with the smaller gradient noise variance, as illustrated in Figure 1. It is a significant challenge to analyze the impact of the gradient noise (both the magnitude and the structure) on generalization and use this insight to improved generalization in practice.

In the present paper, we introduce the concept of nonuniformity, characterized by both the magnitude and the structure of the gradient noise. By decomposing the stochastic gradient of SGD into the full-batch gradient and gradient noise, we establish the connections between the generalization and the nonuniformity. Furthermore, for high-dimensional sparse regression, incorporating the structure of gradient noise into the optimization helps the algorithm find sparser solutions and achieves the superior generalization.

*Corresponding author.

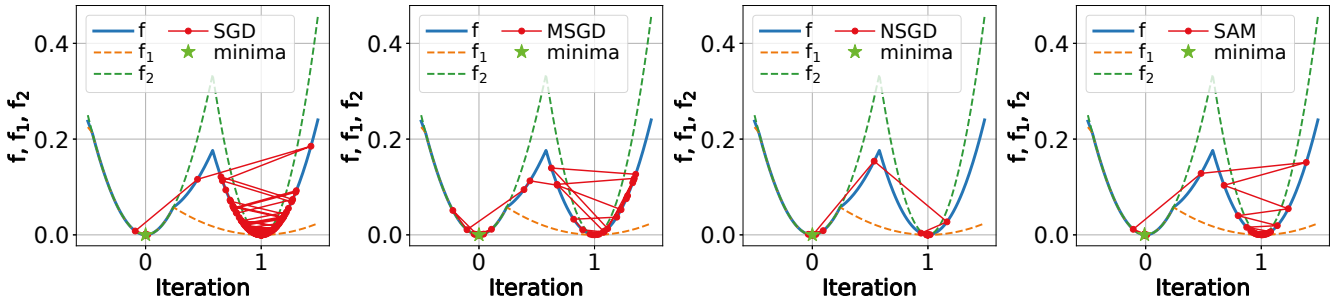

Figure 1: We visualize the trajectories of SGD, Momentum Stochastic Gradient Descent (MSGD), Nesterov Stochastic Gradient Descent (NSGD), and SAM on a 1D toy problem ($f = \frac{1}{2}(f_1 + f_2)$, where $f_1 = \min\{x^2, 0.1(x-1)^2\}$, $f_2 = \min\{x^2, 1.9(x-1)^2\}$) with two global minima of equal flatness but different nonuniformity. All methods consistently escape the higher-nonuniformity minimum at $x_1 = (1, 0)$ and converge to the lower-nonuniformity solution at $x_2 = (0, 0)$.

Finally, we verify our theoretical results through extensive experiments. Our contributions are summarized as follows:

- We establish the expected generalization gap for SGD that is bounded by nonuniformity. This finding suggests that nonuniformity, beyond flatness, is a crucial factor for generalization. Next, we empirically compare the relationship of sharpness with generalization and that of nonuniformity with generalization in ResNets and Vision Transformers (ViTs). The results show that for model ResNets, both metrics accurately reflect generalization. However, for ViTs, sharpness fails, as sharp and flat minima yield similar generalization, while nonuniformity remains a reliable indicator.
- Moreover, we theoretically show that embedding the anisotropic structure of gradient noise in SAM can encourage it to find the sparser solution and enhances generalization in high-dimensional sparse regression.
- We embed nonuniformity into the optimization process (Algorithm 1 with the convergence rate $\log 1/\sqrt{T}$) to guide the algorithm toward lower-nonuniformity solutions. The experiments, including standard training, transfer learning, hyperparameter sensitivity, and label noise robustness, show that this method effectively improve the generalization.

2 Preliminaries

2.1 Notations

Let $\mathcal{X}$ and $\mathcal{Y}$ be the sample space and the label space, respectively. We denote $\mathcal{D}$ as the underlying training distribution on $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n$ denote the training dataset with n data-points drawn independently from $\mathcal{D}$. Let $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ represent the parameters of the model. We use $\|\cdot\|$ to denote the Euclidean norm of a vector and $\mathrm{Tr}(\cdot)$ to denote the trace of a matrix. In addition, we define $B(\boldsymbol{\theta}, \rho)$ as the open ball of the radius $\rho > 0$ centered at the point $\boldsymbol{\theta}$ in the Euclidean space, i.e., $B(\boldsymbol{\theta}, \rho) = \{\boldsymbol{\theta}' : \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \rho\}$. The population loss is defined as $L_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\boldsymbol{\theta}, x, y)]$. In general, the supervised learning usually involves minimizing the empirical loss over the training data $\mathcal{S}$, i.e., $L_{\mathcal{S}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{(x_i,y_i)\in\mathcal{S}} \ell(\boldsymbol{\theta}, x_i, y_i) = \frac{1}{n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta})$, where $\ell_i(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}, x_i, y_i)$ represents the loss function for the i-th sample, and $(x_i, y_i) \in \mathcal{S}$ represents the i-th sample in $\mathcal{S}$.

Let $\nabla L_{\mathcal{S}}(\boldsymbol{\theta})$ and $\nabla^2 L_{\mathcal{S}}(\boldsymbol{\theta})$ denote the gradient vector and the Hessian matrix of the loss function $L_{\mathcal{S}}(\boldsymbol{\theta})$. Additionally, we use $|A|$ to represent the cardinality (i.e., the number of elements) of a set A.

2.2 Stochastic Gradient Descent

We consider Stochastic Gradient Descent (SGD) with the learning rate η_t, as defined by the update rule

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla L_{\mathcal{B}}(\boldsymbol{\theta}_t), \tag{1}$$

where the mini-batch stochastic gradients $\nabla L_{\mathcal{B}}(\boldsymbol{\theta}_t) = \frac{1}{b}\sum_{(x_i,y_i)\in\mathcal{B}} \nabla\ell(\boldsymbol{\theta}_t, x_i, y_i)$ arise when we consider a mini-batch $\mathcal{B} \subseteq \mathcal{S}$ of the size $b = |\mathcal{B}| \leq n$ of random indices drawn uniformly from $\{1, \ldots, n\}$. This provides an unbiased estimate of the gradient $\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)$. Then we can rewrite the SGD (1) as

$$\begin{aligned}\boldsymbol{\theta}_{t+1} =& \boldsymbol{\theta}_t - \eta_t \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t) + \eta_t \left[\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t) - \nabla L_{\mathcal{B}}(\boldsymbol{\theta}_t)\right] \\ =& \boldsymbol{\theta}_t - \eta_t \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t) + \eta_t \boldsymbol{\epsilon}(\boldsymbol{\theta}_t),\end{aligned} \tag{2}$$

where $\boldsymbol{\epsilon}(\boldsymbol{\theta}_t) = \eta_t \left[\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t) - \nabla L_{\mathcal{B}}(\boldsymbol{\theta}_t)\right]$ is the gradient noise. The conditional covariance (without replacement) of the gradient noise is calculated by the train set $\mathcal{S}$, i.e., gradient covariance matrix (GCM)

$$\mathrm{Cov}(\boldsymbol{\epsilon}(\boldsymbol{\theta}_t)|\boldsymbol{\theta}_t) := C(\boldsymbol{\theta}_t) \approx$$
$$\hat{\alpha}\left[\frac{1}{n}\sum_{i=1}^n \nabla\ell_i(\boldsymbol{\theta}_t)\nabla\ell_i(\boldsymbol{\theta}_t)^\top - \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)^\top\right],$$

where $\hat{\alpha} = \frac{(n-b)}{b(n-1)}$.

2.3 Sharpness-Aware Minimization

The authors of (Foret et al. 2020) proposed Sharpness-Aware Minimization (SAM), which is an effective procedure that improves model generalization by penalizing sharpness. In their work, the sharpness defined by the robustness of the parameters

$$\max_{\boldsymbol{\theta}'\in B(\boldsymbol{\theta},\rho)} L_{\mathcal{S}}(\boldsymbol{\theta}') - L_{\mathcal{S}}(\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}$ is the perturbation of the parameters and $\rho > 0$ is the perturbation radius.

2.4 Algorithm Stability

Let $\mathcal{A}$ be a randomized optimization algorithm. We are interested in the expected generalization gap of $\mathcal{A}$ when training with n samples from a distribution $\mathcal{D}$ which is given by,

$$\text{gap}(\mathcal{D}, n) = \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n} \mathop{\mathbb{E}}_{\boldsymbol{\theta}} [L_{\mathcal{D}}(\mathcal{A}(\boldsymbol{\theta})) - L_{\mathcal{S}}(\mathcal{A}(\boldsymbol{\theta}))],$$

where $\mathcal{A}(\boldsymbol{\theta})$ denotes the model obtained by running $\mathcal{A}$ on $\mathcal{S}$ and $\boldsymbol{\theta}$ denotes the sequence of coin tosses used by $\mathcal{A}(\boldsymbol{\theta})$ in a given run. The definition of the algorithm stability (Chatterjee and Zielinski 2022) is as follows.

Definition 1 (Algorithm Stability)**.** *Let $\mathcal{S}' = \{(x_i', y_i')\}_{i=1}^n$ be a second dataset of n sample drawn $i.i.d.$ from $\mathcal{D}$. The expected stability of $\mathcal{A}$ is given by*

$$\begin{aligned}
&\text{stable}(\mathcal{D}, n) \\
&= \mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n} \mathop{\mathbb{E}}_{\mathcal{S}'\sim\mathcal{D}^n} \mathop{\mathbb{E}}_{\boldsymbol{\theta}} [\frac{1}{n} \sum_{(x_i,y_i)\in\mathcal{S}} [\ell(\mathcal{A}_{\boldsymbol{\theta}}(\mathcal{S}^{(i)}), x_i, y_i) \\
&\quad - \ell(\mathcal{A}_{\boldsymbol{\theta}}(\mathcal{S}), x_i, y_i)]]
\end{aligned}$$

where $\mathcal{S}^{(i)} = \{(x_1, y_1), \ldots, (x_i', y_i'), \ldots, (x_n, y_n)\}$, that is, $\mathcal{S}^{(i)}$ is obtained by replacing the i-th sample (x_i, y_i) in $\mathcal{S}$ with (x_i', y_i').

Proposition 1. *(Stability equals generalization (Chatterjee and Zielinski 2022)). When training with n samples from a distribution $\mathcal{D}$, we have,*

$$\text{gap}(\mathcal{D}, n) = \text{stable}(\mathcal{D}, n).$$

We will establish the expected generalization gap for SGD in Theorem 1 by Proposition 1.

3 Nonuniformity and Generalization

In this section, we propose the concept of nonuniformity and study how it influences the optimization dynamics and the generalization performance.

Nonuniformity (hessian-based), as first introduced in (Wu, Ma, and Weinan 2018), is a concept that characterizes the stability properties of global minima (see Definition 2).

Definition 2. *Let $H = \frac{1}{n}\sum_{i=1}^n H_i$, $\Sigma = \frac{1}{n}\sum_{i=1}^n H_i^2 - H^2$ and $H_i = \nabla^2 \ell_i(\boldsymbol{\theta}^*)$ where $\boldsymbol{\theta}^*$ is a local minima. The nonuniformity is defined as $s = \lambda_{\max}(\Sigma^{1/2})$ to be the nonuniformity.*

As defined in Definition 2, nonuniformity involves the largest eigenvalue of Σ, making it computationally expensive. Moreover, since it depends on the local curvature near minima, it cannot capture nonuniformity along the full optimization trajectory. In this paper, nonuniformity (gradient-based) is defined by two factors: (1) **the magnitude of the gradient noise;** (2) **the structure of the gradient noise.**

Our definition of non-uniformity, compared to Definition 2, is computationally efficient and can be utilized to improve generalization performance.

3.1 The Magnitude of the Gradient Noise

Definition 3. *For any $\boldsymbol{\theta} \in \Theta$, we define the magnitude of the gradient noise $\boldsymbol{\epsilon}(\boldsymbol{\theta})$ at point $\boldsymbol{\theta}$ as*

$$\text{Tr}\,[C(\boldsymbol{\theta})] = \hat{\alpha}[\frac{1}{n}\sum_{i=1}^n \|\nabla \ell_i(\boldsymbol{\theta})\|^2 - \|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2]. \quad (4)$$

The gradient noise $\boldsymbol{\epsilon}(\boldsymbol{\theta})$ arises from the stochasticity of mini-batch sampling and reflects the deviation of mini-batch gradients from the full-batch gradient. Smaller noise implies better directional alignment, leading to more stable updates and improved generalization. To explicitly promote gradient consistency, we can directly regularize $\text{Tr}\,[C(\boldsymbol{\theta})]$, which quantifies the directional variance of stochastic gradients. Notably, 1-SAM corresponds to a specific case of this regularization. We give the extreme versions of the b-SAM as follows

$$\text{1-SAM: } \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{n}\sum_{i=1}^n \nabla \ell_i(\boldsymbol{\theta}_t + \rho' \nabla \ell_i(\boldsymbol{\theta}_t)), \quad (5)$$

$$n\text{-SAM: } \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t + \rho' \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)), \quad (6)$$

where ρ' in (5) and (6) is the perturbation radius of the parameters. $\|\nabla \ell_i(\boldsymbol{\theta}_t)\|$ and $\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)\|$ is not necessary for improving generalization, so we will omit it from our theoretical analysis (Andriushchenko and Flammarion 2022).

Next, we directly set $\text{Tr}\,[C(\boldsymbol{\theta})]$ and the gradient norm as the penalty term, i.e., $\tilde{L}_{\mathcal{S}}(\boldsymbol{\theta}) = L_{\mathcal{S}}(\boldsymbol{\theta}) + \lambda\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\| + \lambda \text{Tr}\,[C(\boldsymbol{\theta})]$, where λ is the penalty coefficient. Then, we approximate the gradient $\tilde{L}_{\mathcal{S}}(\boldsymbol{\theta})$ using the Hessian-vector product approximation:

$$\begin{aligned}
&\nabla \tilde{L}_{\mathcal{S}}(\boldsymbol{\theta}) = \nabla L_{\mathcal{S}}(\boldsymbol{\theta}) + \lambda \nabla \|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2 + \lambda \nabla \text{Tr}(\text{Cov}(\boldsymbol{\theta})) \\
&\approx (1 - \frac{\lambda}{\rho'}) \nabla L_{\mathcal{S}}(\boldsymbol{\theta}) + \frac{\lambda}{\rho'}(1 - \hat{\alpha}) \nabla L_{\mathcal{S}}(\boldsymbol{\theta} + \rho \nabla L_{\mathcal{S}}(\boldsymbol{\theta})) \\
&+ \frac{\lambda}{\rho'} \hat{\alpha} \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(\boldsymbol{\theta} + \rho \nabla \ell_i(\boldsymbol{\theta})), \quad (7)
\end{aligned}$$

If $\frac{\lambda}{\rho'} = 1$ and $\hat{\alpha} = 1$, we can obtain the update rule of 1-SAM. Importantly, the smaller values of the parameter b lead to better generalization in b-SAM (Andriushchenko and Flammarion 2022). This aligns with our analysis: 1-SAM implicitly regularizes $\text{Tr}[C(\boldsymbol{\theta})]$, thereby reducing gradient noise. These findings emphasize the role of gradient noise in generalization.

Nonuniformity and Sharpness. Nonuniformity (Definition 4) is defined over the sample space and reflects the variability of mini-batch gradients, while sharpness characterizes the local curvature of the loss in the parameter space. Despite their distinction, they may coincide in special cases, as shown by a simple example.

Example 1. *Let $\mathcal{S} = \{x_i \times y_i \in \mathbb{R} \times \mathbb{R}\}_{i=1}^n$ is a dataset, $\boldsymbol{\theta} = (\theta_1, \theta_2)^T \in \mathbb{R}^2$ are parameters of network. We definite the loss is $L_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^n (\theta_1\theta_2 x_i - y_i)^2$.*

Algorithms based on sharpness metrics, such as SAM, implicitly minimize the trace of the Hessian matrix (Wen, Ma, and Li 2023). Thus, We calculate the trace of $H(\boldsymbol{\theta})$ and

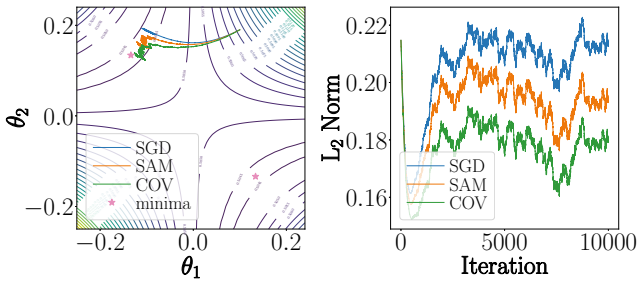

Figure 2: We visualize the optimization trajectories (left) and solution norms (right) of SGD, SAM ($\rho = 0.005$), and COV ($\lambda = 0.3$) on $L_{\mathcal{S}}(\boldsymbol{\theta})$ using synthetic data $X \sim \mathcal{N}(0, I)$, $y = \theta_1^* \theta_2^* X + \mathcal{N}(0, I)$, with $\eta = 0.01$, $b = 100$, and $k = 10^4$.

$C(\boldsymbol{\theta})$ as $\mathrm{Tr}(H(\boldsymbol{\theta})) = (\frac{1}{n}\sum_{i=1}^n x_i^2) \cdot \|\boldsymbol{\theta}\|^2$ and $\mathrm{Tr}(C(\boldsymbol{\theta})) = (\frac{1}{n}\sum_{i=1}^n r_i^2 x_i^2 - \frac{1}{n^2}(\sum_{i=1}^n r_i x_i)^2) \cdot \|\boldsymbol{\theta}\|^2$ where the residual term is $r_i = (\theta_1\theta_2 x_i - y_i)$. The conclusion show that minimizing sharpness or nonuniformity is equivalent to obtaining a minimum norm solution. Detailed proof are provided in Appendix A.7.

We used SAM to reduce sharpness and added $\mathrm{Tr}(C(\boldsymbol{\theta}))$ as a regularizer to minimize nonuniformity (COV). As shown in Fig. 2, all three methods converge to a connected region composed of $\boldsymbol{\theta}$ satisfying $\nabla L_{\mathcal{S}}(\boldsymbol{\theta}) = \mathbf{0}$. Within this region, SGD nearly stops updating, while SAM and COV continue exploring to find minimum-norm solutions. This difference arises from SAMs focus on loss sharpness and COVs on gradient nonuniformity. Notably, COV tends to find solutions closer to the minimum-norm solution (i.e., the flatter solution under the setting of example 1) than SAM. However, when the two properties differ, nonuniformity guided solutions may not be flat, providing insight into why sharp minima can still generalize well.

Before presenting the main theorem, we make some standard assumptions in the stochastic optimization (Andriushchenko and Flammarion 2022; Jiang et al. 2023; Karimi, Nutini, and Schmidt 2016; Ghadimi and Lan 2013).

Assumption 1. *We assume that $\ell(\cdot, x_i, y_i)$ is γ-Lipschitz for every $(x_i, y_i) \in \mathcal{S}$, i.e.,*

$$\left|\ell(\boldsymbol{\theta}, x_i, y_i) - \ell(\boldsymbol{\theta}', x_i, y_i)\right| \leq \gamma \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|,$$

and $\ell(\cdot, x_i, y_i)$ is β-smooth, i.e.,

$$\|\nabla \ell(\boldsymbol{\theta}, x_i, y_i) - \nabla \ell(\boldsymbol{\theta}', x_i, y_i)\| \leq \beta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Assumption 2. *There exists a constant $M > 0$ for any data batch $\mathcal{B}$ such that*

$$\mathbb{E}\left[\|\nabla L_{\mathcal{B}}(\boldsymbol{\theta}) - \nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2\right] \leq M,$$

and exists $G > 0$ such that $\mathbb{E}\,\|\nabla \ell(\boldsymbol{\theta}, x_i, y_i)\| \leq G$.

Generalization. Now, we utilize Theorem 1 to demonstrate how nonuniformity impact generalization.

Theorem 1. *Under Assumption 1 and 2, if stochastic gradient descent is run for T steps on the training set consisting of n examples drawn from the distribution $\mathcal{D}$, we have the following expected algorithm stability bound for the models trained by SGD*

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\mathcal{D}^n} \mathop{\mathbb{E}}_{(x_i', y_i')\sim\mathcal{D}} \mathop{\mathbb{E}}_{\boldsymbol{\theta}} \left|\ell(A_{\boldsymbol{\theta}}(\mathcal{S}^{(i)}), x_i, y_i) - \ell(A_{\boldsymbol{\theta}}(\mathcal{S}), x_i, y_i)\right|$$
$$\leq \sum_{t\in[T]} c_T \eta_t \sqrt{\mathrm{Tr}\left[C(\boldsymbol{\theta}_{t-1})\right]}, \tag{8}$$

where c_T is the constant.

Remark 1. *Combining Theorem 1 (the proof provided in Appendix A.1) and Proposition 1, we can obtain*

$$|\mathrm{gap}| \leq \sum_{t\in[T]} c_T \eta_t \sqrt{\mathrm{Tr}\left[C(\boldsymbol{\theta}_{t-1})\right]}.$$

The conclusion shows that generalization is bounded by nonuniformity: lower nonuniformity (Definition 3) implies better generalization.

Correlation. Theorem 1 suggests that lower nonuniformity leads to better generalization. To verify this, we visualize the correlation between nonuniformity, sharpness, and test error (1 - test accuracy) using scatter plots. In models like ResNet, both metrics correlate positively with generalization, with nonuniformity showing stronger alignment (Fig. A.8 in Appendix A.8). However, in ViTs, the correlation between sharpness and generalization is less pronounced: low test error can occur at both high and low sharpness (see Fig. 3 and Fig. A.7 in Appendix A.8), suggesting that both sharp and flat minima can generalize well. The experimental settings are detailed in the Appendix A.8.

3.2 The Structure of the Gradient Noise

The papers (Mori et al. 2022; Xie, Sato, and Sugiyama 2020; Liu, Ziyin, and Ueda 2020; Wojtowytsch 2021) study the anisotropic structure of gradient noise. Specifically, GCM tends to align with Hessian, with larger variances corresponding to sharper directions. This alignment helps optimizers escape sharp minima. (Peng et al. 2025) shows that the learned models generalize better if the large-variance directions of the final weight covariance have small local curvatures. This motivates perturbing parameters along high-variance directions. We formalize this idea by defining the noise energy along the direction $\boldsymbol{v}$.

Definition 4. *For a fixed direction $\boldsymbol{v}$, the noise energy along the direction $\boldsymbol{v}$ is defined by*

$$\mathbb{E}\left[(\boldsymbol{\epsilon}(\boldsymbol{\theta})^\top \boldsymbol{v})^2\right] = \boldsymbol{v}^\top C(\boldsymbol{\theta}) \boldsymbol{v}. \tag{9}$$

Definition 4 quantifies the magnitude of nonuniformity along a given direction $\boldsymbol{v}$ via $\boldsymbol{v}^\top C(\boldsymbol{\theta})\boldsymbol{v}$. Thus, the geometry of the ellipsoid is determined by the anisotropic structure of the GCM. By utilizing the anisotropic structure of the GCM, we introduce perturbations within the noise ellipsoid characterized by this quadratic form. Building on this, we demonstrate from both theoretical and experimental perspectives that incorporating the nonuniformity can further enhance generalization.

Definition 5. *Considering the perturbation in the noise ellipsoid defined in (9), the perturbation can be measured by*

$$\max_{\bar{\boldsymbol{\epsilon}}^T \boldsymbol{C}(\boldsymbol{\theta})\bar{\boldsymbol{\epsilon}} \leq \rho^2} L_{\mathcal{S}}(\boldsymbol{\theta} + \bar{\boldsymbol{\epsilon}}) - L_{\mathcal{S}}(\boldsymbol{\theta}).$$

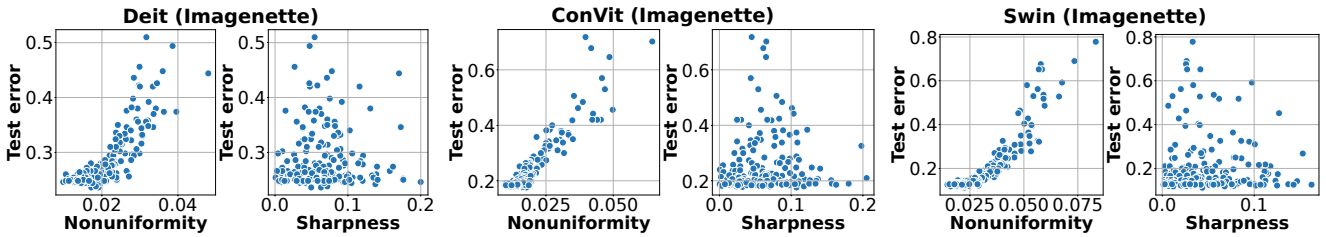

Figure 3: The correlation between nonuniformity, sharpness, and generalization during optimization in ViTs. The Pearson correlation coefficient is given in Appendix A.8.

We apply the Lagrange multipliers method to find the optimal perturbation $\bar{\epsilon}^*$ which is given by Lemma 1.

Lemma 1. *According to Definition 5, the optimal perturbation is calculated by*

$$\bar{\epsilon}_t^* = \rho \frac{C(\boldsymbol{\theta}_t)^{-1}\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)}{\sqrt{\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)^T \boldsymbol{C}(\boldsymbol{\theta}_t)^{-1}\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)}}. \tag{10}$$

The proof can be found in Appendix A.2. According to Lemma 1, we embed the structure of the GCM in the standard SAM update, called SAMGCM, i.e.,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t + \bar{\epsilon}_t^*) \tag{11}$$

where $\bar{\epsilon}^*$ is updated by (10).

To demonstrate that incorporating nonuniformity can enhance generalization, we theoretically explore the implicit bias of SAMGCM (11) in the diagonal linear network for the high sparse regression problem introduced by (Woodworth et al. 2020).

Model. We consider the following two-layer linear diagonal network, where a linear predictor $\langle \boldsymbol{\theta}, x \rangle$ can be parametrized via $\boldsymbol{\theta} = \boldsymbol{\theta}_+^2 - \boldsymbol{\theta}_-^2$. In this setting, we investigate the over-parametrized sparse regression problem, where the ground truth that $\boldsymbol{\theta}^*$ is a sparse vector and the loss function used is the squared loss, i.e.,

$$L(\boldsymbol{\theta}) := \frac{1}{4n}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_+^2 - \boldsymbol{\theta}_-^2)x_i - y_i\right)^2. \tag{12}$$

Before presenting the result of SAMGCM's implicit bias, we introduce the key implicit regularizer $\phi_{\boldsymbol{\alpha}}$ (see Appendix A.3 for details) interpolates between the ℓ_1 and ℓ_2 norms in (Woodworth et al. 2020). The initialization scale determines the implicit bias of the gradient flow. Specifically, the algorithm, starting from $\boldsymbol{\alpha}$, converges to the minimum ℓ_1-norm interpolator for the small $\boldsymbol{\alpha}$ and to the minimum ℓ_2-norm interpolator for the large $\boldsymbol{\alpha}$. A smaller $\boldsymbol{\alpha}$ indicates a sparser solution, which often correlates with better generalization. To simplify the analysis, we assume that the GCM is diagonal, i.e., $C(\boldsymbol{\theta}) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$, where λ_i is the i-th eigenvalue of GCM. We now present the implicit regularization result for SAMGCM in the following theorem.

Theorem 2. *Suppose $C(\boldsymbol{\theta}) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$. If the solution of SAMGCM (11) $\boldsymbol{\theta}_{\mathrm{cov},\infty}$ started from initial values $\boldsymbol{\theta}_+ = \boldsymbol{\theta}_- = \boldsymbol{\alpha} \in \mathbb{R}^d_{>0}$, for the squared parameter problem (12) satisfies $X\boldsymbol{\theta}_\infty = y$, then*

$$\boldsymbol{\theta}_{\mathrm{cov},\infty} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \phi_{\boldsymbol{\alpha}_{\mathrm{cov}}}(\boldsymbol{\theta}) \quad s.t. \quad X\boldsymbol{\theta} = y,$$

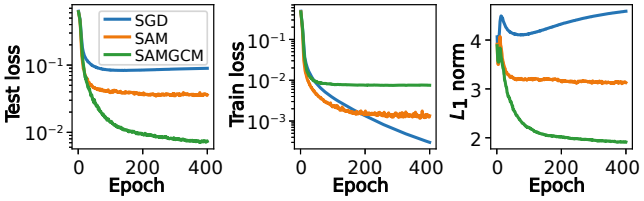

Figure 4: (Left) The test loss of SGD, SAM and SAMGCM. (Middle) The train loss of the three algorithms. (Right) The ℓ_1-norm of the solutions for the three algorithms. We generate data $X \sim \mathcal{N}(0, I)$, initialize the parameters $\boldsymbol{\theta}_0 \sim \mathcal{N}(0, 1) * 0.2$. The learning rate is $\eta = 0.05$ and the radius is $\rho = 0.1$ for SAM and $\rho = 0.005$ for SAMGCM.

where $\boldsymbol{\alpha}_{\mathrm{cov}} \approx \boldsymbol{\alpha} \odot \exp\left[-\frac{\rho}{n^2}\int_0^t \boldsymbol{C}(\boldsymbol{\theta}(s))^{-1}\left(X^T\boldsymbol{r}(s)\right)^2 \mathrm{ds}\right]$ and $\boldsymbol{r}(s) = X^\top\boldsymbol{\theta}(s) - y$ is the residual term.

Remark 2. *The proof can be found Appendix A.3. Under the same settings, (Andriushchenko and Flammarion 2022) shows that the term $\boldsymbol{a}_{sam}$ of SAM is $\boldsymbol{\alpha}_{\mathrm{sam}} \approx \boldsymbol{\alpha} \odot \exp\left[-\frac{\rho}{n^2}\int_0^t \left(X^T\boldsymbol{r}(s)\right)^2 \mathrm{ds}\right]$, which can be derived from $\boldsymbol{\alpha}_{\mathrm{cov}}$ by setting $\boldsymbol{C}(\boldsymbol{\theta}) = I$. In deep neural networks, GCM is highly anisotropic, i.e., it has the distinguished top eigenvalues (Zhu et al. 2018; Xie, Sato, and Sugiyama 2020; Wu et al. 2019). Therefore, $\boldsymbol{C}(\boldsymbol{\theta})^{-1}$ has mostly large eigenvalues and a few small eigenvalues. Let $J_{\mathrm{cov}} = \int_0^t \boldsymbol{C}(\boldsymbol{\theta}(s))^{-1}\left(X^T\boldsymbol{r}(s)\right)^2 \mathrm{ds}$ in $\boldsymbol{\alpha}_{\mathrm{cov}}$. We discretize J_{cov} as $J_{\mathrm{cov}} \approx \sum_{\tau_i\in(0,t)}\sum_{j=1}^d \lambda_j^{-1}(\tau_i)\left(X^T\boldsymbol{r}(\tau_i)\right)_j^2$, where $\tau_i \in (0,t)$ is a partition of the interval $(0,t)$, and j denotes the j-th element of the vector. Since λ_j^{-1} is generally larger than 1 for most j, we have $J_{\mathrm{cov}} > \sum_{\tau_i\in(0,t)}\sum_{j=1}^d \left(X^T\boldsymbol{r}(\tau_i)\right)_j^2 \approx J_{sam}$, where $J_{sam} = \int_0^t \left(X^T\boldsymbol{r}(s)\right)^2 \mathrm{ds}$. Then we can obtain that $\|\boldsymbol{a}_{\mathrm{cov}}\|_1 < \|\boldsymbol{a}_{\mathrm{sam}}\|_1$. Consequently, the value of $\boldsymbol{\alpha}$ in SAMGCM is lower than that in SAM, leading to the sparser solution in SAMGCM, which enhances generalization.*

Empirical evidence for SAMGCM (11). Fig. 4 show that SAMGCM produces the sparsest solutions, significantly outperforming SAM and SGD in terms of generalization. Furthermore, SAMGCM shows superior performance in preventing overfitting compared to SAM and SGD.

4 Optimization Guided by Nonuniformity

In Sec 3.1 and 3.2, we theoretically analyzed two key contributors to nonuniformity: its magnitude and structure. Utilizing the structural properties of the GCM in practice requires costly matrix inversion, making efficient approximation a key open challenge for future work. This section investigates noise magnitude and empirically demonstrates its superiority over sharpness. While Theorem 1 implies that minimizing $\mathrm{Tr}[C(\boldsymbol{\theta})]$ is a natural objective, its practical implementation is computationally expensive because computing the gradient of per-sample gradient norms involves multiple backpropagations and Hessian computations. To reduce the computational overhead, we approximate $\mathrm{Tr}[C(\boldsymbol{\theta})]$ by using mini-batch gradients. According to the Cauchy-Schwarz inequality, we known that $\frac{1}{n}\sum_{i=1}^{n}\|\nabla \ell_i(\boldsymbol{\theta})\|^2 \geq \frac{b}{n}\sum_{i=1}^{n/b}\|\nabla L_{\mathcal{B}_i}(\boldsymbol{\theta})\|^2$. Thus, the number of backpropagation steps is reduced from n to n/b. Since each update is based on a mini-batch, we can compute the deviation between mini-batch and full-batch gradients incrementally. After a full pass over the training data (i.e., one epoch), this approach accumulates all such deviations, yielding a more tractable surrogate for $\mathrm{Tr}[C(\boldsymbol{\theta})]$. That is,

$$\Sigma(\boldsymbol{\theta}) = \|\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\|^2 - \|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2. \tag{13}$$

According to (13), we can obtain the following constrained optimization problem to minimize the worst-case loss,

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\mathcal{B}}[L_{\mathcal{B}}(\boldsymbol{\theta}+\tilde{\boldsymbol{\epsilon}})], \ \text{ s.t. } \tilde{\boldsymbol{\epsilon}} = \arg\max_{\|\tilde{\boldsymbol{\epsilon}}\|_2\leq\rho} \Sigma(\boldsymbol{\theta}+\tilde{\boldsymbol{\epsilon}}). \tag{14}$$

We now apply the Lagrange multipliers method to find the optimal perturbation $\tilde{\boldsymbol{\epsilon}}^*(\boldsymbol{\theta})$.

Lemma 2. *The optimal perturbation $\tilde{\boldsymbol{\epsilon}}^*$, which is defined in (14) can be computed by*

$$\tilde{\boldsymbol{\epsilon}}^*(\boldsymbol{\theta}) = \rho\frac{\nabla\|\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\|^2 - \nabla\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2}{\|\nabla\|\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\|^2 - \nabla\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta})\|^2\|}.$$

The proof can be found in Appendix A.4. We note that computing $\tilde{\boldsymbol{\epsilon}}^*(\boldsymbol{\theta})$ involves addressing two key challenges: (1) Calculating the Hessian, i.e., $\nabla\|\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\|^2 = 2\nabla^2 L_{\mathcal{B}}(\boldsymbol{\theta})\nabla L_{\mathcal{B}}(\boldsymbol{\theta})$. (2) Calculating the gradient of the overall loss $\nabla\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)\|$. Firstly, for the challenge (1), we can approximate $\nabla\|\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\|$ by using Hessian-vector products

$$\begin{aligned}\nabla^2 L_{\mathcal{B}}(\boldsymbol{\theta})\nabla L_{\mathcal{B}}(\boldsymbol{\theta}) \approx& [\nabla L_{\mathcal{B}}\left(\boldsymbol{\theta}+\rho\nabla L_{\mathcal{B}}(\boldsymbol{\theta})\right) - \nabla L_{\mathcal{B}}(\boldsymbol{\theta})]/\rho \\ :=& (\boldsymbol{g}(\boldsymbol{\theta},\mathcal{B},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{B}))/\rho.\end{aligned}$$

This approach leverages finite differences to approximate the gradient without explicitly computing the Hessian. Thus, the optimal perturbation can be calculate as follows

$$\tilde{\boldsymbol{\epsilon}}^*(\boldsymbol{\theta}) \approx \frac{\boldsymbol{g}(\boldsymbol{\theta},\mathcal{B},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{B}) - [\boldsymbol{g}(\boldsymbol{\theta},\mathcal{S},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{S})]}{\|\boldsymbol{g}(\boldsymbol{\theta},\mathcal{B},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{B}) - [\boldsymbol{g}(\boldsymbol{\theta},\mathcal{S},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{S})]\|}.$$

For the challenge (2), we estimate $\boldsymbol{g}(\boldsymbol{\theta},\mathcal{S},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{S})$ using an exponentially moving average (EMA) of the historical mini-batch gradients, i.e., $\boldsymbol{m}_t = \alpha\boldsymbol{m}_{t-1} + (1-\alpha)\left[\boldsymbol{g}(\boldsymbol{\theta},\mathcal{B},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{B})\right]$, where $0 < \alpha < 1$ is a hyperparameter. EMA allows us to approximate the full gradient $\boldsymbol{g}(\boldsymbol{\theta},\mathcal{S},\rho) - \boldsymbol{g}(\boldsymbol{\theta},\mathcal{S})$ with minimal additional computational overhead. We further demonstrate the effectiveness of this approximation in Theorem 3, which shows that $\boldsymbol{m}_t$ is a reliable estimate of the full gradient.

Algorithm 1

Input: Batch size b, Learning rate η_t, Perturbation radius ρ_{sam}, ρ_{cov}, Trade-off coefficient $\alpha > 0$, Small constant ξ.
Parameter: Optional list of parameters
Output: Trained weight $\boldsymbol{\theta}_t$

Let $t = 0$.
while $\boldsymbol{\theta}_t$ not converged **do**
  $\boldsymbol{g}_{1,t} = \nabla\hat{L}_{B_t}(\boldsymbol{\theta}_t)$
  $\boldsymbol{g}_{2,t} = \nabla\hat{L}_{B_t}\left(\boldsymbol{\theta}_t + \rho_{\text{sam}}\frac{\boldsymbol{g}_{1,t}}{\|\boldsymbol{g}_{1,t}\|+\xi}\right)$
  $\boldsymbol{m_t} = \alpha\boldsymbol{m_{t-1}} + (1-\alpha)(\boldsymbol{g}_{2,t} - \boldsymbol{g}_{1,t})$
  Compute adversarial perturbation:

$$\tilde{\boldsymbol{\epsilon}}_t = \rho_{cov}\boldsymbol{d}_t/\|\boldsymbol{d}_t\| \ \text{ where } \ \boldsymbol{d}_t = \boldsymbol{g}_{2,t} - \boldsymbol{g}_{1,t} - \boldsymbol{m}_t.$$

  Compute the gradient: $\boldsymbol{g}_t = \nabla\hat{L}_{B_t}(\boldsymbol{\theta}_t + \rho_{cov}\tilde{\boldsymbol{\epsilon}}_t)$
  Update $\boldsymbol{\theta}$ using gradient descent: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{g}_t$
  $t \leftarrow t+1$
end while
return $\boldsymbol{\theta}_t$

Theorem 3. *Suppose Assumptions 1, 2 hold. Assume that Algorithm 1 uses SGD as the base optimizer with a learning rate $\eta = \mathcal{O}(1/\sqrt{T})$ and $\rho = \mathcal{O}(1/\sqrt{T})$ to update the model parameter. Then by setting $\alpha = 1 - C'(\eta\beta\rho + 2\rho + 2\eta)^{\frac{2}{3}}$, $\boldsymbol{g}_{\text{full}}(\boldsymbol{\theta}_T,\mathcal{S}) = \boldsymbol{g}(\boldsymbol{\theta}_T,\mathcal{S},\rho) - \boldsymbol{g}(\boldsymbol{\theta}_T,\mathcal{S})$, and after a sufficiently large number of iterations T, with probability $1-\delta$,*

$$\|\boldsymbol{m}_T - \boldsymbol{g}_{\text{full}}(\boldsymbol{\theta}_T,\mathcal{S})\| \leq \mathcal{O}(M^{\frac{1}{3}}(\log(\frac{1}{\delta}))^{\frac{1}{3}}G^{\frac{1}{3}}\beta^{\frac{1}{3}}T^{-\frac{1}{6}}).$$

where C' is the universal constants.

The proof is provided in Appendix A.5. Theorem 3 show that the EMA approximation is reliable for large T (see Figure A.6 in Appendix A.8). We summarize the algorithmic steps in Algorithm 1. Furthermore, we analyze the convergence of Algorithm 1 in the non-convex setting.

Theorem 4. *Assume Assumptions 1 and 2 hold. Choosing $\eta_t = \eta_0/\sqrt{T} \leq 1/\beta$ and $\rho_t = \rho_0/\sqrt{T}$, then for any $\alpha \in (0,1)$, Algorithm 1 ensures that*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L_{\mathcal{S}}(\boldsymbol{\theta}_t)\|^2 \leq& \frac{2\mathbb{E}L_{\mathcal{S}}(\boldsymbol{\theta}_0) - 2\mathbb{E}L_{\mathcal{S}}(\boldsymbol{\theta}_T)}{\eta_0\sqrt{T}} \\ &+ \frac{2\eta_0\beta M}{\sqrt{T}} + \frac{2\eta_0\beta}{\sqrt{T}} + \frac{2\beta^2\rho_0}{\sqrt{T}},\end{aligned}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_0$ are a optimal solution and a initial value.

We provide the proof in Appendix A.6. Theorem 4 shows that Algorithm 1 achieves the convergence rate $\mathcal{O}(1/\sqrt{T})$.

5 Empirical Validation of Nonuniformity

Standard Training. To evaluate the effectiveness of nonuniformity, we compare Algorithm 1 with flatness-based methods to assess its impact on generalization. The results shown in Tab. 1 and experimental details are see Appendix A.8.

Robustness to Perturbation Radius. One notable limitation of SAM is its sensitivity to ρ, needing dataset-specific

MNIST	SAM	ASAM	**Ours**
FCN-3	$98.65_{\pm 0.01}$	$98.66_{\pm 0.01}$	$\mathbf{98.82}_{\pm 0.02}$
CIFAIR-10	SAM	ASAM	**Ours**
Resnet-20	$93.57_{\pm 0.16}$	$93.96_{\pm 0.19}$	$\mathbf{94.13}_{\pm 0.12}$
Resnet-56	$95.11_{\pm 0.04}$	$95.10_{\pm 0.07}$	$\mathbf{95.17}_{\pm 0.12}$
Resnet-28-6	$96.85_{\pm 0.15}$	$96.73_{\pm 0.12}$	$\mathbf{97.14}_{\pm 0.08}$
Resnet-58-6	$96.93_{\pm 0.03}$	$96.95_{\pm 0.05}$	$\mathbf{97.19}_{\pm 0.12}$
Pyramidnet110	$96.59_{\pm 0.1}$	$96.61_{\pm 0.14}$	$\mathbf{96.77}_{\pm 0.2}$
DenseNet100	$87.63_{\pm 0.18}$	$87.16_{\pm 0.16}$	$\mathbf{88.19}_{\pm 0.11}$
CIFAIR-100	SAM	ASAM	**Ours**
Resnet-20	$71.38_{\pm 0.18}$	$\mathbf{71.45}_{\pm 0.11}$	$71.32_{\pm 0.17}$
Resnet-56	$75.75_{\pm 0.09}$	$76.10_{\pm 0.13}$	$\mathbf{76.33}_{\pm 0.17}$
Resnet-28-6	$81.73_{\pm 0.12}$	$82.65_{\pm 0.17}$	$\mathbf{82.97}_{\pm 0.16}$
Resnet-58-6	$82.31_{\pm 0.16}$	$82.28_{\pm 0.18}$	$\mathbf{82.73}_{\pm 0.17}$
Pyramidnet110	$81.17_{\pm 0.12}$	$81.72_{\pm 0.17}$	$\mathbf{81.90}_{\pm 0.11}$
ViT-Small	SAM	ASAM	**Ours**
CIFAR-10	$82.23_{\pm 0.04}$	$82.68_{\pm 0.05}$	$\mathbf{83.03}_{\pm 0.02}$

Table 1: Comparison of test accuracy (%).

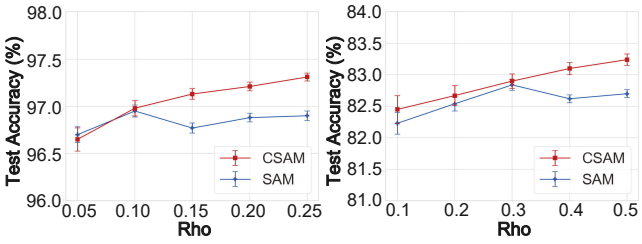

Figure 5: Results under different perturbation radii ρ.

tuning. In contrast, Algorithm 1 enforces gradient nonuniformity, resulting in more stable performance and greater robustness to ρ, particularly at larger values (see Fig. 5).

Robustness to Label Noise. We introduce the symmetric label noise by random flipping on CIFAR-10 using ResNet-28-6. The results show that Algorithm 1 consistently improves the performance from SAM, confirming its improved generalization (see Tab. 2).

Transfer Learning. Transfer learning shows the generalization of models when trained on sufficient labeled data and finetuned on a novel dataset. We use SGD, ASAM and Algorithm 1 to train ViTs (pre-trained on ImageNet). For the perturbation radii of ASAM, we performed grid searches over {2.0,1.0,0.5,0.1,0.05} for ASAM to identify the optimal values (see Tab. 5 in Appendix A.8). We present the search results to demonstrate the reasonableness of the radii choices and experimental details are see Appendix A.8. The results of transfer learning are shown in Tab. 3.

6 Discussion and Conclusion

In this paper, we introduce nonuniformity, a new metric beyond sharpness to explain generalization, defined via the magnitude and structure of gradient noise. Furthermore, we validate the effectiveness of nonuniformity through theoretical analysis and extensive experiments. However, there are still some interesting phenomena and key issues should be studied.

Network Architecture and Data Distribution. (Andriushchenko et al. 2023) shows a effective metric of generalization relates to factors like data distribution and model family. We empirically study the relation between nonuniformity and network architecture (such as depth and width) as well as data distribution (Appendix A.8).

The inverse of GCM. Since the $C(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$, where d represents the number of parameters in the neural network, a key open problem is efficiently approximating the inverse of the GCM. We offer a preliminary discussion with further details in Appendix A.9. This could lead to better generalization via structure- and magnitude-aware optimization.

Rates	20%	40%	60%
SAM	$92.45_{\pm 0.2}$	$89.52_{\pm 0.21}$	$85.55_{\pm 0.25}$
Ours	$\mathbf{93.48}_{\pm 0.05}$	$\mathbf{90.61}_{\pm 0.29}$	$\mathbf{86.15}_{\pm 0.23}$

Table 2: Results under label noise.

CIFAR-10	SGD	ASAM	**Ours**
ViT-Tiny	$97.46_{\pm 0.01}$	$97.82_{\pm 0.01}$	$\mathbf{98.03}_{\pm 0.03}$
ViT-Small	$98.51_{\pm 0.01}$	$98.65_{\pm 0.01}$	$\mathbf{98.72}_{\pm 0.01}$
ConViT-Tiny	$97.17_{\pm 0.02}$	$97.41_{\pm 0.02}$	$\mathbf{97.82}_{\pm 0.02}$
ConViT-Small	$98.06_{\pm 0.01}$	$98.40_{\pm 0.01}$	$\mathbf{98.61}_{\pm 0.01}$
DeiT-Tiny	$96.67_{\pm 0.02}$	$97.26_{\pm 0.02}$	$\mathbf{97.52}_{\pm 0.03}$
DeiT-Small	$97.81_{\pm 0.03}$	$98.13_{\pm 0.02}$	$\mathbf{98.33}_{\pm 0.02}$
CIFAR-100	SGD	ASAM	**Ours**
ViT-Tiny	$85.99_{\pm 0.08}$	$86.67_{\pm 0.05}$	$\mathbf{87.78}_{\pm 0.1}$
ViT-Small	$90.76_{\pm 0.02}$	$91.55_{\pm 0.02}$	$\mathbf{91.87}_{\pm 0.03}$
ConViT-Tiny	$83.91_{\pm 0.02}$	$84.40_{\pm 0.12}$	$\mathbf{84.96}_{\pm 0.11}$
ConViT-Small	$87.86_{\pm 0.03}$	$88.51_{\pm 0.04}$	$\mathbf{88.95}_{\pm 0.02}$
DeiT-Tiny	$83.07_{\pm 0.08}$	$83.29_{\pm 0.07}$	$\mathbf{84.03}_{\pm 0.08}$
DeiT-Small	$86.18_{\pm 0.06}$	$86.71_{\pm 0.02}$	$\mathbf{87.33}_{\pm 0.02}$
Food-101	SGD	ASAM	**Ours**
ViT-Tiny	$83.45_{\pm 0.2}$	$84.31_{\pm 0.15}$	$\mathbf{84.91}_{\pm 0.2}$
ViT-Small	$89.96_{\pm 0.09}$	$90.45_{\pm 0.04}$	$\mathbf{90.60}_{\pm 0.05}$
ConViT-Tiny	$84.44_{\pm 0.08}$	$85.09_{\pm 0.09}$	$\mathbf{85.41}_{\pm 0.05}$
ConViT-Small	$87.44_{\pm 0.03}$	$88.03_{\pm 0.05}$	$\mathbf{88.23}_{\pm 0.03}$
DeiT-Tiny	$82.07_{\pm 0.2}$	$82.25_{\pm 0.1}$	$\mathbf{82.45}_{\pm 0.1}$
DeiT-Small	$86.48_{\pm 0.1}$	$86.97_{\pm 0.04}$	$\mathbf{87.16}_{\pm 0.03}$
Tiny-Imagenet	SGD	ASAM	**Ours**
ViT-Tiny	$78.84_{\pm 0.07}$	$79.17_{\pm 0.02}$	$\mathbf{80.39}_{\pm 0.07}$
ViT-Small	$85.85_{\pm 0.03}$	$86.76_{\pm 0.02}$	$\mathbf{87.23}_{\pm 0.02}$
ConViT-Tiny	$78.06_{\pm 0.07}$	$78.63_{\pm 0.02}$	$\mathbf{79.31}_{\pm 0.03}$
ConViT-Small	$86.98_{\pm 0.02}$	$87.29_{\pm 0.01}$	$\mathbf{88.07}_{\pm 0.03}$
DeiT-Tiny	$77.72_{\pm 0.06}$	$78.20_{\pm 0.07}$	$\mathbf{78.40}_{\pm 0.06}$
DeiT-Small	$85.87_{\pm 0.02}$	$87.13_{\pm 0.04}$	$\mathbf{87.84}_{\pm 0.07}$

Table 3: Results on transfer learning by fine-tuning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62472082).

References

Andriushchenko, M.; Croce, F.; Mueller, M.; Hein, M.; and Flammarion, N. 2023. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning*.

Andriushchenko, M.; and Flammarion, N. 2022. Towards Understanding Sharpness-Aware Minimization. In *International Conference on Machine Learning*, volume 162, 639–668.

Chatterjee, S.; and Zielinski, P. 2022. On the Generalization Mystery in Deep Learning. *ArXiv*, abs/2203.10036.

Chen, H.; Yeh, C.-C. M.; Fan, Y.; Zheng, Y.; Wang, J.; Lai, V.; Das, M.; and Yang, H. 2023. Sharpness-Aware Graph Collaborative Filtering. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chen, X.; Hsieh, C.-J.; and Gong, B. 2022. When Vision Transformers Outperform ResNets without Pretraining or Strong Data Augmentations. In *The Tenth International Conference on Learning Representations*.

Du, J.; Yan, H.; Feng, J.; Zhou, J. T.; Zhen, L.; Goh, R. S. M.; and Tan, V. Y. F. 2021. Efficient Sharpness-aware Minimization for Improved Training of Neural Networks. *International Conference on Learning Representations*, abs/2110.03141.

Du, J.; Zhou, D.; Feng, J.; Tan, V. Y. F.; and Zhou, J. T. 2022. Sharpness-Aware Training for Free. *Neural Information Processing Systems*, abs/2205.14083.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-Aware Minimization for Efficiently Improving Generalization. *International Conference on Learning Representations*, abs/2010.01412.

Ghadimi, S.; and Lan, G. 2013. Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. *SIAM J. Optim.*, 23: 2341–2368.

HaoChen, J. Z.; Wei, C.; Lee, J.; and Ma, T. 2020. Shape Matters: Understanding the Implicit Bias of the Noise Covariance. *Annual Conference Computational Learning Theory*, abs/2006.08680.

Hochreiter, S.; and Schmidhuber, J. 1994. Simplifying Neural Nets by Discovering Flat Minima. In *Neural Information Processing Systems*.

Jiang, W.; Yang, H.; Zhang, Y.; and Kwok, J. 2023. An Adaptive Policy to Employ Sharpness-Aware Minimization. In *International Conference on Learning Representations*, volume abs/2304.14647.

Karimi, H.; Nutini, J.; and Schmidt, M. W. 2016. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-ojasiewicz Condition. In *ECML/PKDD*.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2016. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *International Conference on Learning Representations*, abs/1609.04836.

Kwon, J.; Kim, J.; Park, H.; and Choi, I. 2021. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. *International Conference on Machine Learning*, 139: 5905–5914.

Li, H.; Xu, Z.; Taylor, G.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. *Neural Information Processing Systems*, 6391–6401.

Li, T.; Zhou, P.; He, Z.; Cheng, X.; and Huang, X. 2024. Friendly Sharpness-Aware Minimization. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5631–5640.

Liu, K.; Ziyin, L.; and Ueda, M. 2020. Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent. In *International Conference on Machine Learning*, volume 139, 7045–7056.

Liu, Y.; Mai, S.; Chen, X.; Hsieh, C.-J.; and You, Y. 2022. Towards Efficient and Scalable Sharpness-Aware Minimization. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12350–12360.

Mi, P.; Shen, L.; Ren, T.; Zhou, Y.; Sun, X.; Ji, R.; and Tao, D. 2022. Make Sharpness-Aware Minimization Stronger: A Sparsified Perturbation Approach. *Neural Information Processing Systems*, abs/2210.05177.

Mori, T.; Ziyin, L.; Liu, K.; and Ueda, M. 2022. Power-Law Escape Rate of SGD. In *International Conference on Machine Learning*, volume 162, 15959–15975.

Peng, Z.; Zhang, J.; Wang, Y.; Qi, L.; Shi, Y.; and Gao, Y. 2025. Leveraging Flatness to Improve Information-Theoretic Generalization Bounds for SGD. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Smith, S. L.; Dherin, B.; Barrett, D. G. T.; and De, S. 2021. On the Origin of Implicit Regularization in Stochastic Gradient Descent. *International Conference on Learning Representations*, abs/2101.12176.

Wang, Y.; ya Zhou, K.; Liu, N.; Wang, Y.; and Wang, X. 2024. Efficient Sharpness-Aware Minimization for Molecular Graph Transformer Models. *ArXiv*, abs/2406.13137.

Wen, K.; Ma, T.; and Li, Z. 2023. Sharpness Minimization Algorithms Do Not Only Minimize Sharpness To Achieve Better Generalization. *Neural Information Processing Systems*, abs/2307.11007.

Wojtowytsch, S. 2021. Stochastic Gradient Descent with Noise of Machine Learning Type Part II: Continuous Time Analysis. *Journal of Nonlinear Science*, 34: 1–45.

Woodworth, B. E.; Gunasekar, S.; Lee, J. D.; Moroshko, E.; Savarese, P.; Golan, I.; Soudry, D.; and Srebro, N. 2020. Kernel and Rich Regimes in Overparametrized Models. In Abernethy, J. D.; and Agarwal, S., eds., *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, 3635–3673. PMLR.

Wu, J.; Hu, W.; Xiong, H.; Huan, J.; Braverman, V.; and Zhu, Z. 2019. On the Noisy Gradient Descent that Generalizes as SGD. In *International Conference on Machine Learning*.

Wu, L.; Ma, C.; and Weinan, E. 2018. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In *Neural Information Processing Systems*.

Wu, L.; and Su, W. J. 2023. The Implicit Regularization of Dynamical Stability in Stochastic Gradient Descent. In *International Conference on Machine Learning*, volume abs/2305.17490.

Wu, L.; Wang, M.; and Su, W. J. 2022. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In *Neural Information Processing Systems*.

Xie, Z.; Sato, I.; and Sugiyama, M. 2020. A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima. *International Conference on Learning Representations*.

Yue, X.; Nouiehed, M.; and Kontar, R. A. 2020. SALR: Sharpness-Aware Learning Rate Scheduler for Improved Generalization. *IEEE transactions on neural networks and learning systems*, PP.

Zhang, R.; Fan, Z.; Yao, J.; Zhang, Y.; and Wang, Y. 2024. Domain-Inspired Sharpness-Aware Minimization Under Domain Shifts. *International Conference on Learning Representations*, abs/2405.18861.

Zhang, X.; Xu, R.; Yu, H.; Zou, H.; and Cui, P. 2023. Gradient Norm Aware Minimization Seeks First-Order Flatness and Improves Generalization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20247–20257.

Zhang, Z.; Luo, R.; Su, Q.; and Sun, X. 2022. GA-SAM: Gradient-Strength based Adaptive Sharpness-Aware Minimization for Improved Generalization. In *Conference on Empirical Methods in Natural Language Processing*.

Zheng, Y.; Zhang, R.; and Mao, Y. 2020. Regularizing Neural Networks via Adversarial Model Perturbation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8152–8161.

Zhong, Q.; Ding, L.; Shen, L.; Mi, P.; Liu, J.; Du, B.; and Tao, D. 2022. Improving Sharpness-Aware Minimization with Fisher Mask for Better Generalization on Language Models. *Conference on Empirical Methods in Natural Language Processing*, abs/2210.05497.

Zhu, Z.; Wu, J.; Yu, B.; Wu, L.; and Ma, J. 2018. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. In *International Conference on Machine Learning*.

Zhuang, J.; Gong, B.; Yuan, L.; Cui, Y.; Adam, H.; Dvornek, N. C.; Tatikonda, S. C.; Duncan, J. S.; and Liu, T. 2022. Surrogate Gap Minimization Improves Sharpness-Aware Training. *International Conference on Learning Representations*, abs/2203.08065.