Unpacking the Implicit Norm Dynamics of Sharpness-Aware Minimization in Tensorized Models

Tianxiao Cao, Kyohei Atarashi, Hisashi Kashima

Graduate School of Informatics, Kyoto University, Japan
cao.tianxiao.65w@st.kyoto-u.ac.jp, {atarashi,kashima}@i.kyoto-u.ac.jp

Abstract

Sharpness-Aware Minimization (SAM) has been proven to be an effective optimization technique for improving generalization in overparameterized models. While prior works have explored the implicit regularization of SAM in simple two-core scale-invariant settings, its behavior in more general tensorized or scale-invariant models remains underexplored. In this work, we leverage scale-invariance to analyze the norm dynamics of SAM in general tensorized models. We introduce the notion of *Norm Deviation* as a global measure of core norm imbalance, and derive its evolution under SAM using gradient flow analysis. We show that SAM's implicit control of Norm Deviation is governed by the covariance between core norms and their gradient magnitudes. Motivated by these findings, we propose a simple yet effective method, *Deviation-Aware Scaling (DAS)*, which explicitly mimics this regularization behavior by scaling core norms in a data-adaptive manner. Our experiments across tensor completion, noisy training, model compression, and parameter-efficient fine-tuning confirm that DAS achieves competitive or improved performance over SAM, while offering reduced computational overhead.

Introduction

The remarkable success of overparameterized deep neural networks has highlighted a central challenge in modern machine learning: generalization. These models possess enough capacity to memorize the entire training dataset, yet they often perform exceptionally well on unseen data (Zhang et al. 2021). Understanding the mechanisms that prevent overfitting and promote generalization is a fundamental pursuit (Belkin et al. 2019; Ishida et al. 2020; Bisla, Wang, and Choromanska 2022).

A leading explanation involves the geometry of the loss landscape, where flatter minima are empirically associated with better generalization. Sharpness-Aware Minimization (SAM) (Foret et al. 2021) is a widely adopted optimization method that materializes this idea by minimizing the worst-case loss within a small perturbation neighborhood. SAM has demonstrated robust performance across a variety of tasks (Kwon et al. 2021; Bahri, Mobahi, and Tay 2022), and its theoretical properties have been explored in depth (Wen, Ma, and Li 2022; Baek, Kolter, and Raghunathan 2024; Andriushchenko and Flammarion 2022; Andriushchenko et al. 2023), though mostly for standard, dense architectures.

Concurrently, there has been a long-standing interest in parameter-efficient modeling approaches, where structure is imposed on weights to reduce memory and computation for efficient and eco-friendly machine learning (Memmel et al. 2024). These include tensor-decomposed layer (Novikov et al. 2015; Hrinchuk et al. 2020) or low-rank adaptation (LoRA) modules (Yang et al. 2024; Yaras et al. 2024; Si et al. 2025; Veeramacheneni et al. 2025), which offer significant compression and computational savings by modeling parameters as combinations of small tensor cores. Such structured parameterizations introduce scale-invariance and inter-dependencies that can interact non-trivially with optimization dynamics.

While prior work has shown that SAM implicitly promotes norm balancing in simple matrix factorization setups (Li, Zhang, and He 2024), its behavior in general multi-core tensorized models, where factor norms can differ drastically and influence training stability, is far less understood. This gap motivates our central research question: ***How does the implicit regularization of SAM manifest in general, multi-core, scale-invariant tensorized models?*** To answer the question, we investigate the implicit norm dynamics of both SGD and SAM applied to the general scale-invariant problem. We propose a global measure Norm Deviation Q (Definition 1) to capture the norm imbalance. We present theoretical (Theorem 3) and empirical analysis (Fig. 1) and show that Q is governed by the covariance between core norms and gradient norms, and that this effect is amplified by data noise. Inspired by the theoretical findings, we proposed a computationally efficient alternative of SAM, Deviation-Aware Scaling, by mimicking the norm dynamics through core scaling without computing the adversarial perturbation. We experimented on comprehensive tasks related to tensor-based parameterization. To summarize, our contributions are as follows:

- Norm Deviation Q is proposed as a global proxy for imbalance among tensor cores, with its behavior analyzed under different optimizers. We prove conditions for local norm shrinkage and study its data-dependent dynamics.
- We propose Deviation-Aware Scaling (DAS), a novel optimizer that avoids the adversarial perturbation step of

SAM by distilling its norm effects into explicit scaling.

- We empirically validate our findings across various domains, including tensorized neural networks, few-shot language model finetuning with low-rank adapters, and noisy-label learning. Our results show that SAM consistently improves performance in these settings, and that DAS inherits many of these benefits while reducing computational overhead.

Related Work

SAM and mechanism of SAM. With the success of SAM, there emerged a line of extending works, including addressing the computational efficiency (Du et al. 2022a,b; Ji et al. 2024; Xie, Pethick, and Cevher 2024; Deng et al. 2025) and improving the performance (Kwon et al. 2021; Liu et al. 2022). To explain the success of SAM, great efforts are made to understand SAM, mainly through simplified modeling (Wen, Ma, and Li 2022; Bartlett, Long, and Bousquet 2023), empirical observations (Andriushchenko et al. 2023), and proxy measures (Li, Zhang, and He 2024).

Tensor decomposition in model compression. Multi-core structured models or tensor decompositions have long been a promising direction for model compression. Existing works utilize decompositions including Tensor-Train (Novikov et al. 2015; Yin et al. 2021), Tensor-Ring (Wang et al. 2018; Cao et al. 2024), Tucker (Phan et al. 2020), and even arbitrary tensor networks (Hayashi et al. 2019), achieving a practical reduction in storage through structured parameterization.

Tensor-based LoRA. As large pre-trained models have become the standard, fine-tuning them for downstream tasks has become computationally prohibitive. LoRA (Hu et al. 2022) injects low-rank trainable matrices into the layers of a Transformer, drastically reducing the number of trainable parameters. Recent advanced LoRA variants explore high-order tensor decomposition, such as CP (Veeramacheneni et al. 2025), Tensor-Train (Yang et al. 2024; Anjum et al. 2024), Tucker (Si et al. 2025), and deep matrix factorization (Yaras et al. 2024).

Problem Statement

Notations. A scalar and a tensor (including matrices as second-order tensors) are denoted by x and $\mathcal{X}$, respectively, unless otherwise specified. For two tensors $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, we define the *Frobenius inner product* as $\langle \mathcal{A}, \mathcal{B} \rangle_F := \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} \mathcal{A}_{i_1,\ldots,i_d} \cdot \mathcal{B}_{i_1,\ldots,i_d}$. We denote by $\|\mathcal{G}\|_F^2 = \langle \mathcal{G}, \mathcal{G} \rangle_F$, the squared Frobenius norm of a tensor $\mathcal{G}$, *i.e.*, the sum of squares of all entries of $\mathcal{G}$. For a positive integer n, we denote $[n] := \{1, 2, \ldots, n\}$.

We consider a class of general scale-invariant models. The parameters consist of a set of core tensors $\{\mathcal{G}_k\}_{k=1}^K$. These cores are composed via a multilinear reconstruction function $\Phi(\mathcal{G}_1, \ldots, \mathcal{G}_K)$ that produces the full tensor $\mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. The problem of interest is to find the solution to the following optimization problem:

$$\min_{\mathcal{G}_1,\ldots,\mathcal{G}_K} f(\mathcal{T}) = f(\Phi(\mathcal{G}_1, \ldots, \mathcal{G}_K)), \tag{1}$$

Algorithm 1: SAM for Problem (2) on tensorized models

Input: Tensor cores $\{\mathcal{G}_k^{(0)}\}$, step size$\{\eta^{(t)}\}$, radius of perturbation ρ, and number of iterations T.

Output: Final tensor cores $\{\mathcal{G}_k^{(T)}\}_{k=1}^K$

for $t = 0, \ldots, T-1$ **do**
  Compute $g_k = \nabla_{\mathcal{G}_k} f(\Phi(\mathcal{G}_1^{(t)}, \ldots, \mathcal{G}_K^{(t)})), \forall k \in [K]$
  for $k = 1, \ldots, K$ **do**
    $\tilde{\mathcal{G}}_k^{(t)} = \mathcal{G}_k^{(t)} + \rho \frac{g_k}{\sqrt{\sum_{j=1}^K \|g_j\|_F^2}}$
    $\tilde{g}_k = \nabla_{\mathcal{G}_k} f(\Phi(\tilde{\mathcal{G}}_1^{(t)}, \ldots, \tilde{\mathcal{G}}_K^{(t)}))$
    Update $\mathcal{G}_k^{(t+1)} = \mathcal{G}_k^{(t)} - \eta^{(t)} \tilde{g}_k$ via Adam or SGD
  end for
end for

where $f(\cdot)$ is a scalar function. The key structural property is that the reconstruction function Φ is *multilinear* in each of cores. That is, for any core index $k \in \{1, \ldots, K\}$, and any scalar $\lambda \in \mathbb{R}$, we have:

$$\Phi(\mathcal{G}_1, \ldots, \lambda \mathcal{G}_k, \ldots, \mathcal{G}_K) = \lambda \cdot \Phi(\mathcal{G}_1, \ldots, \mathcal{G}_K).$$

Specifically, scale-invariance refers to that $\{c_k \mathcal{G}_k\}_{k=1}^K$ have the same reconstructed tensor and objective value $\forall c_k \in \mathbb{R}, \quad \prod_{k=1}^K c_k = 1$. Such reconstruction functions arise naturally in a broad class of tensor decomposition models and tensor networks, where the output tensor $\mathcal{T}$ is constructed via a sequence of tensor contractions[1] over shared modes.

Sharpness-Aware Minimization

Sharpness-Aware minimization (SAM) (Foret et al. 2021) has emerged as a powerful technique for improving the generalization of various machine learning models by seeking flat solutions. SAM optimizes the worst-case loss over a neighborhood of the parameters, with which Problem (1) can be reformulated as:

$$\min_{\mathcal{G}_1,\ldots,\mathcal{G}_K} \max_{\|\Delta\| \le \rho} f\left(\Phi(\mathcal{G}_1 + \Delta_1, \ldots, \mathcal{G}_K + \Delta_K)\right), \tag{2}$$

where $\|\Delta\| := (\sum_{k=1}^K \|\Delta_k\|_F^2)^{1/2}$ and ρ is the radius of perturbation. A practical implementation of SAM, directly adapted from (Foret et al. 2021) is given in Algorithm 1.

Mechanism understanding with norm. Analyzing the implicit bias of optimization algorithms on structured models by examining the norm of each structured component is becoming a powerful tool, as in the matrix factorization case (Liu et al. 2021; Li, Zhang, and He 2024). In the context of SAM, *balancedness* (Li, Zhang, and He 2024), *i.e.*, the difference in the squared Frobenius norm of two factors in the matrix factorization case, is introduced as a global metric to characterize the implicit regularization of SAM. Motivated by this, we aim to examine the norm behaviors of cores in different optimizers and provide theoretical insights.

[1]This formulation encompasses common structures such as Tucker, CP, Tensor-Train, and Tensor-Ring, as well as tensorized neural network layers where the weight tensor is parameterized via structured factorization.

Analysis of Norm Dynamics for General Scale-Invariant Problems

Due to space constraints, we provide the proofs for all theoretical results in the Appendix. Following previous works on implicit regularization (Arora et al. 2019; Li, Zhang, and He 2024), we consider gradient flow with infinitesimal stepsize $\eta \to 0$ for Problem (1):

$$\begin{aligned} g_k^{(t)} &= \nabla_{\mathcal{G}_k} f(\Phi(\mathcal{G}_1^{(t)}, \dots, \mathcal{G}_K^{(t)})), \\ \mathcal{G}_k^{(t+1)} &= \mathcal{G}_k^{(t)} - \eta g_k^{(t)}, \quad \forall k \in [K]. \end{aligned} \tag{3}$$

In the infinitesimal stepsize limit $\eta \to 0$, time t is treated as continuous. We show that the dynamics of the squared norm among all cores are equivalent.

Theorem 1 (Norm Dynamics of SGD)**.** *Applying the update step Equation* (3) *with infinitesimal stepsize $\eta \to 0$ to the Problem* (1)*, the time derivatives of the squared Frobenius norms of all cores are equal:*

$$\frac{d}{dt}\|\mathcal{G}_1^{(t)}\|_F^2 = \cdots = \frac{d}{dt}\|\mathcal{G}_K^{(t)}\|_F^2.$$

Next, we tackle the dynamics of the squared Frobenius norm in SAM. The update step of SAM in Algorithm 1 can be rewritten as:

$$\begin{aligned} g_k^{(t)} &= \nabla_{\mathcal{G}_k} f(\Phi(\mathcal{G}_1^{(t)}, \dots, \mathcal{G}_K^{(t)})), \\ \tilde{\mathcal{G}}_k^{(t)} &= \mathcal{G}_k^{(t)} + \rho u^{(t)} g_k^{(t)}, \\ \tilde{g}_k^{(t)} &= \nabla_{\mathcal{G}_k} f(\Phi(\tilde{\mathcal{G}}_1^{(t)}, \dots, \tilde{\mathcal{G}}_K^{(t)})), \\ \mathcal{G}_k^{(t+1)} &= \mathcal{G}_k^{(t)} - \eta \tilde{g}_k^{(t)}, \quad \forall k \in [K], \end{aligned} \tag{4}$$

where $u^{(t)} = (\sum_{j=1}^K \|g_j^{(t)}\|_F^2)^{-1/2}$ is the normalization factor, and $\rho > 0$ is the radius of the perturbation.

We use a standard assumption on the Lipschitz smoothness of $f(\cdot)$ following previous works on analyzing SAM (Andriushchenko and Flammarion 2022; Wen, Ma, and Li 2022; Li, Zhang, and He 2024) and optimization (Ge et al. 2015; Wang et al. 2024):

Assumption 1 (Smoothness)**.** *There exists $L > 0$ such that for any $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, it holds that*

$$\|\nabla f(\mathcal{X}) - \nabla f(\mathcal{Y})\|_F \le L\|\mathcal{X} - \mathcal{Y}\|_F.$$

Li, Zhang, and He analyzed the dynamics of *difference* between the squared Frobenius norms of two cores in the matrix factorization case. Under the multi-core setting in our paper, it is natural to derive the dynamics for the difference between the squared norms of two cores, as follows:

Theorem 2 (Pairwise Norm Dynamics under SAM)**.** *Applying the update steps* (4) *with infinitesimal stepsize $\eta \to 0$, the gradient flow of SAM satisfies that $\forall i, j \in [K]$ with $i \neq j$:*

$$\begin{aligned} \frac{d}{dt}\left(\|\mathcal{G}_i^{(t)}\|_F^2 - \|\mathcal{G}_j^{(t)}\|_F^2\right) = 2\rho u^{(t)} \left(\|g_i^{(t)}\|_F^2 - \|g_j^{(t)}\|_F^2\right) \\ + O(\rho^2 L). \end{aligned}$$

Theorems 1 and 2 can be viewed as non-trivial multi-core generalizations of the results in (Li, Zhang, and He 2024), which focused on matrix factorization. Specifically, Theorem 1 shows that under gradient flow, standard SGD preserves the difference in squared Frobenius norms between any two cores, *i.e.*, $\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2$ remains constant. In contrast, Theorem 2 reveals that SAM introduces a dynamic regulation mechanism: the rate of change in $\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2$ is proportional to the difference in squared gradient norms, $\|g_i\|_F^2 - \|g_j\|_F^2$. These results provide important insights into how individual core norms evolve relative to each other. In the multi-core setting under SAM, however, pairwise norm differences are local and not sufficient to characterize the overall norm dynamics of the whole system. To fully understand the norm behavior of all cores under SGD and SAM, we need a new measure to analyze the *global* norm dynamics to capture the collective norm dynamics of the model.

A global measure. While directly using pairwise norm differences is appealing, summing or linearly combining them leads to cancellation and fails to provide a meaningful global measure. For instance, with three cores where their squared norms are $\|\mathcal{G}_1\|_F^2 = 2$, $\|\mathcal{G}_2\|_F^2 = 10$, and $\|\mathcal{G}_3\|_F^2 = 18$, a circular sum of differences $(\|\mathcal{G}_1\|_F^2 - \|\mathcal{G}_2\|_F^2) + (\|\mathcal{G}_2\|_F^2 - \|\mathcal{G}_3\|_F^2) + (\|\mathcal{G}_3\|_F^2 - \|\mathcal{G}_1\|_F^2)$ yields $(-8) + (-8) + (16) = 0$, masking the significant underlying imbalance. To address this, we introduce a global measure that captures the collective norm dynamics of all cores.

Definition 1 (Global Squared Norm Deviation)**.** *For cores $\{\mathcal{G}_k\}_{k=1}^K$, the global squared norm deviation is defined as:*

$$Q := \sum_{k=1}^K \left(\|\mathcal{G}_k\|_F^2 - \frac{1}{K}\sum_{i=1}^K \|\mathcal{G}_i\|_F^2\right)^2. \tag{5}$$

For brevity, we will refer to this quantity as the Norm Deviation *when the context is clear.*

The Norm Deviation Q captures the spread of the squared Frobenius norms of all cores. A larger Q indicates greater imbalance among the cores. All cores have the same Frobenius norm if and only if $Q = 0$. Also, it can be verified that the Norm Deviation Q has an alternative expression:

$$Q = \frac{1}{2K}\sum_{i,j=1}^K \left(\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2\right)^2. \tag{6}$$

This shows that Q aggregates all the pairwise imbalances in the squared Frobenius norms of the cores. Next, we derive the dynamics of the Norm Deviation Q under SGD and SAM. The Norm Deviation Q of SGD is conserved, from a direct application of Theorem 1:

Corollary 1 (Norm Deviation Dynamics under SGD)**.** *Applying the update steps* (3) *with infinitesimal stepsize $\eta \to 0$ to the problem* (1)*,*

$$\frac{dQ}{dt} = 0.$$

The dynamics of Q under SAM is as follows:

Theorem 3 (Norm Deviation Dynamics under SAM)**.** *Applying the update steps* (4) *with infinitesimal stepsize $\eta \to 0$, the gradient flow of SAM satisfies:*

$$\frac{dQ}{dt} = 4\rho u^{(t)} K \cdot \mathrm{Cov}\left(\|\mathcal{G}_k^{(t)}\|_F^2, \|g_k^{(t)}\|_F^2\right) + O(\rho^2 L),$$

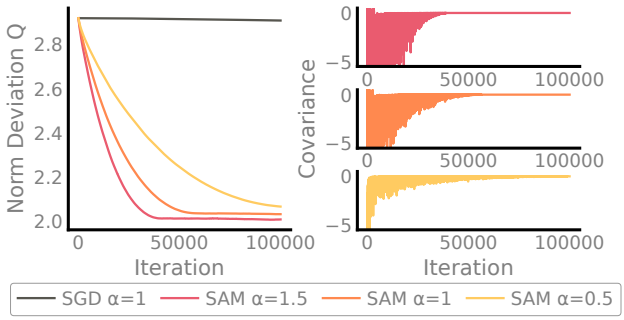

Figure 1: Implicit regularization of SAM on norm deviation (Definition 1). We optimize a toy Tucker-2 model to fit a target tensor with additive noise, where α controls the noise strength (see Appendix B for setup). Left: Norm Deviation Q vs. iterations. Right: $\mathrm{Cov}(\|\mathcal{G}_k\|_F^2, \|g_k\|_F^2)$ vs. iterations.

where $\mathrm{Cov}(x_k, y_k) := \frac{1}{K}\sum_{k=1}^{K}(x_k - \bar{x})(y_k - \bar{y})$ *denotes the empirical covariance, with* $\bar{x}, \bar{y}$ *the means over* $k \in [K]$.

Theorem 3 shows that the Norm Deviation Q under SAM evolves according to the covariance between the squared Frobenius norms of the cores and their corresponding gradient norms. Specifically, if $\|g_k\|_F^2$ is large when $\|\mathcal{G}_k\|_F^2$ is small (negative covariance), SAM grows small cores faster and helps all cores to encourage equalization in their norms. In contrast, positive covariance means that large cores will grow faster, leading to a larger imbalance among the cores.

Toward norm balancing tendencies under SAM. Our theoretical analysis reveals that SAM exhibits strong norm-balancing tendencies at a local level. Specifically, the pairwise norm difference, *i.e.*, $\left|\|\mathcal{G}_i\|_F^2 - \|\mathcal{G}_j\|_F^2\right|, \forall i, j \in [K]$, shrinks *locally* when being large (Proposition A.1 in Appendix). This local corrective dynamic for every pair suggests a global effect; one can intuit from Equation (6) that an accumulation of these pairwise shrinkages would lead to an overall decrease in the Norm Deviation Q. While our proof formally establishes this balancing effect at a local level, our experiments confirm its global impact. Empirical observations consistently show that Q decreases during training (see Fig. 1), supporting the conjecture that SAM promotes norm balancing through an accumulation of these local shrinkage effects. In the controlled Tucker-2 experiment of Fig. 1, we introduce additive Gaussian noise with varying magnitude controlled by a scalar α. As α increases, the gradients become noisier, and the covariance between core norms and their corresponding gradient norms grows. This larger covariance corresponds to a faster reduction in Q, reflecting a stronger balancing effect under SAM. This phenomenon illustrates how the implicit regularization strength of SAM adapts to the noise level, aligning with the "data-responsive" interpretation observed in the two-factor setting of (Li, Zhang, and He 2024). The benefits of norm balancing in gradient-based optimization are well-documented for both tensor decomposition (Du, Hu, and Lee 2018; Razin, Maman, and Cohen 2022; Hariz et al. 2022) and neural networks (Neyshabur et al. 2017), suggesting that the norm-regulating dynamics induced by SAM are potentially favorable for optimization.

Extension to multi-layer models. Our analysis naturally extends models that consist of multiple scale-invariant models across layers, such as tensorized neural networks, SAM governs the Norm Deviation layer-wisely. See Theorem A.2 in the Appendix as a multi-layer extension of Theorem 3.

Mimicking the Implicit Regularization of SAM with Explicit Control

The theoretical insights suggest that SAM induces a tendency to control the core norms, particularly when the covariance between core and gradient magnitudes is high. Prior works turn or enhance implicit effects of optimization with explicit regularizers, such as (Barrett and Dherin 2021; Li, Zhang, and He 2024). Motivated by this, this section aims to address the following question: ***Can we design a method that mimics the implicit regularization to utilize its beneficial impact?*** By an explicit method that replicates the implicit regularization of SAM, it is possible to **(1)** reduce the extra gradient calculation of SAM (line 2 in Algorithm 1) if we use a clever design, and **(2)** decouple the norm control from the optimization process, which allows us to control the norm of cores independently.

Deviation-Aware Scaling

In this section, we gain insights from theory and introduce a novel method `Deviation-Aware Scaling` (DAS), which address the above question by explicitly controlling the Norm Deviation Q through a scaling approach. We adopt a simple exemplar that scales the cores by a factor in each iteration to control norms, following the implementation of weight decay (Loshchilov and Hutter 2019):

$$\mathcal{G}_k^{(t+\frac{1}{2})} = (1 + \lambda_k^{(t)})\mathcal{G}_k^{(t)}, \tag{7}$$

$$\mathcal{G}_k^{(t+1)} = \mathcal{G}_k^{(t+\frac{1}{2})} - \eta g_k^{(t)}, \tag{8}$$

for each core $k \in [K]$ at iteration t, where η is a finite stepsize and $\lambda_k^{(t)} \in \mathbb{R}$ for core k at iteration t. Based on Corollary 1, we can assume only the scaling step (7) controls the Norm Deviation Q. We can derive a closed-form expression for the scaling factor $\lambda_k^{(t)}$ to control the Norm Deviation Q to match the implicit regularization of SAM using only SGD.

Deriving λ_k. With small $\lambda_k^{(t)}$, the change in norm caused by the scaling step (7) is:

$$\Delta\|\mathcal{G}_k\|_F^2 := \|\mathcal{G}_k^{(t+\frac{1}{2})}\|_F^2 - \|\mathcal{G}_k^{(t)}\|_F^2 \approx 2\lambda_k^{(t)}\|\mathcal{G}_k^{(t)}\|_F^2.$$

We would like to set the $\lambda_k^{(t)}$ so that the dynamics in the Norm Deviation Q caused by the scaling step (7) matches the dynamics of SAM in Theorem 3. With a small enough stepsize η, the change in Q under SAM is approximately:

$$\begin{aligned}\Delta Q_{\mathrm{SAM}} &\approx \eta \cdot \left(\frac{dQ}{dt}\right)_{\mathrm{SAM}} \\ &\approx 4\eta\rho u^{(t)} K \cdot \mathrm{Cov}\left(\|\mathcal{G}_k^{(t)}\|_F^2, \|g_k^{(t)}\|_F^2\right),\end{aligned}$$

Algorithm 2: Deviation-Aware Scaling (DAS)

Input: Tensor cores $\{\mathcal{G}_k^{(0)}\}$, stepsize $\{\eta^{(t)}\}$, coefficient $\{\alpha^{(t)}\}$, and number of iterations T.

Output: Final tensor cores $\{\mathcal{G}_k^{(T)}\}$.

for $t = 0, \dots, T-1$ **do**
  Compute $g_k^{(t)} = \nabla_{\mathcal{G}_k} f(\Phi(\mathcal{G}_1^{(t)}, \dots, \mathcal{G}_K^{(t)})), \forall k \in [K]$
  for $k = 1, \dots, K$ **do**
    $\lambda_k^{(t)} = \frac{\eta^{(t)}\alpha^{(t)}u^{(t)}}{\|\mathcal{G}_k^{(t)}\|_F^2} \cdot \left(\|g_k^{(t)}\|_F^2 - \bar{g}\right)$,
    where $\bar{g} = \frac{1}{K}\sum_{i=1}^K \|g_i^{(t)}\|_F^2$, $u^{(t)} = (K \cdot \bar{g})^{-1/2}$
    Update $\mathcal{G}_k^{(t+1)} = (1 + \lambda_k^{(t)})\mathcal{G}_k^{(t)} - \eta g_k^{(t)}$
    via Adam or SGD
  end for
end for

Method	HOOI	ADAM	SAM	DAS
R^2 score	0.9268	0.9482	0.9485**	0.9484*

Table 1: Results of Tucker on COVID dataset. Best and second-best results per row are marked with ** and *, respectively. This notation applies to Tables 1-3.

where the term $O(\rho^2 L)$ is ignored for small ρ. At the same time, the change in Q caused by the scaling step (7) is:

$$\Delta Q \approx 4\sum_{k=1}^{K}\left(\|\mathcal{G}_k^{(t)}\|_F^2 - \frac{1}{K}\sum_{i=1}^{K}\|\mathcal{G}_i^{(t)}\|_F^2\right)\cdot\lambda_k^{(t)}\|\mathcal{G}_k^{(t)}\|_F^2,$$

following the definition of Q (Definition 1). We would like to match the two changes in Q, *i.e.*, $\Delta Q = \Delta Q_{\mathrm{SAM}}$. This leads to the following closed-form expression for $\lambda_k^{(t)}$:

$$\lambda_k^{(t)} = \frac{\rho u^{(t)} \cdot \eta}{\|\mathcal{G}_k^{(t)}\|_F^2} \cdot \left(\|g_k^{(t)}\|_F^2 - \bar{g}\right), \tag{9}$$

where $\bar{g} = \frac{1}{K}\sum_{i=1}^K \|g_i^{(t)}\|_F^2$ is the average squared gradient norm over all cores at iteration t. However, since this expression is derived from approximations of the true dynamics, there is no guarantee that SAM's original perturbation radius ρ is the optimal choice for controlling the strength of this explicit scaling. We therefore replace ρ with a new hyperparameter, α, which directly controls the strength of the scaling. This decouples our method from SAM's settings and leads to the final Deviation-Aware Scaling (DAS) algorithm, summarized in Algorithm 2.

Experiments

We conduct a comprehensive set of experiments to show the effectiveness of SAM and DAS for various tensor-based models, including tensor completion, training tensorized neural networks, and improving tensor-based parameter-efficient fine-tuning (PEFT) for large language model adaptation. See Appendix for more experiment details, available at https://github.com/ctxGou/Tensor-SAM.

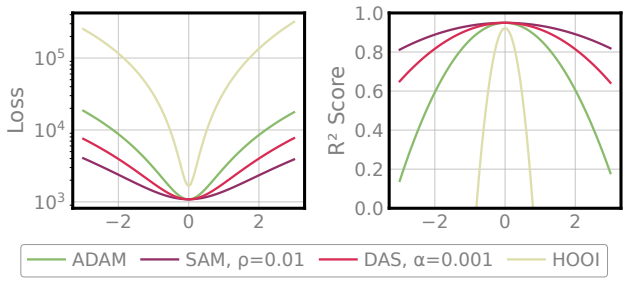

Figure 2: 1-D visualization of train loss and evaluation metric vs. perturbations on the COVID dataset. Left: Training loss. Right: Evaluation R^2 score. The x-axis is the size of a fixed directional perturbation applied to model parameters.

Tucker Decomposition for Tensor Completion

We perform experiments on the Tucker decomposition for tensor completion, using the real-world data COVID dataset from the library tensorly (Kossaifi et al. 2019). We randomly mask 70% of the entries as the evaluation set to be recovered, and use the remaining 30% as the training set. We follow the setup of (Hariz et al. 2024) and use a three-order Tucker with multilinear rank $(8, 6, 8)$ and an MSE loss function optimized with ADAM (Kingma and Ba 2015).

We use SAM and DAS with a base optimizer ADAM to compare with the vanilla ADAM. Tucker-HOOI (Kolda and Bader 2009) is also included as a baseline, available in tensorly. The results averaged from 3 dataset splits are shown in Table 1 and Fig. 2. In the table, we report the coefficient of determination R^2 score, which measures the proportion of variance explained by the model. SAM and DAS show marginal improvements over ADAM, while all three gradient-based methods substantially outperform the traditional HOOI baseline. The loss curves in Fig. 2 reveal an important insight: while SAM and DAS achieve only marginal improvements in the final R^2 score, they find flatter minima compared to vanilla ADAM, which is precisely the intended effect of SAM's sharpness-aware optimization. Importantly, DAS successfully captures this flatness property through explicit norm control rather than SAM's direct perturbation-based approach, validating our theoretical analysis that the norm dynamics are a key driver of SAM's beneficial effects.

Applications to Tensorized Neural Networks

Our setting of general scale-invariant problems applies to a wide range of multi-layer tensorized neural networks.

Training Tensorized Neural Networks from Scratch. We evaluate the effectiveness of SAM and DAS on CIFAR-10 (Krizhevsky, Hinton et al. 2009) using ResNet-20 (He et al. 2016) parameterized by tensor decompositions. Each convolution layer is replaced with a set of tensor cores, which reconstruct the full weight tensor. We experiment with both CP (Kolda and Bader 2009) and Tensor-Ring (TR) (Zhao et al. 2016) decompositions, with five different ranks for each method. The models are trained from scratch using SGD, SAM, and DAS (using SGD as the base opti-

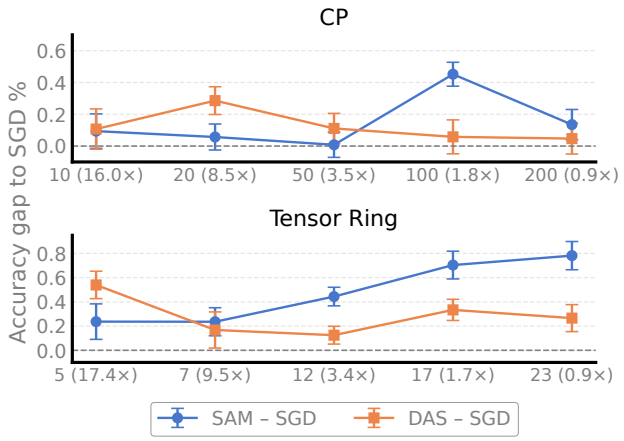

Figure 3: Accuracy improvements over SGD when training tensorized ResNet-20 on CIFAR-10 with CP (top) and Tensor-Ring (bottom) decompositions. The x-axis indicates the chosen rank for each decomposition, with the corresponding model compression ratio shown in parentheses (*e.g.*, $a\times$ denotes the ratio of original to compressed model parameters). Error bars indicate standard error over 10 runs.

mizer). For both SAM and DAS, the corresponding hyperparameters ρ and α are tuned independently over the shared search space $\{0.001, 0.005, 0.01, 0.05, 0.1\}$ using a validation split from the training set.

Figure 3 reports the mean and standard error of accuracy improvements (over SGD) from 10 independent runs, after training for 180 epochs. Both SAM and DAS consistently outperform SGD across all decomposition types and ranks. Interestingly, the performance gains from SAM tend to increase with higher tensor ranks, particularly in the Tensor-Ring setting, suggesting that SAM is especially beneficial for larger tensorized models. DAS also provides consistent, though slightly smaller, improvements across configurations. This highlights that DAS explicitly captures and benefits from the norm dynamics that underpin SAM's implicit regularization, offering an efficient alternative.

Tensorized Neural Networks under Label Noise. It is shown in (Foret et al. 2021; Kwon et al. 2021) that SAM is robust to label noise on general deep models, without specific designs for label noise. We test the robustness of SAM and DAS on a tensorized neural network, using a compact version of ResNet-32, where all convolution layers are re-parameterized using Tensor-Ring decomposition with ranks set as in Appendix. We use a symmetric label noise (Jiang et al. 2020) with corruption rates $40\%, 60\%, 80\%$ on CIFAR-10 training set and a clean test set. Table 2 shows the test accuracies of TR-ResNet-32 on CIFAR-10 with label noise. We compare the performance of SGD, SAM, and DAS (with SGD as the base optimizer). As shown in Table 2, both SAM and DAS improve robustness to label noise over SGD, with SAM achieving significantly better test accuracy under higher noise levels. While DAS does not match SAM's robustness, it still outperforms SGD, indicating partial resistance to label corruption. Prior work (Baek, Kolter, and Raghunathan 2024) attributes SAM's robustness to its influence mainly on network Jacobian effects that DAS does not replicate. This suggests that DAS may derive its robustness from its explicit control over parameter norm dynamics, leading to better optimization behavior than SGD.

Noise rate	SGD	SAM	DAS
40%	$83.76_{\pm 0.24}$	$86.59^{**}_{\pm 0.16}$	$84.35^{*}_{\pm 0.40}$
60%	$77.41_{\pm 1.08}$	$82.03^{**}_{\pm 0.25}$	$77.73^{*}_{\pm 0.27}$
80%	$57.33_{\pm 0.44}$	$67.13^{**}_{\pm 1.70}$	$60.74^{*}_{\pm 0.57}$
Runtime (s)	0.039^{**}	0.090	0.054^{*}

Table 2: Test accuracies and runtime for SGD, SAM, and DAS on CIFAR-10 with label noise using TR-ResNet-32. Means and standard deviations are computed over 3 runs.

	SGD	SAM	DAS
Top-1	$65.47_{\pm 0.14}$	$66.27^{**}_{\pm 0.07}$	$66.16^{*}_{\pm 0.21}$
Top-5	$86.54_{\pm 0.14}$	$87.12^{**}_{\pm 0.05}$	$86.96^{*}_{\pm 0.05}$
Runtime (s)	0.254^{**}	0.425^{*}	0.254^{**}

Table 3: Top1, Top5, and runtime for SGD, SAM, and DAS fine-tuned on ImageNet using compressed TT-ResNet-18.

Finetuning Pre-trained Models after Compression. We consider a practical scenario where a pre-trained model is compressed using tensor decomposition and then fine-tuned to restore the performance lost due to compression, following a paradigm similar to (Phan et al. 2020). We compress all convolution layers in a pre-trained ResNet-18 using Tensor-Train (TT) decomposition, with ranks suggested in (Yin et al. 2021). This results in a TT-ResNet-18 with 4.4M parameters, compared to the original ResNet-18 with 11.2M parameters. The compressed TT-ResNet-18 is fine-tuned on the ImageNet-1k (Deng et al. 2009). We compare the performance of SGD, SAM, and DAS (with SGD as the base optimizer). We consider a lightweight fine-tuning with 15 epochs. Table 3 shows the mean and standard deviation of Top-1 and Top-5 accuracies over 5 runs. Note that the compressed model before fine-tuning achieves $36.74\%/64.32\%$ Top-1/-5 accuracy, which is significantly lower than the original ResNet-18. We observe that both SAM and DAS outperform SGD, with SAM achieving the best performance. DAS uses approximately the same runtime as SGD, but its performance is close to SAM, showing its effectiveness.

Tensor-based Parameter-Efficient Fine-Tuning

We evaluate SAM and DAS on two recent tensor-based low-rank adaptation methods, FLoRA (Si et al. 2025) and LoRETTA (Yang et al. 2024), for fine-tuning language models. See Appendix for details of FLoRA and LoRETTA.

Improving FLoRA on RoBERTa-large. FLoRA utilizes Tucker decomposition to parameterize the incremental up-

RoBERTa		SST-2	SST-5	SNLI	MNLI	RTE	TREC	Avg.(↑)
Zero-Shot†		79.0	35.5	50.2	48.8	51.4	32.0	49.5
Full fine-tuning		$94.01^{**}_{\pm 0.46}$	$56.84^{**}_{\pm 0.86}$	$88.38^{**}_{\pm 0.43}$	$84.32^{**}_{\pm 0.72}$	$83.16_{\pm 1.67}$	$96.92^{*}_{\pm 0.55}$	83.94^{*}
LoRA ($r = 8$)		$90.53_{\pm 1.10}$	$51.60_{\pm 0.77}$	$83.58_{\pm 1.24}$	$76.38_{\pm 2.99}$	$77.55_{\pm 4.32}$	$95.72_{\pm 0.41}$	79.23
FLoRA	ADAM	$93.99^{*}_{\pm 0.26}$	$56.70_{\pm 0.80}$	$87.70_{\pm 0.54}$	$83.96_{\pm 0.67}$	$83.61_{\pm 1.30}$	$96.64_{\pm 0.34}$	83.78
($r = 8$)	SAM	$93.76_{\pm 0.48}$	$56.76_{\pm 1.39}$	$88.16^{*}_{\pm 0.67}$	$83.56_{\pm 0.73}$	$83.68^{*}_{\pm 1.62}$	$97.08^{**}_{\pm 0.27}$	83.83
	DAS	$93.94_{\pm 0.56}$	$56.82^{*}_{\pm 0.90}$	$87.90_{\pm 0.57}$	$84.28^{*}_{\pm 0.63}$	$83.97^{**}_{\pm 0.87}$	$96.80_{\pm 0.42}$	83.95^{**}

Table 4: Results on RoBERTa-large fine-tuned on GLUE in the low-data regime across 5 runs with different sampled data. Results marked with † are reported by (Malladi et al. 2023). Best and second-best results per column are marked with ** and *, respectively. This notation applies to Tables 4 and 5.

OPT-6.7B		Params	CB	BoolQ	WSC	COPA	ReCoRD	SQuAD	DROP	Avg.(↑)
Zero-Shot			60.7	65.9	37.5	80	76.9	69.5	26.4	59.56
Full fine-tuning		6658.47M	71.4	68.7	63.5^{**}	82^{**}	78.7^{**}	81.8	29.2	67.89
LoRA ($r = 16$)		8.39M	87.5^{**}	77.8^{*}	63.5^{**}	81^{*}	77.1	85.9	32.7^{**}	72.21^{**}
LoRETTA	ADAM		80.4	76.6	58.7	80	77.3^{*}	86.7^{*}	32.1	70.25
($r = 16$)	SAM	0.96M	85.7^{*}	78.6^{**}	63.5^{**}	77	76.8	88.7^{**}	31.8	71.72^{*}
	DAS		82.1	77.2	63.5^{**}	81^{*}	77.2	86.2	32.7^{**}	71.41

Table 5: Results on OPT-6.7B fine-tuned on SuperGLUE and generation tasks in the low-data regime.

	ADAM	SAM	DAS
Runtime (↓)	1×	2×	1.04×

Table 6: Normalized runtime on OPT-6.7B using LoRETTA.

date for low-rank adaptation. We fine-tune RoBERTa-large, a pre-trained language model with 355M parameters, following the setting in (Malladi et al. 2023). We use a challenging few-shot learning setting, sampling 512 examples per class. Results are summarized in Table 4. Both SAM and DAS improve over the FLoRA baseline. Notably, DAS slightly outperforms both SAM and full fine-tuning. We hypothesize that this performance gap arises because SAM, despite its sharpness-minimization objective, may exhibit unintended side effects in low-rank subspace adaptation—such as converging to sharp minima due to perturbations outside the update subspace (Li et al. 2025). In contrast, DAS isolates and distills the beneficial norm dynamics of SAM while avoiding such drawbacks.

Improving LoRETTA on OPT-6.7B. LoRETTA utilizes TT decomposition to parameterize the incremental update for low-rank adaptation. We fine-tune OPT-6.7B (Zhang et al. 2022), an autoregressive language model with 6.7B parameters on the SuperGLUE tasks (Wang et al. 2019) and generation tasks including SQuAD (Rajpurkar et al. 2016) and DROP (Dua et al. 2019) using LoRETTA. We follow the setup in (Yang et al. 2024) to use a challenging low-data setting with 1000/500/1000 examples for training/validation/testing, using a prompt-based fine-tuning suggested in (Malladi et al. 2023). Results of SAM and DAS on OPT-6.7B are summarized in Table 5. As shown in the zero-shot and full fine-tuning results, the distribution shift between pre-trained data and fine-tuned data is large and can be overfitted by full fine-tuning. SAM improves LoRETTA by a significant margin, and DAS achieves a similar improvement. We also compare against LoRA (Hu et al. 2022) with rank 16, using approximately $8\times$ trainable parameters of LoRETTA. SAM and DAS enhance the much lighter LoRETTA to achieve a more competitive performance compared to LoRA, demonstrating the effectiveness of the proposed methods. Moreover, we compared the runtime of ADAM, SAM, and DAS. In Table 6, DAS saves more than 90% runtime of SAM but still achieves a competitive performance compared to SAM, suggesting scaling as a strong yet efficient alternative for the extra gradient of SAM.

Conclusion

In this work, we investigated the implicit regularization of SAM in general, multi-core tensorized models. Our theoretical analysis reveals that a key mechanism of SAM is the norm dynamics, governed by the covariance between the norms of tensor cores and gradient magnitudes. We distilled this insight into a simple yet effective method, DAS, explicitly mimicking this regularization. Extensive experiments on tasks including tensor completion, model compression, and parameter-efficient fine-tuning validate the effectiveness of SAM and DAS as a computationally efficient alternative. Future work includes extending our framework to heavy-ball regimes for momentum-based optimizers and developing discrete gradient-flow analyses to better bridge theory with practical training dynamics.

Acknowledgements

This research was supported by Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology CREST Program, Grant Number JPMJCR21.

References

Andriushchenko, M.; Croce, F.; Müller, M.; Hein, M.; and Flammarion, N. 2023. A Modern Look at the Relationship between Sharpness and Generalization. In *International Conference on Machine Learning*. PMLR.

Andriushchenko, M.; and Flammarion, N. 2022. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*. PMLR.

Anjum, A.; Eren, M. E.; Boureima, I.; Alexandrov, B.; and Bhattarai, M. 2024. Tensor train low-rank approximation (tt-lora): Democratizing ai with accelerated llms. In *2024 International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*.

Baek, C.; Kolter, J. Z.; and Raghunathan, A. 2024. Why is SAM Robust to Label Noise? In *The Twelfth International Conference on Learning Representations*.

Bahri, D.; Mobahi, H.; and Tay, Y. 2022. Sharpness-Aware Minimization Improves Language Model Generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Barrett, D.; and Dherin, B. 2021. Implicit Gradient Regularization. In *International Conference on Learning Representations*.

Bartlett, P. L.; Long, P. M.; and Bousquet, O. 2023. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *Journal of Machine Learning Research*.

Belkin, M.; Hsu, D.; Ma, S.; and Mandal, S. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*.

Bisla, D.; Wang, J.; and Choromanska, A. 2022. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Cao, T.; Sun, L.; Nguyen, C. H.; and Mamitsuka, H. 2024. Learning low-rank tensor cores with probabilistic L0-regularized rank selection for model compression. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee.

Deng, J.; Pang, J.; Zhang, B.; and Guo, G. 2025. Asymptotic Unbiased Sample Sampling to Speed Up Sharpness-Aware Minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Du, J.; Yan, H.; Feng, J.; Zhou, J. T.; Zhen, L.; Goh, R. S. M.; and Tan, V. 2022a. Efficient Sharpness-aware Minimization for Improved Training of Neural Networks. In *International Conference on Learning Representations*.

Du, J.; Zhou, D.; Feng, J.; Tan, V.; and Zhou, J. T. 2022b. Sharpness-aware training for free. *Advances in Neural Information Processing Systems*.

Du, S. S.; Hu, W.; and Lee, J. D. 2018. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2021. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*.

Ge, R.; Huang, F.; Jin, C.; and Yuan, Y. 2015. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*. PMLR.

Hariz, K.; Kadri, H.; Ayache, S.; Moakher, M.; and Artieres, T. 2022. Implicit regularization with polynomial growth in deep tensor factorization. In *International Conference on Machine Learning*. PMLR.

Hariz, K.; Kadri, H.; Ayache, S.; Moakher, M.; and Artières, T. 2024. Implicit regularization in deep tucker factorization: Low-rankness via structured sparsity. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Hayashi, K.; Yamaguchi, T.; Sugawara, Y.; and Maeda, S.-i. 2019. Exploring unexplored tensor network decompositions for convolutional neural networks. *Advances in Neural Information Processing Systems*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hrinchuk, O.; Khrulkov, V.; Mirvakhabova, L.; Orlova, E.; and Oseledets, I. 2020. Tensorized Embedding Layers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Hu, E. J.; yelong shen; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ishida, T.; Yamane, I.; Sakai, T.; Niu, G.; and Sugiyama, M. 2020. Do We Need Zero Training Loss After Achieving Zero Training Error? In *International Conference on Machine Learning*. PMLR.

Ji, J.; Li, G.; Fu, J.; Afghah, F.; Guo, L.; Yuan, X.; and Ma, X. 2024. A single-step, sharpness-aware minimization is all you need to achieve efficient and accurate sparse training. *Advances in Neural Information Processing Systems*.

Jiang, L.; Huang, D.; Liu, M.; and Yang, W. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*. PMLR.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Kolda, T. G.; and Bader, B. W. 2009. Tensor decompositions and applications. *SIAM review*.

Kossaifi, J.; Panagakis, Y.; Anandkumar, A.; and Pantic, M. 2019. Tensorly: Tensor learning in python. *Journal of Machine Learning Research*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kwon, J.; Kim, J.; Park, H.; and Choi, I. K. 2021. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*. PMLR.

Li, B.; Zhang, L.; and He, N. 2024. Implicit regularization of sharpness-aware minimization for scale-invariant problems. *Advances in Neural Information Processing Systems*.

Li, T.; He, Z.; Li, Y.; Wang, Y.; Shang, L.; and Huang, X. 2025. Flat-LoRA: Low-Rank Adaptation over a Flat Loss Landscape. In *Forty-second International Conference on Machine Learning*.

Liu, T.; Li, Y.; Wei, S.; Zhou, E.; and Zhao, T. 2021. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Liu, Y.; Mai, S.; Cheng, M.; Chen, X.; Hsieh, C.-J.; and You, Y. 2022. Random sharpness-aware minimization. *Advances in Neural Information Processing Systems*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Malladi, S.; Gao, T.; Nichani, E.; Damian, A.; Lee, J. D.; Chen, D.; and Arora, S. 2023. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*.

Memmel, E.; Menzen, C.; Schuurmans, J.; Wesel, F.; and Batselier, K. 2024. Position: Tensor Networks are a Valuable Asset for Green AI. In *International Conference on Machine Learning*. PMLR.

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*.

Novikov, A.; Podoprikhin, D.; Osokin, A.; and Vetrov, D. P. 2015. Tensorizing neural networks. *Advances in Neural Information Processing Systems*.

Phan, A.-H.; Sobolev, K.; Sozykin, K.; Ermilov, D.; Gusak, J.; Tichavskỳ, P.; Glukhov, V.; Oseledets, I.; and Cichocki, A. 2020. Stable low-rank tensor decomposition for compression of convolutional neural network. In *European Conference on Computer Vision*. Springer.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Razin, N.; Maman, A.; and Cohen, N. 2022. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. In *International Conference on Machine Learning*. PMLR.

Si, C.; Wang, X.; Yang, X.; Xu, Z.; Li, Q.; Dai, J.; Qiao, Y.; Yang, X.; and Shen, W. 2025. Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning. In *The Thirteenth International Conference on Learning Representations*.

Veeramacheneni, L.; Wolter, M.; Kuehne, H.; and Gall, J. 2025. Canonical Rank Adaptation: An Efficient Fine-Tuning Strategy for Vision Transformers. In *Forty-second International Conference on Machine Learning*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*.

Wang, A.; Qiu, Y.; Bai, M.; Jin, Z.; Zhou, G.; and Zhao, Q. 2024. Generalized tensor decomposition for understanding multi-output regression under combinatorial shifts. *Advances in Neural Information Processing Systems*.

Wang, W.; Sun, Y.; Eriksson, B.; Wang, W.; and Aggarwal, V. 2018. Wide compression: Tensor ring nets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wen, K.; Ma, T.; and Li, Z. 2022. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*.

Xie, W.; Pethick, T.; and Cevher, V. 2024. Sampa: Sharpness-aware minimization parallelized. *Advances in Neural Information Processing Systems*.

Yang, Y.; Zhou, J.; Wong, N.; and Zhang, Z. 2024. LoRETTA: Low-Rank Economic Tensor-Train Adaptation for Ultra-Low-Parameter Fine-Tuning of Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Yaras, C.; Wang, P.; Balzano, L.; and Qu, Q. 2024. Compressible Dynamics in Deep Overparameterized Low-Rank Learning & Adaptation. In *International Conference on Machine Learning*. PMLR.

Yin, M.; Sui, Y.; Liao, S.; and Yuan, B. 2021. Towards efficient tensor decomposition-based dnn model compression with optimization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhao, Q.; Zhou, G.; Xie, S.; Zhang, L.; and Cichocki, A. 2016. Tensor ring decomposition. *arXiv preprint arXiv:1606.05535*.