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

Tianxiao Cao, Kyohei Atarashi, Hisashi Kashima

Graduate School of Informatics, Kyoto University

京都大学 KYOTO UNIVERSITY

See arXiv

The 40th Annual AAAI Conference on Artificial Intelligence
January 20–27, 2026
Singapore

## Takeaway Message

In scale-invariant tensorized models, SAM implicitly regulates core norm imbalance, and this regulation is governed by a covariance term between core norms and gradient magnitudes. We developed DAS to mimic that regulation, without doubled gradient calculation. We validated the effectiveness of SAM and DAS.

## Background

- **Sharpness-Aware Minimization (SAM)**: Flat minima in objective functions are related to good generalization. SAM wants to solve

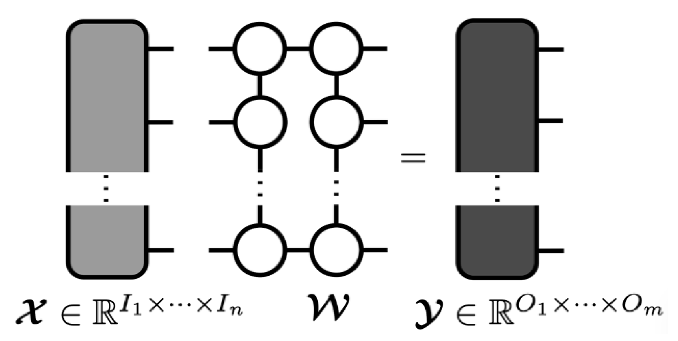
$$\min_{\theta} \max_{\|\epsilon\|_2 \le \rho} f(\theta + \epsilon).$$

SAM solves approximately by a gradient ascent-descent iteration:

$$\theta^{(t+\frac{1}{2})} = \theta^{(t)} + \rho \cdot \frac{\nabla f(\theta^{(t)})}{\|\nabla f(\theta^{(t)})\|_2},$$
$$\theta^{(t+1)} = \theta^{(t)} - \eta \cdot \nabla f(\theta^{(t+\frac{1}{2})}).$$

SAM is an empirically successful optimizer for deep learning [1].

- **Tensorized Models:** Tensor decomposition/networks are efficient representations for matrices/tensors. Example: Tensor networks for linear layer parameterization by folding dimensions.




Useful for model compression (tensorized neural network), tensor completion, and tensor-based low-rank adaptation, ....

## Analyzing SAM on Tensorized Models

Tensorized models are 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 is

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

Following the scheme of [2], **we analyze and compare the implicit dynamics of core tensors under gradient descent and SAM to see how SAM is special**. We study continuous gradient flow and assume Lipschitz-smooth $f(\cdot)$, and trace the following **Norm Deviation** among core tensors:

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

## Theoretical Results (Informal)

1. For SGD, $\frac{dQ}{dt} = 0$, Norm Deviation $Q$ is unchanged.
2. For SAM, $\frac{dQ}{dt} = 4\rho u^{(t)} K \cdot \mathrm{Cov}\left( \|\mathcal{G}_k^{(t)}\|_F^2, \|g_k^{(t)}\|_F^2 \right) + O(\rho^2 L)$, where $g_k^{(t)}$ is the gradient of $k$−th core, and $u^{(t)}$ is a normalization coefficient.

We study how this **implicit norm dynamics** matters.

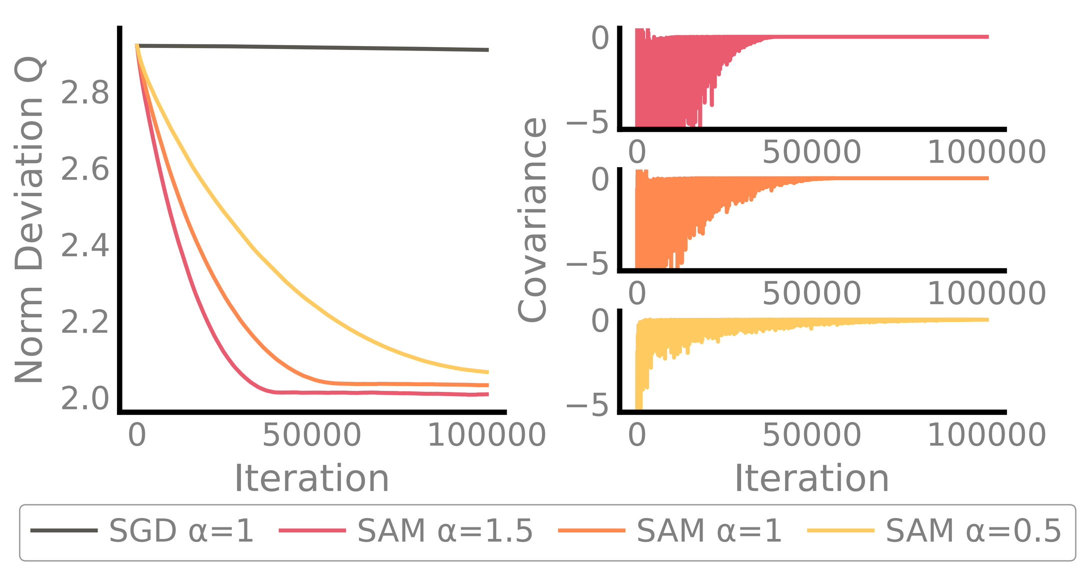



Figure 1. Implicit regularization of SAM on a toy Tucker-2 model to fit a target tensor.

## DAS: Mimicking $\left(\frac{dQ}{dt}\right)_{\text{SAM}}$ at the cost of SGD

We use a weight decay-like scaling scheme for optimization:

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

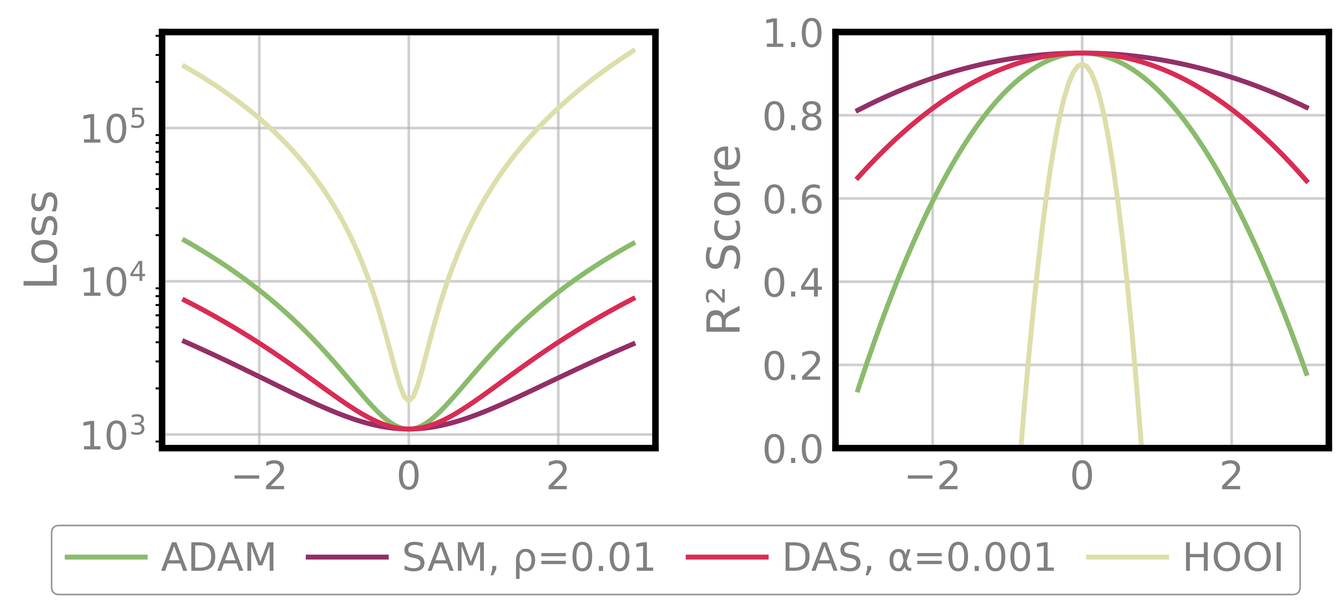
We propose Deviation-Aware Scaling (DAS) by setting

$$\lambda_k^{(t)} = \frac{\rho u^{(t)} \cdot \eta}{\|\mathcal{G}_k^{(t)}\|_F^2} \cdot \left( \|g_k^{(t)}\|_F^2 - \bar{g} \right). \tag{3}$$

This yields a greedy and local match $\left(\frac{dQ}{dt}\right)_{\text{DAS}} \approx \left(\frac{dQ}{dt}\right)_{\text{SAM}}$. DAS is efficient and does not require an extra gradient calculation as SAM.

## Experimental Results

**Tensor Completion with Tucker.** The x-axis is the size of a fixed directional perturbation applied to model parameters to show loss flatness. Both SAM and DAS find flatter minima than the base optimizer ADAM.




**Fine-tune compressed Tensor-train ResNets.** We compress ResNet using Tensor-train and fine-tune the compressed TT-ResNet on ImageNet.

| | 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** |

**Tensor-based LoRA.** LoRETTA [3] is a LoRA variant using Tensor-train. We fine-tune OPT using LoRETTA on a few-shot SuperGLUE task, and find both SAM and DAS to be effective, with DAS being the most efficient.

| OPT-6.7B | | Params | Avg.(↑) |
|---|---|---|---|
| Zero-Shot | | | 59.56 |
| Full fine-tuning | | 6658.47M | 67.89 |
| LoRA ($r = 16$) | | 8.39M | 72.21** |
| LoRETTA ($r = 16$) | ADAM | 0.96M | 70.25 |
| | SAM | | 71.72* |
| | DAS | | 71.41 |

## References

[1] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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

[3] Y. Yang, J. Zhou, N. Wong, and Z. Zhang. 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*, 2024.

## Acknowledgement

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