Revisiting Sharpness-Aware Minimization: A More Faithful and Effective Implementation

Jianlong Chen, Zhiming Zhou*

Shanghai University of Finance and Economics

April, ICLR 2026

Sharpness-Aware Minimization (SAM)

SAM has attained significant attention for its potential to enhance the generalization of machine learning models, *in a direct optimization manner.*

The Objective and Classical Approximation of SAM

- The formal objective:

$$\min_{\theta} \max_{\|\delta\| \le \rho} L(\theta + \delta)$$

Exactly solving the inner maximization is computationally expensive!

Main Approximation 1: SAM approximates best perturbation δ^* by gradient ascent(s). Formally,

$$\delta^* = \vartheta_k - \theta = \sum_{i=0}^{k-1} \rho_i \frac{g_i}{\|g_i\|}, \qquad \text{where } k \ge 1.$$

- After such approximation of δ^*, the SAM objective reduces to:

$$\min_{\theta} L(\theta + \delta^*)$$

Directly computing its gradient requires expensive higher-order derivatives!

Main Approximation 2: SAM approximates $\nabla_\theta L(\theta + \delta^*)$ using $\nabla_{\vartheta_k} L(\vartheta_k)$. Formally,

$$\nabla_\theta L(\theta + \delta^*) = \nabla_\theta L(\vartheta_k) = \nabla_{\vartheta_k} L(\vartheta_k) \cdot \underbrace{\nabla_\theta(\vartheta_k)}_{\text{Approximated as identity matrix } I} \approx \nabla_{\vartheta_k} L(\vartheta_k).$$

- The resulting algorithm essentially applies the gradient at the final ascent point ϑ_k to θ:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \nabla_{\vartheta_k} L(\vartheta_k).$$

Motivation

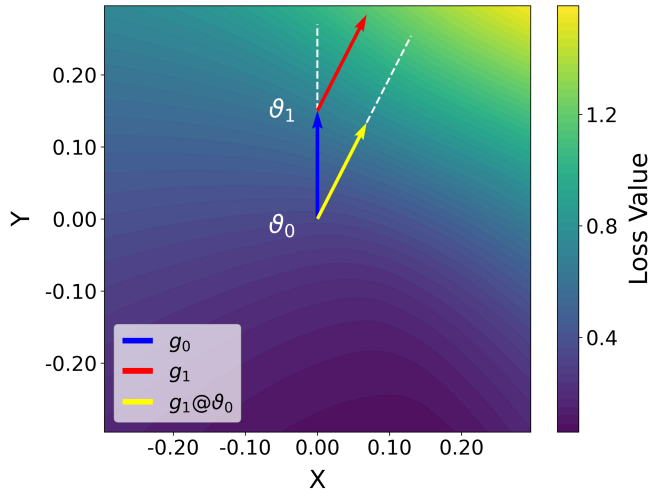
- After several approximations, SAM achieves empirical successes across various tasks using this implementation.
- A body of research has sought to demystify SAM after such approximations.
- However, a direct and intuitive understanding of why applying the nonlocal gradient at the ascent point to update the current parameter works superiorly is still lacking.

$$\theta_{t+1} = \theta_t - \eta_t \cdot \nabla_{\vartheta_k} L(\vartheta_k).$$

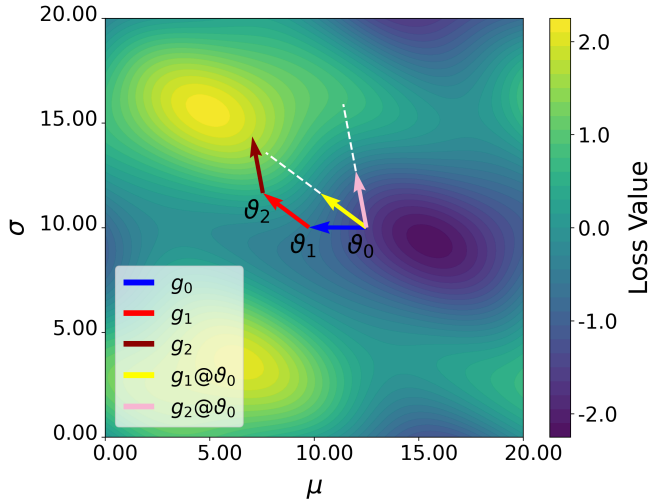
This gap necessitates a deeper investigation into SAM's fundamental mechanisms, which motivates our work.

Novel Interpretation of SAM's Mechanism

Empirical Observations



(a)



(b)

Key Mechanism

The gradient g_1 at the single-step ascent point ϑ_1, when applied to the current parameters ϑ_0, provides a better approximation of the direction from ϑ_0 toward the maximum within the neighborhood than gradient g_0.

Theoretical Confirmation

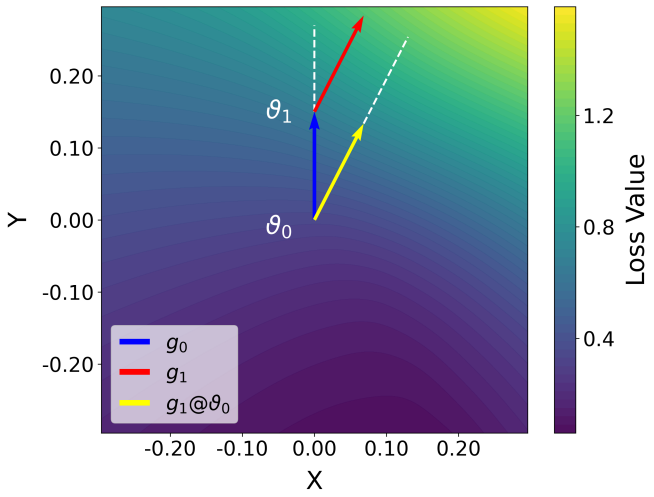
Proposition. Under some assumptions, there exists $\rho_0 > 0$ such that for all $\rho_m > \rho_0$:

SAM better approximates the direction toward the maximum in the vicinity than SGD

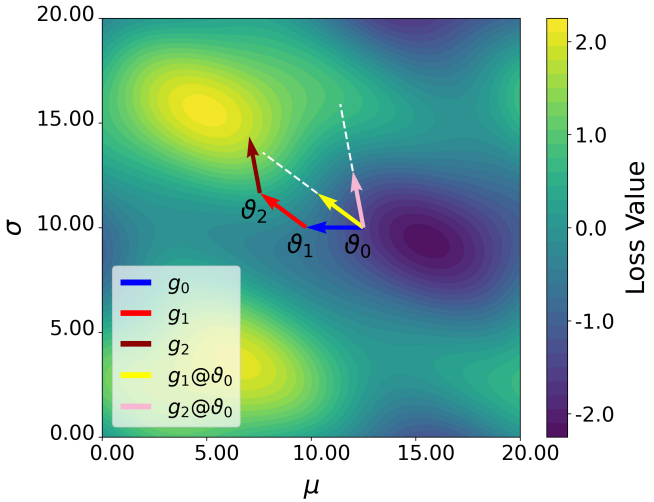
$$L\left(\vartheta_0 + \rho_m \frac{g_1}{\|g_1\|}\right) > L\left(\vartheta_0 + \rho_m \frac{g_0}{\|g_0\|}\right);$$

SAM's Limitations

Empirical Observations



(a)



(b)

Limitations

- L1: The approximation by the gradient at the single-step ascent point is rough, inaccurate and unstable.
- L2: The approximation quality may degrade as the number of ascent steps increases.

Theoretical Confirmation

Proposition. Under some assumptions, there exists $\rho_0 > 0$ such that for all $\rho_m > \rho_0$:

There exist better approximations than SAM there exists $\alpha \in \mathbb{R}$ such that

$$L\left(\vartheta_0 + \rho_m \frac{g_\alpha}{\|g_\alpha\|}\right) > L\left(\vartheta_0 + \rho_m \frac{g_1}{\|g_1\|}\right),$$

$$g_\alpha = \alpha g_1 + (1-\alpha) g_0.$$

So, how to improve?

EXplicit Sharpness-Aware Minimization (XSAM)

Takeaway: XSAM explicitly and dynamically estimates the maximum direction during training.

Main steps of XSAM

- **Explicit direction probing**: Find the direction of the maximum in the local neighborhood; (line 11)
  - ▷ Constrain the probe to a 2D hyperplane spanned by g_k at ϑ_k and the vector from ϑ_0 to ϑ_k; (line 8)
- **Dynamic adaptation**: Update direction during training. (line 10)
- **Guided optimization**: Steer the optimization using the explicitly estimated direction. (line 16)

▶ Full algorithm on the right

Algorithm 1: XSAM

Input: θ_0, T, $k \geq 1$, $\{\rho_i\}$, ρ_m, T_α, $\{\eta_t\}$
Output: θ_T

1: **for** $t = 0$ **to** $T - 1$ **do**
2: $\quad \vartheta_0 = \theta_t$
3: $\quad$ **for** $i = 0$ **to** $k - 1$ **do** ▷ *Single-step:* $k = 1$
4: $\quad\quad g_i = \nabla_{\vartheta_i} L(\vartheta_i)$
5: $\quad\quad \vartheta_{i+1} = \vartheta_i + \rho_i \frac{g_i}{\|g_i\|}$
6: $\quad$ **end for**
7: $\quad g_k = \nabla_{\vartheta_k} L(\vartheta_k)$
8: $\quad v_0 = \frac{\vartheta_k - \vartheta_0}{\|\vartheta_k - \vartheta_0\|}, \quad v_1 = \frac{g_k}{\|g_k\|}$
9: $\quad \psi = \arccos(v_0 \cdot v_1)$
10: $\quad$ **if** $t \bmod T_\alpha = 0$ **then**
11: $\quad\quad \alpha_t^* = \arg\max_\alpha L(\vartheta_0 + \rho_m \cdot v(\alpha)),$
12: $\quad\quad$ where $v(\alpha) = \frac{\sin((1-\alpha)\psi)}{\sin(\psi)} v_0 + \frac{\sin(\alpha\psi)}{\sin(\psi)} v_1$
13: $\quad$ **else**
14: $\quad\quad \alpha_t^* = \alpha_{t-1}^*$
15: $\quad$ **end if**
16: $\quad \theta_{t+1} = \theta_t - \eta_t \cdot v(\alpha_t^*) \cdot \|g_k\|$
17: **end for**

Experimental Results of Single-step Setting

XSAM consistently improves SAM across different datasets and architectures.

CIFAR-10

Dataset	CIFAR-10		
Model	VGG-11	ResNet-18	DenseNet-121
SGD	$93.19_{\pm 0.11}$	$96.15_{\pm 0.05}$	$96.34_{\pm 0.11}$
SAM	$93.83_{\pm 0.06}$	$96.59_{\pm 0.06}$	$96.97_{\pm 0.02}$
XSAM	$\mathbf{94.25}_{\pm 0.14}$	$\mathbf{96.74}_{\pm 0.04}$	$\mathbf{97.15}_{\pm 0.03}$

CIFAR-100

Dataset	CIFAR-100		
Model	VGG-11	ResNet-18	DenseNet-121
SGD	$71.46_{\pm 0.17}$	$78.55_{\pm 0.20}$	$81.78_{\pm 0.06}$
SAM	$74.01_{\pm 0.05}$	$80.93_{\pm 0.11}$	$83.81_{\pm 0.02}$
XSAM	$\mathbf{74.21}_{\pm 0.14}$	$\mathbf{81.24}_{\pm 0.07}$	$\mathbf{83.96}_{\pm 0.10}$

Tiny-ImageNet

Dataset	Tiny-ImageNet		
Model	VGG-11	ResNet-18	DenseNet-121
SGD	$47.44_{\pm 0.33}$	$57.02_{\pm 0.42}$	$61.93_{\pm 0.10}$
SAM	$51.96_{\pm 0.26}$	$62.81_{\pm 0.09}$	$66.31_{\pm 0.09}$
XSAM	$\mathbf{52.58}_{\pm 0.38}$	$\mathbf{63.82}_{\pm 0.23}$	$\mathbf{66.81}_{\pm 0.08}$

ImageNet, Transformer, and ViT-Ti

	ImageNet ResNet-50	Transformer IWSLT2014	ViT-Ti CIFAR-100
SAM	$77.04_{\pm 0.09}$	$35.30_{\pm 0.04}$	$67.80_{\pm 0.22}$
XSAM	$\mathbf{77.22}_{\pm 0.07}$	$\mathbf{35.63}_{\pm 0.13}$	$\mathbf{68.32}_{\pm 0.18}$

Experimental Results of Multi-step Setting

XSAM achieves the best performance among SAM variants.

- **Multi-step SAMs on CIFAR-100 with ResNet-18**

Methods	$k = 1$	$k = 2$	$k = 4$
SAM	$80.93_{\pm 0.11}$	$80.91_{\pm 0.10}$	$80.65_{\pm 0.26}$
LSAM	$80.93_{\pm 0.11}$	$80.94_{\pm 0.09}$	$80.74_{\pm 0.18}$
LSAM+	$80.61_{\pm 0.20}$	$80.83_{\pm 0.11}$	$80.41_{\pm 0.03}$
MSAM	$80.93_{\pm 0.11}$	$81.18_{\pm 0.06}$	$81.01_{\pm 0.09}$
MSAM+	$80.83_{\pm 0.05}$	$80.86_{\pm 0.34}$	$80.77_{\pm 0.08}$
XSAM	$\mathbf{81.27}_{\pm 0.07}$	$\mathbf{81.44}_{\pm 0.09}$	$\mathbf{81.37}_{\pm 0.24}$

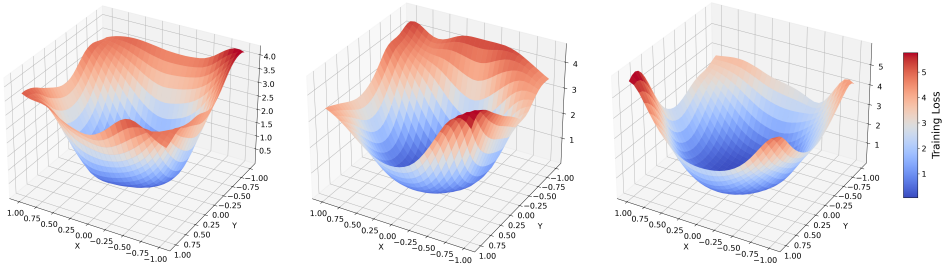
- **Multi-step ($k = 3$) SAMs on CIFAR-100 using ResNet-18 across different ρ**

Method	$\rho = 0.04$	$\rho = 0.08$	$\rho = 0.12$
SAM	$80.79_{\pm 0.41}$	$80.75_{\pm 0.27}$	$79.72_{\pm 0.33}$
LSAM	$81.00_{\pm 0.21}$	$81.20_{\pm 0.24}$	$81.16_{\pm 0.04}$
LSAM+	$80.56_{\pm 0.20}$	$80.77_{\pm 0.04}$	$80.21_{\pm 0.27}$
MSAM	$81.04_{\pm 0.06}$	$81.12_{\pm 0.17}$	$80.93_{\pm 0.11}$
MSAM+	$80.72_{\pm 0.16}$	$81.16_{\pm 0.05}$	$81.16_{\pm 0.05}$
XSAM	$\mathbf{81.23}_{\pm 0.06}$	$\mathbf{81.36}_{\pm 0.08}$	$\mathbf{81.29}_{\pm 0.06}$

Additional Experimental Results

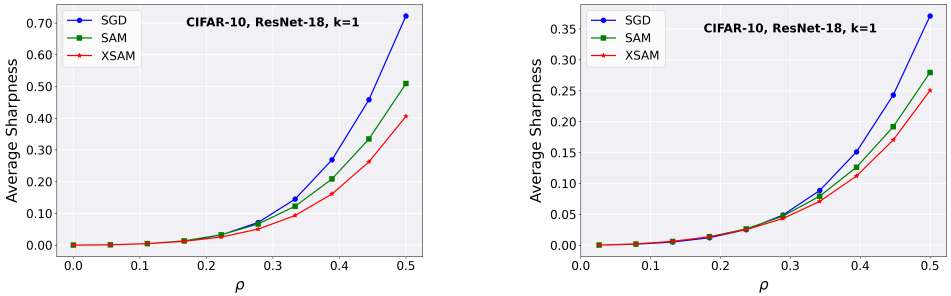
Flatness Analysis

- Visualization of loss landscape



SGD SAM XSAM

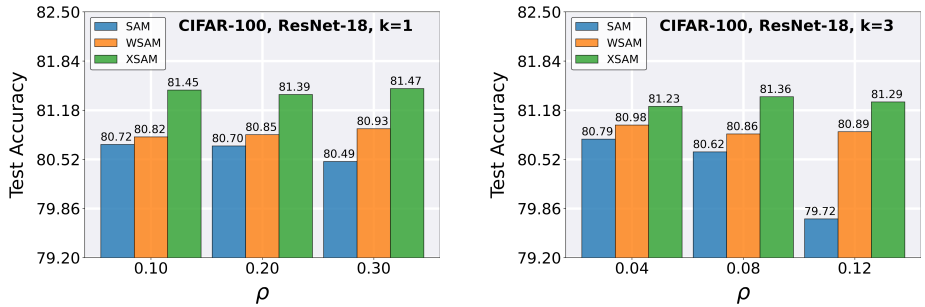
- Average Sharpness (perturbation)



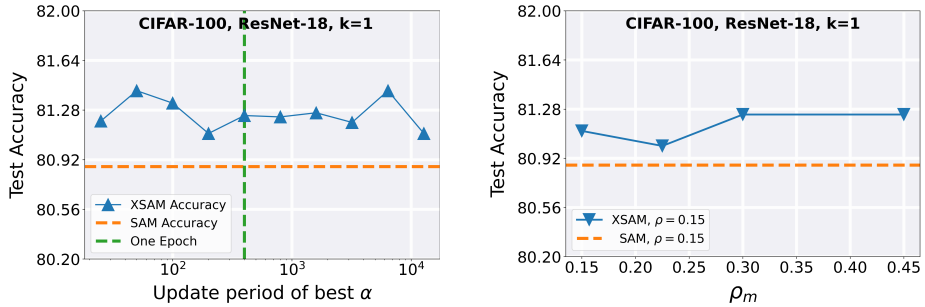
Element-wise Filter-wise

Robustness Analysis

- Robustness to ρ



- Robustness to T_α and ρ_m



Summary

Takeaway: More accurately estimate the direction toward the maximum and then move away along it for improved sharpness-aware minimization.

- SAM better approximates the direction toward the maximum in the vicinity than SGD.
- The classic SAM gradient cannot accurately approximate the direction of the maximum, and may get worse as the number of ascent steps increases.
- XSAM explicitly and dynamically estimate the direction of the maximum, leading to a more faithful and effective sharpness-aware minimization.
- XSAM features a unified formulation that applies to both single-step and multi-step settings and only incurs negligible computational overhead.

Thank you for your attention!

If you have any questions, please feel free to contact me.