Unveiling m-Sharpness Through the Structure of Stochastic Gradient Noise

Haocheng Luo Mehrtash Harandi Dinh Phung Trung Le
Monash University, Australia

Motivation

Sharpness-Aware Minimization (SAM) [1] seeks flat minima by minimizing the loss under adversarial perturbations:

$$\min_x f(x+\rho\epsilon), \quad \epsilon = \arg\max_{\|\epsilon\|\le 1} \langle \nabla f(x), \epsilon \rangle.$$

m-SAM splits each mini-batch into micro-batches of size m, computes perturbations and gradients on them, and then aggregates the updates. This design also enables efficient multi-GPU parallelism, as each device computes its own perturbation **locally** without additional inter-GPU communication.

m-sharpness phenomenon [1,2]: smaller $m \Rightarrow$ better performance!

Contributions

- **Two-parameter SDE framework:** jointly tracks learning rate η and perturbation radius ρ to *arbitrary* orders, and derives continuous-time dynamics for several SAM/unnormalized SAM (USAM) variants.
- **Theoretical explanation of m-sharpness:** shows how *stochastic gradient noise (SGN)* induces an implicit **variance/sharpness regularization** term in the drift, whose strength increases as m decreases.
- **Reweighted SAM (RW-SAM):** uses the magnitude of SGN as importance weights to **mimic the generalization benefits of m-SAM while remaining parallelizable**.

Key Theory: SDE for USAM Variants

For all three USAM variants, the drift term can be written in closed form under our two-parameter SDE framework.

full-batch USAM (n-USAM):

$$dX_t = -\nabla\left(f(X_t) + \frac{\rho}{2}\|\nabla f(X_t)\|^2\right) dt + \sqrt{\eta}\,\Sigma_{\text{n-USAM}}\, dW_t.$$

Mini-batch USAM:

$$dX_t = -\nabla\left(f(X_t) + \frac{\rho}{2}\|\nabla f(X_t)\|^2 + \frac{\rho}{2|\gamma|}\operatorname{tr} V(X_t)\right) dt + \sqrt{\eta}\,\Sigma_{\text{USAM}}\, dW_t.$$

m-USAM:

$$dX_t = -\nabla\left(f(X_t) + \frac{\rho}{2}\|\nabla f(X_t)\|^2 + \frac{\rho}{2m}\operatorname{tr} V(X_t)\right) dt + \sqrt{\frac{m\eta}{|\gamma|}}\,\Sigma_{\text{m-USAM}}\, dW_t.$$

Here, $|\gamma|$ denotes the mini-batch size, $V(x)$ denotes the SGN covariance.

Key insight: The trace of the SGN covariance $V(x)$ appears in the drift. Its coefficient increases from $\rho/(2|\gamma|)$ (mini-batch USAM) to $\rho/(2m)$ (m-USAM), while it disappears entirely in n-USAM. Thus,

smaller m $\Rightarrow$ stronger variance regularization $\Rightarrow$ better generalization.

These insights extend to the normalized (vanilla) SAM variants as well, which follow the same noise-induced pattern but *do not admit closed-form drift expressions*.

Visualizing the m-Sharpness Effect and SGN Structure

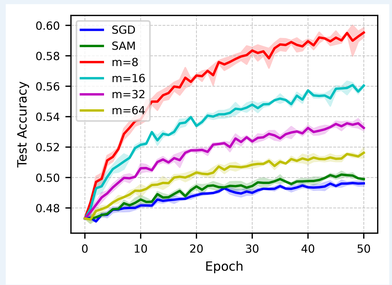

Fig. 1: Speed of escape from poor minima

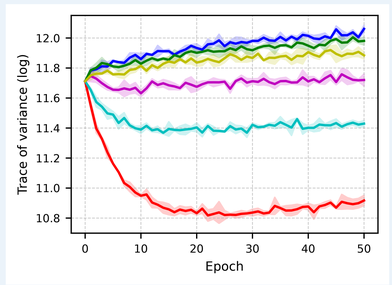

Fig. 2: Trace of SGN covariance

Parallelism Constraint of m-SAM

When $m < b$ (per-device batch size), the perturbations for the micro-batches must be computed **sequentially** on each device, creating a local dependency that prevents parallel execution within the device and introduces substantial extra computation.

Proposed Method: Reweighted SAM (RW-SAM)

Goal: **strengthen SGN-induced regularization while retaining full parallelism**.

We assign importance weights p_i within a mini-batch and maximize

$$\max_{P\in\Delta} \sum_{i\in\gamma} p_i \|\nabla f_i(x)\| + \frac{1}{\lambda}\mathbb{H}(P),$$

yielding Gibbs weights

$$p_i^\star = \frac{\exp(\lambda\|\nabla f_i(x)\|)}{\sum_{j\in\gamma}\exp(\lambda\|\nabla f_j(x)\|)}.$$

The weights are then used to form a reweighted gradient

$$\tilde{g}(x) = \sum_{i\in\gamma} p_i^\star \nabla f_i(x),$$

and the RW-SAM perturbation is computed as

$$\epsilon_{\text{RW}} = \rho\frac{\tilde{g}(x)}{\|\tilde{g}(x)\|}.$$

Practical notes:

- Estimate per-sample gradient norms via finite differences and Monte Carlo with only an extra forward pass:

$$\|\nabla f_i(x)\| \approx \sqrt{\left(\frac{f_i(x+\delta z) - f_i(x)}{\delta}\right)^2}.$$

- Use Rademacher noise z to reduce estimator variance.
- Training cost is about $\sim 1/6$ more than SAM, but much lower than m-SAM with small m.

CIFAR-10/100: Test Accuracy Comparison

(a) CIFAR-10 Test Accuracy

Model	SGD	SAM	RW-SAM
ResNet-18	95.62 ± 0.03	95.99 ± 0.07	**96.24** ± 0.05
ResNet-50	95.64 ± 0.37	96.06 ± 0.04	**96.34** ± 0.04
WideResNet	96.47 ± 0.03	96.91 ± 0.02	**97.11** ± 0.05

(b) CIFAR-100 Test Accuracy

Model	SGD	SAM	RW-SAM
ResNet-18	78.91 ± 0.18	78.90 ± 0.27	**79.31** ± 0.28
ResNet-50	79.55 ± 0.16	80.31 ± 0.35	**80.83** ± 0.05
WideResNet	81.55 ± 0.15	83.25 ± 0.07	**83.52** ± 0.08

RW-SAM consistently outperforms SAM and SGD across all models.

ImageNet-1K and Fine-Tuning Results

(a) ImageNet-1K — ResNet-50 Test Accuracy

Model	SGD	SAM	RW-SAM
ResNet-50	76.67 ± 0.05	77.16 ± 0.04	**77.37** ± 0.05

(b) ViT-B/16 Fine-Tuning on CIFAR-10/100

Dataset	SGD	SAM	RW-SAM
CIFAR-10	98.24 ± 0.05	98.40 ± 0.02	**98.58** ± 0.02
CIFAR-100	78.91 ± 0.18	89.63 ± 0.12	**89.89** ± 0.09

Reference

[1] Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). Sharpness-aware minimization for efficiently improving generalization. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net.

[2] Andriushchenko, M. and Flammarion, N. (2022). Towards understanding sharpness-aware minimization. In International Conference on Machine Learning, pages 639–668. PMLR.

Contact

- Paper QR:
- Email: haocheng.luo@monash.edu