A Unified Stability Analysis of SAM vs SGD: Role of Data Coherence and Emergence of Simplicity Bias

Wei-Kai Chang • Purdue University

Rajiv Khanna • Purdue University

NeurIPS 2025

Motivation

- Why does SAM find flatter / simpler solutions than SGD?
- Current theory: sharpness local, ignores data geometry.
- Goal: explain stability using data coherence & alignment.

Key Idea / Intuition

- Stability ≠ just curvature — depends on gradient and loss curvature coherence.
- Coherence matrix eigenvalues determine stability region.
- Incorporation of data into scope of optimization.
- SAM tightens stability → forces aligned, low-rank features.

Problem Setup

- Study linearized dynamics around solution.

$$w_{t+1} = w_t - \eta H_t w_t = (I - \eta H_t) w_t = \hat{J}_t w_t,$$

- Coherence matrix captures inter-sample directional alignment:

Definition 1. ***Coherence measure*** *[Dexter et al., 2024].For a collection of per-example Hessians* $\{H_i\}_{i=1}^n$*, define the coherence matrix* $S \in \mathbb{R}^{n \times n}$ *with entries* $S_{ij} = \|H_i^{1/2} H_j^{1/2}\|_F = \sqrt{\mathrm{Tr}(H_i H_j)}$*. The coherence measure* σ *is defined as follows:*

$$\sigma = \frac{\lambda_{\max}(S)}{\max_{i \in n} \lambda_{\max}(H_i)} \tag{2}$$

- We want to study the convergence or divergence through the norm of weight:

$$\mathbb{E}[\|w_k\|^2] = \mathbb{E}[w_0^\top \hat{J}_1^\top \cdots \hat{J}_k^\top \hat{J}_k \cdots \hat{J}_1 w_0] = \mathbb{E}[\mathrm{Tr}(\hat{J}_k \cdots \hat{J}_1 w_0 w_0^\top \hat{J}_1^\top \cdots \hat{J}_k^\top)]$$

Theory: Random perturbation

- For random perturbation algorithm as follows:

$$\begin{aligned} w_{t+1} &= w_t - \eta \nabla_S l(w + \delta_t) \\ &= w_t - \eta H_t w_t - \eta H_t \delta_t \end{aligned}$$

1. *Sufficient condition for divergence is as follows:*

$$\eta \geq \frac{\sigma}{\lambda_1}\left(\frac{n}{b} - 1\right)^{-\frac{1}{2}}$$

2. *(Comparative Divergence Speed) Suppose* $\mathrm{Tr}[J^{2k}] \leq C_0 \alpha^k$ *for some constants* C_0 *and* α_k*, then the divergence rate of the random perturbation method is asymptotically within a constant factor of that of standard SGD:*

$$\lim_{k \to \infty} \frac{E[\|w_k\|^2]_{\textit{Random, lower bound}}}{E[\|w_k\|^2]_{\textit{SGD, lower bound}}} = \mathcal{O}(1)$$

3. *Suppose the step size satisfies the convergence criterion established in prior stability analyses (e.g., Dexter et al. [2024]). Then, under the random perturbation update (3), the expected squared norm of the iterates remains bounded as* $k \to \infty$*:*

$$\lim_{k \to \infty} E[w_k^T w_k]_{\textit{upper bound}} = \mathcal{O}(1)$$

Theory: Sharpness-aware-Minimization(SAM)

- For SAM algorithm as follows: $w_{t+1} = \left(I - \eta H_t \left(I + \frac{\rho}{\alpha} H\right)\right) w_t$

1. ***Divergence criterion.*** *SAM diverges if the largest eigenvalue of the Hessian exceeds the following threshold:*

$$\lambda_{\max}(H) \geq \frac{\sigma}{\eta} \left(\frac{n}{B} - 1\right)^{-1/2} \left(1 + \frac{\rho}{\alpha} \lambda_{\min}(H)\right)^{-1}.$$

Compared to SGD, this condition is stricter due to the additional curvature-dependent term in the denominator, implying that SAM escapes sharp minima more aggressively.

2. ***Convergence criterion.*** *If there exists* $\epsilon \in (0, 1)$ *such that*

$$\frac{\epsilon}{\eta} \leq \lambda_i + \frac{\rho}{\alpha} \lambda_i^2 \leq \frac{2 - \epsilon}{\eta}, \quad \forall i \in [d],$$

and the accumulated noise decays sufficiently fast (see Appendix C.9), then the iterates converge in expectation:

$$\lim_{k \to \infty} \mathbb{E}[\|w_k\|^2] = 0.$$

2-layer ReLu Network under linear stability

- We study the coherence under the following solution with linear stability framework:

Definition 2. *((C, r)-generalizing solution) Let $\{a_1, a_2, \ldots, a_C\} \in \{0, 1\}^C$. We construct W_1 such that each hidden unit encodes a pattern of the form:*

$$W_{1,j} = r \cdot [(-1)^{a_1}, (-1)^{a_2}, \ldots, (-1)^{a_C}, 0, \ldots, 0],$$

with j indexing a binary encoding of the a_i's i.e. $j = 1 + \sum 2^{i-1} a_i$. We set $W_2[j] = \frac{1}{r}(-1)^{a_1+a_2}$ to match. For $k > C$, $W_{1,k} = 0$, $W_2[k] = 0, b[k] = 0$.

2-layer ReLu Network under linear stability

- Coherence for SGD and SAM:

Theorem 3.5 (SGD Stability of (C, r)-Generalizing Solutions)**.** *Fix* $r = (d+1)^{1/4}$. *Then, with probability at least* $1-\delta$*, for a randomly drawn dataset of size* n*, the top eigenvalue of the coherence matrix under a* (C, r)*-generalizing solution satisfies:*

$$\lambda_{\max}(S) = \mathcal{O}\left(\frac{n}{2^C}(d+1)^{1/2}\right),$$

while $\max_i \lambda_{\max}(H_i) = 2(d+1)^{\frac{1}{2}}$.

Theorem 3.6 (SAM Stability of (C, r)-Generalizing Solutions)**.** *Under a* (C, r)*-generalizing solution for a randomly iid drawn dataset of size n, the top eigenvalue of the SAM-induced coherence matrix satisfies:*

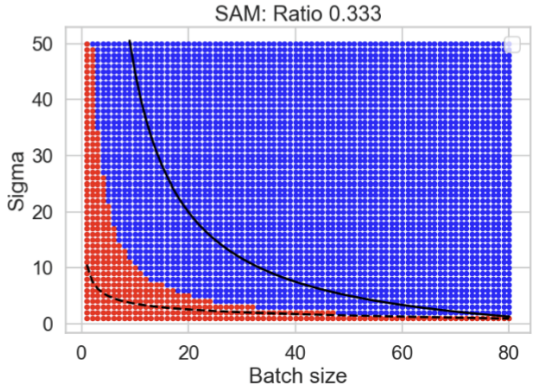
$$\lambda_{\max}(S^{\mathrm{SAM}}) = \mathcal{O}\left(\frac{n}{2^c}(d+1)^{\frac{1}{2}}\sqrt{(1+\frac{\rho}{\alpha}\frac{2(d+1)^{\frac{1}{2}}}{2^c})^2 + \frac{\rho^2}{\alpha^2}(\frac{1}{n}(\frac{1}{2^c}-\frac{1}{2^{2c}}))4(d+1)}\right),$$

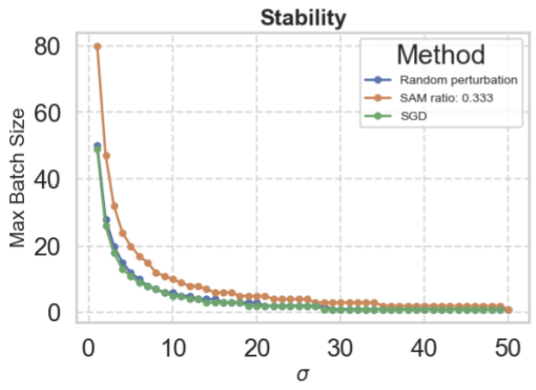
and,

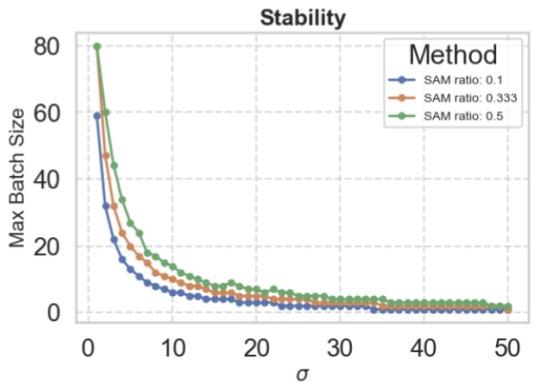
$$\max_i \lambda_{\max}((I+\frac{\rho}{\alpha}H)H_i) = 2(d+1)^{\frac{1}{2}}(1+\frac{\rho}{\alpha}\frac{1}{2^C}2(d+1)^{\frac{1}{2}})$$

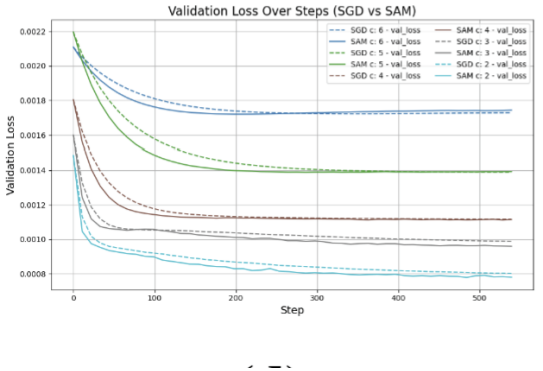
Experiments results

- Synthetic data and 2-layer ReLu:









- CIRAR-10 and ResNet-18:

Optimizer	Rank	Coherence
SGD	148.33 ± 3.22	1.0045 ± 0.0020
SAM (0.05)	157.67 ± 9.29	1.0052 ± 0.0060
SAM (0.1)	144.33 ± 7.10	1.0771 ± 0.0680
SAM (0.2)	128.67 ± 5.51	1.0907 ± 0.0890

Table 2: Effect of Optimizer and SAM radius on feature rank and coherence. We record the rank and coherence on subset of CIFAR10 after training for 200 epochs.

Summary

- Data coherence can play a central role in optimization compared to classic optimization which only investigate the largest Hessian direction.

- Contributions:
  - Unified Linear Stability Analysis.
  - Emergence of Simplicity Bias for SGD through data coherence.
  - SAM intensifies the Simplicity Bias.

Thank You