uai2025

Critical Influence of Overparameterization on Sharpness-aware Minimization

Sungbin Shin[1*], Dongyeop Lee[1*], Maksym Andriushchenko[2] and Namhoon Lee[1]

[1]Pohang University of Science and Technology, [2]EPFL

Close/extensive empirical/theoretical analysis reveals critical influence of overparameterization on the effectiveness of SAM.

Sharpness-Aware Minimization

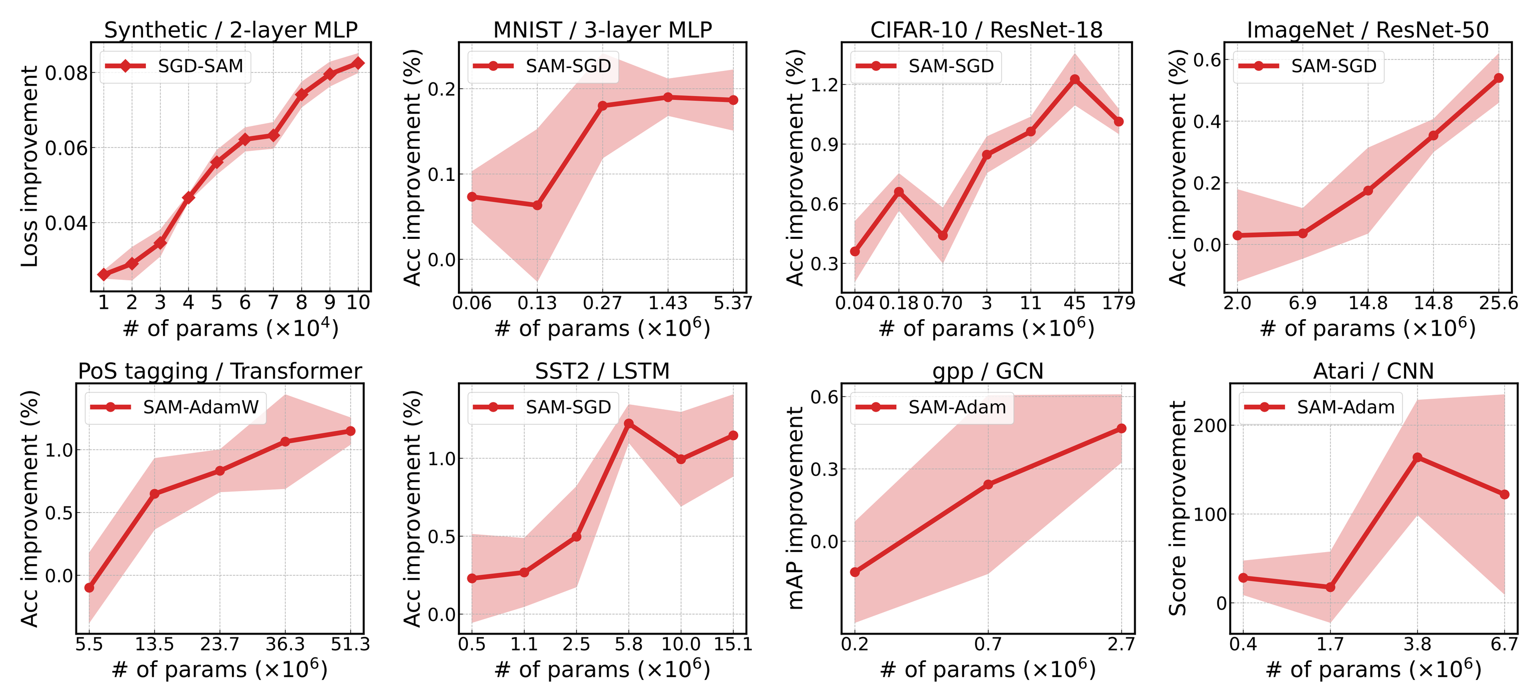
- Prior observations: sharpness of the minima $\propto$ generalization error (Keskar et al., 2017; Jiang et al., 2020)
- Foret et al. (2021): suggests SAM to explicitly minimize sharpness to improve generalization

$$\min_x \max_{\|\epsilon\|_2 \le \rho} f(x+\epsilon) \quad \rightarrow \quad x_{t+1} = x_t - \eta \nabla f\left(x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|_2}\right)$$

- Overlooked assumption: there exist sufficiently many solutions with large variations of sharpness/flatness for SAM to exploit.
- Overparameterization is usually believed to provide such conditions $\rightarrow$ eludes possibility of its critical influence

SAM depends on overparameterization

Workload #	Domain	Task	Dataset	Architecture	Model
1	Synthetic	Regression	Synthetic	MLP	Two-layer MLP
2	Vision	Image classification	MNIST	MLP	LeNet-300-100
3	Vision	Image classification	CIFAR-10	CNN	ResNet-18
4	Vision	Image classification	ImageNet	CNN	ResNet-50
5	Language	PoS tagging	Universal Dependencies	Transformer	Encoder-only Transformer
6	Language	Sentiment classification	SST-2	RNN	LSTM
7	Chemistry	Graph property prediction	ogbg-molpcba	GNN	GCN
8	Game	Proximal policy optimization	Atari Breakout	CNN	Five-layer CNN



- number of parameters $\uparrow \implies$ benefit of SAM $\uparrow$: overparameterization helps SAM
- number of parameters $\downarrow \implies$ benefit of SAM $\downarrow \approx 0$: SAM does not work without overparameterization

How Overparameterization influences SAM

1. Enlarged solution space allows finding simpler/flatter solutions

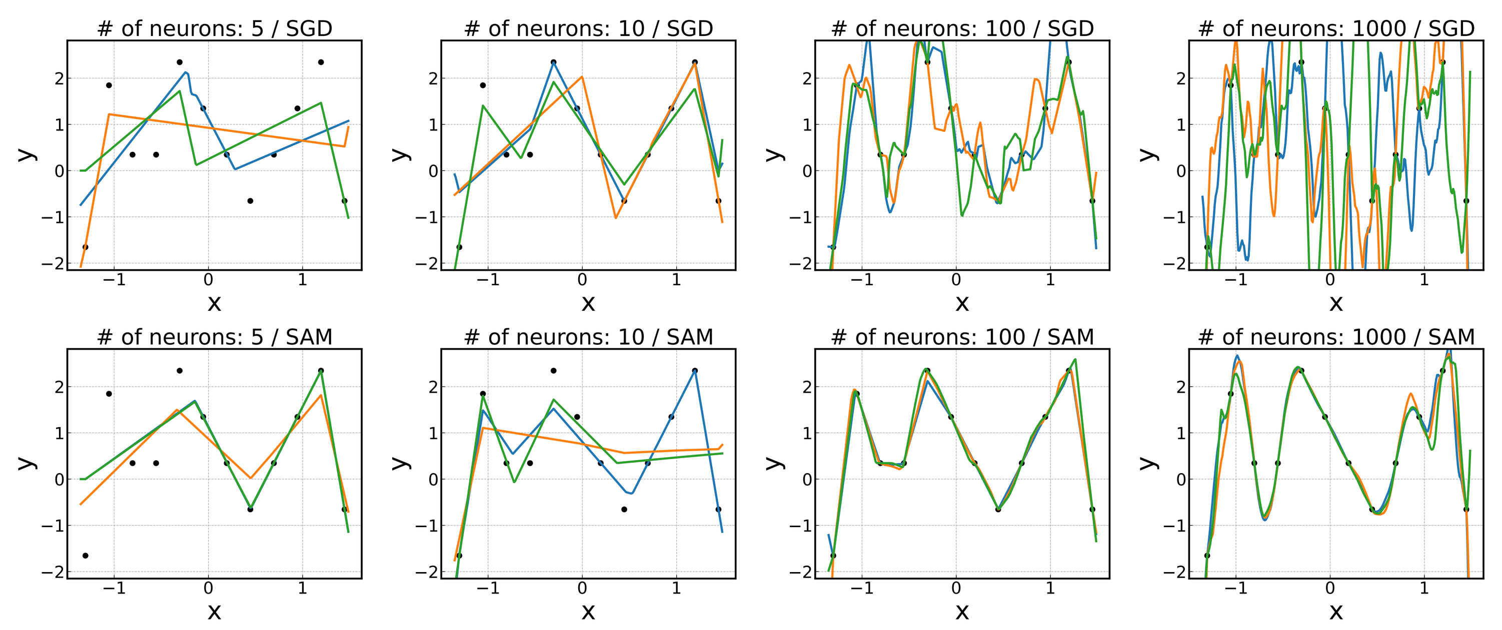


Fig. 2: Solutions found by SGD (top) and SAM (bottom)

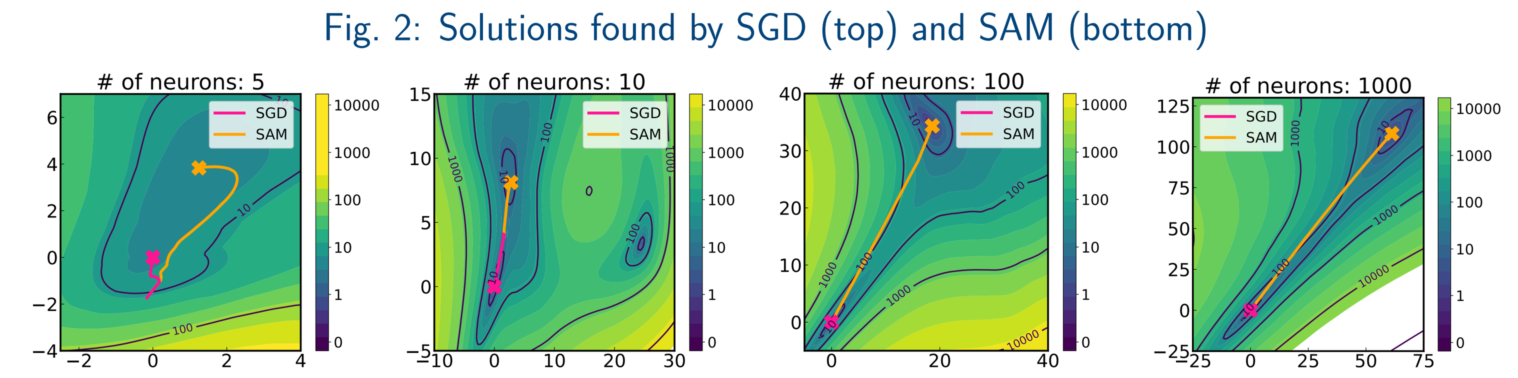


Fig. 3: Optimization trajectories of SGD and SAM

2. Allows stronger implicit bias

SDE modeling by Compagnoni et al. (2023)

$$\tilde{f}(x) := f(x) + \rho \mathbb{E}\|\nabla f_\gamma(x)\|_2$$

: larger perturbation bound $\rho \uparrow \rightarrow$ stronger implicit regularization

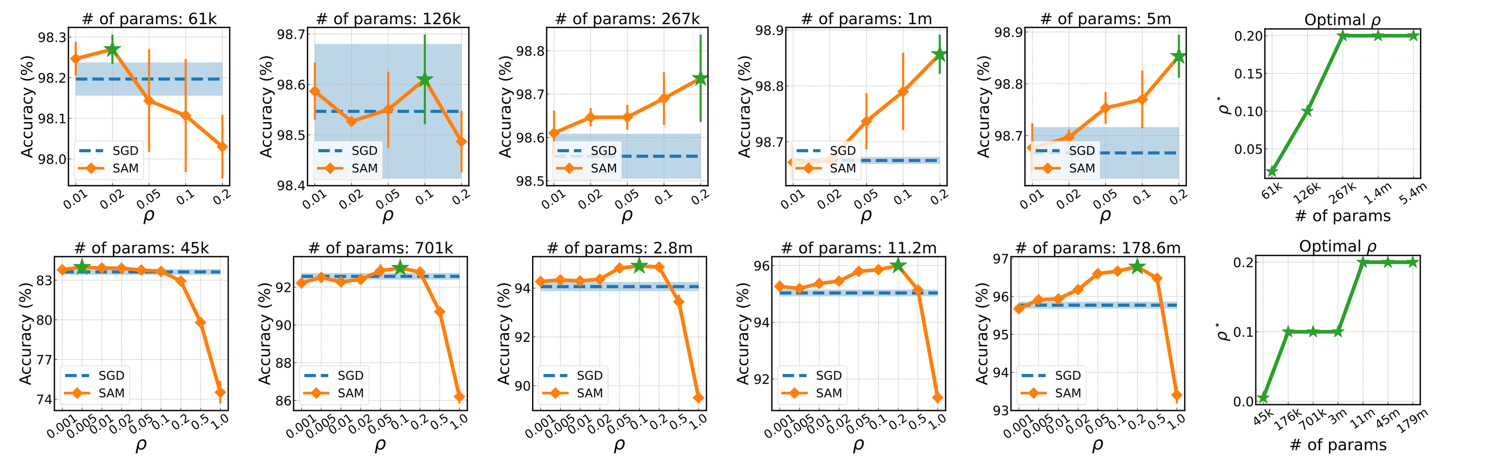


Fig. 4: Larger models prefer larger ρ

Further merits and caveats

Effect of label noise, sparsity, and regularization

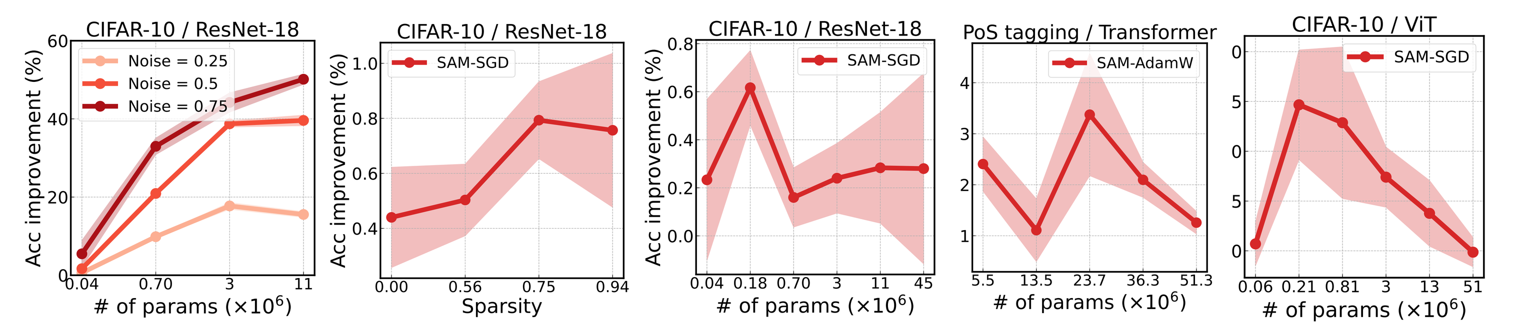


Fig. 5: Label noise Fig. 6: Sparsity Fig. 7: w/o weight decay Fig. 8: w/o early stop. Fig. 9: w/o induc. bias

- (Fig. 5) The benefit of SAM is more pronounced with a higher noise level.
- (Fig. 6) The improvement by SAM tends to increase in large sparse models compared to their small dense counterparts.
- (Fig. 7-9) SAM does not always benefit from overparameterization without sufficient regularization.

Other effects of overparameterization: Theoretical aspects

Definition 1. *(Interpolation) Let $f(x) = \sum_{i=1}^n f_i(x)$. There exists $x^\star$ s.t. $f_i(x^\star) = 0$ and $\nabla f_i(x^\star) = 0$ for $i = 1, \dots, n$.*

SAM escapes sharp minima with non-uniform Hessian

Definition 2. *(Linear stability) A minimizer $x^\star$ is linearly stable if there exists a constant C such that $\mathbb{E}[\|\tilde{x}_t - x^\star\|^2] \le C\|\tilde{x}_0 - x^\star\|^2$ for all $t > 0$ under $\tilde{x}_{t+1} = \tilde{x}_t - \nabla G(x^\star)(\tilde{x}_t - x^\star)$.*

Necessary condition of Linear stability

$x^\star$ is a linearly stable minima of SAM if

$$\underbrace{0 \le a(1+\rho a) \le \frac{2}{\eta}}_{\text{bounded sharpness}}, \quad \underbrace{0 \le s_2^2 \le \frac{1}{\eta(\eta - 2\rho)}, \quad 0 \le s_3^3 \le \frac{1}{2\eta^2\rho}, \quad 0 \le s_4^4 \le \frac{1}{\eta^2\rho^2}}_{\text{bounded Hessian non-uniformity}}$$

where $a = \lambda_{\max}(H)$, $s_k = \lambda_{\max}((\mathbb{E}_i[H_i^k] - H^k)^{1/k})$.

- This is stricter than SGD, i.e., $0 \le a \le 2/\eta$ (Wu et al., 2018)

Stochastic SAM converges much faster with overparameterization

Theorem 6 (Linear convergence of Stochastic SAM under overparameterization)

Suppose that f_i is β-smooth, f is λ-smooth and α-PL, and interpolation holds. For any $\rho \le \frac{1}{(\beta/\alpha + 1/2)\beta}$, a stochastic SAM that runs for t iterations with step size $\eta^\star \overset{\text{def}}{=} \frac{\alpha - (\beta + \alpha/2)\beta\rho}{2\lambda\beta(\beta\rho+1)^2}$ gives the following convergence guarantee:

$$\mathbb{E}_{x_t}[f(x_t)] \le \left(1 - \frac{\alpha - (\beta + \alpha/2)\beta\rho}{2}\eta^\star\right)^t f(x_0).$$

Contact: {ssbin4, dylee23, namhoonlee}@postech.ac.kr, maksym@andriushchenko.me