The Effects of Overparameterization on Sharpness-Aware Minimization: An Empirical and Theoretical Analysis

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

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

Training an overparameterized neural network can yield minimizers of the same level of training loss and yet different generalization capabilities. With evidence of correlation between sharpness of minima and their generalization errors, increasing efforts were made to develop an optimization method to find flat minima as more generalizable solutions. This sharpness-aware minimization (SAM) strategy, however, has not been studied much yet as to how overparameterization can actually affect its behavior. In this work, we analyze SAM under varying degrees of overparameterization and present both empirical and theoretical results that suggest a critical influence of overparameterization on SAM.

Our contributions

- We prove that SAM can achieve a linear convergence rate under overparameterization in a stochastic setting using standard techniques in optimization.
- We also show based on a stability analysis that the solutions found by SAM are indeed flatter and have more uniformly distributed Hessian moments compared to those of SGD.
- These results are corroborated with our experiments that reveal a consistent trend that the generalization improvement made by SAM continues to increase as the model becomes more overparameterized.
- We further present that sparsity can open up an avenue for effective overparameterization in practice.

Sharpness-Aware Minimization

- Based on recent observations that indicate a correlation between the sharpness of empirical risk $f := \Sigma_{i=1}^n f_i$ at a minimum and its generalization error (Keskar et al., 2017; Jiang et al., 2020), Foret et al. (2021) suggest the min-max problem of the following form
$$\min_x \max_{\|\epsilon\|_2 \leq \rho} f(x+\epsilon)$$
where ϵ and ρ denote some perturbation added to x and its bound, respectively; thus, the goal is now to seek x that minimizes f in its entire ϵ-neighborhood, such that the objective landscape becomes flat.
- Taking the first-order Talyor approximation of f at x and solving for optimal $\epsilon^\star$ gives the following update rule for SAM:
$$x_{t+1} = x_t - \eta \nabla f\left(x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|_2}\right).$$

Overparameterization

- A neural network can be called overparameterized if it has a sufficient number of parameters to interpolate the whole training data, i.e., it achieves zero training loss.
- From a stochastic optimization perspective, this means that there exists some point that is stationary for all the risk for individual data point f_i. We formalize these observations as follows:

Definition 1. *(Interpolation) There exists $x^\star$ such that $f_i(x^\star) = 0$ and $\nabla f_i(x^\star) = 0$ for all $i = 1, \dots, n$, where n is the number of training data points.*

Stochastic SAM achieves linear convergence

Definition of relevant assumptions

Definition 2. *(Smoothness) f is β-smooth if there exists $\beta > 0$ s.t. $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x-y\|$ for all $x, y \in \mathbb{R}^d$.*

Definition 3. *(Polyak-Lojasiewicz) f is α-PL if there exists $\alpha > 0$ s.t. $\|\nabla f(x)\|^2 \geq \alpha(f(x) - f(x^\star))$ for all $x \in \mathbb{R}^d$.*

Lemmas

These two lemmas essentially show that the stochastic SAM gradient aligns well with and scales to the standard stochastic gradient, i.e., how similar SAM is to SGD.

Lemma 4. *Suppose that f_i is β-smooth. Then*
$$\langle \nabla f_i(x + \rho \nabla f_i(x)), \nabla f(x) \rangle \geq \langle \nabla f_i(x), \nabla f(x) \rangle - \frac{\beta\rho}{2}\|\nabla f_i(x)\|^2 - \frac{\beta\rho}{2}\|\nabla f(x)\|^2.$$

Lemma 5. *Suppose that f_i is β-smooth. Then*
$$\|\nabla f_i(x_t + \rho \nabla f_i(x_t))\|^2 \leq (\beta\rho+1)^2 \|\nabla f_i(x_t)\|^2.$$

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

This result shows that a stochastic SAM converges at a linear rate under overparameterization, which we corroborate through the following empirical results.

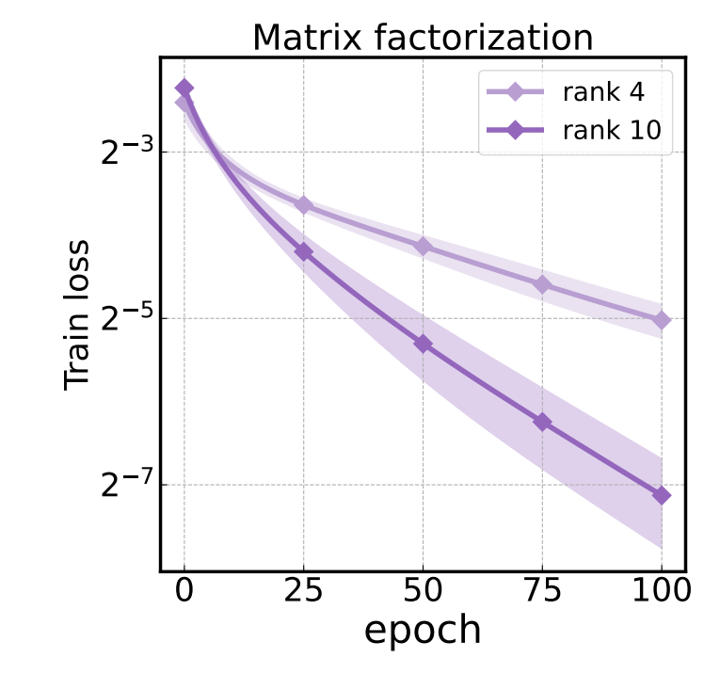

Fig. 1: Matrix Factorization

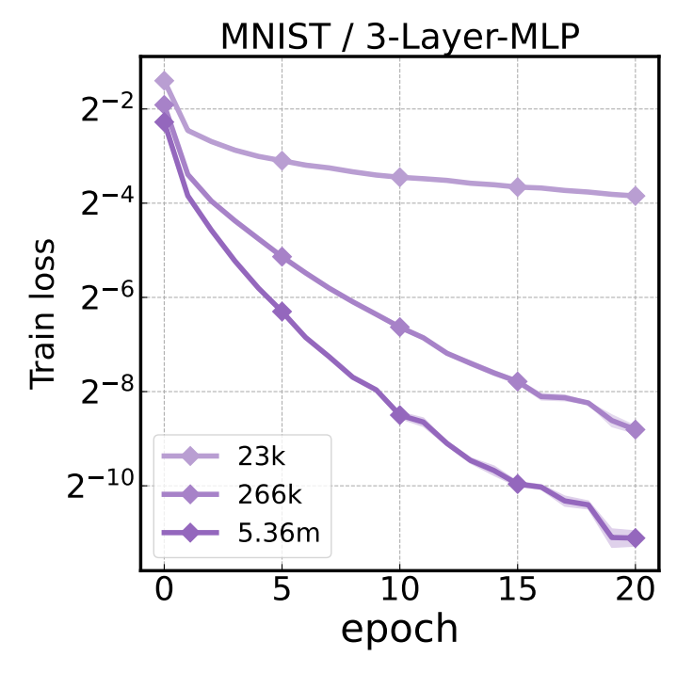

Fig. 2: MNIST

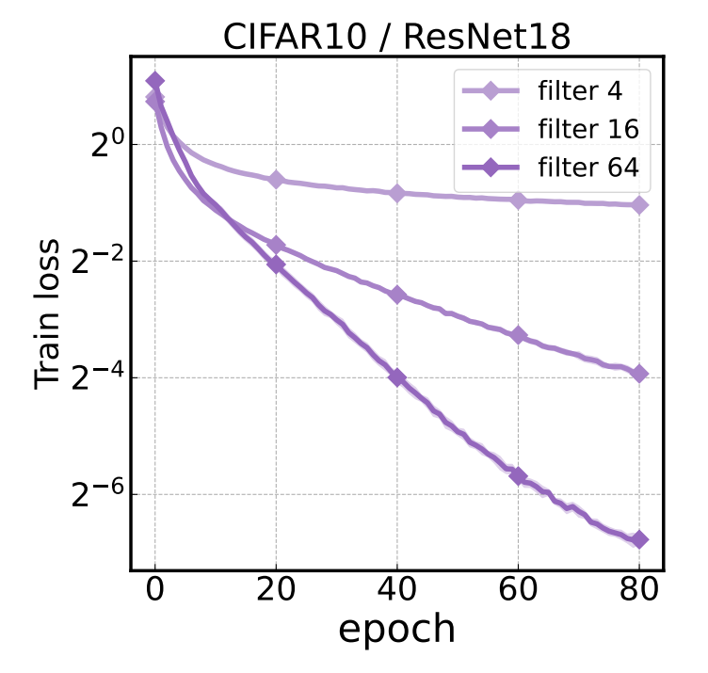

Fig. 3: CIFAR-10

Linear stability analysis of SAM

Definition of Linear stability

Definition 6. *(Linear stability) A minimizer $x^\star$ is linearly stable if there exists a constant C such that $\mathbb{E}[\|\tilde{x}_t - x^\star\|^2] \leq C\|\tilde{x}_0 - x^\star\|^2$ 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

SAM requires bounded sharpness for linear stability.

Let $a = \lambda_{\max}(H)$ be the sharpness. Then,
$$0 \leq a(1+\rho a) \leq \frac{2}{\eta}.$$
Here, the sharpness decreases as ρ increases. Comparing with the necessary condition for SGD in Wu et al. (2018), i.e., $0 \leq a \leq 2/\eta$, this result indicates that SAM selects flatter minima than SGD in the overparameterized regime.

SAM requires bounded Hessian non-uniformity for linear stability.

Let $s_k = \lambda_{\max}((\mathbb{E}_i[H_i^k] - H^k)^{1/k})$ be the non-uniformity of the Hessian measured with the k-th moment. Then,
$$0 \leq s_2^2 \leq \frac{1}{\eta(\eta-2\rho)}, \quad 0 \leq s_3^3 \leq \frac{1}{2\eta^2\rho}, \quad 0 \leq s_4^4 \leq \frac{1}{\eta^2\rho^2}.$$
We find that SAM puts additional constraints on s_3 and s_4 whereas SGD only upper bounds s_2. We also remark that a larger ρ makes the bounds on s_3 and s_4 smaller, which may lead to more uniform Hessian moments.

To corroborate our result, we evaluate the empirical sharpness and non-uniformity of Hessian on an overparameterized MLP network for MNIST with squared loss. All models are trained to reach near-zero loss.

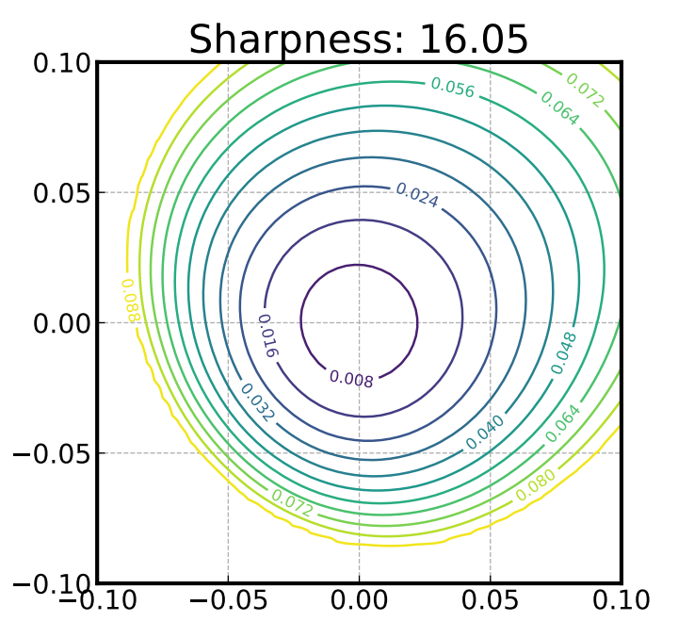

Fig. 4: Loss landscape (SGD)

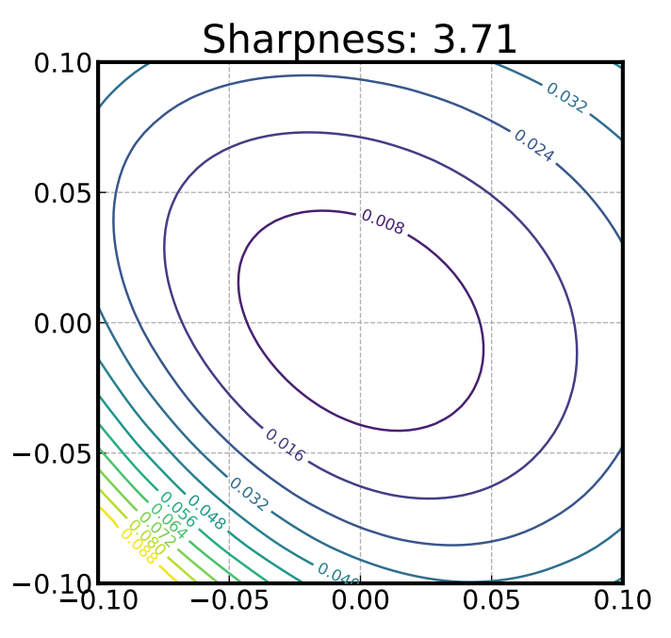

Fig. 5: Loss landscape (SAM)

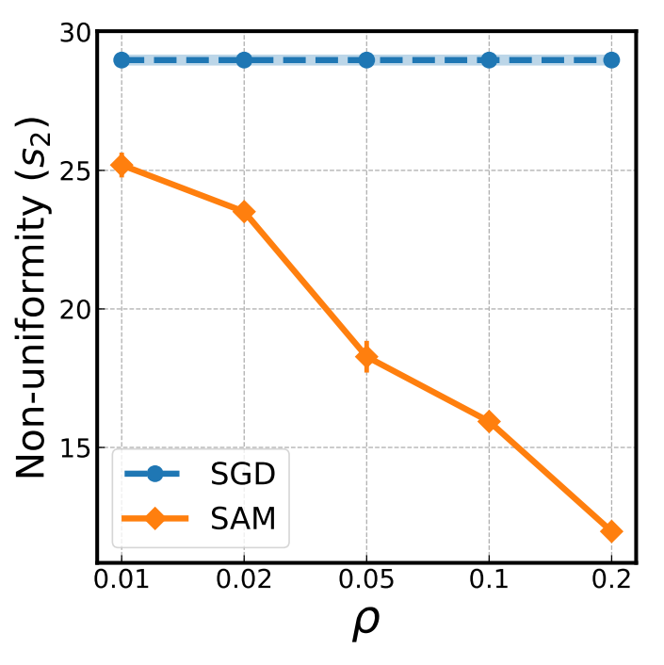

Fig. 6: Hessian non-uniformity

Overparameterization benefits SAM improvement

Here, we evaluate the effect of overparameterization on the generalization improvement of SAM, i.e., the gap of validation accuracy between SAM and SGD.

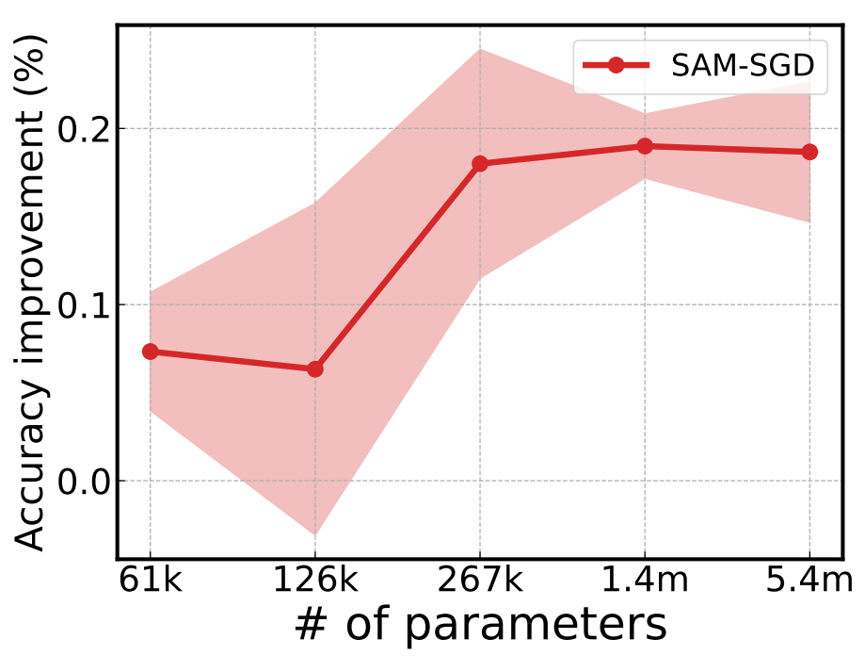

Fig. 7: MNIST/3-layer MLP

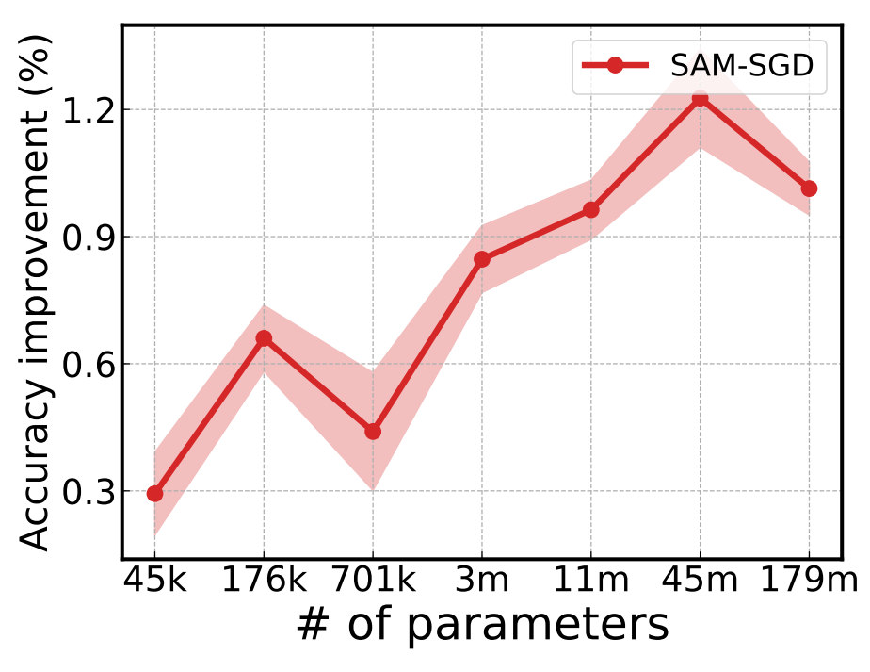

Fig. 8: CIFAR10 / Resnet18

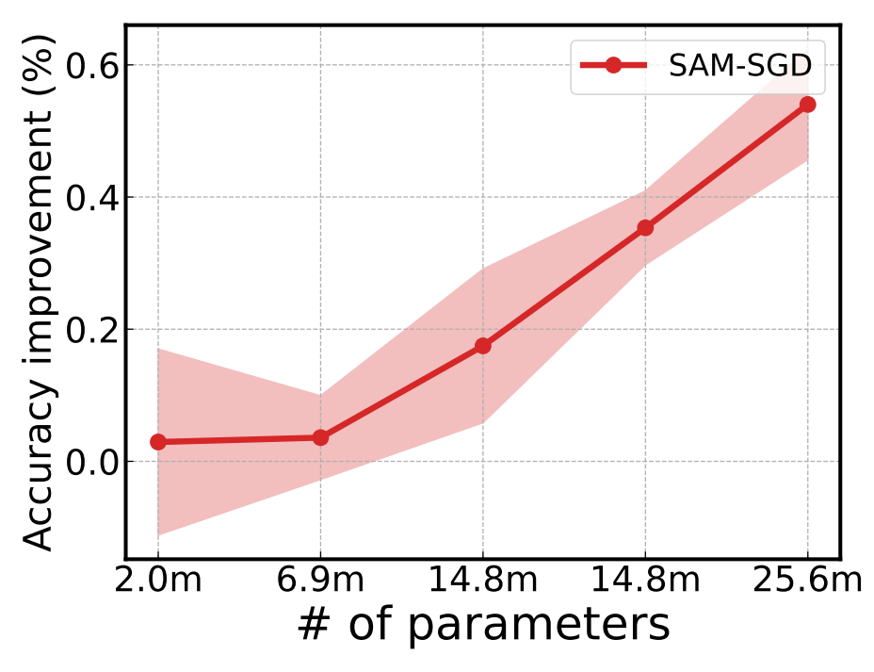

Fig. 9: Imagenet / Resnet50

- The generalization benefit of SAM increases with an increasing number of parameters across all settings; in other words, SAM outperforms SGD with a larger margin in overparameterized regimes.
- Specifically for Cifar-10 with ResNet18, the accuracy gap between SAM and SGD increases from around 0.3% for networks with 45 thousand parameters to 1.0% for those with more than 11 million parameters.
- Overall, the increased generalization performance of SAM with more parameters renders a promising avenue since modern neural network models are often heavily overparameterized (Zhang et al., 2022; Dehghani et al., 2023).

Sparse overparameterization for SAM

- In this section, we train several models of varying sparsity levels from scratch using SGD and SAM and compare their generalization performances.
- Here we sparsify an overparameterized model such that the number of parameters matches the original model.

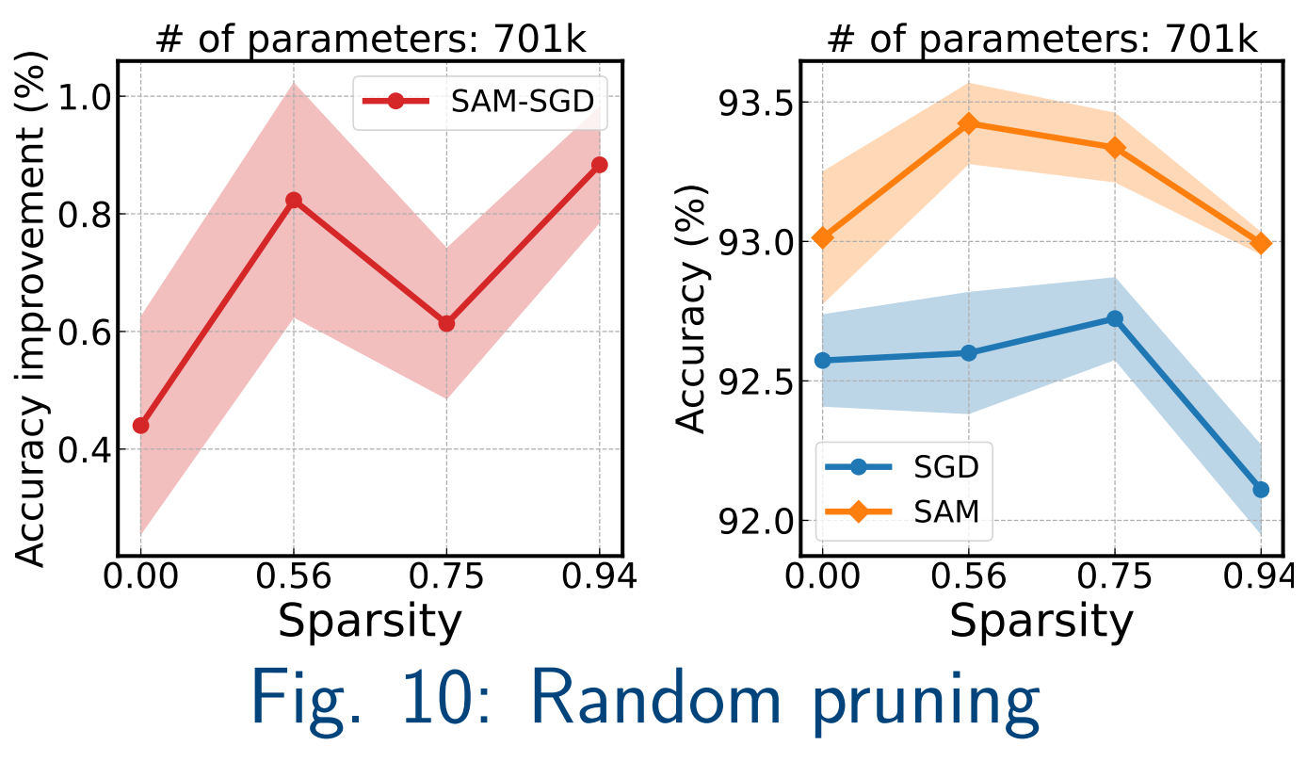

Fig. 10: Random pruning

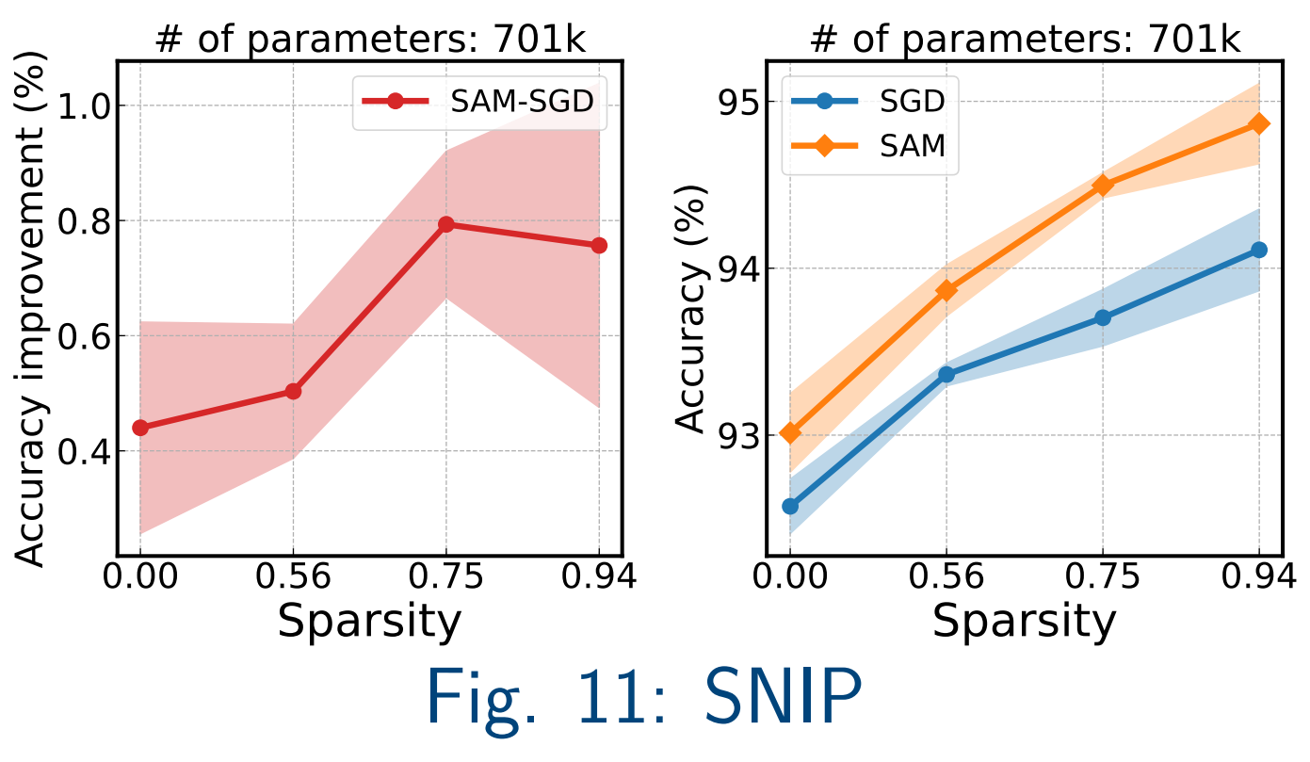

Fig. 11: SNIP

- We first observe that the generalization improvement of SAM from SGD tends to increase as the model becomes more sparsely overparameterized, which suggests that one can consider taking sparsification more actively when using SAM.
- We also see that the trend seems more evident with SNIP which preserves the trainability of sparse models better than the naive random sparsity.

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