An Adaptive Policy to Employ Sharpness-Aware Minimization

Weisen Jiang◇† Hansi Yang† Yu Zhang◇ James Kwok†

◇Southern University of Science and Technology †Hong Kong University of Science and Technology

Introduction

- Empirical Risk Minimization (ERM) $\min_{\mathbf{w}} \mathcal{L}(\mathcal{D};\mathbf{w})$ and its update rule is

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \nabla \mathcal{L}(\mathcal{B}_t; \mathbf{w}_{t-1}).$$

- $\mathcal{L}(\mathcal{D};\mathbf{w})$ is non-convex and has many local minima with **poor generalization**.
- **Sharpness-Aware Minimization** (SAM) [1] seeks flat minima by solving a min-max optimization $\min_{\mathbf{w}} \max_{\|\epsilon\|\le\rho} \mathcal{L}(\mathcal{D};\mathbf{w}+\epsilon)$ and its update rule is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla\mathcal{L}(\mathcal{B}_t; \mathbf{w}_t + \rho \nabla \mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)).$$

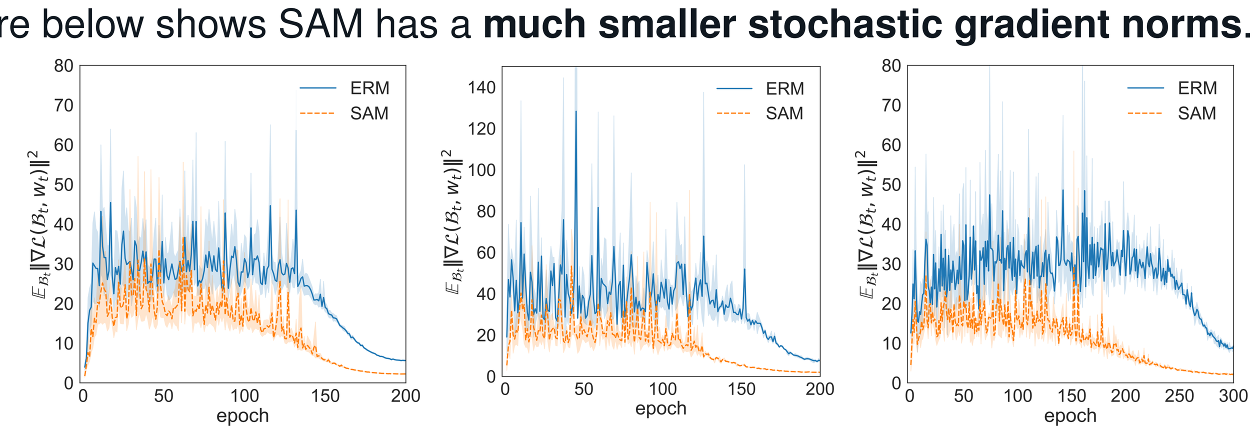
- Though generalizing better, each SAM update consists of two gradient computations: one for computing the perturbation and the other for the actual update direction, thus, is computationally expensive.
- **Prior works** on improving the efficiency of SAM:
  - ESAM [2] uses **fewer** samples to compute gradients and updates **fewer** parameters, but still requires **two** gradient computations
  - LookSAM [3] switches SAM and ERM **periodically**
  - SS-SAM [4] **randomly** selects SAM or ERM according to a Bernoulli trial
- **Research GAP**: Though more efficient, the random or periodic use of SAM is sub-optimal as it is **not geometry-aware**
- Intuitively, **SAM is more useful in sharp regions than in flat regions** In this paper, we propose an **adaptive** policy to employ SAM based on loss landscape geometry.

A Sharpness Measure

- Introduce **a sharpness measure** $\mathbb{E}_{\mathcal{B}_t}\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2$.
  - $\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2$ = trace of $\mathrm{diag}\left([\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)]^2\right)$ (a Hessian approximation).
  - $\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2$ is related to gradient variance

$$\mathrm{Var}(\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)) = \mathbb{E}_{\mathcal{B}_t}\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2 - \underbrace{\|\nabla\mathcal{L}(\mathcal{D};\mathbf{w}_t)\|^2}_{\approx 0 \text{ when algorithm converges}}$$

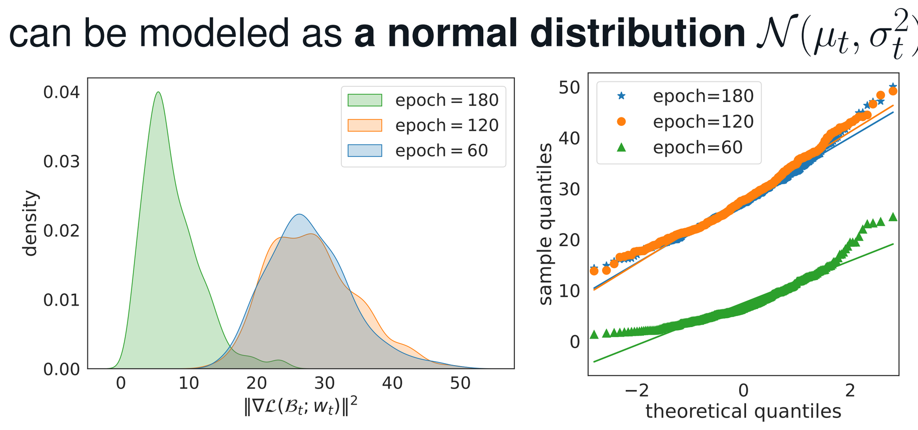
  which is **positively correlated** with the generalization gap [5].
- Figure below shows SAM has a **much smaller stochastic gradient norms**.



(a): *ResNet-18* (b): *WRN-28-10*. (c): *PyramidNet-110*.

Figure 1: Squared stochastic gradient norms $\mathbb{E}_{\mathcal{B}}\|\nabla\mathcal{L}(\mathcal{B};\mathbf{w}_t)\|^2$ on *CIFAR-100*.

- Computing $\mathbb{E}_{\mathcal{B}_t}\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2$ for each iteration is **expensive**.
- $\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2$ can be modeled as **a normal distribution** $\mathcal{N}(\mu_t,\sigma_t^2)$



(a): Distributions (b): Q-Q plots.

Figure 2: Stochastic gradient norms $\{\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2 : \mathcal{B}_t \sim \mathcal{D}\}$ of *ResNet-18* on *CIFAR-100*.

- Use exponential moving average (EMA) to estimate μ_t and σ_t^2 ($\delta = 0.9$):

$$\mu_t = \delta\mu_{t-1} + (1-\delta)\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2,\ \sigma_t^2 = \delta\sigma_{t-1}^2 + (1-\delta)(\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2 - \mu_t)^2.$$

Adaptive Policy to Employ SAM

- **An adaptive policy**: Employ SAM only at $\mathbf{w}_t$ where loss landscape is locally sharp:
  - when $\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2 \ge \mu_t + c_t\sigma_t$, SAM is used;
  - otherwise, ERM is used.
- Note that when $c_t \to -\infty$, it reduces to SAM; when $c_t \to \infty$, it reduces to ERM.
- SAM update is more effective towards the end of training [6], thus, we design a schedule $c_t \equiv g_{\lambda_1,\lambda_2}(t) = \frac{t}{T}\lambda_1 + (1-\frac{t}{T})\lambda_2$ (decrease c_t from λ_2 to λ_1 linearly).
- **The policy can be combined with any SAM variant, e.g., AE-LookSAM for LookSAM.**

Proposed Algorithms

Algorithm 1 AE-SAM and AE-LookSAM.

Require: $\mathcal{D}$, stepsize η, radius ρ; λ_1 and λ_2 for $g_{\lambda_1,\lambda_2}(t)$; α for AE-LookSAM;
for $t = 0, \dots, T-1$ **do**
  sample a mini-batch data $\mathcal{B}_t$ from $\mathcal{D}$;
  compute $\mathbf{g} = \nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)$, update μ_t and σ_t^2 by EMA;
  compute $c_t = g_{\lambda_1,\lambda_2}(t)$;
  if $\|\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t)\|^2 \ge \mu_t + c_t\sigma_t$ **then**
    $\mathbf{g}_s = \nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t + \rho\nabla\mathcal{L}(\mathcal{B}_t;\mathbf{w}_t))$;
    if AE-LookSAM: decompose $\mathbf{g}_s$ as $\mathbf{g}_v = \mathbf{g}_s - \frac{\mathbf{g}^\top\mathbf{g}_s}{\|\mathbf{g}\|^2}\mathbf{g}$;
  else:
    if AE-SAM: $\mathbf{g}_s = \mathbf{g}$;
    if AE-LookSAM: $\mathbf{g}_s = \mathbf{g} + \alpha\frac{\|\mathbf{g}\|}{\|\mathbf{g}_v\|}\mathbf{g}_v$;
  end if
  $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{g}_s$;
end for
return $\mathbf{w}_T$.

Let $\mathcal{A}$ be an algorithm whose update in each iteration can be either SAM or ERM.
Theorem. Under smoothness and bounded variance assumptions, $\mathcal{A}$ satisfies

$$\min_{0\le t\le T-1} \mathbb{E}\|\nabla\mathcal{L}(\mathcal{D};\mathbf{w}_t)\|^2 \le \frac{32\beta\left(\mathcal{L}(\mathcal{D};\mathbf{w}_0) - \mathbb{E}\mathcal{L}(\mathcal{D};\mathbf{w}_T)\right)}{\sqrt{T}\,(7-6\zeta)} + \frac{(1+\zeta+5\beta^2\zeta)\sigma^2}{b\sqrt{T}\,(7-6\zeta)},$$

where $\zeta = \frac{1}{T}\sum_{t=0}^{T-1}\xi_t \in [0,1]$ is the fraction of SAM updates.
Remarks: (i) A larger ζ leads to a larger upper bound; (ii) $\zeta = 1$ recovers SAM.

Experiments on *CIFAR-10*, *CIFAR-100*, *ImageNet*

Table 1: Testing accuracy and fraction of SAM updates (%SAM).

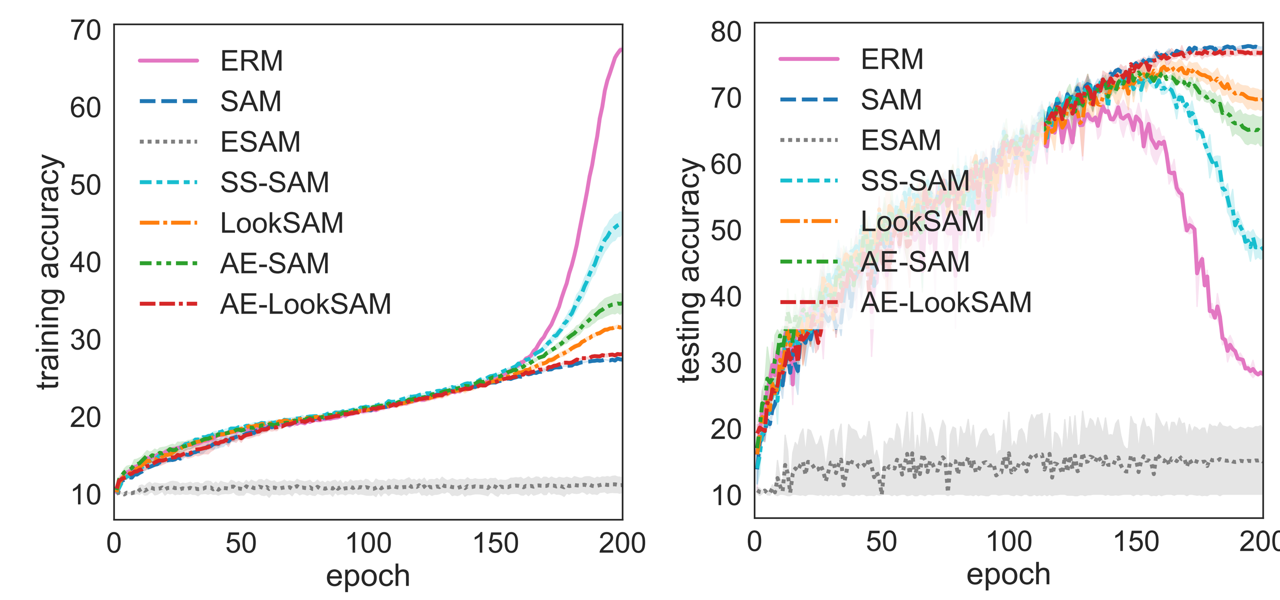
	CIFAR-10		*CIFAR-100*		*ImageNet*	
	Accuracy	%SAM	Accuracy	%SAM	Accuracy	%SAM
ERM	95.41 ±0.03	0.0 ±0.0	78.17 ±0.05	0.0 ±0.0	77.11 ±0.14	0.0 ±0.0
SAM	96.52 ±0.12	100.0 ±0.0	80.17 ±0.15	100.0 ±0.0	**77.47** ±0.12	100.0 ±0.0
ESAM	96.56 ±0.08	100.0 ±0.0	80.41 ±0.10	100.0 ±0.0	77.25 ±0.75	100.0 ±0.0
SS-SAM	96.40 ±0.16	50.0 ±0.0	80.10 ±0.16	50.0 ±0.0	77.38 ±0.06	50.0 ±0.0
AE-SAM	**96.63** ±0.04	50.1 ±0.1	**80.48** ±0.11	49.8 ±0.0	77.43 ±0.06	49.4 ±0.0
LookSAM	96.32 ±0.12	20.0 ±0.0	79.89 ±0.29	20.0 ±0.0	77.13 ±0.09	20.0 ±0.0
AE-LookSAM	96.56 ±0.21	20.0 ±0.1	80.29 ±0.37	20.0 ±0.0	77.29 ±0.08	20.3 ±0.0

- Using **only 50% of SAM updates**, AE-SAM performs better than SAM on *CIFAR-10* and *CIFAR-100*.
- The proposed adaptive policy is **more effective** than the random or periodic policy:
  - AE-SAM performs better than SS-SAM (with about 50% SAM);
  - AE-LookSAM is better than LookSAM (with about 20% SAM).
- Like SAM, AE-SAM has much smaller stochastic gradient norm and variance than ERM.

Experiments on *CIFAR-10* with Label Noise

Table 2: Testing accuracy and fraction of SAM updates on *CIFAR-10* with different levels of label noise.

	noise $=20\%$		noise $=40\%$		noise $=60\%$		noise $=80\%$	
	accuracy	%SAM	accuracy	%SAM	accuracy	%SAM	accuracy	%SAM
ERM	87.92	0.0	70.82	0.0	49.61	0.0	28.23	0.0
SAM	**94.80**	100.0	91.50	100.0	**88.15**	100.0	**77.40**	100.0
ESAM	94.19	100.0	91.46	100.0	81.30	100.0	15.00	100.0
SS-SAM	90.62	50.0	77.84	50.0	61.18	50.0	47.32	50.0
AE-SAM	92.84	50.0	84.17	50.0	73.54	49.9	65.00	50.0
LookSAM	92.72	50.0	88.04	50.0	72.26	50.0	69.72	50.0
AE-LookSAM	94.34	49.9	**91.58**	50.0	87.85	50.0	76.90	50.0



(a): Training accuracy. (b): Testing accuracy

Figure 3: Accuracies with #epochs on *CIFAR-10* (with 80% noise labels).

- AE-LookSAM achieves comparable performance with SAM but is **faster**, showing AE-LookSAM has **the same high level of robustness** as SAM.
- AE-LookSAM performs **better** than ESAM, SS-SAM, and LookSAM.

Summary

- Study the problem of improving SAM's efficiency
- Introduce **a sharpness measure**: squared stochastic gradient norm
- Design an **adaptive** policy: use SAM in sharp regions, while use ERM in flat regions
- Propose two **efficient** algorithms: AE-SAM and AE-LookSAM
- Results on *CIFAR-10*, *CIFAR-100*, and *ImageNet* show the **efficiency** and **effectiveness** of the adaptive policy
- Results on *CIFAR-10* with label noise show the **robustness** of AE-LookSAM

Reference

[1] Sharpness-aware minimization for efficiently improving generalization. ICLR 2021
[2] Efficient sharpness-aware minimization for improved training of neural networks. ICLR 2022
[3] Towards efficient and scalable sharpness-aware minimization. CVPR 2022
[4] SS-SAM: Stochastic scheduled sharpness-aware minimization for efficiently training deep neural networks. Preprint arXiv:2203.09962, 2022
[5] Fantastic generalization measures and where to find them, ICLR 2020
[6] Towards understanding sharpness-aware minimization, ICML 2022