香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Improving Sharpness-Aware Minimization by Lookahead

Runsheng Yu[1], Youzhi Zhang[2], James T. Kwok[1]

1 Department of Computer Science and Engineering, Hong Kong University of Science and Technology

2 Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, CAS.

Background

Sharpness-Aware Minimization (SAM)

Flat minima often imply better generalization(Chatterji et al., 2020; Jiang et al., 2020).

SAM (Foret et al., 2021) is designed to locate flat minima:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|\leq\rho} L(\boldsymbol{w}+\boldsymbol{\epsilon}),$$

where $\boldsymbol{w}$ is the model parameter, $\boldsymbol{\epsilon}$ is the perturbation whose magnitude is bounded ρ.

Background

Update Scheme for Sharpness-Aware Minimization (SAM)

To solve:

First-order approximation on the objective, the optimal $\boldsymbol{\epsilon}$ for the maximization sub-problem: $\boldsymbol{\epsilon}^*(\boldsymbol{w}) = \frac{\rho \nabla_{\boldsymbol{w}} L(\boldsymbol{w})}{\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\|}$.

Then, the update scheme is:

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1} + \boldsymbol{\epsilon}_{t-1}), \boldsymbol{\epsilon}_{t-1} = \boldsymbol{\epsilon}^* \left(\boldsymbol{w}_{t-1}\right)$$

However, SAM is easy to be trapped into saddle points (Kim et al., 2023; Compagnoni et al., 2023).

Background: Extra-Gradient and Optimistic Gradient

Consider the minimax problem: $\min_{x\in\mathbb{R}^m} \max_{y\in\mathbb{R}^n} f(x,y).$

$$\text{Let } z := [x,y]^\top \text{ and } F(z) := [\nabla_x f(x,y), -\nabla_y f(x,y)]^\top$$

Gradient Descent-Ascent (GDA) : $z_{t+1} = z_t - \eta_t F(z_t)$ Unstability (Gidel et al., 2019).

To ensure stability:

Extra-Gradient (Korpelevich, 1976):

$$\bar{z}_t = z_t - \eta_t F(z_t)$$
$$z_{t+1} = z_t - \eta_t F(\bar{z}_t)$$

Extra extrapolation step, looking one step ahead

Optimistic-Gradient (Popov, 1980):

$$\bar{z}_t = \bar{z}_{t-1} - \eta_t F(\bar{z}_{t-1})$$
$$z_{t+1} = z_t - \eta_t F(\bar{z}_t)$$

Reuse gradient at t-1, Reduce computation

Method

To address the issue that SAM is easy to be trapped saddle points,

Direct EG and OG to SAM objective: $\min_{\boldsymbol{w}\in\mathbb{R}^n} \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|\leq\rho} L(\boldsymbol{w}+\boldsymbol{\epsilon}),$

EG

$$
\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \Pi\left(\boldsymbol{\epsilon}_{t-1} + \nabla_{\boldsymbol{\epsilon}_{t-1}} L\left(\boldsymbol{w}_{t-1}+\boldsymbol{\epsilon}_{t-1}\right)\right) \\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\boldsymbol{w}_{t-1}} L\left(\boldsymbol{w}_{t-1}+\boldsymbol{\epsilon}_{t-1}\right), \\
\boldsymbol{\epsilon}_t &= \Pi\left(\boldsymbol{\epsilon}_{t-1} + \nabla_{\hat{\boldsymbol{\epsilon}}_t} L\left(\hat{\boldsymbol{w}}_t+\hat{\boldsymbol{\epsilon}}_t\right)\right) \\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L\left(\hat{\boldsymbol{w}}_t+\hat{\boldsymbol{\epsilon}}_t\right).
\end{aligned}
$$

OG

$$
\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \Pi\left(\boldsymbol{\epsilon}_{t-1} + \eta_t' \nabla_{\hat{\boldsymbol{\epsilon}}_{t-1}} L\left(\hat{\boldsymbol{w}}_{t-1}+\hat{\boldsymbol{\epsilon}}_{t-1}\right)\right), \\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L\left(\hat{\boldsymbol{w}}_{t-1}+\hat{\boldsymbol{\epsilon}}_{t-1}\right), \\
\boldsymbol{\epsilon}_t &= \Pi\left(\boldsymbol{\epsilon}_{t-1} + \nabla_{\hat{\boldsymbol{\epsilon}}_t} L\left(\hat{\boldsymbol{w}}_t+\hat{\boldsymbol{\epsilon}}_t\right)\right) \\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L\left(\hat{\boldsymbol{w}}_t+\hat{\boldsymbol{\epsilon}}_t\right).
\end{aligned}
$$

$\Pi(\cdot)$ is projection function.

However, directly using both EG and OG which converges at a $O(T^{-\frac{1}{4}})$, which is much slower than the $O(1/\sqrt{T})$ rate of SAM.

Method

A faster convergence by using approximated closed form solution for the max problem.

Lookahead-SAM:

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}^* \left(\boldsymbol{w}_{t-1}\right) \leftarrow \text{closed-form of } \boldsymbol{\epsilon} \text{ w.r.t. } \boldsymbol{w}.$$

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\boldsymbol{w}_{t-1}} L\left(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t\right) \leftarrow \text{Lookahead step.}$$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L\left(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t\right) \leftarrow \text{Update } w \text{ using the } \hat{\boldsymbol{w}}_t \text{ with pertubation.}$$

Intuition:

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L\left(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t\right)$$

Reduces the perturbation.

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\boldsymbol{w}_{t-1}} L\left(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t\right)$$

As larger perturbations is prone to being trapped in saddle points (Kim et al., 2023; Compagnoni et al., 2023).

Method

Lookahead-SAM

$$
\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \boldsymbol{\epsilon}^* \left(\boldsymbol{w}_{t-1}\right) \\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\boldsymbol{w}_{t-1}} L \left(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t\right) \\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L \left(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t\right)
\end{aligned}
$$

Optimistic Lookahead-SAM (Opt-SAM):
to further save computation time by reuse gradient

Reduces the perturbation.

$$
\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \boldsymbol{\epsilon}^* \left(\boldsymbol{w}_{t-1}\right) \\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\hat{\boldsymbol{w}}_{t-1}} L \left(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}\right), \\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L \left(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t\right)
\end{aligned}
$$

However, Opt-SAM still has to compute the gradient in each iteration, and can be expensive.

Background: AE-SAM (Jiang et al., 2023)

A SAM variant to reduce computation.

Empirically, $\|\nabla L(\boldsymbol{w}_t)\|^2 \sim \mathcal{N}(\mu_t, \sigma_t)$

$$\mu_t = \delta \mu_{t-1} + (1-\delta)\|\nabla L(\boldsymbol{w}_t)\|^2$$
$$\sigma_t^2 = \delta \sigma_{t-1}^2 + (1-\delta)\left(\|\nabla L(\boldsymbol{w}_t)\|^2 - \mu_t\right)^2$$

← Exponential Moving Average

$\|\nabla L(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t \sigma_t$ SAM is chosen; Otherwise ERM.

$c_t = \frac{t}{T}\kappa_1 + \left(1 - \frac{t}{T}\right)\kappa_2$, κ_1 and κ_2 are two constants.

Method: Adaptive Lookahead-SAM

Use AE-SAM technique to help reduce computation.

If $\|\frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t\sigma_t$ we use Opt-SAM. Otherwise, SGD (i.e., ERM) is used instead.

Algorithm 2: Adaptive Optimistic SAM (AO-SAM).

Input: Training set S, number of epochs T, batch size b, $\boldsymbol{w}_0$, $\epsilon_0 = 0$, $\mu_0 = 0$, and $\sigma_0 = e^{-10}$.

for $t = 1, 2, \ldots, T$ **do**
- sample a minibatch I_t from S with size b;
- $\mathbf{g}_t = \frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)$;
- update μ_t and σ_t as in AE-SAM (5);
- **if** $\|\frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t\sigma_t$ **then**
  - $\hat{\boldsymbol{\epsilon}}_t = \frac{\rho_t \nabla_{\boldsymbol{w}_{t-1}} \frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} \frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w}_{t-1})\|}$;
  - $\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta'_t\mathbf{g}_{t-1}$;
  - $\mathbf{g}_t = \nabla_{\hat{\boldsymbol{w}}_t}\left[\frac{1}{b}\sum_{i\in I_t}\ell_i(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t)\right]$;
- $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t\mathbf{g}_t$;

return $\boldsymbol{w}_T$.

Analysis: Region of attraction (ROA)

ROA (informal): The Region of Attraction is all the starting points from which the system will eventually settle into this stable state.

Objective: $\min_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{H} \boldsymbol{w}$

ODE SAM:

$$d\boldsymbol{w}_\tau = -\boldsymbol{H}\left(\boldsymbol{w}_\tau + \frac{\rho \boldsymbol{H}\boldsymbol{w}_\tau}{||\boldsymbol{H}\boldsymbol{w}_\tau||}\right) d\tau$$

ROA for ODE SAM:

$$\left\{ \boldsymbol{w}_\tau \,\middle|\, \rho \geq -\frac{\|\boldsymbol{H}\boldsymbol{w}_\tau\|}{\lambda_{\min}} \right\}$$

$$d\mathbf{w}_\tau = -\boldsymbol{H}\left(\mathbf{w}_\tau - \eta'_\tau \boldsymbol{H}\left(\mathbf{w}_\tau + \frac{\rho \boldsymbol{H}\mathbf{w}_\tau}{\|\boldsymbol{H}\mathbf{w}_\tau\|}\right) + \frac{\rho \boldsymbol{H}\mathbf{w}_\tau}{\|\boldsymbol{H}\mathbf{w}_\tau\|}\right) d\tau$$

ROA for ODE Lookahead-SAM:

$$\left\{ \boldsymbol{w}_\tau \,\middle|\, (1+\eta'_\tau \lambda_{\min})\,\rho \geq \frac{1}{\lambda_{\min}} \|\boldsymbol{H}\boldsymbol{w}_\tau\| \,(\eta'_\tau \lambda_{\min} - 1) \right\}$$

$\lambda_{\min}$ is the minimum eigenvalue of $\boldsymbol{H}$.

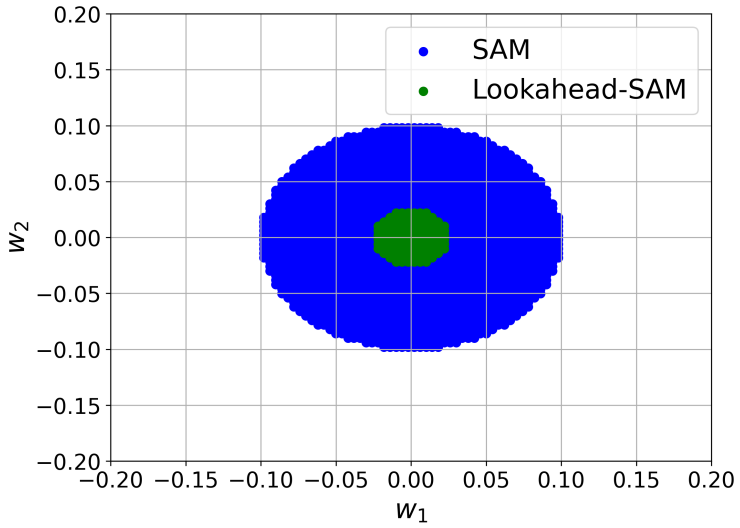
Analysis: Region of attraction (ROA)

ROA for ODE SAM:

$$\left\{ \boldsymbol{w}_\tau \,\middle|\, \rho \geq -\frac{\|\boldsymbol{H}\boldsymbol{w}_\tau\|}{\lambda_{\min}} \right\}$$

ROA for ODE Lookahead-SAM:

$$\left\{ \boldsymbol{w}_\tau \,\middle|\, (1+\eta'_\tau \lambda_{\min})\,\rho \geq \frac{1}{\lambda_{\min}} \|\boldsymbol{H}\boldsymbol{w}_\tau\| \,(\eta'_\tau \lambda_{\min} - 1) \right\}$$



ROAs for SAM and Lookahead-SAM at saddle point

Analysis: Convergence Analysis

Theorem (Informal): under mild assumptions, both Lookead-SAM, Opt-SAM, and AO-SAM satisfies

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\boldsymbol{w}_t} L(\boldsymbol{w}_t)\|^2 = O\left(\frac{1}{\sqrt{Tb}}\right)$$

.

Lookahead-SAM, Opt-SAM and AO-SAM have the same $O(\frac{1}{\sqrt{Tb}})$ convergence rate as SAM and its variant AESAM.

Experiments: Comparison with EG,OG and the proposed methods

Table 1: Testing accuracy and fraction of SAM updates (%SAM) on *CIFAR-10* using *ResNet-18*. The best accuracy is in bold.

	accuracy	%SAM
SAM	96.52 ±0.12	100.0 ±0.0
EG	96.45 ±0.05	200.0 ±0.0
OG	96.52 ±0.03	100.0 ±0.0
Lookahead-SAM	96.81 ±0.01	150.0 ±0.0
Opt-SAM	96.79 ±0.02	100.0 ±0.0
AO-SAM	**96.82** ±0.04	61.1 ±0.0

Table 2: Testing accuracy and fraction of SAM updates (%SAM) on *CIFAR-100* using *ResNet-18*. The best accuracy is in bold.

	accuracy	%SAM
SAM	80.17 ±0.05	100.0 ±0.0
EG	79.91 ±0.16	200.0 ±0.0
OG	79.92 ±0.08	100.0 ±0.0
Lookahead-SAM	**80.79** ±0.13	150.0 ±0.0
Opt-SAM	80.76 ±0.15	100.0 ±0.0
AO-SAM	80.70 ±0.14	61.2 ±0.0

Lookahead-SAM has the highest accuracy on CIFAR-100 and the second highest on CIFAR-10, it also has a higher %SAM.

Opt-SAM is as fast as SAM w.r.t. %SAM but is more accurate.

AO-SAM is as accurate as Opt-SAM, but is even faster.

Experiments: Comparison with SAM variants in CIFAR

Table 3: Testing accuracies (mean and standard deviation) and fractions of SAM updates on *CIFAR-10* and *CIFAR-100*. Methods with similar %SAM's are grouped together for easier comparison. Results of ERM, SAM, and ESAM are from (Jiang et al., 2023), while the other baseline results are obtained with the corresponding authors' codes. The best accuracy is in bold. $*$ means the improvements over the second-best baseline are statistically significant (achieving a p-value of less than 0.05 in t-test).

		CIFAR-10		*CIFAR-100*	
		Accuracy	% SAM	Accuracy	% SAM
ResNet-18	ERM	95.41 ±0.03	0.0 ±0.0	78.17 ±0.05	0.0 ±0.0
	SAM (Foret et al., 2021)	96.52 ±0.12	100.0 ±0.0	80.17 ±0.15	100.0 ±0.0
	ESAM (Du et al., 2022a)	96.56 ±0.08	100.0 ±0.0	80.41 ±0.10	100.0 ±0.0
	ASAM (Kwon et al., 2021)	96.55 ±0.14	100.0 ±0.0	80.52 ±0.13	100.0 ±0.0
	GSAM (Zhuang et al., 2022)	96.70 ± 0.01	100.0 ±0.0	80.48 ± 0.11	100.0 ±0.0
	Opt-SAM	96.79 ±0.02	100.0 ±0.0	**80.76*** ±0.15	100.0 ±0.0
	SS-SAM (Zhao, 2022)	96.64 ±0.02	60.0 ±0.0	80.49 ±0.10	60.0 ±0.0
	AE-SAM (Jiang et al., 2023)	96.66 ±0.02	61.3 ±0.1	79.96 ±0.08	61.3 ±0.0
	AO-SAM	**96.82*** ±0.04	61.1 ±0.0	80.70 ±0.14	61.2 ±0.0
WideResNet-28-10	ERM	96.34 ±0.12	0.0 ±0.0	81.56 ±0.14	0.0 ±0.0
	SAM (Foret et al., 2021)	97.27 ±0.11	100.0 ±0.0	83.42 ±0.05	100.0 ±0.0
	ESAM (Du et al., 2022a)	97.29 ±0.11	100.0 ±0.0	84.51 ±0.02	100.0 ±0.0
	ASAM (Kwon et al., 2021)	97.38 ±0.09	100.0 ±0.0	84.48 ±0.10	100.0 ±0.0
	GSAM (Zhuang et al., 2022)	97.44 ± 0.07	100.0 ±0.0	84.50 ± 0.12	100.0 ±0.0
	Opt-SAM	**97.56*** ±0.03	100.0 ±0.0	84.74 ±0.02	100.0 ±0.0
	SS-SAM (Zhao, 2022)	97.32 ±0.03	60.0 ±0.0	84.39 ±0.04	60.0 ±0.0
	AE-SAM (Jiang et al., 2023)	97.37 ±0.08	61.3 ±0.0	84.23 ±0.08	61.3 ±0.0
	AO-SAM	97.49 ±0.02	61.2 ±0.0	**84.80*** ±0.11	61.2 ±0.0

Opt-SAM
and AO-SAM are consistently more accurate than SAM and its variants on all datasets and backbones

Experiments: Comparison with SAM variants in CIFAR

		CIFAR-10		*CIFAR-100*	
		Accuracy	% SAM	Accuracy	% SAM
PyramidNet-110	ERM	96.62 ±0.10	0.0 ±0.0	81.89 ±0.15	0.0 ±0.0
	SAM (Foret et al., 2021)	97.30 ±0.10	100.0 ±0.0	84.46 ±0.05	100.0 ±0.0
	ESAM (Du et al., 2022a)	97.81 ±0.01	100.0 ±0.0	85.56 ±0.05	100.0 ±0.0
	ASAM (Kwon et al., 2021)	97.71 ±0.09	100.0 ±0.0	85.55 ±0.11	100.0 ±0.0
	GSAM (Zhuang et al., 2022)	97.74 ± 0.02	100.0 ±0.0	85.25 ± 0.11	100.0 ±0.0
	Opt-SAM	97.79 ±0.04	100.0 ±0.0	**85.74*** ±0.14	100.0 ±0.0
	SS-SAM (Zhao, 2022)	97.62 ±0.03	60.0 ±0.0	85.41 ±0.11	60.0 ±0.0
	AE-SAM (Jiang et al., 2023)	97.52 ±0.07	61.4 ±0.1	85.43 ±0.08	61.4 ±0.1
	AO-SAM	**97.87*** ±0.02	61.2 ±0.0	85.60 ±0.07	61.2 ±0.12
ViT-S16	ERM	86.69 ±0.11	0.0 ±0.0	62.42 ±0.22	0.0 ±0.0
	SAM (Foret et al., 2021)	87.37 ±0.09	100.0 ±0.0	63.23 ±0.25	100.0 ±0.0
	ESAM (Du et al., 2022a)	84.27 ±0.11	100.0 ±0.0	62.11 ±0.15	100.0 ±0.0
	ASAM (Kwon et al., 2021)	82.25 ±0.41	100.0 ±0.0	63.26 ±0.18	100.0 ±0.0
	GSAM (Zhuang et al., 2022)	83.62 ± 0.11	100.0 ±0.0	59.82 ± 0.12	100.0 ±0.0
	Opt-SAM	87.91 ±0.26	100.0 ±0.0	63.78 ±0.22	100.0 ±0.0
	SS-SAM (Zhao, 2022)	83.36 ± 0.04	60.0 ±0.0	54.04 ±5.09	60.0 ±0.0
	AE-SAM (Jiang et al., 2023)	77.37 ±0.07	61.4 ±0.0	57.13 ±2.87	61.3 ±0.0
	AO-SAM	**88.27*** ±0.12	61.3 ±0.0	**64.45*** ±0.23	61.2 ±0.0

Opt-SAM
and AO-SAM are consistently more accurate than SAM and its variants on all datasets and backbones

Experiments: Comparison with SAM variants in CIFAR with noise

Table 8: Testing accuracies and fractions of SAM updates on *CIFAR-10* with different levels of label noise. Results of ERM, SAM, and ESAM with *ResNet-18* and *ResNet-32* are from (Jiang et al., 2023) (standard derivations for some baselines are not provided in (Jiang et al., 2023)), while the other baseline results are obtained with the authors' codes. The best accuracy is in bold.

		noise = 20%		noise = 40%		noise = 60%		noise = 80%	
		accuracy	%SAM	accuracy	%SAM	accuracy	%SAM	accuracy	%SAM
ResNet-18	ERM	87.92	0.0	70.82	0.0	49.61	0.0	28.23	0.0
	SAM (Foret et al., 2021)	94.80	100.0	91.50	100.0	88.15	100.0	77.40	100.0
	ESAM (Du et al., 2022a)	94.19	100.0	91.46	100.0	81.30	100.0	15.00	100.0
	ASAM (Kwon et al., 2021)	91.17 ± 0.19	100.0	87.38 ± 0.61	100.0	83.22 ± 0.41	100.0	71.03 ± 0.88	100.0
	GSAM (Zhuang et al., 2022)	94.54 ± 0.18	100.0	91.72 ± 0.05	100.0	87.70 ± 0.02	100.0	24.70 ± 10.69	100.0
	Opt-SAM	**95.12** ± 0.12	100.0	92.16 ± 0.35	100.0	88.45 ± 0.53	100.0	77.47 ± 0.65	100.0
	SS-SAM (Zhao, 2022)	94.61 ± 0.16	60.0	91.81 ± 0.13	60.0	78.67 ± 0.42	60.0	62.94 ± 1.01	60.0
	AE-SAM (Jiang et al., 2023)	92.13 ± 0.14	61.4	86.02 ± 0.62	61.4	75.95 ± 1.30	61.4	67.28 ± 1.66	61.4
	AO-SAM	95.02 ± 0.04	61.2	**92.62** ± 0.18	61.3	**89.36** ± 0.12	61.2	**78.12** ± 0.38	61.2
ResNet-32	ERM	87.43	0.0	70.82	0.0	46.26	0.0	29.00	0.0
	SAM (Foret et al., 2021)	95.08	100.0	91.01	100.0	88.90	100.0	77.32	100.0
	ESAM (Du et al., 2022a)	93.42	100.0	91.63	100.0	82.73	100.0	10.09	100.0
	ASAM (Kwon et al., 2021)	92.04 ± 0.09	100.0	88.83 ± 0.11	100.0	83.90 ± 0.56	100.0	75.64 ± 0.75	100.0
	GSAM (Zhuang et al., 2022)	94.12 ± 0.09	100.0	91.74 ± 0.05	100.0	89.23 ± 0.06	100.0	31.16 ± 2.77	100.0
	Opt-SAM	95.25 ± 0.04	100.0	**92.11** ± 0.07	100.0	88.36 ± 0.22	100.0	77.61 ± 0.39	100.0
	SS-SAM (Zhao, 2022)	95.03 ± 0.23	60.0	90.59 ± 0.30	60.0	87.22 ± 0.46	60.0	48.89 ± 1.02	60.0
	AE-SAM (Jiang et al., 2023)	92.04 ± 0.27	61.3	86.83 ± 0.49	61.3	73.90 ± 0.44	61.2	67.64 ± 1.34	61.3
	AO-SAM	**95.32** ± 0.12	61.2	91.73 ± 0.65	61.2	**89.40** ± 0.44	61.2	**77.78** ± 0.84	61.2

AO-SAM and Opt-SAM outperform all baselines at all label noise ratios.

Experiments: Comparison with SAM variants in ImageNet

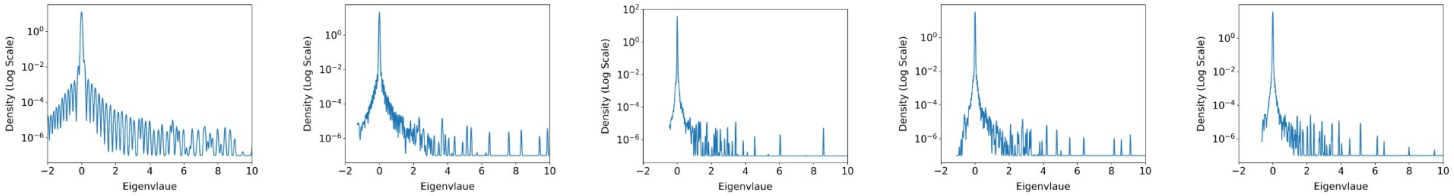
Table 7: Testing accuracies and fractions of SAM updates (%SAM) on *ImageNet*. Results of ERM, SAM and ESAM on *ResNet-50* are from (Jiang et al., 2023), ASAM is from (Kwon et al., 2021), GSAM is from (Zhuang et al., 2022), while the other baseline results are obtained by the corresponding authors' codes. The best accuracy is in bold. † means that the original papers do not provide standard deviation. We do not report ASAM on *ResNet-101* and *Vit-S/32*, and GSAM on *Vit-S/32* because they are not provided in the original papers.

		Accuracy	%SAM
ResNet-50	ERM	77.11 ± 0.14	0.0
	SAM	77.47 ± 0.12	100.0
	ESAM	77.25 ± 0.75	100.0
	ASAM	76.63 ± 0.18	100.0
	GSAM	$77.2^{\dagger}$	100.0
	AO-SAM	$\mathbf{77.68 \pm 0.04}$	61.1

AO-SAM again outperforms all the baselines.

		Accuracy	%SAM
ResNet-101	ERM	$77.80^{\dagger}$	0.0
	SAM	$78.90^{\dagger}$	100.0
	ESAM	$79.09^{\dagger}$	100.0
	GSAM	$78.9^{\dagger}$	100.0
	AO-SAM	$\mathbf{79.38 \pm 0.10}$	61.2
ViT-S/32	ERM	$67.0^{\dagger}$	0.0
	SAM	$69.1^{\dagger}$	100.0
	ESAM	$66.1^{\dagger}$	100.0
	AO-SAM	$\mathbf{69.38 \pm 0.24}$	61.6

Experiments: Flat Minima



(a) ERM. (b) SAM. (c) Lookahead-SAM. (d) Opt-SAM. (e) AO-SAM.

Figure 4: Hessian spectra obtained by ERM, SAM, Lookahead-SAM, Opt-SAM, and AO-SAM on *CIFAR-10* with *ResNet18*.

Table 5: Eigenvalues of the Hessian on *CIFAR-10* with *ResNet18* backbone. The smallest is in bold.

	λ_1	λ_1/λ_5
ERM	88.8	3.3
SAM	29.6	3.3
Lookahead-SAM	**10.2**	**1.8**
Opt-SAM	13.1	2.0
AO-SAM	11.1	**1.8**

The eigenvalues of Lookahead-SAM, Opt-SAM, and AO-SAM are smaller than ERM and SAM.

This indicates the loss landscapes at the converged solutions of these SAM variants are flatter.

Conclusion

1. Incorporate the idea of extrapolation into SAM to gain more information about the landscape, and thus help convergence.

2. Develop a method that combines SAM's approximate maximizer to its inner optimization subproblem with lookahead.

3. Provide theoretical guarantees that they converge to stationary points at the same rate as SAM, and are not easily trapped at saddle points..