Improving Sharpness-Aware Minimization by Lookahead

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

Abstract

Sharpness-Aware Minimization (SAM), which performs gradient descent on adversarially perturbed weights, can improve generalization by identifying flatter minima. However, recent studies have shown that SAM may suffer from convergence instability and oscillate around saddle points, resulting in slow convergence and inferior performance. To address this problem, we propose the use of a lookahead mechanism to gather more information about the landscape by looking further ahead, and thus find a better trajectory to converge. By examining the nature of SAM, we simplify the extrapolation procedure, resulting in a more efficient algorithm. Theoretical results show that the proposed method converges to a stationary point and is less prone to saddle points. Experiments on standard benchmark datasets also verify that the proposed method outperforms the SOTAs, and converge more effectively to flat minima.

1. Introduction

Deep learning models have demonstrated remarkable success in various real-world applications (LeCun et al., 2015). However, highly over-parameterized neural networks may suffer from overfitting and poor generalization (Zhang et al., 2021). Hence, reducing the performance gap between training and testing is an important research topic (Neyshabur et al., 2017). Recently, there have been a number of works exploring the close relationship between loss geometry and generalization performance. In particular, it has been observed that flat minima often imply better generalization (Chatterji et al., 2020; Jiang et al., 2020; Chaudhari et al., 2019; Dziugaite & Roy, 2017; Petzka et al., 2021).

To locate flat minima, a popular approach is based on Sharpness-Aware Minimization (SAM) (Foret et al., 2021). Recently, a number of variants have also been proposed (Kwon et al., 2021; Zhuang et al., 2022; Du et al., 2022b;a; Jiang et al., 2023; Liu et al., 2022). Their main idea is to first add a (adversarial) perturbation to the weights and then perform gradient descent there. However, these methods are myopic as they only update their parameters based on the gradient of the adversarially perturbed parameters. Consequently, the model may converge slowly as it lacks good information about the loss landscape. In particular, recent research has found that SAM can suffer from convergence instability and may be easily trapped in a saddle point (Kim et al., 2023; Compagnoni et al., 2023; Kaddour et al., 2022; Tan et al., 2024).

To mitigate this problem, one possibility is to encourage the model to gather more information about the landscape by looking further ahead, and thus find a better trajectory to converge (Leng et al., 2018; Wang et al., 2022). In game theory, two popular methods that can encourage the agent to look ahead are the method of extra-gradient (Korpelevich, 1976; Gidel et al., 2019; Lee et al., 2021) and its approximate cousin, the method of optimistic gradient (Popov, 1980; Gidel et al., 2019; Daskalakis & Panageas, 2018; Daskalakis et al., 2018; Mokhtari et al., 2020). Their key idea is to first perform an extrapolation step that looks one step ahead into the future, and then perform gradient descent based on the extrapolation result (Bohm et al., 2022). Besides game theory, similar ideas have also been proven successful in deep learning optimization (Zhou et al., 2021; Zhang et al., 2019; Lin et al., 2020b), and reinforcement learning (Liu et al., 2023). As SAM is formulated as a minimax optimization problem (Foret et al., 2021), this also inspires us to leverage an extrapolation step for better convergence.

In this paper, we introduce the look-ahead mechanism to SAM. Our main contributions are fourfold:

(i) We incorporate the idea of extrapolation into SAM so that the model can gain more information about the landscape, and thus help convergence. We also discuss a concrete example on how extrapolation reduces the perturbation and thus helps escape saddle point.

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology [2]Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, CAS. Correspondence to: Runsheng Yu <runshengyu@gmail.com>, Youzhi Zhang <youzhi.zhang@cair-cas.org.hk>, James T. Kwok <jamesk@cse.ust.hk>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

(ii) By studying the SAM update scheme, we develop a method that combines SAM's approximate maximizer to its inner optimization subproblem with lookahead. We further propose a method that reduces the computational cost by removing some steps from a straightforward application of extra-gradient or optimistic gradient ascent.

(iii) We provide theoretical guarantees that they converge to stationary points at the same rate as SAM, and are not easily trapped at saddle points.

(iv) Experimental results show that the proposed method has better performance and converge to a flatter minimum.

2. Background

2.1. Sharpness-Aware Minimization (SAM)

SAM (Foret et al., 2021) attempts to improve generalization by finding flat minima. This is achieved by minimizing the worst-case loss within some perturbation radius. Mathematically, it is formulated as the following minimax optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^n} \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|\leq\rho} L(\boldsymbol{w}+\boldsymbol{\epsilon}), \tag{1}$$

where L is the loss, $\boldsymbol{w}$ is the model parameter, and $\boldsymbol{\epsilon}$ is the perturbation whose magnitude is bounded by ρ. The loss on the ith sample is denoted $\ell_i(\boldsymbol{w}_t)$. By taking first-order approximation on the objective, the optimal $\boldsymbol{\epsilon}$ for the maximization sub-problem can be obtained as:

$$\boldsymbol{\epsilon}^*(\boldsymbol{w}) = \frac{\rho\nabla_{\boldsymbol{w}} L(\boldsymbol{w})}{\|\nabla_{\boldsymbol{w}} L(\boldsymbol{w})\|}. \tag{2}$$

Problem (1) is then solved by performing gradient descent (with learning rate η) at each iteration t:

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1}+\boldsymbol{\epsilon}_{t-1}), \tag{3}$$

where

$$\boldsymbol{\epsilon}_{t-1} = \boldsymbol{\epsilon}^*(\boldsymbol{w}_{t-1}). \tag{4}$$

As SAM requires two forward-backward calculations in each iteration, it is computationally more expensive than standard Empirical Risk Minimization (ERM). Recently, a number of variants (e.g., AE-SAM (Jiang et al., 2023), LookSAM (Liu et al., 2022), SS-SAM (Zhao, 2022), ESAM (Du et al., 2022a), GSAM (Zhuang et al., 2022)) have been proposed to reduce this cost by using SAM only in iterations that it is likely to be useful and use ERM otherwise. For example, Jiang et al. (2023) proposes the AE-SAM, which uses SAM only when the loss landscape is locally sharp, with sharpness being approximated efficiently by $\|\nabla L(\boldsymbol{w}_t)\|^2$. It is shown that $\|\nabla L(\boldsymbol{w}_t)\|^2$ can be modeled empirically with a normal distribution $\mathcal{N}(\mu_t, \sigma_t^2)$, in which μ_t and σ_t^2 are estimated in an online manner by exponential moving average as:

$$\begin{aligned}\mu_t &= \delta\mu_{t-1} + (1-\delta)\|\nabla L(\boldsymbol{w}_t)\|^2,\\ \sigma_t^2 &= \delta\sigma_{t-1}^2 + (1-\delta)(\|\nabla L(\boldsymbol{w}_t)\|^2 - \mu_t)^2,\end{aligned} \tag{5}$$

where $\delta\in(0,1)$ controls the forgetting rate. When the loss landscape is locally sharp (i.e., $\|\nabla L(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t\sigma_t$), SAM is used; otherwise, ERM is used. The threshold c_t is decreased linearly from κ_2 to κ_1 according to the schedule:

$$c_t = \frac{t}{T}\kappa_1 + \left(1-\frac{t}{T}\right)\kappa_2, \tag{6}$$

where T is the total number of iterations.

Besides reducing the training cost, there are recent attempts on improving the generalization performance of SAM. For example, ASAM (Kwon et al., 2021) introduces adaptive sharpness, and GSAM (Zhuang et al., 2022) uses a new surrogate loss that focuses more on sharpness.

2.2. Extra-Gradient (EG)

Consider the minimax problem:

$$\min_{x\in\mathbb{R}^m} \max_{y\in\mathbb{R}^n} f(x,y). \tag{7}$$

The method of extra-gradient (EG) (Korpelevich, 1976) performs gradient descent-ascent (GDA), i.e., gradient ascent $\nabla_y f(x,y)$ on y and gradient descent $-\nabla_x f(x,y)$ on x. Specifically, let $z := [x,y]^\top$ and $F(z) := [\nabla_x f(x,y), -\nabla_y f(x,y)]^\top$. At the tth iteration, the EG update can be written as:

$$\overline{z}_t = z_t - \eta_t F(z_t), \tag{8}$$

$$z_{t+1} = z_t - \eta_t F(\overline{z}_t), \tag{9}$$

where η_t is the learning rate at epoch t. Note that (8) is an extra extrapolation step, which avoids the shortsightedness of both players (x and y) by looking one step ahead into the future (Gidel et al., 2019; Bohm et al., 2022; Jelassi et al., 2020; Pethick et al., 2022). EG has been widely used in two-player zero-sum games (Fudenberg & Tirole, 1991). In machine learning, this has been used in the training of generative adversarial networks (Gidel et al., 2019) and poker games (Lee et al., 2021).

As shown in (8) and (9), each EG iteration requires computing the gradients w.r.t. x and y twice. To reduce the cost, the method of optimistic gradient (OG) (Popov, 1980) stores the past gradient $F(\overline{z}_{t-1})$ and reuses it in the next extrapolation step. The update in $\overline{z}_t$ is thus changed to:

$$\overline{z}_t = z_t - \eta F(\overline{z}_{t-1}). \tag{10}$$

Hence, the gradients w.r.t. x and y only need to be computed once in each iteration. It can be shown that OG enjoys a similar convergence rate as EG (Gidel et al., 2019), and has been commonly used in solving differentiable minimax games (Gidel et al., 2019; Liang & Stokes, 2019; Daskalakis & Panageas, 2018; Daskalakis et al., 2018).

3. Lookahead in SAM

Recently, it is observed that SAM can have convergence instability near a saddle point (Kim et al., 2023; Compagnoni et al., 2023), leading to slow convergence and poor performance. As an illustration, consider minimizing the following quadratic objective as in (Compagnoni et al., 2023):

$$\min_{\boldsymbol{w}\in\mathbb{R}^2} \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w}, \tag{11}$$

where $\boldsymbol{H} \equiv \text{diag}(-1, 1)$. The saddle point is at $[0, 0]$. We run SAM with an initial $\boldsymbol{w}_0 = [0.03, 0.03]$, and SGD optimizer with a learning rate of 0.005. In every SGD step t, we add Gaussian noise from $\mathcal{N}(0, 0.01)$ to the gradient as in (Compagnoni et al., 2023). Figure 1a shows the trajectory, and Figure 1b shows the objective with number of epochs. As can be seen, SAM is trapped in the saddle point.

Inspired by the method of extra-gradient (EG), we propose in the following a number of lookahead mechanisms to alleviate the convergence instability problem of SAM.

3.1. Direct Adaptation of EG and Its Variant on SAM

First, we consider the direct adaptation of EG on SAM's minimax problem (1), which leads to the following update:

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \Pi(\boldsymbol{\epsilon}_{t-1} + \nabla_{\boldsymbol{\epsilon}_{t-1}} L(\boldsymbol{w}_{t-1} + \boldsymbol{\epsilon}_{t-1})), && (12)\\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1} + \boldsymbol{\epsilon}_{t-1}), && (13)\\
\boldsymbol{\epsilon}_t &= \Pi(\boldsymbol{\epsilon}_{t-1} + \nabla_{\hat{\boldsymbol{\epsilon}}_t} L(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t)) && (14)\\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t). && (15)
\end{aligned}$$

Here, $\Pi(\boldsymbol{\epsilon})$ is the projection $\arg\min_{\boldsymbol{\epsilon}':\|\boldsymbol{\epsilon}'\|\le\rho} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\| = \frac{\boldsymbol{\epsilon}}{\max(1, \|\boldsymbol{\epsilon}\|/\rho)}$, and η'_t is a learning rate. Note that the learning rates in (12) and (14) are set to 1, as is commonly used in SAM and its variants.

As can be seen, the update requires four gradient computations. This can be expensive, particularly for large deep networks. By using the optimistic gradient (OG) approach (Popov, 1980) (Section 2.2), we replace (12) and (13) by:

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \Pi(\boldsymbol{\epsilon}_{t-1} + \eta'_t \nabla_{\hat{\boldsymbol{\epsilon}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})), && (16)\\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}), && (17)
\end{aligned}$$

respectively. Since $\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ and $\nabla_{\hat{\boldsymbol{\epsilon}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ have already been computed at iteration $t-1$, we only need to compute the gradient w.r.t. $\boldsymbol{w}$ and $\boldsymbol{\epsilon}$ once in every epoch.

However, both EG and OG perform gradient descent ascent (GDA) on (1), which converges at a $O(T^{-\frac{1}{4}})$ rate (where T is the number of epochs) on non-convex strongly-concave problems (Lin et al., 2020a; Mahdavinia et al., 2022) and even slower on non-convex non-concave problems (Mahdavinia et al., 2022). This is much slower than the $O(1/\sqrt{T})$ rate of SAM (Andriushchenko & Flammarion, 2022).

3.2. Lookahead-SAM and Its Variants

3.2.1. Lookahead-SAM

Recall that the maximization subproblem in (1) has an approximated solution (2). We can directly use this approximate maximizer instead of performing gradient descent in (12) and (14), leading to:

$$\hat{\boldsymbol{\epsilon}}_t = \boldsymbol{\epsilon}^*(\boldsymbol{w}_{t-1}) \text{ and } \boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}^*(\hat{\boldsymbol{w}}_t). \tag{18}$$

To further reduce gradient computation, we remove the update of $\boldsymbol{\epsilon}_t$ in (14) and replace $\boldsymbol{\epsilon}_{t-1}$ in (13) by $\hat{\boldsymbol{\epsilon}}_t$ in (18). The whole update rule is then:[1]

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}_t &= \boldsymbol{\epsilon}^*(\boldsymbol{w}_{t-1}), && (19)\\
\hat{\boldsymbol{w}}_t &= \boldsymbol{w}_{t-1} - \eta'_t \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t), && (20)\\
\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t). && (21)
\end{aligned}$$

In other words, we first compute the perturbation $\hat{\boldsymbol{\epsilon}}_t$ in (19), take a lookahead step $\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t$ in (20), and then update $\boldsymbol{w}_{t-1}$ using $L(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t)$ via (21). This procedure, which will be called Lookahead-SAM, is shown in Algorithm 1. In Sections 4.1 and 4.2, we will show theoretically that it can better avoid saddle points than SAM while having the same convergence rate.

Note that the update (19)-(21) can also be rewritten as:

$$\boldsymbol{w_t}=\boldsymbol{w}_{t-1}-\eta_t \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1}-\eta'_t \perp[\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1}+\hat{\boldsymbol{\epsilon}}_t)]+\hat{\boldsymbol{\epsilon}}_t),$$

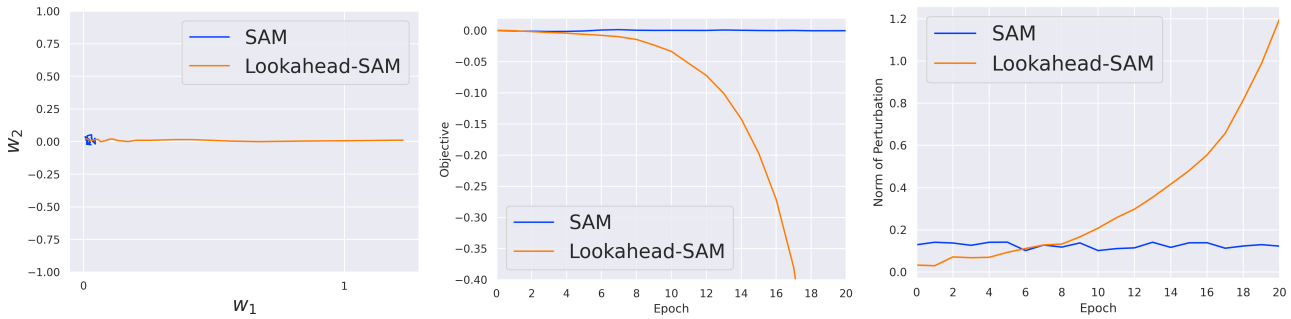
where $\perp$ is the stop-gradient operator[2] Compared to SAM, it reduces the perturbation from $\hat{\boldsymbol{\epsilon}}_t$ to$\hat{\boldsymbol{\epsilon}}_t - \eta'_t \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t)$. Intuitively, this is desirable as larger perturbations can make the model more prone to being trapped in saddle points (Kim et al., 2023; Compagnoni et al., 2023). Figures 1a and 1b empirically demonstrate this on the toy problem in (11). Figure 1c shows the norms of the perturbations. As can be seen, when the iterate is close to the saddle point (before epoch 7), the perturbations from Lookahead-SAM are smaller than those from SAM.

3.2.2. Optimistic Lookahead-SAM

Similar to EG, Lookahead-SAM has to compute the gradient w.r.t. $\boldsymbol{w}$ twice in each iteration, which can be expen-

[1] Note that we also allow the learning rates in (20) and (21) to be different.

[2] In other words, $\nabla_x \perp[g(x)] \equiv 0$ for any differentiable g.



(a) Trajectory. (b) Objective. (c) Norm of perturbation.

Figure 1: Example showing that SAM can be trapped in a saddle point.

Algorithm 1: Lookahead SAM and Optimistic Lookahead-SAM.

Input: Training set S, number of epochs T, batch size b, $\boldsymbol{w}_0$, $\epsilon_0 = 0$.

Sample a minibatch I_t from S with size b;
for $t = 1, 2, \ldots, T$ **do**
  $\hat{\boldsymbol{\epsilon}}_t = \frac{\rho_t \nabla_{\boldsymbol{w}_{t-1}} \frac{1}{b}\sum_{i\in I_t} \ell_i(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} \frac{1}{b}\sum_{i\in I_t} \ell_i(\boldsymbol{w}_{t-1})\|}$;
  if *Lookahead SAM* **then**
    $\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\boldsymbol{w}_{t-1}} \left[\frac{1}{b}\sum_{i\in I_t} \ell_i(\boldsymbol{w}_{t-1} + \hat{\boldsymbol{\epsilon}}_t)\right]$;
  else if *Optimistic Lookahead-SAM* **then**
    $\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \mathbf{g}_{t-1}$;
  $\mathbf{g}_t = \nabla_{\hat{\boldsymbol{w}}_t} \left[\frac{1}{b}\sum_{i\in I_t} \ell_i(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t)\right]$;
  $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t \mathbf{g}_t$;
return $\boldsymbol{w}_T$.

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 $\boldsymbol{w}_T$.

sive for large deep networks. Following the idea of optimistic gradient (Popov, 1980), we reuse the past gradient $\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ in (20), which then becomes:

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}), \qquad (22)$$

and the gradient is only computed once. Note that $\hat{\boldsymbol{w}}_t$ in (22) is also equal to $\arg\min_{\boldsymbol{w}} \left\{\eta_t'(\boldsymbol{w} \cdot \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})) + \frac{\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|_2^2}{2}\right\}$. Hence, update (22) can be interpreted as optimistic mirror descent (Chiang et al., 2012; Wei et al., 2021), which improves performance by leveraging information from the past gradient (Rakhlin & Sridharan, 2013), and has been widely used in online learning (Chiang et al., 2012) and game theory (Wei et al., 2021). The procedure (again based on stochastic gradient), called Optimistic Lookahead-SAM (Opt-SAM), and is also shown in Algorithm 1.

3.2.3. Adaptive Lookahead-SAM

Opt-SAM still has to compute the gradient in each iteration, and can be expensive. To alleviate this issue, we further integrate the adaptive policy in AE-SAM (Jiang et al., 2023) with Opt-SAM. Assume that $\|\frac{1}{b}\sum_{i\in I_t} \nabla_{\boldsymbol{w}_t} \ell_i(\boldsymbol{w}_t)\|^2$ follows the normal distribution $\mathcal{N}(\mu_t, \sigma_t^2)$ with mean μ_t and variance σ_t^2. If $\|\frac{1}{b}\sum_{i\in I_t} \nabla_{\boldsymbol{w}_t} \ell_i(\boldsymbol{w}_t)\|^2$ is large (i.e., $\geq \mu_t + c_t\sigma_t$, where c_t is varied as in (6)), we use Opt-SAM. Otherwise, SGD (i.e., ERM) is used instead.

Recall that in (22), we need to access $\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ at iteration t. If $\|\frac{1}{b}\sum_{i\in I_{t-1}} \nabla_{\boldsymbol{w}_{t-1}} \ell_i(\boldsymbol{w}_{t-1})\|^2 < \mu_{t-1} + c_{t-1}\sigma_{t-1}$ in iteration $t-1$, SAM is not used, and $\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ is not computed. In that case, we replace $\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1})$ with $\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})$. The whole procedure, which will be called Adaptive lookahead-SAM (AO-SAM), is shown in Algorithm 2.

4. Theoretical Analysis

4.1. Region of attraction (ROA)

In this section, we show theoretically that the proposed method is less likely than SAM to be trapped in a saddle

point. We consider the following minimization problem which has been widely used in the theoretical analysis of SAM and SGD (Compagnoni et al., 2023; Kim et al., 2023):

$$\min_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{H} \boldsymbol{w}. \tag{23}$$

Recall that the ordinary differential equation (ODE) of SAM is (Compagnoni et al., 2023): $d\boldsymbol{w}_\tau = -\boldsymbol{H}\left(\boldsymbol{w}_\tau + \frac{\rho \boldsymbol{H}\boldsymbol{w}_\tau}{||\boldsymbol{H}\boldsymbol{w}_\tau||}\right) d\tau$, where τ is the time. For a given $\boldsymbol{w'}$, its region of attraction (ROA) is the set of $\boldsymbol{w}$ such that any trajectory starting inside it converges to $\boldsymbol{w'}$ (Mao, 2007). The ROA of SAM is given by the following.

Proposition 4.1. *(Compagnoni et al., 2023) For a non-singular $\boldsymbol{H}$, the ROA for SAM is $\left\{\boldsymbol{w}_\tau | \rho \geq -\frac{\|\boldsymbol{H}\ \boldsymbol{w}_\tau\|}{\lambda_{\min}}\right\}$, where $\lambda_{\min}$ is the minimum eigenvalue of $\boldsymbol{H}$.*

The following Proposition provides the ODE and ROA of Lookahead-SAM. Proof is in Appendix A.1.

Proposition 4.2. *The ODE for Lookahead-SAM is:*

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

where η'_τ is a function of τ. For a non-singular $\boldsymbol{H}$, when $\eta'_\tau > 0, \forall \tau$, the ROA of Lookahead-SAM is: $\left\{\boldsymbol{w}_\tau | (1+\eta'_\tau \lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}} ||\boldsymbol{H}\boldsymbol{w}_\tau|| (\eta'_\tau \lambda_{\min} - 1)\right\}$.

The following Corollary shows that Lookahead-SAM has a smaller ROA than SAM. In other words, Lookahead-SAM has a smaller chance of being trapped in a saddle point.

Corollary 4.2.1. *For non-singular $\boldsymbol{H}$, Lookahead-SAM has a smaller ROA than SAM.*

As an illustration, Fig. 2 compares the ROAs of SAM and Lookahead-SAM at the saddle point $[0, 0]$ on the toy function in (11).

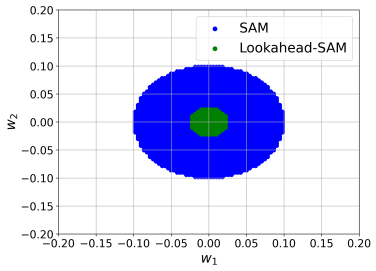


Figure 2: ROAs for SAM and Lookahead-SAM at saddle point.

4.2. Convergence Analysis

In this section, we study the convergence properties of Lookahead-SAM, Opt-SAM, and AO-SAM. Note that our analysis is different from those in the literature on extra-gradient (EG) (Gidel et al., 2019; Bohm et al., 2022; Jelassi et al., 2020; Pethick et al., 2022; Gorbunov et al., 2022; Cai et al., 2022) and optimistic gradient (OG) (Gidel et al., 2019; Liang & Stokes, 2019; Daskalakis & Panageas, 2018; Daskalakis et al., 2018; Mahdavinia et al., 2022). EG and OG assume that f in (7) is (strongly) convex w.r.t. x, and (strongly) concave, (strongly) monotonic or co-coercive w.r.t. y (Gorbunov et al., 2022; Cai et al., 2022; Mahdavinia et al., 2022). In the context of SAM optimization, x corresponds to $\boldsymbol{w}$, and y corresponds to $\boldsymbol{\epsilon}$. Obviously, these assumptions do not hold for deep networks.

On the other hand, the following analysis does not need to assume convex loss, and only uses the common assumptions in smooth and non-convex analysis for stochastic gradient methods. Specifically, Assumptions 4.3 and 4.4 below are from (Andriushchenko & Flammarion, 2022; Bottou et al., 2018; Cutkosky & Orabona, 2019), while Assumption 4.5 follows (Bottou et al., 2018; Hazan & Kale, 2014; Huang et al., 2021). Assumptions 4.3, 4.4 and 4.5 are employed collectively in the convergence analysis of SAM (Mi et al., 2022; Zhang et al., 2023b; Yue et al., 2023; Mueller et al., 2023; Si & Yun, 2023; Zhang et al., 2023a).

Assumption 4.3. *(Bounded variance) There exists $\sigma \geq 0$ s.t. $\mathbb{E}_{i\sim\mathcal{U}([1,n])}\|\nabla \ell_i(\boldsymbol{w}) - \nabla L(\boldsymbol{w})\|^2 \leq \sigma^2$. Here, $\mathcal{U}([1,n])$ is the uniform distribution over $\{1, 2, \ldots, n\}$, and n is the number of samples.*

Assumption 4.4. *(β-smoothness) There exists $\beta \geq 0$ s.t. $\|\nabla \ell_i(\boldsymbol{w}) - \nabla \ell_i(\boldsymbol{v})\| \leq \beta \|\boldsymbol{w} - \boldsymbol{v}\|$ for all $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^m$ and $i = 1, 2, \ldots, n$.*

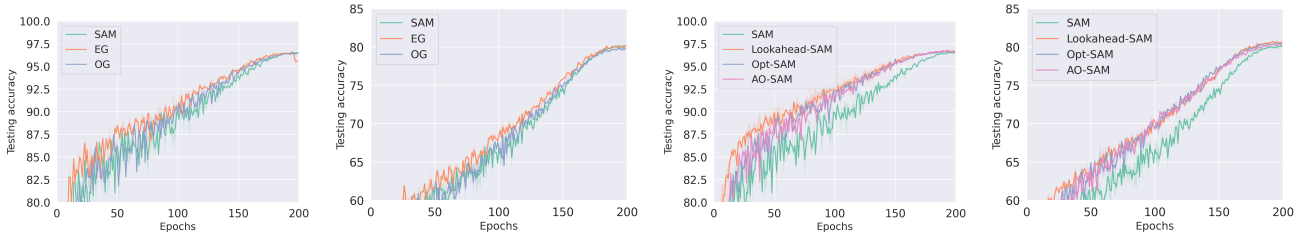
Assumption 4.5. *(Uniformly Bounded Gradient) There exists $G \geq 0$ s.t. $\mathbb{E}_{i\sim\mathcal{U}([1,n])}\|\nabla \ell_i(\boldsymbol{w})\|^2 \leq G^2$.*

The following Theorems provide convergence rates on Lookahead-SAM, Opt-SAM, and AO-SAM. Proofs are in Appendices A.3, A.4 and A.5 respectively. Note that we set ρ_t related to T in our proofs, which is a necessary step as shown in (Si & Yun, 2023).

Theorem 4.6. *In Algorithm 1, set $\rho_t = \frac{1}{\sqrt{T}}$ and $\eta_t = \eta'_t = \min\left(\frac{1}{\sqrt{T}}, \frac{1}{2\beta}\right)$. Lookahead-SAM satisfies $\frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\|\nabla_{\boldsymbol{w}_t} L(\boldsymbol{w}_t)\|^2 = O\left(\frac{1}{\sqrt{Tb}}\right)$.*

Theorem 4.7. *In Algorithm 1, set $\rho_t = \min\left(\frac{1}{\beta}, \frac{1}{\sqrt{T}}\right)$ and $\eta_t = \eta'_t = \min\left(\frac{1}{\sqrt{T}}, \frac{1}{\beta}\right)$. Opt-SAM satisfies $\frac{1}{T}\sum_{t=0}^{T} \mathbb{E}\|\nabla_{\boldsymbol{w}_t} L(\boldsymbol{w}_t)\|^2 = O\left(\frac{1}{\sqrt{Tb}}\right)$.*

Theorem 4.8. *In Algorithm 2, set $\rho_t = \min\left(\frac{1}{\beta}, \frac{1}{\sqrt{T}}\right)$*



(a) *CIFAR-10*. (b) *CIFAR-100*. (c) *CIFAR-10*. (d) *CIFAR-100*.

Figure 3: Convergence on *CIFAR-10* and *CIFAR-100* (with *ResNet-18* backbone).

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

and $\eta_t = \eta_t' = \min\left(\frac{1}{\sqrt{T}}, \frac{1}{2\beta}\right)$. *AO-SAM satisfies* $\frac{1}{T}\sum_{t=0}^{T} E\|\nabla_{\boldsymbol{w}_t} L(\boldsymbol{w}_t)\|^2 = O\left(\frac{1}{\sqrt{Tb}}\right)$.

In summary, Lookahead-SAM, Opt-SAM, and AO-SAM have the same $O(\frac{1}{\sqrt{Tb}})$ convergence rate as SAM (Andriushchenko & Flammarion, 2022) and its variant AESAM (Jiang et al., 2023), and is faster than the $O(\log T/\sqrt{T})$ rate of GSAM (Zhuang et al., 2022) and SSAM (Mi et al., 2022).

5. Experiments

In this section, we empirically demonstrate the performance of the proposed methods on a number of standard benchmark datasets.

Recall that for SAM-based algorithms, the training speed is mainly determined by how often the SAM update is used. As in (Jiang et al., 2023), we evaluate efficiency by measuring the fraction of SAM updates used: $\%\text{SAM} \equiv 100 \times (\sum_{t=1}^{T} \#\{\text{SAMs}\} \text{ used at epoch } t)/T$, where T is the number of epochs and is the same for all methods. Note that as EG takes two SAM steps ((12), (13) and (14), (15)) in every epoch, its %SAM is 200. Similarly, for Lookahead-SAM, its %SAM is 150.

5.1. *CIFAR-10* and *CIFAR-100*

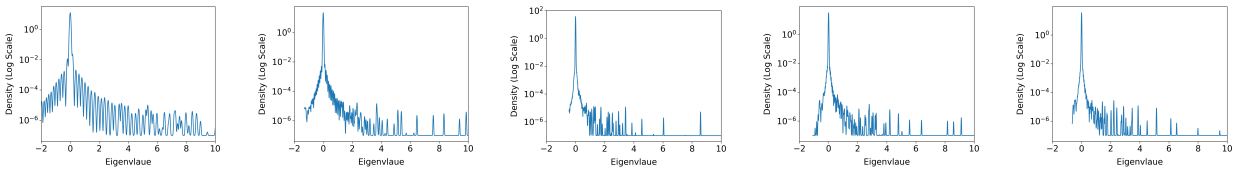
In this experiment, we use the popular image classification datasets *CIFAR-10* and *CIFAR-100* (Krizhevsky et al., 2009). 10% of the training set is used for validation.

First, we compare SAM with the direct adaptation of EG and its OG variant (Section 3.1), and the proposed Lookahead-SAM, Opt-SAM, AO-SAM. Experiment is performed on the *ResNet-18* backbone, and repeated 5 times with different random seeds.

Following the setup in (Jiang et al., 2023; Foret et al., 2021), we use batch size 128, initial learning rate 0.1, cosine learning rate schedule (Loshchilov & Hutter, 2017), and SGD optimizer. Learning rate η_t' is always set to η_t. The number of training epochs is 200. For the proposed methods, we select $\rho \in \{0.01, 0.05, 0.08, 0.1, 0.5, 0.8, 1, 1.5, 1.8, 2\}$ by using *CIFAR-10*'s validation set on *ResNet-18*. The selected ρ is then directly used on *CIFAR-100* and the other backbones. For the c_t schedule in (6), since different SAM variants yield different %SAM's, we vary the hyper-parameters (κ_1, κ_2) so that the %SAM obtained by AO-SAM matches their %SAM values. Hyper-parameters for the other baselines are the same as their original papers.

5.1.1. Comparison with Direct Use of EG and OG

Figure 3 shows the convergence of the testing accuracy with number of epochs. First, we focus on SAM and the direct adaptations of EG and OG. As can be seen from Figures 3a and 3b, on *CIFAR-10*, EG and OG only converge slightly faster than SAM. On *CIFAR-100*, EG is only slightly faster than SAM, while OG is even slower. This is because, EG



(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*.

and OG belong to the GDA family (as discussed in Section 3.2.1). In contrast, as can be seen from Figures 3c and 3d, Lookahead-SAM, Opt-SAM, and AO-SAM converge faster than SAM on both datasets.

Tables 1 and 2 show the testing accuracies versus %SAM for *CIFAR-10* and *CIFAR-100*, respectively. Though Lookahead-SAM has the highest accuracy on *CIFAR-100* and the second highest on *CIFAR-10*, it also has a higher %SAM. On the other hand, 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. Hence, we will only focus on OPT-SAM and AO-SAM in the sequel.

5.1.2. Comparison with SAM Variants

Following (Jiang et al., 2023), we compare AO-SAM and Opt-SAM with: (i) ERM; (ii) SAM (Foret et al., 2021); and its variants (iii) ESAM (Du et al., 2022a), (iv) ASAM (Kwon et al., 2021), (v) SS-SAM (Zhao, 2022), (vi) AE-SAM (Jiang et al., 2023), and (vii) GSAM (Zhuang et al., 2022). As different SAM variants yield different %SAM's, we vary the (κ_1, κ_2) values in (6) for AE-SAM and AO-SAM so as to attain comparable %SAM values for fairer comparison. Following the SAM literature (Jiang et al., 2023; Mi et al., 2022; Kwon et al., 2021), we use the commonly-used *ResNet-18* (He et al., 2016), and *WideResNet-28-10* (Zagoruyko & Komodakis, 2016) as backbones. We also include deeper network models including the *PyramidNet-110* (Han et al., 2017) and vision transformer *Vit-S16* (Dosovitskiy et al., 2021).

Results are shown in Table 3. As can be seen, Opt-SAM and AO-SAM are consistently more accurate than SAM and its variants on all datasets and backbones. Moreover, The improvements obtained over the second-best baseline are statistically significant (achieving a p-value of less than 0.05 in t-test) across all network architectures.

To study the method's robustness to label noise, it is common to use datasets with label noise in the SAM literature (Jiang et al., 2023; Kwon et al., 2021; Foret et al., 2021; Yue et al., 2023; Zhang et al., 2023b). We randomly flip a certain fraction (20%, 40%, 60% and 80%) of the training labels in *CIFAR-10* and *CIFAR-100*.

Results are shown in Tables 8 and 9 in Appendix B. As can be seen, on both datasets, AO-SAM and Opt-SAM outperform all baselines at all label noise ratios. Moreover, the accuracy improvement gets larger as the label noise ratio increases. This demonstrates the superiority of AO-SAM and Opt-SAM, particularly in difficult learning environments.

5.1.3. Time and Memory.

Table 4 compares the total training time and GPU memory for the proposed SAM variants with SAM and ERM (i.e., SGD), using *CIFAR-10* with the *ResNet-18* backbone.

As can be seen, ERM is the fastest, while Lookahead-SAM is the slowest. Opt-SAM is comparable to SAM, while AO-SAM is faster than SAM. These are also consistent with the observations in Table 3 based on the %SAM, verifying that %SAM is a useful metric. For GPU memory usage, SAM and the proposed variants are very similar.

5.1.4. Flat Minima

In this section, we demonstrate the abilities of Lookahead-SAM, Opt-SAM and AO-SAM to avoid saddle points, using *CIFAR-10* with the *ResNet-18* backbone.

Following (Mi et al., 2022; Foret et al., 2021), Figure 4 shows the eigenvalue spectra of the Hessians at the converged solutions of the various methods. As can be seen, the Hessian's eigenvalues of Lookahead-SAM, Opt-SAM, and AO-SAM are smaller than those of ERM and SAM, indicating that the loss landscapes at the converged solutions of these SAM variants are flatter compared to SAM and ERM.

As in (Foret et al., 2021; Mi et al., 2022), Table 5 shows the largest eigenvalue of the Hessian (λ_1) and the ratio λ_1/λ_5 (where λ_5 is the 5th largest eigenvalue). As can be seen, Lookahead-SAM, Opt-SAM, and AO-SAM have smaller λ_1 and λ_1/λ_5 than ERM and SAM, again indicating that they have flatter minima than SAM and ERM.

5.2. *ImageNet*

In this experiment, we perform experiments on the *ImageNet* dataset using *ResNet-50* (He et al., 2016), *ResNet-100* (He et al., 2016), and *ViT-S/32* (Dosovitskiy et al., 2021) back-

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
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

Table 4: Comparison on training time (seconds) and GPU memory usage (GB) on *CIFAR-10* with *ResNet-18* backbone.

	ERM	SAM	Lookahead-SAM	Opt-SAM	AO-SAM
training time	3,630	6,780	10,946	6,994	4,704
GPU memory	2.7	2.7	2.8	2.8	2.8

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**

Table 6: Performance on *MRPC* with *Bert-Large*.

	Accuracy	F1-score	%SAM
ERM	87.3	91.1	0.0
SAM	87.9	91.4	100.0
Opt-SAM	**89.1**	**92.2**	100.0
AO-SAM	88.7	91.9	60.8

bones. The batch size is 512, and the number of training epochs is 90. The other experimental setup is the same as in Section 5.1. Hyper-parameters for the other baselines are the same as their original papers. The experiment is repeated 3 times with different random seeds. Table 7 shows the testing accuracy and %SAM. As can be seen, the proposed AO-SAM again outperforms all the baselines.

5.3. NLP Paraphrase Identification

Following (Zhong et al., 2022), we perform NLP paraphrase identification using the pre-trained *Bert-Large* (Devlin et al., 2018) on the Microsoft Research Paraphrase Corpus (*MRPC*) dataset (Dolan & Brockett, 2005). The learning rate is 2×10^{-5}, batch size is 16, and the number of epochs is 10. The other experiment setup follows (Zhong et al., 2022).

Table 6 shows the accuracy, F1-score and %SAM for ERM, SAM, Opt-SAM and AO-SAM. As can be seen, Opt-SAM is the best, while AO-SAM still outperforms SAM and ERM and has a smaller %SAM than Opt-SAM.

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
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

6. Conclusion

In this paper, we integrate lookahead into SAM. The lookahead mechanism has been proven effective in game theory and optimization. It enables the model to gain more information about the loss landscape, thus alleviating the problem of convergence instability in SAM's minimax optimization process. Theoretical results show that the proposed method can converge to a stationary point and is not easy to be trapped in saddle points. Experiments on standard benchmark datasets also verify that the proposed method outperforms the SOTAs and converges more effectively to flat minima.

In the future, we will study the performance of the proposed methods in scenarios with distribution shift.

Impact Statement

This paper advances the machine learning field. While it may have societal consequences, we believe specific highlights are not necessary here.

Acknowledgement

This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China (Grants 16200021 and HKU C7004-22G). Youzhi Zhang is supported by the InnoHK Fund.

References

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *ICML*, 2022.

Apostol, T. M. *Calculus, Volume 1*. John Wiley & Sons, 1991.

Bohm, A., Sedlmayer, M., Csetnek, E. R., and Bot, R. I. Two steps at a time – taking GAN training in stride with Tseng's method. *SIMODS*, 2022.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 2018.

Butcher, J. C. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, 2016.

Cai, Y., Oikonomou, A., and Zheng, W. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. In *NeurIPS*, 2022.

Chatterji, N. S., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. In *ICLR*, 2020.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019.

Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *COLT*, 2012.

Compagnoni, E. M., Biggio, L., Orvieto, A., Proske, F. N., Kersting, H., and Lucchi, A. An SDE for modeling SAM: Theory and insights. In *ICML*, 2023.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *NeurIPS*, 2019.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *NeurIPS*, 2018.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. In *ICLR*, 2018.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint arXiv:1810.04805, 2018.

Dolan, B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *International Workshop on Paraphrasing*, 2005.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. Efficient sharpness-aware minimization for improved training of neural networks. In *ICLR*, 2022a.

Du, J., Zhou, D., Feng, J., Tan, V., and Zhou, J. T. Sharpness-aware training for free. In *NeurIPS*, 2022b.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*, 2017.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

Fudenberg, D. and Tirole, J. *Game Theory*. MIT press, 1991.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *ICLR*, 2019.

Gorbunov, E., Taylor, A., and Gidel, G. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *NeurIPS*, 2022.

Han, D., Kim, J., and Kim, J. Deep pyramidal residual networks. In *CVPR*, 2017.

Hazan, E. and Kale, S. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Huang, F., Li, J., and Huang, H. Super-ADAM: faster and universal framework of adaptive gradients. In *NeurIPS*, 2021.

Jelassi, S., Domingo-Enrich, C., Scieur, D., Mensch, A., and Bruna, J. Extragradient with player sampling for faster Nash equilibrium finding. In *NeurIPS*, 2020.

Jiang, W., Yang, H., Zhang, Y., and Kwok, J. An adaptive policy to employ sharpness-aware minimization. In *ICLR*, 2023.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *ICLR*, 2020.

Kaddour, J., Liu, L., Silva, R., and Kusner, M. J. When do flat minima optimizers work? *NeurIPS*, 35:16577–16595, 2022.

Kim, H., Park, J., Choi, Y., and Lee, J. Stability analysis of sharpness-aware minimization. Preprint arXiv:2301.06308, 2023.

Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kwon, J., Kim, J., Park, H., and Choi, I. K. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *ICML*, 2021.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 2015.

Lee, C.-W., Kroer, C., and Luo, H. Last-iterate convergence in extensive-form games. In *NeurIPS*, 2021.

Leng, C., Dou, Z., Li, H., Zhu, S., and Jin, R. Extremely low bit neural network: Squeeze the last bit out with ADMM. In *AAAI*, 2018.

Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, 2019.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *ICML*, 2020a.

Lin, T., Kong, L., Stich, S., and Jaggi, M. Extrapolation for large-batch training in deep learning. In *ICML*, 2020b.

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. Towards efficient and scalable sharpness-aware minimization. In *CVPR*, 2022.

Liu, Z., Li, S., Lee, W. S., Yan, S., and Xu, Z. Efficient offline policy optimization with a learned model. In *ICLR*, 2023.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Mahdavinia, P., Deng, Y., Li, H., and Mahdavi, M. Tight analysis of extra-gradient and optimistic gradient methods for nonconvex minimax problems. In *NeurIPS*, 2022.

Mao, X. *Stochastic Differential Equations and Applications*. Elsevier, 2007.

Mi, P., Shen, L., Ren, T., Zhou, Y., Sun, X., Ji, R., and Tao, D. Make sharpness-aware minimization stronger: A sparsified perturbation approach. In *NeurIPS*, 2022.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, 2020.

Mueller, M., Vlaar, T., Rolnick, D., and Hein, M. Normalization layers are all that sharpness-aware minimization needs. In *NeurIPS*, 2023.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *NeurIPS*, 2017.

Pethick, T., Patrinos, P., Fercoq, O., Cevherå, V., and Latafat, P. Escaping limit cycles: Global convergence for constrained nonconvex-nonconcave minimax problems. In *ICLR*, 2022.

Petzka, H., Kamp, M., Adilova, L., Sminchisescu, C., and Boley, M. Relative flatness and generalization. In *NeurIPS*, 2021.

Popov, L. D. A modification of the Arrow-Hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 1980.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *NeurIPS*, 2013.

Si, D. and Yun, C. Practical sharpness-aware minimization cannot converge all the way to optima. In *NeurIPS*, 2023.

Tan, C., Zhang, J., Liu, J., Wang, Y., and Hao, Y. Stabilizing sharpness-aware minimization through a simple renormalization strategy. Preprint arXiv:2401.07250, 2024.

Wang, B., Nguyen, T., Sun, T., Bertozzi, A. L., Baraniuk, R. G., and Osher, S. J. Scheduled restart momentum for accelerated stochastic gradient descent. *SIAM Journal on Imaging Sciences*, 15(2):738–761, 2022.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 2021.

Yue, Y., Jiang, J., Ye, Z., Gao, N., Liu, Y., and Zhang, K. Sharpness-aware minimization revisited: Weighted sharpness as a regularization term. Preprint arXiv:2305.15817, 2023.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *BMVC*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. In *Communications of the ACM*, 2021.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. Lookahead optimizer: k steps forward, 1 step back. In *NeurIPS*, 2019.

Zhang, X., Xu, R., Yu, H., Dong, Y., Tian, P., and Cui, P. Flatness-aware minimization for domain generalization. In *ICCV*, pp. 5189–5202, 2023a.

Zhang, X., Xu, R., Yu, H., Zou, H., and Cui, P. Gradient norm aware minimization seeks first-order flatness and improves generalization. In *CVPR*, 2023b.

Zhao, Y. Randomized sharpness-aware training for boosting computational efficiency in deep learning. Preprint arXiv:2203.09962, 2022.

Zhong, Q., Ding, L., Shen, L., Mi, P., Liu, J., Du, B., and Tao, D. Improving sharpness-aware minimization with fisher mask for better generalization on language models. In *Findings of EMNLP*, 2022.

Zhou, P., Yan, H., Yuan, X., Feng, J., and Yan, S. Towards understanding why lookahead generalizes better than SGD and beyond. In *NeurIPS*, 2021.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N. C., s Duncan, J., Liu, T., et al. Surrogate gap minimization improves sharpness-aware training. In *ICLR*, 2022.

A. Proofs

A.1. Proof of Proposition 4.2

Lemma A.1. *The Euler's discretization of*

$$d\mathbf{w}_\tau = -\boldsymbol{H}\left(\mathbf{w}_\tau - \eta'_\tau \boldsymbol{H}\left(\mathbf{w}_\tau + \frac{\rho \boldsymbol{H}\mathbf{w}_\tau}{\|\boldsymbol{H}\mathbf{w}_\tau\|}\right) + \frac{\rho \boldsymbol{H}\mathbf{w}_\tau}{\|\boldsymbol{H}\mathbf{w}_\tau\|}\right) d\tau \tag{24}$$

is (21).

Proof. Given an ODE $d\boldsymbol{y} = f(\boldsymbol{y}(\tau))d\tau$, recall that the main process of Euler's discretization (Sec. 2.1, (Butcher, 2016)) is to first set $\boldsymbol{y}_0 = \boldsymbol{y}(\tau_0)$, choose h as the step size along the τ-axis, and set $\tau_{t+1} = \tau_t + h$. Then, we have the discretized version of $\frac{d\boldsymbol{y}}{d\tau}$: $\boldsymbol{y}_{t+1} = \boldsymbol{y}_t + hf(\boldsymbol{y}_t)$.

For (24), we take the same approach and set $h = \eta_{t-1}$. We then have

$$\boldsymbol{w}_t = \mathbf{w}_{t-1} - \eta_{t-1}\, \boldsymbol{H}\left(\boldsymbol{w}_{t-1} - \eta'_{t-1}\boldsymbol{H}\left(\boldsymbol{w}_{t-1} + \frac{\rho \boldsymbol{H}\boldsymbol{w}_{t-1}}{\|\boldsymbol{H}\boldsymbol{w}_{t-1}\|}\right) + \frac{\rho \boldsymbol{H}\boldsymbol{w}_{t-1}}{\|\boldsymbol{H}\boldsymbol{w}_{t-1}\|}\right). \tag{25}$$

Note that when $L(\boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{H}\boldsymbol{w}$, we have the following equation:

$$\begin{aligned} & L\left(\boldsymbol{w}_t - \eta'_{t-1} \perp (\nabla_{\boldsymbol{w}_{t-1}} L\left(\boldsymbol{w}_{t-1} + \frac{\rho \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})\|}\right) + \frac{\rho \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})\|})\right) \\ = \quad & \boldsymbol{H}\left(\boldsymbol{w}_t - \eta'_{t-1}\boldsymbol{H}\left(\boldsymbol{w}_{t-1} + \frac{\rho \boldsymbol{H}\boldsymbol{w}_{t-1}}{\|\boldsymbol{H}\boldsymbol{w}_{t-1}\|}\right) + \frac{\rho \boldsymbol{H}\boldsymbol{w}_{t-1}}{\|\boldsymbol{H}\boldsymbol{w}_{t-1}\|}\right). \end{aligned}$$

Therefore, (25) is exactly (21). We obtain the desired result. □

Next, we provide the proof of Lookahead-SAM ODE:

Lemma A.2. *(Lookahead-SAM ODE) For (23), if* $(1+\eta'_\tau\lambda_i)\rho \geq \frac{1}{\lambda_i}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_i - 1), \forall i, \tau$*, and* $\boldsymbol{H}$ *is non-singular, the ROA for Lookahead-SAM is:* $\left\{\boldsymbol{w}_\tau | (1+\eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_{\min} - 1)\right\}$*, where* $\lambda_{\min}$ *is the smallest eigenvalue of* $\boldsymbol{H}$*.*

Proof. Let $V(\boldsymbol{w}_\tau) := \frac{\boldsymbol{w}_\tau^\top \boldsymbol{K}\boldsymbol{w}_\tau}{2}$ be the Lyapunov function of Lookahead-SAM, where $\boldsymbol{K}$ is a diagonal matrix with positive diagonal entries $(k_1, \ldots, k_d)$. We have

$$V(\boldsymbol{w}_\tau) = \frac{1}{2}\sum_{i=1}^{d} k_i w_{i,\tau}^2 > 0,$$

and

$$\begin{aligned} \dot{V}(\boldsymbol{w}_\tau) &:= \frac{dV(\boldsymbol{w}_\tau)}{d\tau} \\ &= \sum_{i=1}^{d} k_i w_{i,\tau}\frac{dw_{i,\tau}}{d\tau} \\ &\overset{(a)}{=} \sum_{i=1}^{d} q_{ii}^2 k_i w_{i,\tau}\left(-\lambda_i\left(w_{i,\tau} - \eta'_\tau\lambda_i w_{i,\tau} + \frac{\rho\eta'_\tau\lambda_i^2 w_{i,\tau}}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i w_{i,\tau}}{||H\boldsymbol{w}_\tau||}\right)\right) \\ &= \sum_{i=1}^{d} q_{ii}^2 k_i(-\lambda_i)\left(1 - \eta'_\tau\lambda_i + \frac{\rho\eta'_\tau\lambda_i^2}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i}{||H\boldsymbol{w}_\tau||}\right)(w_{i,\tau})^2\, d\tau. \end{aligned}$$

(a) holds because $\boldsymbol{H}$ is symmetric and can be decomposed as $\boldsymbol{H} = \boldsymbol{Q}^\top \Lambda \boldsymbol{Q}$ (Theorem 5.11 in (Apostol, 1991)), where $\boldsymbol{Q}$ is an orthogonal matrix (with elements $[q_{ij}]$), and Λ is diagonal matrix containing the eigenvalues of $\boldsymbol{H}$. Then, $\frac{dw_{i,\tau}}{d\tau}$ can be written as $\frac{dw_{i,\tau}}{d\tau} = q_{ii}^2 k_i(-\lambda_i)\left(1 - \eta'_\tau\lambda_i + \frac{\rho\eta'^2_\tau\lambda_i^2}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i}{||\boldsymbol{H}\boldsymbol{w}_\tau||}\right)(w_{i,\tau})^2$.

For the term $q_{ii}^2 k_i(-\lambda_i)\left(1 - \eta'_\tau\lambda_i + \frac{\rho\eta'^2_\tau\lambda_i^2}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i}{||\boldsymbol{H}\boldsymbol{w}_\tau||}\right)(w_{i,\tau})^2$, when $\lambda_i < 0$, , we use $1 - \eta'_\tau\lambda_i + \frac{\rho\eta'^2_\tau\lambda_i^2}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i}{||\boldsymbol{H}\boldsymbol{w}_\tau||} \leq 0, \forall i$ to ensure $\dot{V}(\boldsymbol{w}_\tau) \leq 0$, and thus $(1 + \eta'_\tau\lambda_i)\rho \geq \frac{1}{\lambda_i}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_i - 1)$.

Similarly, when $\lambda_i > 0$, we have $1 - \eta'_\tau\lambda_i + \frac{\rho\eta'^2_\tau\lambda_i^2}{||\boldsymbol{H}\boldsymbol{w}_\tau||} + \frac{\rho\lambda_i}{||\boldsymbol{H}\boldsymbol{w}_\tau||} \geq 0$ to ensure $\dot{V}(\boldsymbol{w}_\tau) \leq 0$, and thus we also have $(1 + \eta'_\tau\lambda_i)\rho \geq \frac{1}{\lambda_i}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_i - 1)$.

Therefore, when $(1 + \eta'_\tau\lambda_i)\rho \geq \frac{1}{\lambda_i}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_i - 1), \forall i$, we have $V(\boldsymbol{w}_\tau) > 0$ and $\dot{V}(\boldsymbol{w}_\tau) \leq 0$. Using Theorem 1.1 (Mao, 2007), the dynamics of $\boldsymbol{w}_\tau$ is bounded inside this set and cannot diverge if $V(\boldsymbol{w}_\tau) > 0$ and $\dot{V}(\boldsymbol{w}_\tau) \leq 0$.

Since $(1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}||H\boldsymbol{w}_\tau||(\eta'_\tau\lambda_{\min} - 1)$ implies $(1 + \eta'_\tau\lambda_i)\rho \geq \frac{1}{\lambda_i}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_i - 1), \forall i$, by the definition of ROA, we conclude that the ROA for Lookahead-SAM is: $\left\{\boldsymbol{w}_\tau | (1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}||H\boldsymbol{w}_\tau||(\eta'_\tau\lambda_{\min} - 1)\right\}$.

□

Now, we state the proof of Proposition 4.2.

Proof. This can be directly obtained by using the results of lemma A.2 and lemma A.1. □

A.2. Proof of Corollary 4.2.1

Proof. Recall that a non-singular saddle point has both positive and negative eigenvalues (Theorem 9.6. in (Apostol, 1991)). Note that for the minimum eigenvalue $\lambda_{\min}$ of $\boldsymbol{H}$, we have

$$\begin{aligned}
& (1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}\|\boldsymbol{H}\mathbf{w}_\tau\|(\eta'_\tau\lambda_{\min} - 1) \\
\Rightarrow\quad & (\lambda_{\min} + \eta'_\tau\lambda_{\min}^2)\rho \leq \|\boldsymbol{H}\mathbf{w}_\tau\|(\eta'_\tau\lambda_{\min} - 1) \\
\Rightarrow\quad & -\lambda_{\min}\rho \geq \|\boldsymbol{H}\mathbf{w}_\tau\|(1 - \eta'_\tau\lambda_{\min}) + \eta'_\tau\lambda_{\min}^2\rho \\
\Rightarrow\quad & -\lambda_{\min}\rho \geq \quad \|\boldsymbol{H}\mathbf{w}_\tau\|(1 - \eta'_\tau\lambda_{\min}) + \eta'_\tau\lambda_{\min}^2\rho \geq \|\boldsymbol{H}\mathbf{w}_\tau\| \\
\Rightarrow\quad & -\lambda_{\min}\rho \geq \|\boldsymbol{H}\mathbf{w}_\tau\|.
\end{aligned}$$

The second last inequality is due to the fact that $-\|\boldsymbol{H}\mathbf{w}_\tau\|\eta'_\tau\lambda_{\min} + \eta'_\tau\lambda_{\min}^2\rho \geq 0$.

Using Proposition 4.1, the ROA for SAM is: $\left\{\boldsymbol{w}_\tau | \rho \geq -\frac{\|\boldsymbol{H}\,\boldsymbol{w}_\tau\|}{\lambda_{\min}}\right\}$. The ROA for Lookahead-SAM is: $\left\{\boldsymbol{w}_\tau | (1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_{\min} - 1)\right\}$, using the results of Lemma A.2. Since $(1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}\|\boldsymbol{H}\mathbf{w}_\tau\|(\eta'_\tau\lambda_{\min} - 1)$ implies $-\lambda_{\min}\rho \geq \|\boldsymbol{H}\mathbf{w}_\tau\|$, we have $\left\{\boldsymbol{w}_\tau | (1 + \eta'_\tau\lambda_{\min})\rho \geq \frac{1}{\lambda_{\min}}||\boldsymbol{H}\boldsymbol{w}_\tau||(\eta'_\tau\lambda_{\min} - 1)\right\} \subset \left\{\boldsymbol{w}_\tau | \rho \geq -\frac{\|\boldsymbol{H}\,\boldsymbol{w}_\tau\|}{\lambda_{\min}}\right\}$, which implies that Lookahead-SAM has a smaller ROA than SAM. □

A.3. Proof of Theorem 4.6

Let $L_t(\boldsymbol{w}) := \frac{1}{b}\sum_{i \in I_t}\ell_i(\boldsymbol{w})$ be the mini-batch version of $L(\boldsymbol{w})$ at epoch t, where $\boldsymbol{w} \in \mathbb{R}^m$, I_t is the mini-batch, and $\ell_i(\boldsymbol{w})$ is the loss for sample i. Let $\boldsymbol{w}_{t-1/2} := \hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t = \boldsymbol{w}_{t-1} - \eta'_t\nabla_{\boldsymbol{w}_{t-1}}L\left(\boldsymbol{w}_{t-1} + \frac{\rho\nabla_{\boldsymbol{w}_{t-1}}L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L(\boldsymbol{w}_{t-1})\|}\right) + \frac{\rho\nabla_{\boldsymbol{w}_{t-1}}L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L(\boldsymbol{w}_{t-1})\|}$. Note that the update of Lookahead-SAM in (21) can be rewritten as: $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t\nabla_{\boldsymbol{w}_{t-1/2}}L_t(\boldsymbol{w}_{t-1/2})$. In the following, we assume that $\rho, \eta, \beta > 0$ [3]. We define $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$ as the inner product of vectors $\boldsymbol{A}$ and $\boldsymbol{B}$.

Lemma A.3.

$$\left\langle \nabla L\left(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|} - \eta\nabla L(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})\right), \nabla L(\boldsymbol{w})\right\rangle \geq -\beta\rho\|\nabla L(\boldsymbol{w})\| - \frac{\eta^2}{2}G^2 + \frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2.$$

[3] To simplify notations, we drop the subscript t from η_t and ρ_t.

Proof.

$$
\begin{aligned}
&\left\langle \nabla L\left(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}\right),\nabla L(\boldsymbol{w})\right\rangle \\
&= \frac{\|\nabla L(\boldsymbol{w})\|}{\rho}\left\langle \nabla L\left(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}\right)-\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})),\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}\right\rangle \\
&\quad +\langle\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})),\nabla L(\boldsymbol{w})\rangle \\
&\geq -\beta\|\nabla L(\boldsymbol{w})\|\rho||\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}||^2+\langle\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}))-\nabla L(\boldsymbol{w}),\nabla L(\boldsymbol{w})\rangle \\
&= -\beta\rho\|\nabla L(\boldsymbol{w})\|+\langle\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}))-\nabla L(\boldsymbol{w}),\nabla L(\boldsymbol{w})\rangle \\
&\geq -\beta\rho\|\nabla L(\boldsymbol{w})\|-\frac{\eta^2}{2}G^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2.
\end{aligned}
$$

The first equation is due to the fact that $\langle\nabla L(w),\nabla L(w)\rangle=\|\nabla L(w)\|^2$, while the first inequality uses the property of the smooth function L, i.e., $\langle\nabla L\left(w_1\right)-\nabla L\left(w_2\right),w_1-w_2\rangle\geq-\beta||w_1-w_2||^2$ for smooth L (Lemma 7 in (Andriushchenko & Flammarion, 2022)). The last inequality is based on the following inequalities:

$$
\begin{aligned}
&\langle\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})),\nabla L(\boldsymbol{w})\rangle \\
&= \langle\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}))-\nabla L(\boldsymbol{w}),\nabla L(\boldsymbol{w})\rangle+\|\nabla L(\boldsymbol{w})\|^2 \\
&\geq -\frac{1}{2}||\nabla L(\boldsymbol{w}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}))-\nabla L(\boldsymbol{w})||^2-\frac{1}{2}||\nabla L(\boldsymbol{w})||^2+\|\nabla L(\boldsymbol{w})\|^2 \\
&\geq -\frac{\eta^2}{2}||\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})||^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2,
\end{aligned}
$$

where the first inequality based on the Young's inequality. □

Lemma A.4.

$$
\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|})\right),\nabla L(\boldsymbol{w})\right\rangle \geq -\beta\rho\|\nabla L(\boldsymbol{w})\|-\frac{\eta^2}{2}G^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2-\frac{\beta^2\sigma^2\eta^2}{2b}-\rho^2\beta^2.
$$

Proof. Note that

$$
\begin{aligned}
&\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|})\right),\nabla L(\boldsymbol{w})\right\rangle \\
&= \left\langle \nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|})\right)-\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|})\right),\nabla L(\boldsymbol{w})\right\rangle \\
&\quad -\left\langle -\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|})\right),\nabla L(\boldsymbol{w})\right\rangle.
\end{aligned}
$$

In the following, we bound the first term and second term of the RHS separately. For the first term,

$$\begin{aligned}
& -\mathbb{E}\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right)\right. \\
& \left. -\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right), \nabla L(\boldsymbol{w})\right\rangle \\
\leq\ & \frac{1}{2}\mathbb{E}\left\|\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right)\right. \\
& \left. -\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right)\right\|^2 \\
& +\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 \\
\leq\ & \frac{\beta^2}{2}\mathbb{E}\left\|\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right. \\
& \left. -\left(\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}\right)\right)\right\|^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 \\
\overset{(b)}{\leq}\ & \frac{\beta^2\sigma^2\eta^2}{2b}+\rho^2\beta^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2.
\end{aligned}$$

(a) is based on the Young's inequality, (b) is using the triangle inequality and Assumption 4.4.
For the second term, on using Lemma A.3,

$$\begin{aligned}
& \left\langle \nabla L\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})\right), \nabla L(\boldsymbol{w})\right\rangle \\
\geq\ & -\beta\rho\|\nabla L(\boldsymbol{w})\|-\frac{\eta^2}{2}G^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2.
\end{aligned}$$

Combining them together, we obtain the result. □

Lemma A.5. *With assumptions 4.3, 4.4 and 4.5, and also assume $\eta<\frac{1}{2\beta}$, we have*

$$\eta(1-\eta\beta)\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2\leq \mathbb{E}L(\boldsymbol{w}_t)-\mathbb{E}L(\boldsymbol{w}_{t+1})+\eta^2\beta^3\rho^2+G^2\eta^3\beta^3+\beta\eta\rho\|\nabla L(\boldsymbol{w})\|+\frac{\eta^2}{2}G^2+\eta^2\beta\frac{\sigma^2}{b}.$$

Proof. Using the property of smooth function L (Assumption (4.4)), we have:

$$\begin{aligned}
\mathbb{E}L(\boldsymbol{w}_{t+1}) &\leq \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L_t(\boldsymbol{w}_{t-1/2})\|^2 \\
&\leq \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L_t(\boldsymbol{w}_t)-\nabla L(\boldsymbol{w}_{t-1/2})\|^2 \\
&\quad +\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2 \\
&\leq \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\eta^2\beta\frac{\sigma^2}{2b}+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2.
\end{aligned}$$

The first inequality is using the property of smooth function (Assumption (4.4)) and take the expectation on both sides.

Then,

$$\begin{aligned}
\mathbb{E}L(\boldsymbol{w}_{t+1}) &\overset{(a)}{\leq} \mathbb{E}L(\boldsymbol{w}_t) - \frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 \\
&\quad -\eta(1-\eta\beta)\mathbb{E} < \nabla L(\boldsymbol{w}_{t-1/2}), \nabla L(\boldsymbol{w}_t) > + \eta^2\beta\frac{\sigma^2}{2b} \\
&\overset{(b)}{\leq} \mathbb{E}L(\boldsymbol{w}_t) - \frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{\eta^2\beta^3}{2}\mathbb{E}\|\boldsymbol{w}_{t-1/2} - \boldsymbol{w}_t\|^2 \\
&\quad -\eta(1-\eta\beta)\left[-\beta\rho\|\nabla L(\boldsymbol{w})\| - \frac{\eta^2}{2}G^2 + \frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 - \frac{\beta^2\sigma^2\eta^2 G}{2b} - \rho^2\beta^2 G\right] \\
&\quad +\eta^2\beta\frac{\sigma^2}{2b} \\
&\overset{(c)}{\leq} \mathbb{E}L(\boldsymbol{w}_t) - \frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \frac{1}{2}\eta^2\beta^3\rho^2 + \frac{1}{2}G^2\eta^3\beta^3 \\
&\quad -\left(\eta^2\beta - \eta\right)\beta\rho\|\nabla L(\boldsymbol{w})\| - \left(\eta^2\beta - \eta\right)\frac{\eta^2}{2}G^2 + \left(\eta^2\beta - \eta\right)\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 \\
&\quad +\eta^2\left(\eta^2\beta - \eta\right)\frac{\beta^2\sigma^2}{2b}G + \rho^2\beta^2\left(\eta^2\beta - \eta\right)G + \eta^2\beta\frac{\sigma^2}{b} \\
&\leq \mathbb{E}L(\boldsymbol{w}_t) - \frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \eta^2\beta^3\rho^2 + G^2\eta^3\beta^3 + (1-\eta\beta)\beta\eta\rho\|\nabla L(\boldsymbol{w})\| - (1-\eta\beta)\frac{\eta^2}{2}G^2 \\
&\quad +\left(\eta^2\beta - \eta\right)\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 + \eta^2\beta\frac{\sigma^2}{b}.
\end{aligned}$$

(a) is by using $\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2 = -\|\nabla L(w_t)\|^2 + \|\nabla L(\boldsymbol{w}_{t-1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 + 2\langle\nabla L(\boldsymbol{w}_{t-1/2}), \nabla L(\boldsymbol{w}_t)\rangle$. (b) is the smooth property, the assumption $1 > 2\eta\beta > \eta\beta$, and the Lemma A.4. (c) is by using the trick: $\mathbb{E}\|\boldsymbol{w}_{t-1/2} - \boldsymbol{w}_t\|^2 = \mathbb{E}||\eta\nabla L(\boldsymbol{w}_t + \rho\frac{\nabla L(\boldsymbol{w}_t)}{\|\nabla L(\boldsymbol{w}_t)\|}) + \rho\frac{\nabla L(\boldsymbol{w}_t)}{\|\nabla L(\boldsymbol{w}_t)\|}||^2 \leq \rho^2 + \eta G^2$. The last inequality is by the assumption that $1 > 2\eta\beta > \eta\beta$, which implies $\eta^2\beta - \eta < 0$. Finally, after simplification, we obtain the result. □

Now, we state the proof for Theorem 4.6.

Proof. First, note that $1 - 2\eta\beta > 0$. Then from Lemma A.5, there exists a positive constant c such that

$$c\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2 \leq \frac{1}{\eta}[\mathbb{E}L(\boldsymbol{w}_t) - \mathbb{E}L(\boldsymbol{w}_{t+1})] + \eta\beta^3\rho^2 + G^2\eta^2\beta^3 + \beta\rho\|\nabla L(\boldsymbol{w})\| + \frac{\eta}{2}G^2 + \eta\beta\frac{\sigma^2}{b}.$$

Set $\rho = \frac{1}{\sqrt{T}}$ and $\eta_t = \min\left(\frac{1}{2\beta}, \frac{1}{\sqrt{T}}\right)$. By telescoping from $t = 1$ to T , and divide by T, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 &\leq \frac{1}{c}[\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbb{E}L(\boldsymbol{w}_t) - \mathbb{E}L(\boldsymbol{w}_{t+1})}{\eta} + \eta\beta^2\rho^2 \\
&\quad +\eta\beta^3\rho^2 + G^2\eta^2\beta^3 + \beta\rho G + \frac{\eta}{2}G^2 + \eta\beta\frac{\sigma^2}{b}] \\
&\leq \frac{\mathbb{E}L(\boldsymbol{w}_0)}{\sqrt{T}c} + \frac{\beta^2}{cT\sqrt{T}} + \frac{\beta^3}{cT\sqrt{T}} + \frac{\beta^3 G^2}{cT\sqrt{T}} + \frac{\beta G}{\sqrt{T}c} + \frac{G^2}{2\sqrt{T}c} + \frac{2\beta\sigma^2 c}{c\sqrt{T}b} \\
&= O\left(\frac{1}{\sqrt{T}b}\right).
\end{aligned}$$

The second inequality uses the fact that $2\eta\beta\mathbb{E}\left\langle\nabla L(\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|} - \eta\nabla L(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})), \nabla L(\boldsymbol{w})\right\rangle \leq 2\eta\beta G^2$.

□

A.4. Proof of Theorem 4.7

Recall the update scheme of Opt-SAM in Alg. 1:

$$\hat{\boldsymbol{\epsilon}}_t = \frac{\rho \nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})\|}, \tag{26}$$

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t' \nabla_{\hat{\boldsymbol{w}}_{t-1}} L\left(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}\right) \tag{27}$$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L(\hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t). \tag{28}$$

Recall from Section A.3 that $L_t(\boldsymbol{w}) := \frac{1}{b}\sum_{i\in I_t} \ell_i(\boldsymbol{w})$. In the following, we assume $\rho, \eta, \beta > 0$, and setting $\eta_t' = \eta_t, \forall t$. Also, we use assumptions 4.3, 4.4 and 4.5.

Lemma A.6. *The update scheme of Opt-SAM is equivalent to*

$$\begin{aligned}\hat{\boldsymbol{w}}_t = \hat{\boldsymbol{w}}_{t-1} &- 2\eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L\left(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}\Big|_{\hat{\boldsymbol{\epsilon}}_{t-1}=\frac{\rho\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})}{\|\nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\boldsymbol{w}_{t-1})\|}}\right) \\ &+\eta_{t-1}\nabla_{\hat{\boldsymbol{w}}_{t-2}} L\left(\hat{\boldsymbol{w}}_{t-2} + \hat{\boldsymbol{\epsilon}}_{t-2}\Big|_{\hat{\boldsymbol{\epsilon}}_{t-2}=\frac{\rho\nabla_{\boldsymbol{w}_{t-2}} L(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}} L(\boldsymbol{w}_{t-2})\|}}\right).\end{aligned}$$

Proof. Recall that in Opt-SAM, we have:

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}), \tag{29}$$

and

$$\boldsymbol{w}_{t-1} = \boldsymbol{w}_{t-2} - \eta_t \nabla_{\hat{\boldsymbol{w}}_t} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}). \tag{30}$$

Substitute $\boldsymbol{w}_{t-1}$ in (29) by (30), we have

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-2} - 2\eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1}). \tag{31}$$

Also, note that in Opt-SAM,

$$\hat{\boldsymbol{w}}_{t-1} = \boldsymbol{w}_{t-2} - \eta_{t-1} \nabla_{\hat{\boldsymbol{w}}_{t-2}} L(\hat{\boldsymbol{w}}_{t-2} + \hat{\boldsymbol{\epsilon}}_{t-2}). \tag{32}$$

Substitute (32) into (31), we have:

$$\begin{aligned}\hat{\boldsymbol{w}}_t = \hat{\boldsymbol{w}}_{t-1} &- 2\eta_t \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1} + \hat{\boldsymbol{\epsilon}}_{t-1} |\ \hat{\boldsymbol{\epsilon}}_{t-1} = \frac{\rho\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L(\boldsymbol{w}_{t-1})\|}) \\ &+\eta_{t-1}\nabla_{\hat{\boldsymbol{w}}_{t-2}} L(\hat{\boldsymbol{w}}_{t-2} + \hat{\boldsymbol{\epsilon}}_{t-2} | \hat{\boldsymbol{\epsilon}}_{t-2} = \frac{\rho\nabla_{\boldsymbol{w}_{t-2}} L(\boldsymbol{w}_{t-2})}{\|\nabla_{\hat{\boldsymbol{w}}_{t-2}} L(\boldsymbol{w}_{t-2})\|}),\end{aligned}$$

which is the desired result. □

Lemma A.7.

$$\left\langle \nabla L\left(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}\right), \nabla L(\boldsymbol{w})\right\rangle \geq \|\nabla L(\boldsymbol{w})\|^2 - \beta\rho\|\nabla L(\boldsymbol{w})\|.$$

Proof.

$$\begin{aligned}&\left\langle \nabla L(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}), \nabla L(\boldsymbol{w})\right\rangle \\ &= \left\langle \nabla L\left(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}\right) - \nabla L(\boldsymbol{w}), \nabla L(\boldsymbol{w})\right\rangle + \|\nabla L(\boldsymbol{w})\|^2 \\ &= \frac{\|\nabla L(\boldsymbol{w})\|}{\rho}\left\langle \nabla L(\boldsymbol{w} + \rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}) - \nabla L(\boldsymbol{w}), \frac{\rho}{\|\nabla L(\boldsymbol{w})\|}\nabla L(\boldsymbol{w})\right\rangle + \|\nabla L(\boldsymbol{w})\|^2 \\ &\overset{(a)}{\geq} \left(1 - \frac{\beta\rho}{\|\nabla L(\boldsymbol{w})\|}\right)\|\nabla L(\boldsymbol{w})\|^2 \\ &\overset{(b)}{\geq} \|\nabla L(\boldsymbol{w})\|^2 - \beta\rho\|\nabla L(\boldsymbol{w})\|.\end{aligned}$$

(a) is based on Lemma 7 in (Andriushchenko & Flammarion, 2022). (b) is by simple calculation. □

Lemma A.8.

$$\begin{aligned}&\mathbb{E}\left[\left\langle \nabla_{\boldsymbol{w}_{t-1}}[L(\boldsymbol{w}_{t-1})], \nabla_{\boldsymbol{w}_{t-1}} L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})\|}\right)\right\rangle\right]\\ \geq\ & \frac{1}{2}\|\nabla L(\boldsymbol{w}_{t-1})\|^2-\beta\rho\|\nabla L(\boldsymbol{w}_{t-1})\|-\rho^2\beta^2.\end{aligned}$$

Proof. Similar to the proof of Lemma A.14, note that

$$\begin{aligned}&\left\langle \nabla L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{\|\nabla L_t(\boldsymbol{w}_{t-1})\|}\right), \nabla L(\boldsymbol{w}_{t-1})\right\rangle\\ =\ & \left\langle \nabla L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{\|\nabla L_t(\boldsymbol{w}_{t-1})\|}\right)-\nabla L_t(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}), \nabla L(\boldsymbol{w}_{t-1})\right\rangle\\ &-\left\langle -\nabla L_t(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}), \nabla L(\boldsymbol{w}_{t-1})\right\rangle.\end{aligned}$$

To bound the RHS,

$$\begin{aligned}&-\mathbb{E}\left\langle \nabla L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{\|\nabla L_t(\boldsymbol{w}_{t-1})\|}\right)-\nabla L_t(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}), \nabla L(\boldsymbol{w}_{t-1})\right\rangle\\ \leq\ & \frac{1}{2}\mathbb{E}\|\nabla L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{\|\nabla L_t(\boldsymbol{w}_{t-1})\|}\right)-\nabla L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}\right)\|^2+\frac{1}{2}\|\nabla L(\boldsymbol{w}_{t-1})\|^2\\ \leq\ & \frac{\beta^2}{2}\mathbb{E}\left\|\rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{\|\nabla L_t(\boldsymbol{w}_{t-1})\|}-\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}\right\|^2+\frac{1}{2}\|\nabla L(\boldsymbol{w}_{t-1})\|^2\\ \leq\ & \rho^2\beta^2+\frac{1}{2}\|\nabla L(\boldsymbol{w}_{t-1})\|^2.\end{aligned}$$

The first inequality is by the Young's inequality. Also, it has been proven in Lemma A.7 that

$$\left\langle \nabla L_t(\boldsymbol{w}_{t-1}+\rho\frac{\nabla L(\boldsymbol{w}_{t-1})}{\|\nabla L(\boldsymbol{w}_{t-1})\|}), \nabla L(\boldsymbol{w}_{t-1})\right\rangle \geq \|\nabla L(\boldsymbol{w}_{t-1})\|^2-\beta\rho\|\nabla L(\boldsymbol{w}_{t-1})\|.$$

Combining the two inequalities together, we obtain the desired result. □

Lemma A.9.

$$\begin{aligned}&\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}} L(\hat{\boldsymbol{w}}_{t-1}),\right.\right.\\ &\left.\left.\nabla_{\hat{\boldsymbol{w}}_{t-2}}\left[L_{t-1}\left(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})\|}\right)-\nabla_{\boldsymbol{w}_{t-1}}\left[L_t\left(\boldsymbol{w}_{t-1}+\rho\frac{\nabla_{\hat{\boldsymbol{w}}_{t-1}} L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})\|}\right)\right]\right\rangle\right]\\ \leq\ & G\left(\beta^2\rho^2+\frac{5\beta^2\eta^2G^2}{2}\right)^{\frac{1}{2}}.\end{aligned}$$

Proof.

$$\begin{aligned}&\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], \nabla_{\hat{\boldsymbol{w}}_{t-2}} L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})\|})-\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})\|})]\right\rangle\right]\\ \leq\ & \left(\mathbb{E}\|\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})]\|^2\right)^{\frac{1}{2}}\left(\mathbb{E}\|\nabla_{\hat{\boldsymbol{w}}_{t-2}} L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})\|})-\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}} L_t(\boldsymbol{w}_{t-1})\|})]\|^2\right)^{\frac{1}{2}}\\ \leq\ & G\left(\left(\beta^2\rho^2+\frac{\beta^2}{2}\mathbb{E}\left[\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}} L_{t-1}(\boldsymbol{w}_{t-2})\|})-\eta\nabla_{\hat{\boldsymbol{w}}_{t-3}}[L_{t-2}(\hat{\boldsymbol{w}}_{t-3}+\rho\frac{\nabla_{\boldsymbol{w}_{t-3}} L_{t-2}(\boldsymbol{w}_{t-3})}{\|\nabla_{\boldsymbol{w}_{t-3}} L_{t-2}(\boldsymbol{w}_{t-3})\|})\|^2\right]\right)^2\right)^{\frac{1}{2}}\\ \leq\ & G\left(\beta^2\rho^2+\frac{5\beta^2\eta^2G^2}{2}\right)^{\frac{1}{2}}.\end{aligned}$$

The first inequality uses the Cauchy-Schwartz inequality. The last inequality uses the triangle inequality. □

Lemma A.10. *With assumptions 4.3, 4.4 and 4.5,*

$$\begin{aligned}&\mathbb{E}\left\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})-\eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\|^2\\ \leq\ & \frac{4\eta^2\beta^2\rho^2}{b}+\frac{4\eta^2\sigma^2}{b}+\eta^2G^2.\end{aligned}$$

Proof. This can be done by using the triangle inequality:

$$\begin{aligned}&\mathbb{E}\left\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})-\eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\|^2\\ \leq\ & \mathbb{E}\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})\|^2+\mathbb{E}\|\eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\|^2\\ \leq\ & 4\eta^2\mathbb{E}\left\|\nabla_{\hat{\boldsymbol{w}}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1})\|})-\nabla_{\boldsymbol{w}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1})\|^2+\eta^2\mathbb{E}\|\nabla_{\hat{\boldsymbol{w}}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1})-\nabla_{\hat{\boldsymbol{w}}_{t-1}}L(\hat{\boldsymbol{w}}_{t-1})\right\|^2\\ &+\eta^2G^2\\ \leq\ & \frac{4\eta^2\beta^2\mathbb{E}\left[\|\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|}\|^2\right]}{b}+\frac{4\eta^2\sigma^2}{b}+\eta^2G^2\\ \leq\ & \frac{4\eta^2\beta^2\rho^2}{b}+\frac{4\eta^2\sigma^2}{b}+\eta^2G^2.\end{aligned}$$

□

Lemma A.11. *If* $\rho<\frac{1}{2\beta}$,

$$\mathbb{E}[L(\boldsymbol{w}_{t-1})]\leq\mathbb{E}[L(\hat{\boldsymbol{w}}_t)]+\frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}||^2+\eta_tG^2.$$

Proof. Based on the update scheme of Opt-SAM, $\boldsymbol{w}_{t-1}=\hat{\boldsymbol{w}}_t+\eta_t\mathbf{g}_{t-1}$. By the smoothness assumption on Assumption 4.4,

$$L(\boldsymbol{w}_{t-1})-L(\hat{\boldsymbol{w}}_t)-\eta_t\langle\nabla L(\hat{\boldsymbol{w}}_t),\nabla L_t(\boldsymbol{w}_t)\rangle\leq\frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}||^2+\eta_tG^2.$$

Also, based on the Young's inequality and assumption 4.5,

$$\mathbb{E}\left[\langle\nabla L(\boldsymbol{w}_{t-1}+\eta_t\mathbf{g}_{t-1}),\nabla L_t(\boldsymbol{w}_{t-1})\rangle\right]\leq G^2$$

Therefore,

$$\mathbb{E}[L(\boldsymbol{w}_{t-1})]\leq\mathbb{E}[L(\hat{\boldsymbol{w}}_t)]+\frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}||^2+\eta_tG^2.$$

□

Lemma A.12. *With assumptions 4.3, 4.4 and 4.5, we have:*

$$\begin{aligned}\frac{1}{2}\eta\mathbb{E}\|\nabla L(\hat{\boldsymbol{w}}_{t-1})\|^2\ \leq\ & \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})]-\mathbb{E}[L(\hat{\boldsymbol{w}}_t)]+\eta\beta\rho E\|\nabla L(\hat{\boldsymbol{w}}_t)\|+\eta\rho^2\beta^2\\ &+\eta\left[G\left(\beta^2\rho^2+\frac{5\beta^2\eta^2G^2}{2}\right)^{\frac{1}{2}}\right]+\frac{\beta}{2}\left(\frac{4\eta^2\beta^2\rho^2}{b}+4\eta^2\sigma^2+\eta^2G^2\right).\end{aligned}$$

Proof. Using the definition of smoothness on Assumption (4.4) and taking the expectation on both sides, we have:

$$\begin{aligned}
&\mathbb{E}[L(\hat{\boldsymbol{w}}_t)] \leq \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] \\
&-\eta\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], 2\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|}) - \nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\boldsymbol{w}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\rangle\right] \\
&+\frac{\beta}{2}\mathbb{E}\left\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|}) - \eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\|^2 \\
\leq\ & \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] - \eta E\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})\right\rangle\right] \\
&+\eta\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], \nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_t(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|}) - \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\boldsymbol{w}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\hat{\boldsymbol{w}}_{t-1})\|})]\right\rangle\right] \\
&+\frac{\eta^2\beta}{2}\mathbb{E}\left\|2\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|}) - \nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\|^2.
\end{aligned}$$

Using Lemmas A.10, A.9 and A.8 to the above RHS,

$$\begin{aligned}
&\mathbb{E}[L(\hat{\boldsymbol{w}}_t)] \leq \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] - \eta\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})\right\rangle\right] \\
&+\eta\mathbb{E}\left[\left\langle \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L(\hat{\boldsymbol{w}}_{t-1})], \nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|}) - \nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|})]\right\rangle\right] \\
&+\frac{\beta}{2}\mathbb{E}\left\|2\eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}[L_t(\hat{\boldsymbol{w}}_{t-1}+\rho\frac{\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}L_t(\boldsymbol{w}_{t-1})\|}) - \eta\nabla_{\hat{\boldsymbol{w}}_{t-2}}[L_{t-1}(\hat{\boldsymbol{w}}_{t-2}+\rho\frac{\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})}{\|\nabla_{\boldsymbol{w}_{t-2}}L_{t-1}(\boldsymbol{w}_{t-2})\|})\right\|^2 \\
\leq\ & \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] - \eta\mathbb{E}\left[\|\nabla L(\hat{\boldsymbol{w}}_{t-1})\|^2 - \beta\rho\|\nabla L(\hat{\boldsymbol{w}}_{t-1})\| - \rho^2\beta^2\right] \\
&+\eta\mathbb{E}\left[G\left(\beta^2\rho^2 + \frac{5\beta^2\eta^2G^2}{2}\right)^{\frac{1}{2}}\right] + \frac{\beta\eta^2}{2}\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right).
\end{aligned}$$

This can then be simplified as:

$$\begin{aligned}
\frac{1}{2}\eta\mathbb{E}\|\nabla L(\hat{\boldsymbol{w}}_{t-1})\|^2 \quad\leq\quad & \mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] - \mathbb{E}[L(\hat{\boldsymbol{w}}_t)] + \eta\beta\rho\mathbb{E}\|\nabla L(\hat{\boldsymbol{w}}_t)\| + \eta\rho^2\beta^2 \\
&+\eta\left[G\left(\beta^2\rho^2 + \frac{5\beta^2\eta^2G^2}{2}\right)^{\frac{1}{2}}\right] + \frac{\beta}{2}\left(\frac{4\eta^2\beta^2\rho^2}{b} + 4\eta^2\sigma^2 + \eta^2G^2\right),
\end{aligned}$$

which is the desired result. □

Now, we state the proof for Theorem 4.7:

Proof. Using Lemma A.12, and with $\rho_t = \min\left(\frac{1}{\sqrt{T}}, \frac{1}{\beta}\right)$, $\eta_t = \min\left(\frac{1}{\sqrt{T}}, \frac{1}{\beta}\right)$, and the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have

$$\begin{aligned}
\mathbb{E}\|\nabla L(\hat{\boldsymbol{w}}_{t-1})\|^2 \quad\leq\quad & \frac{2[\mathbb{E}[L(\hat{\boldsymbol{w}}_{t-1})] - \mathbb{E}[L(\hat{\boldsymbol{w}}_t)]}{\eta} + 2\beta\rho G + 2\rho^2\beta^2 + \\
& 2\left[\beta\rho + \frac{5\beta\eta G}{2}\right] + 2\eta\beta\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right).
\end{aligned}$$

Telescoping from 1 to T, and substitute η_T to $\frac{1}{\sqrt{T}}$, we have:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(\hat{\boldsymbol{w}}_t)\|^2 \quad\leq\quad & \frac{2(\mathbb{E}[L(\hat{\boldsymbol{w}}_0)] - \mathbb{E}[L(\hat{\boldsymbol{w}}_{t+1})])}{\sqrt{T}\eta} + \frac{2\beta G}{\sqrt{T}} + \frac{2\rho^2\beta^2}{T} \\
& + \frac{2\beta}{\sqrt{T}} + \frac{5\beta G}{\sqrt{T}} + \frac{2\beta\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right)}{T} \\
\leq\quad & O\left(\frac{1}{\sqrt{T}b}\right).
\end{aligned}$$

We obtain:

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

Finally, using Lemma A.11, we have:

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\boldsymbol{w}_t}L(\boldsymbol{w}_t)\|^2 \leq \frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\hat{\boldsymbol{w}}_t}L(\hat{\boldsymbol{w}}_t)\|^2 + \frac{\beta+\sqrt{T}G^2}{T} = O\left(\frac{1}{\sqrt{T}b}\right).$$

□

A.5. Proof of Theorem 4.8

First, we prove the convergence of a surrogate update scheme, named Surrogate Lookahead-SAM, which will be used in the proof of AO-SAM:

$$\hat{\boldsymbol{w}}_t = \boldsymbol{w}_{t-1} - \eta_t\nabla_{\boldsymbol{w}_{t-1}}\frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w}_{t-1}), \tag{33}$$

$$\hat{\boldsymbol{\epsilon}}_t = \frac{\rho\nabla_{\boldsymbol{w}_{t-1}}\frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w}_{t-1})}{\|\nabla_{\boldsymbol{w}_{t-1}}\frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w}_{t-1})\|}, \tag{34}$$

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t\nabla_{\hat{\boldsymbol{w}}_t}\frac{1}{b}\sum_{i\in I_t}\ell_i(\hat{\boldsymbol{w}}_t+\hat{\boldsymbol{\epsilon}}_t). \tag{35}$$

Let $L_t(\boldsymbol{w}) := \frac{1}{b}\sum_{i\in I_t}\ell_i(\boldsymbol{w})$ be the mini-batch version of $L(\boldsymbol{w})$ at epoch t. Let $\boldsymbol{w}_{t-1/2} := \hat{\boldsymbol{w}}_t + \hat{\boldsymbol{\epsilon}}_t = \boldsymbol{w}_{t-1} + \rho\frac{\nabla L_t(\boldsymbol{w}_{t-1})}{||\nabla L_t(\boldsymbol{w}_{t-1})||} - \eta\nabla L_t(\boldsymbol{w}_{t-1})$. Note that the update scheme of Surrogate Lookahead-SAM can be rewritten as: $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta_t\nabla L_t(\boldsymbol{w}_{t-1/2})$. In the following, we assume that $\rho, \eta, \beta > 0$.

Lemma A.13.

$$\left\langle\nabla L\left(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w})\right),\nabla L(\boldsymbol{w})\right\rangle \geq (1+\beta\eta)\|\nabla L(\boldsymbol{w})\|^2 - \beta\rho\|\nabla L(\boldsymbol{w})\|.$$

Proof.

$$\begin{aligned}
&\left\langle\nabla L(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w})),\nabla L(\boldsymbol{w})\right\rangle \\
&= \langle\nabla L(\boldsymbol{w}+(\rho-\eta\|\nabla L(\boldsymbol{w})\|)\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})-\nabla L(\boldsymbol{w}),\nabla L(\boldsymbol{w})\rangle + \|\nabla L(\boldsymbol{w})\|^2 \\
&= \frac{\|\nabla L(\boldsymbol{w})\|}{\rho-\eta\|\nabla L(\boldsymbol{w})\|}\langle\nabla L(\boldsymbol{w}+(\rho-\eta\|\nabla L(\boldsymbol{w})\|)\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|})-\nabla L(\boldsymbol{w}),\frac{(\rho-\eta\|\nabla L(\boldsymbol{w})\|)}{\|\nabla L(\boldsymbol{w})\|}\nabla L(\boldsymbol{w})\rangle \\
&\quad +\|\nabla L(\boldsymbol{w})\|^2 \\
&\overset{(a)}{\geq} (1-\frac{\beta(\rho-\eta\|\nabla L(\boldsymbol{w})\|)}{\|\nabla L(\boldsymbol{w})\|})\|\nabla L(\boldsymbol{w})\|^2 \\
&\geq (1+\eta\beta)\|\nabla L(\boldsymbol{w})\|^2 - \beta\rho\|\nabla L(\boldsymbol{w})\|.
\end{aligned}$$

The first equation is based on the fact that $\langle\nabla L(w),\nabla L(w)\rangle = \|\nabla L(w)\|^2$, while (a) uses Lemma 7 in (Andriushchenko & Flammarion, 2022). □

Lemma A.14.

$$\mathbb{E}\langle\nabla L_t(\boldsymbol{w}+\rho\nabla L_t(\boldsymbol{w})/\|\nabla L_t(\boldsymbol{w})\|-\eta\nabla L_t(\boldsymbol{w})),\nabla L(\boldsymbol{w})\rangle \geq (\frac{1}{2}+\eta\beta)\|\nabla L(\boldsymbol{w})\|^2 - \beta\rho\|\nabla L(\boldsymbol{w})\| - \frac{\beta^2\sigma^2\eta^2}{2b} - \rho^2\beta^2.$$

Proof. Note that

$$\begin{aligned}
&\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\tfrac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w})\right),\nabla L(\boldsymbol{w})\right\rangle\\
&=\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\tfrac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w})\right)-\nabla L_t(\boldsymbol{w}+\rho\tfrac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w})),\nabla L(\boldsymbol{w})\right\rangle\\
&-\left\langle -\nabla L_t(\boldsymbol{w}+\rho\tfrac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w})),\nabla L(\boldsymbol{w})\right\rangle.
\end{aligned}$$

In the following, we bound the first term and second term of the RHS separately. For the first term,

$$\begin{aligned}
&-\mathbb{E}\left\langle \nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w})\right)-\nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(w)\|}-\eta\nabla L(\boldsymbol{w})),\nabla L(w)\right\rangle\\
&\overset{(a)}{\leq}\ \frac{1}{2}\mathbb{E}\|\nabla L_t\left(\boldsymbol{w}+\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w})\right)-\nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w}))\|^2\\
&\quad+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2\\
&\leq\ \frac{\beta^2}{2}\mathbb{E}\|\rho\frac{\nabla L_t(\boldsymbol{w})}{\|\nabla L_t(\boldsymbol{w})\|}-\eta\nabla L_t(\boldsymbol{w})-(\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w}))\|^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2\\
&\overset{(b)}{\leq}\ \frac{\beta^2\sigma^2\eta^2}{2b}+\rho^2\beta^2+\frac{1}{2}\|\nabla L(\boldsymbol{w})\|^2.
\end{aligned}$$

(a) is based on the Young's inequality, (b) is using the triangle inequality and Assumption 4.4.

For the second term, on using Lemma A.13,

$$\begin{aligned}
&\mathbb{E}\left\langle \nabla L_t(\boldsymbol{w}+\rho\frac{\nabla L(\boldsymbol{w})}{\|\nabla L(\boldsymbol{w})\|}-\eta\nabla L(\boldsymbol{w})),\nabla L(\boldsymbol{w})\right\rangle\\
&\geq\ (1+\beta\eta)\|\nabla L(\boldsymbol{w})\|^2-\beta\rho\|\nabla L(\boldsymbol{w})\|.
\end{aligned}$$

Combining them together, we obtain the result. □

Lemma A.15. *With assumptions 4.3, 4.4 and 4.5, and also assume $\eta\leq\frac{1}{2\beta}$, we have:*

$$\begin{aligned}
&\eta\left(\eta\beta-\eta^3\beta^3-2\eta\beta\left(\frac{1}{2}+\eta\beta\right)+\left(\frac{1}{2}+\eta\beta\right)\right)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2\\
&\leq\ \mathbb{E}L(\boldsymbol{w}_t)-\mathbb{E}L(\boldsymbol{w}_{t+1})+\eta^2\beta^3\rho^2+\eta(1-2\eta\beta)\beta\rho G+\eta(1-2\eta\beta)\frac{\beta^2\sigma^2}{2b}\\
&\quad+2\rho^2\beta^2\eta(1-2\eta\beta)+\eta^2\beta\frac{\sigma^2}{b}.
\end{aligned}$$

Proof. Using the property of smooth function L (Assumption (4.4)), we have:

$$\begin{aligned}
\mathbb{E}L(\boldsymbol{w}_{t+1})\ &\leq\ \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L_t(\boldsymbol{w}_{t-1/2})\|^2\\
&\leq\ \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L_t(\boldsymbol{w}_{t-1/2})-\nabla L(\boldsymbol{w}_{t-1/2})\|^2\\
&\quad+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2\\
&\leq\ \mathbb{E}L(\boldsymbol{w}_t)-\eta\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}),\nabla L(\boldsymbol{w}_t)\rangle+\eta^2\beta\frac{\sigma^2}{2b}+\frac{\eta^2\beta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2.
\end{aligned}$$

The first inequality is using the property of smooth function (Assumption (4.4)) and take the expectation on both sides.

Then,

$$\begin{aligned}
\mathbb{E}L(\boldsymbol{w}_{t+1}) &\overset{(a)}{\leq} \mathbb{E}L(\boldsymbol{w}_t) - \eta^2\beta\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \eta^2\beta\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 \\
&\quad -\eta(1-2\eta\beta)\mathbb{E}\langle\nabla L(\boldsymbol{w}_{t-1/2}), \nabla L(\boldsymbol{w}_t)\rangle + \eta^2\beta\frac{\sigma^2}{2b} \\
&\overset{(b)}{\leq} \mathbb{E}L(\boldsymbol{w}_t) - \eta^2\beta\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \eta^2\beta^3\mathbb{E}\|\boldsymbol{w}_{t-1/2} - \boldsymbol{w}_t\|^2 \\
&\quad -\eta(1-2\eta\beta)\left[-\beta\rho\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\| + \left(\frac{1}{2}+\eta\beta\right)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 - \frac{\beta^2\sigma^2\eta^2}{2b} - \rho^2\beta^2\right] \\
&\quad +\eta^2\beta\frac{\sigma^2}{2b} \\
&\leq \mathbb{E}L(\boldsymbol{w}_t) - \eta^2\beta\mathbb{E}\|\nabla L(w_t)\|^2 + \eta^2\beta^3\rho^2 + \eta^4\beta^3\,\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&\quad +\eta(1-2\eta\beta)\beta\rho\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\| - \eta(1-2\eta\beta)\left(\frac{1}{2}+\eta\beta\right)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&\quad +\eta^3(1-2\eta\beta)\frac{\beta^2\sigma^2}{2b} + 2\rho^2\beta^2\eta(1-2\eta\beta) + \eta^2\beta\frac{\sigma^2}{b}.
\end{aligned}$$

(a) is by using the trick: $\|\nabla L(\boldsymbol{w}_{t-1/2})\|^2 = -\|\nabla L(w_t)\|^2 + \|\nabla L(\boldsymbol{w}_{t-1/2}) - \nabla L(\boldsymbol{w}_t)\|^2 + 2\langle\nabla L(\boldsymbol{w}_{t-1/2}), \nabla L(\boldsymbol{w}_t)\rangle$. (b) is by using Lemma A.14. The last inequality is based on the fact that $\beta^2\mathbb{E}\|\boldsymbol{w}_{t-1/2} - \boldsymbol{w}_t\|^2 \leq \beta^2\mathbb{E}\|(\rho - \eta\|\nabla L(\boldsymbol{w}_t)\|)\frac{\nabla L(\boldsymbol{w}_t)}{\|\nabla L(\boldsymbol{w}_t)\|}\|^2 \leq \beta^2\rho^2 + \eta^2\beta^2\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2$, and the assumption $\eta \leq \frac{1}{2\beta}$. Finally, after simplification, we obtain the result. □

Theorem A.16. *Assume that* $\eta_t = \min\left(\frac{1}{2\beta}, \frac{1}{\sqrt{T}}\right)$, $\rho = \frac{1}{\sqrt{T}}$, *then Surrogate Lookahead-SAM satisfies* $\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\|\nabla_{\boldsymbol{w}_t} L(\boldsymbol{w}_t)\|^2 = O\left(\frac{1}{\sqrt{T}b}\right)$.

Proof. With $\eta_t = \min\left(\frac{1}{2\beta}, \frac{2.99}{4\beta}\right) = \frac{1}{2\beta}$, we have $\left(\eta\beta - \eta\beta^3 - 2\eta\beta\left(\frac{1}{2}+\eta\beta\right) + \left(\frac{1}{2}+\eta\beta\right)\right) > 0$. By further using Lemma A.15 and the above analysis, there exists a positive constant C such that:

$$\begin{aligned}
C\eta\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 &\leq \mathbb{E}L(\boldsymbol{w}_t) - \mathbb{E}L(\boldsymbol{w}_{t+1}) + \eta^3(1-2\eta\beta)\frac{\beta^2\sigma^2}{2b} + 2\rho^2\eta(1-2\eta\beta) + \eta^2\beta\frac{\sigma^2}{b} \\
&\quad +\eta^2\beta^3\rho^2 + \eta\beta\rho G.
\end{aligned}$$

Setting $\rho = \frac{1}{\sqrt{T}}$ and $\eta_t = \min\left(\frac{1}{2\beta}, \frac{1}{\sqrt{T}}\right)$. By telescoping from $t = 1$ to T, and divide by T,

$$\begin{aligned}
&\frac{C}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&\leq \frac{L(\boldsymbol{w}_0) - \mathbb{E}L(\boldsymbol{w}_{T+1})}{\eta T} + \eta^2\frac{\beta^2\sigma^2}{2b} + 2\rho^2(1-2\eta\beta) + \eta\beta\frac{\sigma^2}{b} + \eta\beta^3\rho^2 + \beta\rho G \\
&= \frac{L(\boldsymbol{w}_0) - \mathbb{E}L(\boldsymbol{w}_{T+1})}{\sqrt{T}} + \frac{\left(\beta^2\sigma^2 b/\sqrt{T} + 2b/\sqrt{T} + 2\beta\sigma^2 + 2b\beta^2/T + 2\beta G\right)}{2\sqrt{T}b} \\
&= O\left(\frac{1}{\sqrt{T}b}\right).
\end{aligned}$$

□

Now we state the proof of Theorem 4.8:

Proof. Note that for AO-SAM,

1. If $\|\frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)\|^2 < \mu_t + c_t\sigma_t$, then it is the SGD update scheme on epoch t.

2. If $\mathbf{g}_{t-1}$ in step 7 of Algorithm 2 is the gradient obtained by Opt-SAM, and $\|\frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t\sigma_t$, i.e., $g_{t-1} = \nabla_{\hat{\boldsymbol{w}}_{t-1}}\left[\frac{1}{b}\sum_{i\in I_{t-1}}\ell_i\left(\hat{\boldsymbol{w}}_{t-1}+\hat{\boldsymbol{\epsilon}}_{t-1}\right)\right]$, then it is the Opt-SAM update scheme on epoch t.

3. If $\mathbf{g}_{t-1}$ in step 7 of Algorithm 2 is the gradient obtained by SGD, and $\|\frac{1}{b}\sum_{i\in I_t}\nabla_{\boldsymbol{w}_t}\ell_i(\boldsymbol{w}_t)\|^2 \geq \mu_t + c_t\sigma_t$, then it is Surrogate Lookahead-SAM update scheme on epoch t. The reason is that when the last step is SGD, $\mathbf{g}_{t-1} = \nabla_{\boldsymbol{w}_{t-1}}L_t\left(\boldsymbol{w}_{t-1}\right)$, by following steps 7-8 in Algorithm 2, Therefore,

$$\begin{aligned}\boldsymbol{w}_t &= \boldsymbol{w}_{t-1} - \eta_t\mathbf{g}_t \\ &= \boldsymbol{w}_{t-1} - \eta_t\nabla_{\boldsymbol{w}_t}\frac{1}{b}\sum_{i\in I_t}\ell_i\left(\boldsymbol{w}_{t-1} - \eta_t\nabla_{\boldsymbol{w}_{t-1}}L_t\left(\boldsymbol{w}_{t-1}\right) + \frac{\rho\nabla_{\boldsymbol{w}_{t-1}}\frac{1}{b}\sum_{i\in I_t}\ell_i\left(\boldsymbol{w}_{t-1}\right)}{\left\|\nabla_{\boldsymbol{w}_{t-1}}\frac{1}{b}\sum_{i\in I_t}\ell_i\left(\boldsymbol{w}_{t-1}\right)\right\|}\right),\end{aligned}$$

which is exactly Surrogate Lookahead-SAM.

Therefore, our goal is to combine these three different schemes together. Define $\zeta_t^1 \in \{0,1\}$, $\zeta_t^2 \in \{0,1\}$ and $\zeta_t^3 \in \{0,1\}$ to indicate which update scheme is used in epoch t: $\zeta_t^1 = 1$ means that Opt-SAM is used; $\zeta_t^2 = 1$ means that SGD is used; while $\zeta_t^3 = 1$ means that Surrogate Lookahead-SAM is used. Obviously, $\zeta_t^1 + \zeta_t^2 + \zeta_t^3 = 1$.

Recall that in Theorem A.16, we have:

$$\begin{aligned}C\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \quad\leq\quad & \mathbb{E}L(\boldsymbol{w}_t) - \mathbb{E}L(\boldsymbol{w}_{t+1}) + \eta^2(1-2\eta\beta)\frac{\beta^2\sigma^2}{2} + 2\rho^2\eta(1-2\eta\beta) + \eta^2\beta\frac{\sigma^2}{b} \\ & +\eta^2\beta^3\rho^2 + \eta\beta\rho G.\end{aligned}$$

Recall that in Theorem 4.7, we have:

$$\begin{aligned}\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1})\|^2 \quad\leq\quad & \frac{2[\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)]]}{\eta} + 2\beta\rho G + 2\rho^2\beta^2 \\ & 2\left[\beta\rho + \frac{5\beta\eta G}{2}\right] + 2\eta\beta\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right) + \frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}\|^2 + \eta_t G^2.\end{aligned}$$

Also, from (2.9) in Theorem 2.1 of (Ghadimi & Lan, 2013),

$$\left(\eta_t - \frac{L}{2}\eta_t^2\right)\|\nabla L(\boldsymbol{w}_{t-1})\|^2 \leq L(\boldsymbol{w}_{t-1}) - L(\boldsymbol{w}_t) + \frac{\beta}{2}\eta_t^2\sigma^2$$

for SGD. By simple calculation and using assumption 4.5, we have

$$\|\nabla L(\boldsymbol{w}_{t-1})\|^2 \leq \frac{\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)]}{\eta_t} + \frac{\beta}{2}\eta_t\sigma^2 + \frac{\beta}{2}\eta_t G^2.$$

Therefore,

$$\begin{aligned}&\mathbb{E}\|\nabla L(\boldsymbol{w}_{t-1})\|^2 \\ &\leq\ \zeta_t^1\left[\frac{\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)]}{\eta} + 2\beta\rho G + 2\rho^2\beta^2 + 2\left[\beta\rho + \frac{5\beta\eta G}{2}\right] + 2\eta\beta\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right) + \frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}\|^2 + \eta_t G^2\right] \\ &\quad +\zeta_t^2\left[\frac{\mathbb{E}[L(\boldsymbol{w}_{t-1})] - E[L(\boldsymbol{w}_t)]}{\eta} + \frac{\beta}{2}\eta\sigma^2 + \frac{\beta}{2}\eta G^2\right] \\ &\quad +\frac{\zeta_t^3}{C}\left[\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)] + \eta(1-2\eta\beta)\frac{\beta^2\sigma^2}{2} + 2\rho^2\eta^2(1-2\eta\beta) + \eta^2\beta\frac{\sigma^2}{b} + \eta^2\beta^3\rho^2 + G\beta\rho\eta\right] \\ &\leq\ \left[\frac{\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)]}{\eta} + 2\beta\rho G + 2\rho^2\beta^2 + 2\left(\beta\rho + \frac{5\beta\eta G}{2}\right) + 2\eta\beta\left(\frac{4\beta^2\rho^2}{b} + 4\sigma^2 + G^2\right) + \frac{\beta\eta_t^2}{2}||\mathbf{g}_{t-1}\|^2 + \eta_t G^2\right] \\ &\quad +\left[\frac{\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)]}{\eta} + \frac{\beta}{2}\eta\sigma^2 + \frac{\beta}{2}\eta G^2\right] \\ &\quad +\frac{1}{C}\left[\mathbb{E}[L(\boldsymbol{w}_{t-1})] - \mathbb{E}[L(\boldsymbol{w}_t)] + \eta^2(1-2\eta\beta)\frac{\beta^2\sigma^2}{2} + 2\rho^2\eta(1-2\eta\beta) + \eta^2\beta\frac{\sigma^2}{b} + \eta^2\beta^3\rho^2 + G\beta\rho\eta\right].\end{aligned}$$

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 $\pm$ 0.19	100.0	87.38 $\pm$ 0.61	100.0	83.22 $\pm$ 0.41	100.0	71.03 $\pm$ 0.88	100.0
	GSAM (Zhuang et al., 2022)	94.54 $\pm$ 0.18	100.0	91.72 $\pm$ 0.05	100.0	87.70 $\pm$ 0.02	100.0	24.70 $\pm$ 10.69	100.0
	Opt-SAM	**95.12** $\pm$ 0.12	100.0	92.16 $\pm$ 0.35	100.0	88.45 $\pm$ 0.53	100.0	77.47 $\pm$ 0.65	100.0
	SS-SAM (Zhao, 2022)	94.61 $\pm$ 0.16	60.0	91.81 $\pm$ 0.13	60.0	78.67 $\pm$ 0.42	60.0	62.94 $\pm$ 1.01	60.0
	AE-SAM (Jiang et al., 2023)	92.13 $\pm$ 0.14	61.4	86.02 $\pm$ 0.62	61.4	75.95 $\pm$ 1.30	61.4	67.28 $\pm$ 1.66	61.4
	AO-SAM	95.02 $\pm$ 0.04	61.2	**92.62** $\pm$ 0.18	61.3	**89.36** $\pm$ 0.12	61.2	**78.12** $\pm$ 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 $\pm$ 0.09	100.0	88.83 $\pm$ 0.11	100.0	83.90 $\pm$ 0.56	100.0	75.64 $\pm$ 0.75	100.0
	GSAM (Zhuang et al., 2022)	94.12 $\pm$ 0.09	100.0	91.74 $\pm$ 0.05	100.0	89.23 $\pm$ 0.06	100.0	31.16 $\pm$ 2.77	100.0
	Opt-SAM	95.25 $\pm$ 0.04	100.0	**92.11** $\pm$ 0.07	100.0	88.36 $\pm$ 0.22	100.0	77.61 $\pm$ 0.39	100.0
	SS-SAM (Zhao, 2022)	95.03 $\pm$ 0.23	60.0	90.59 $\pm$ 0.30	60.0	87.22 $\pm$ 0.46	60.0	48.89 $\pm$ 1.02	60.0
	AE-SAM (Jiang et al., 2023)	92.04 $\pm$ 0.27	61.3	86.83 $\pm$ 0.49	61.3	73.90 $\pm$ 0.44	61.2	67.64 $\pm$ 1.34	61.3
	AO-SAM	**95.32** $\pm$ 0.12	61.2	91.73 $\pm$ 0.65	61.2	**89.40** $\pm$ 0.44	61.2	**77.78** $\pm$ 0.84	61.2
WideResNet-28-10	ERM	90.07 $\pm$ 0.36	0.0	86.02 $\pm$ 0.33	0.0	80.98 $\pm$ 0.52	0.0	67.67 $\pm$ 0.72	0.0
	SAM (Foret et al., 2021)	94.47 $\pm$ 0.12	100.0	91.74 $\pm$ 0.04	100.0	88.35 $\pm$ 0.21	100.0	71.37 $\pm$ 1.55	100.0
	ESAM (Du et al., 2022a)	95.09 $\pm$ 0.04	100.0	89.16 $\pm$ 0.21	100.0	42.64 $\pm$ 0.55	100.0	20.14 $\pm$ 0.69	100.0
	ASAM (Kwon et al., 2021)	91.25 $\pm$ 0.16	100.0	88.08 $\pm$ 0.07	100.0	83.45 $\pm$ 0.12	100.0	71.44 $\pm$ 0.46	100.0
	GSAM (Zhuang et al., 2022)	95.19 $\pm$ 0.07	100.0	92.04 $\pm$ 0.07	100.0	88.11 $\pm$ 0.33	100.0	57.42 $\pm$ 4.99	100.0
	Opt-SAM	95.31 $\pm$ 0.06	100.0	92.67 $\pm$ 0.13	100.0	88.37 $\pm$ 0.58	100.0	**77.86** $\pm$ 1.83	100.0
	SS-SAM (Zhao, 2022)	94.47 $\pm$ 0.09	60.0	91.90 $\pm$ 0.11	60.0	88.43 $\pm$ 0.37	60.0	74.64 $\pm$ 0.79	60.0
	AE-SAM (Jiang et al., 2023)	93.49 $\pm$ 0.14	61.3	90.36 $\pm$ 0.12	61.3	85.95 $\pm$ 0.47	61.3	71.21 $\pm$ 1.56	61.3
	AO-SAM	**95.52** $\pm$ 0.24	61.1	**92.68** $\pm$ 0.10	61.2	**89.29** $\pm$ 0.28	61.2	77.13 $\pm$ 0.72	61.2

The second equation is due to the fact that $\zeta_1, \zeta_2, \zeta_3 \leq 1$. Note that the summation of RHS from $t = 0$ to T is exactly the summation of Surrogate Lookahead-SAM, Opt-SAM, and SGD together, which has been shown to have $O(\frac{1}{\sqrt{Tb}})$, $O(\frac{1}{\sqrt{Tb}})$, and $O(\frac{1}{\sqrt{T}})$ rates in Theorem A.16, Theorem 4.7, and Theorem 2.1 in (Ghadimi & Lan, 2013), respectively. As the finite summation of $O(\frac{1}{\sqrt{Tb}})$ is still $O(\frac{1}{\sqrt{Tb}})$, the result follows. □

B. Supplementary Experimental Results

Tables 8 and 9 show the testing accuracies on *CIFAR-10* and *CIFAR-100*, respectively, with label noise. As can be seen, AO-SAM and Opt-SAM outperform all baselines at all label noise ratios.

Table 9: Testing accuracy and fraction of SAM updates on *CIFAR-100* with different levels of label noise. All baseline results are obtained with the authors' provided code. 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	66.83 ± 0.21	0.0	54.58 ± 0.96	0.0	47.98 ± 0.36	0.0	26.21 ± 3.40	0.0
	SAM (Foret et al., 2021)	69.60 ± 0.19	100.0	59.85 ± 0.53	100.0	52.50 ± 0.25	100.0	23.79 ± 3.21	100.0
	ESAM (Du et al., 2022a)	75.33 ± 0.19	100.0	67.75 ± 0.83	100.0	4.79 ± 3.58	100.0	1.29 ± 0.10	100.0
	ASAM (Kwon et al., 2021)	67.76 ± 0.86	100.0	57.13 ± 0.06	100.0	48.69 ± 0.04	100.0	29.46 ± 0.10	100.0
	GSAM (Zhuang et al., 2022)	70.30 ± 0.32	100.0	61.15 ± 0.01	100.0	53.08 ± 0.05	100.0	6.42 ± 0.70	100.0
	Opt-SAM	75.45 ± 0.27	100.0	68.01 ± 0.19	100.0	56.63 ± 0.10	100.0	**29.77** ± 1.08	100.0
	SS-SAM (Zhao, 2022)	75.68 ± 0.62	60.0	64.72 ± 0.20	60.0	55.55 ± 1.49	60.0	23.90 ± 5.63	60.0
	AE-SAM (Jiang et al., 2023)	68.69 ± 0.35	61.4	57.35 ± 0.24	61.4	47.95 ± 1.01	61.4	27.11 ± 0.57	61.4
	AO-SAM	**75.69** ± 0.35	61.2	**68.35** ± 0.21	61.3	**56.95** ± 1.00	61.2	29.76 ± 1.21	61.3
ResNet-32	ERM	69.33 ± 0.24	0.0	55.77 ± 0.74	0.0	46.96 ± 0.93	0.0	25.67 ± 0.98	0.0
	SAM (Foret et al., 2021)	70.88 ± 0.32	100.0	60.40 ± 0.07	100.0	53.10 ± 0.36	100.0	10.66 ± 5.56	100.0
	ESAM (Du et al., 2022a)	77.09 ± 0.22	100.0	66.17 ± 1.78	100.0	3.02± 0.26	100.0	1.85± 0.73	100.0
	ASAM (Kwon et al., 2021)	69.64 ± 1.36	100.0	57.88 ± 0.61	100.0	48.79 ± 0.24	100.0	28.06± 1.05	100.0
	GSAM (Zhuang et al., 2022)	71.69 ± 0.04	100.0	63.23 ± 0.04	100.0	54.22 ± 0.51	100.0	3.03 ± 0.88	100.0
	Opt-SAM	78.05 ± 0.23	100.0	66.74 ± 0.19	100.0	**56.06** ± 0.13	100.0	29.55 ± 2.08	100.0
	SS-SAM (Zhao, 2022)	71.34 ± 0.32	60.0	61.45 ± 1.36	60.0	51.76 ± 0.04	60.0	13.96± 3.17	60.0
	AE-SAM (Jiang et al., 2023)	68.94 ± 0.12	61.3	58.41± 1.89	61.3	51.48± 1.08	61.2	28.44 ± 0.76	61.3
	AO-SAM	**78.11** ± 0.14	61.2	**68.71** ± 0.46	61.2	54.58 ± 0.30	61.2	**29.78**± 0.42	61.2
WideResNet-28-10	ERM	74.31 ± 0.61	0.0	62.31 ± 0.41	0.0	48.23 ± 0.92	0.0	29.96 ± 0.21	0.0
	SAM (Foret et al., 2021)	76.04 ± 0.38	100.0	64.65 ± 0.79	100.0	56.03 ± 0.76	100.0	29.48 ± 0.23	100.0
	ESAM (Du et al., 2022a)	80.06 ± 0.12	100.0	72.03 ± 0.79	100.0	9.75 ± 2.12	100.0	1.16 ± 0.08	100.0
	ASAM (Kwon et al., 2021)	74.37 ± 0.18	100.0	62.91 ± 0.71	100.0	51.35 ± 0.31	100.0	33.12 ± 0.16	100.0
	GSAM (Zhuang et al., 2022)	75.90 ± 0.06	100.0	64.57 ± 0.16	100.0	56.80 ± 0.39	100.0	11.72 ± 0.25	100.0
	Opt-SAM	80.14 ± 0.29	100.0	**72.79**± 0.51	100.0	**57.01**± 0.34	100.0	**36.33**± 1.68	100.0
	SS-SAM (Zhao, 2022)	75.48 ± 0.26	60.0	64.72 ± 0.25	60.0	54.83 ± 0.48	60.0	35.88 ± 3.23	60.0
	AE-SAM (Jiang et al., 2023)	75.46 ± 0.36	61.3	63.04 ± 0.68	61.3	52.29 ± 0.63	61.3	33.72 ± 0.62	61.3
	AO-SAM	**80.32** ± 0.07	61.2	72.10 ± 0.48	61.2	56.89 ± 0.30	61.2	36.03 ± 0.59	61.2