2025 Deep Learning Theory

# SAMPa: Sharpness-aware Minimization Parallelized

Yoseop Kim and Joohyuk Park

Department of Electrical Engineering, POSTECH

# Contents

- Introduction
- Background and Challenge of SAM
- SAM Parallelized (SAMPa) & Convergence Analysis
- Experiments

# Contents

# Generalization

- A DNN's proficiency in effectively processing and responding to new, previously unseen data originating from the same distribution as the training dataset
  - Excess risk

$$R(\hat{f}) - R(f_{\mathrm{GT}}) \leq \underbrace{R(\hat{f}) - R_n(\hat{f})}_{\text{Generalization}} + \underbrace{R_n(\hat{f}) - R_n(f_{\mathrm{ERM}})}_{\text{Optimization}} + \underbrace{R_n(f^\star) - R(f^\star)}_{\text{Generalization}} + \underbrace{R(f^\star) - R(f_{\mathrm{GT}})}_{\text{Approximation}}$$

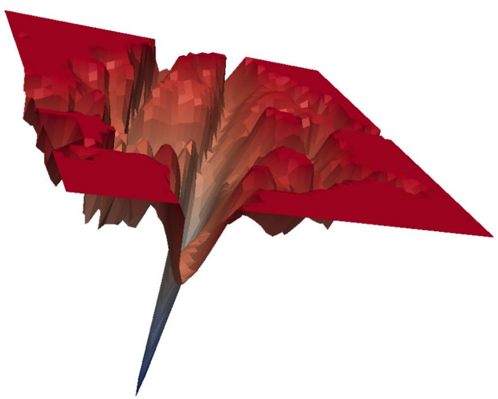
  - Classically handled via the uniform deviation

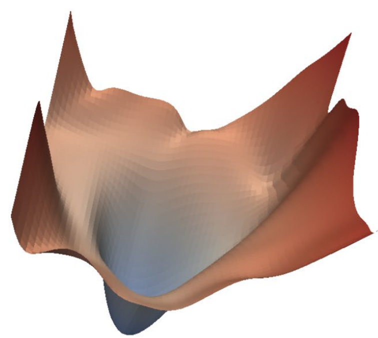
$$R(\hat{f}) - R_n(\hat{f}) + R_n(f^\star) - R(f^\star) \leq 2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)|$$

    - $\mathcal{F}$: Function space (expressible with MLP)

# Generalization

- Recent studies suggest that smoother loss landscapes lead to better generalization [Keskar et al., 2017, Jiang* et al., 2020]
  - Sharpness-aware minimization (SAM) has emerged as a promising optimization approach [Foret et al., 2021, Zheng et al., 2021, Wu et al., 2020b]
  - Seek flat minima by solving a **min-max optimization** problem
    - Inner maximizer quantifies the sharpness ($\nabla R_n(\hat{f})$)
    - Outer minimizer reduces training loss and sharpness





# Contents

# Sharpness-aware minimization (SAM)

- SAM attempts to enforce **small loss around the neighborhood** in the parameter space

$$\min_{x} \max_{\epsilon : \|\epsilon\| \leq \rho} f(x+\epsilon)$$

  - $x$: weight vector
  - $\rho$: radius of considered neighborhood

  - Inner maximization problem can be approximately solved as

$$\epsilon^\star = \arg\max_{\epsilon:\, \|\epsilon\| \leq \rho} f(x+\epsilon) \approx \arg\max_{\epsilon:\, \|\epsilon\| \leq \rho} \left(f(x) + \langle \nabla f(x), \epsilon \rangle\right) = \rho \frac{\nabla f(x)}{\|\nabla f(x)\|}$$

→ First-order Taylor approximation

# Sharpness-aware minimization (SAM)

- The objective function of SAM update

$$\min_{x} \; f\left(x + \rho \frac{\nabla f(x)}{\|\nabla f(x)\|}\right)$$

  - SAM first obtains the perturbed weight $\tilde{x} = x + \epsilon^{\star}$ by this approximated worst-case perturbation and then adopts the gradient of $\tilde{x}$ to update the original weight $x$

$$\tilde{x}_t = x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}, \qquad x_{t+1} = x_t - \eta_t \nabla f(\tilde{x}_t)$$

# Sharpness-aware minimization (SAM)

- Challenges
  - Although SAM and some variants achieve remarkable generalization improvement, they increase the computational overhead of the given base optimizers
  - Two forward-backward computations
    - Computing the perturbation: $\nabla f(x_t)$
    - Computing the update direction: $\nabla f(\tilde{x}_t)$
    - Two computations are not parallelizable
    - SAM doubles the **computational overhead** as well as **training time** compared to base optimizers (e.g., SGD)

$$\tilde{x}_t = x_t + \rho \frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}, \qquad x_{t+1} = x_t - \eta_t \nabla f(\tilde{x}_t)$$

# Contents

# SAM Parallelized (SAMPa)

- To break the sequential nature of SAM, we seek to replace the gradient $\nabla f(x_t)$

With another gradient $\nabla f(y_t)$ computed at some auxiliary sequence $(y_t)_{t\in\mathbb{N}}$

$$\tilde{x}_t = x_t + \rho \frac{\nabla f(y_t)}{\|\nabla f(y_t)\|},$$
$$y_{t+1} = x_t - \eta_t \nabla f(y_t),$$
$$x_{t+1} = x_t - \eta_t \nabla f(\tilde{x}_t)$$

- $\nabla f(\tilde{x}_t)$ and $\nabla f(y_{t+1})$ can be computed in parallel
- How to choose the auxiliary sequence $(y_t)_{t\in\mathbb{N}}$?
  - Difference $\|\nabla f(x_t) - \nabla f(y_t)\|$ can be controlled

# SAM Parallelized (SAMPa)

- Convergence analysis

**Lemma 4.3.** SAMPa satisfies the following descent inequality for $\rho > 0$ and a decreasing sequence $(\eta_t)_{t\in\mathbb{N}}$ with $\eta_t \in (0, \min\{1, c/L\})$ and $c \in (0,1)$

$$\mathcal{V}_{t+1} \leq \mathcal{V}_t - \eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2 + \eta_t^2 \rho^2 C$$

where $\mathcal{V}_t \triangleq f(x_t) + 0.5\,(1 - \eta_t L)\,\|\nabla f(x_t) - \nabla f(y_t)\|^2$ and $C = 0.5(L^2 + L^3 + \frac{1}{1-c^2} L^4)$

- Assumption 1: The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex
- Assumption 2: The operator $\nabla f : \mathbb{R}^d \to \mathbb{R}$ is $L$-Lipschitz with $L \in (0, \infty)$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

# Formulation of Lemma 4.3

**Assumptions:**

- **4.1 (Convexity):** The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex.
- **4.2 ($L$-Smoothness):** The gradient $\nabla f$ is $L$-Lipschitz continuous.

**Lemma 4.3 (Descent Inequality)** Let $\rho > 0$. For step sizes satisfying $\eta_t \in (0, \min\{1, c/L\})$ with $c \in (0, 1)$:

$$\mathcal{V}_{t+1} \leq \mathcal{V}_t - \eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2 + \eta_t^2 \rho^2 C$$

where the potential function is defined as $\mathcal{V}_t := f(x_t) + \frac{1}{2}(1 - \eta_t L) \|\nabla f(x_t) - \nabla f(y_t)\|^2$ and the constant $C$ is given by $C = \frac{1}{2}\left(L^2 + L^3 + \frac{L^4}{1-c^2}\right)$.

# Roadmap of the Proof

We derive the descent inequality in four logical steps:

1. **Step 1: Expansion & Decomposition**
   - Expand using smoothness and isolate the descent term.
   - Identify the problematic "Cross Term".
2. **Step 2: Handling the Cross Term via Auxiliary Sequence**
   - Introduce $y_t$ and apply **Convexity** and **Young's Inequality**.
   - Reverse-engineer $y_t$ to satisfy convergence requirements.
3. **Step 3: Ensuring Telescoping**
   - Design parameter $e$ to telescope the potential function.
   - **Correct the flaw** in the paper's ratio argument.
4. **Step 4: Lyapunov Function Derivation**
   - Combine all inequalities to construct the Lemma.

# Step 1.1: Primary Expansion via Smoothness

We start with the $L$-smoothness inequality and the SAMPa update rule $x_{t+1} = x_t - \eta_t \nabla f(\tilde{x}_t)$:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\ &= f(x_t) + \langle \nabla f(x_t), -\eta_t \nabla f(\tilde{x}_t) \rangle + \frac{L}{2} \|-\eta_t \nabla f(\tilde{x}_t)\|^2 \\ &= f(x_t) - \eta_t \langle \nabla f(x_t), \nabla f(\tilde{x}_t) \rangle + \frac{\eta_t^2 L}{2} \|\nabla f(\tilde{x}_t)\|^2 \end{aligned}$$

This equation depends on the perturbed gradient $\nabla f(\tilde{x}_t)$, which hinders direct convergence analysis.

# Step 1.2: Gradient Decomposition Identity

To isolate the descent direction, we decompose the perturbed gradient:

$$\nabla f(\tilde{x}_t) = \nabla f(x_t) + \underbrace{(\nabla f(\tilde{x}_t) - \nabla f(x_t))}_{\text{Perturbation Error}}$$

Substituting this into the terms from the previous slide:

**1. Norm Squared Expansion:**

$$\|\nabla f(\tilde{x}_t)\|^2 = \|\nabla f(x_t)\|^2 + \|\nabla f(\tilde{x}_t) - \nabla f(x_t)\|^2 + 2\langle \nabla f(x_t), \nabla f(\tilde{x}_t) - \nabla f(x_t)\rangle$$

**2. Inner Product Expansion:**

$$-\eta_t \langle \nabla f(x_t), \nabla f(\tilde{x}_t)\rangle = -\eta_t \|\nabla f(x_t)\|^2 - \eta_t \langle \nabla f(x_t), \nabla f(\tilde{x}_t) - \nabla f(x_t)\rangle$$

# Step 1.3: Isolating Error & Descent Terms

Substituting the identities back into the smoothness inequality yields **Eq (5)**:

$$f(x_{t+1}) \leq f(x_t) \underbrace{-\eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2}_{\text{Descent Term}} + \underbrace{\frac{\eta_t^2 L}{2} \|\nabla f(\tilde{x}_t) - \nabla f(x_t)\|^2}_{\text{Perturbation Error Term}} \underbrace{-\eta_t (1 - \eta_t L) \langle \nabla f(x_t), \nabla f(\tilde{x}_t) - \nabla f(x_t) \rangle}_{\text{Cross Term}} \quad (5)$$

**1. Bounding Perturbation Error Term**

By $L$-smoothness, $\|\nabla f(\tilde{x}_t) - \nabla f(x_t)\|^2 \leq L^2 \rho^2 \implies$ This term is safely bounded by $\frac{1}{2}\eta_t^2 L^3 \rho^2$.

**2. The Challenge with the Cross Term**

Unlike the squared norm, the inner product has an **indefinite sign**: We need to further decompose $\nabla f(x_t)$ using the auxiliary sequence $y_t$.

# Step 2.1: Constructing the Auxiliary Sequence $y_t$

We introduce an **auxiliary sequence** $\{y_t\}$ with initialization $y_0 = x_0$:

$$y_{t+1} = x_t - \eta_t \nabla f(y_t),$$

$$\tilde{x}_t = x_t + \rho \frac{\nabla f(y_t)}{\|\nabla f(y_t)\|}$$

**1. Derivation via Telescoping Constraint**

Treating the convergence guarantee as a hard constraint, the authors explain that $y_t$ was specifically constructed to generate the term $\|\nabla f(x) - \nabla f(y)\|^2$ needed to cancel the cross term.

**2. Parallelism via Decoupling**

Since $y_{t+1}$ depends on $x_t$ (not $\tilde{x}_t$), it serves as a **stable proxy** that enables parallel computation.

# Step 2.2: Handling the Cross Term using Convexity

Recall the **Cross Term** from Eq (5):

$$-\eta_t(1-\eta_t L)\langle \nabla f(x_t), \nabla f(\tilde{x}_t) - \nabla f(x_t)\rangle$$

Let $\Delta_g = \nabla f(\tilde{x}_t) - \nabla f(x_t)$. We decompose the inner product using the auxiliary gradient $\nabla f(y_t)$:

$$\langle \nabla f(x_t), \Delta_g\rangle = \langle \nabla f(x_t) - \nabla f(y_t), \Delta_g\rangle + \underbrace{\langle \nabla f(y_t), \Delta_g\rangle}_{\geq 0}$$

We see that the inner product with the perturbation direction is non-negative, due to $f$ being convex:

$$\langle \nabla f(y_t), \nabla f(\tilde{x}_t) - \nabla f(x_t)\rangle = \frac{\|\nabla f(y_t)\|}{\rho}\langle \tilde{x}_t - x_t, \nabla f(\tilde{x}_t) - \nabla f(x_t)\rangle \geq 0$$

We can now drop the non-negative part to obtain an upper bound to the Cross Term:

$$-\eta_t(1-\eta_t L)\langle \nabla f(x_t), \Delta_g\rangle \leq -\eta_t(1-\eta_t L)\langle \nabla f(x_t) - \nabla f(y_t), \Delta_g\rangle$$

# Step 2.3: From Inner Product to Squared Norms

Since the **inner product** form has an indefinite sign hindering convergence analysis, we transform it into **squared norms** using Polarization Identity. (We will take another upper-bound afterwards)

First, factor out the coefficient $\frac{1}{2}(1-\eta_t L)$ and analyze the core term:

$$-2\eta_t\langle\nabla f(x_t)-\nabla f(y_t),\Delta_g\rangle$$

Then, using $2\langle a,b\rangle=\|a\|^2+\|b\|^2-\|a-b\|^2$ with $a=\nabla f(x_t)-\nabla f(y_t)$ and $b=-\eta_t\Delta_g$:

$$\begin{aligned}-2\eta_t\langle\nabla f(x_t)-\nabla f(y_t),\Delta_g\rangle = \ & \|\nabla f(x_t)-\nabla f(y_t)\|^2 && (\rightarrow \text{Term for } \mathcal{V}_t)\\ & +\eta_t^2\|\Delta_g\|^2 && (\rightarrow \text{Error Part 1})\\ & -\|\nabla f(x_t)-\nabla f(y_t)+\eta_t\Delta_g\|^2 && (\rightarrow \text{Precursor to } \mathcal{V}_{t+1})\end{aligned}$$

We want to relate the negative precursor term to the future state variables.

# Step 2.4: Bounding the Negative Term via Young's Inequality

Rewriting the inside vector to involve the future state $\tilde{x}_t$:

$$\nabla f(x_t) - \nabla f(y_t) + \eta_t \Delta_g = \underbrace{(\nabla f(\tilde{x}_t) - \nabla f(y_t))}_{X} - \underbrace{(1-\eta_t)\Delta_g}_{Y} \quad (\because \Delta_g = \nabla f(\tilde{x}_t) - \nabla f(x_t))$$

**Young's Inequality** states:

$$\|X\|^2 \leq (1+e)\|X-Y\|^2 + (1+\frac{1}{e})\|Y\|^2$$

Rearranging for $-\|X-Y\|^2$, we obtain:

$$-\|X-Y\|^2 \leq -\frac{1}{1+e}\|X\|^2 + \frac{1}{e}\|Y\|^2 \quad (\text{for } e > 0)$$

Substituting $X$ and $Y$:

$$-\|\nabla f(x_t) - \nabla f(y_t) + \eta_t \Delta_g\|^2 \leq \underbrace{-\frac{1}{1+e}\|\nabla f(\tilde{x}_t) - \nabla f(y_t)\|^2}_{\text{Source of } \mathcal{V}_{t+1}} + \underbrace{\frac{(1-\eta_t)^2}{e}\|\Delta_g\|^2}_{\text{Error Part 2}}$$

# Step 2.5: Intermediate Bound

Merging the previous steps, we get:

$$-2\eta_t\langle\nabla f(x_t)-\nabla f(y_t),\Delta_g\rangle \leq -\underbrace{\frac{1}{1+e}\left\|\nabla f(\tilde{x}_t)-\nabla f(y_t)\right\|^2}_{\text{Source of }\mathcal{V}_{t+1}} + \underbrace{\left\|\nabla f(x_t)-\nabla f(y_t)\right\|^2}_{\text{Term for }\mathcal{V}_t} + \underbrace{\left(\eta_t^2+\frac{(1-\eta_t)^2}{e}\right)\left\|\Delta_g\right\|^2}_{\text{Error Part 1 + 2}}$$

Recall from the first slide:

$$\mathcal{V}_t := f(x_t)+\frac{1}{2}(1-\eta_t L)\left\|\nabla f(x_t)-\nabla f(y_t)\right\|^2$$

While we want the "Source of $\mathcal{V}_{t+1}$" term to actually become $\mathcal{V}_{t+1}$ to cancel out with future steps, the coefficients and the state variables ($\tilde{x}_t$ vs $x_{t+1}$) do not match yet.

# Step 3.1: Matching State Variables via $L$-smoothness

Examine the difference between the update rules for $x_{t+1}$ and $y_{t+1}$:

$$x_{t+1} - y_{t+1} = (x_t - \eta_t \nabla f(\tilde{x}_t)) - (x_t - \eta_t \nabla f(y_t)) = \eta_t(\nabla f(y_t) - \nabla f(\tilde{x}_t)).$$

We have

$$\frac{1}{\eta_t^2}\|x_{t+1} - y_{t+1}\|^2 = \|\nabla f(\tilde{x}_t) - \nabla f(y_t)\|^2$$

Then according to the $L$-smoothness of $f$,

$$\|x_{t+1} - y_{t+1}\|^2 \geq \frac{1}{L^2}\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\|^2$$

Thereby obtaining the aforementioned secondary upper-bound, with the matched variables as:

$$-\frac{1}{1+e}\|\nabla f(\tilde{x}_t) - \nabla f(y_t)\|^2 \leq -\frac{1}{(1+e)\eta_t^2 L^2}\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\|^2$$

# Step 3.2: Cross Term Bound with Parameter $e$

We now get the following bound to the **Cross Term** from Eq (5):

$$-\eta_t(1-\eta_t L)\langle \nabla f(x_t), \Delta_g\rangle \le \frac{1}{2}(1-\eta_t L)\left[\underbrace{-\frac{1}{(1+e)\eta_t^2 L^2}\left\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\right\|^2}_{\text{Term for } \mathcal{V}_{t+1}} + \underbrace{\left\|\nabla f(x_t) - \nabla f(y_t)\right\|^2}_{\text{Term for } \mathcal{V}_t} + \underbrace{\left(\eta_t^2 + \frac{(1-\eta_t)^2}{e}\right)\left\|\Delta_g\right\|^2}_{\text{Error Part } 1+2}\right]$$

To enforce perfect cancellation via telescoping, we must select $e$ such that the coefficient of the future term matches the potential function's definition.

# Step 3.3: Designing Parameter $e$ for Cancellation

We choose $e$ to satisfy:

$$\underbrace{\frac{1}{2}(1-\eta_t L)}_{\text{Global Factor}} \times \underbrace{\frac{1}{1+e}}_{\text{Young's Coeff}} \times \underbrace{\frac{1}{\eta_t^2 L^2}}_{\text{Conversion Factor}} = \underbrace{\frac{1}{2}(1-\eta_{t+1} L)}_{\text{Target Coeff for } \mathcal{V}_{t+1}}$$

That is, **if there exists** a valid $e > 0$ to satisfy Young's inequality.
Rearranging for $1+e$, we get:

$$1+e = \frac{1-\eta_t L}{\eta_t^2 L^2 (1-\eta_{t+1} L)}$$

# Step 3.4: The Logical Flaw in the Original Proof

The paper relies solely on the decreasing property of the step size sequence $(\eta_t)_{t\in\mathbb{N}}$ to justify $1 + e > 1$:

**Appendix A, Eq (9)**

To verify that $e > 0$, use that $(\eta_t)_{t\in\mathbb{N}}$ is decreasing to obtain

$$\frac{1-\eta_t L}{1-\eta_{t+1} L} \geq 1 \geq \eta_t^2 L^2$$

However, for a decreasing sequence $\eta_t > \eta_{t+1}$, the inequality actually holds in the **opposite direction**:

$$1-\eta_t L < 1-\eta_{t+1} L \implies \frac{1-\eta_t L}{1-\eta_{t+1} L} < 1$$

Clearly invalidating the paper's justification.

# Step 3.5: The Correction via Magnitude Analysis

To fix this, we utilize the **magnitude** of the step size rather than just the ratio. Analyzing the full expression for $1 + e$ reveals the true source of the bound:

$$1 + e = \underbrace{\frac{1 - \eta_t L}{1 - \eta_{t+1} L}}_{\substack{\approx 1 \\ \text{(Slightly } < 1\text{)}}} \times \underbrace{\frac{1}{\eta_t^2 L^2}}_{\substack{\gg 1 \\ \text{(Dominant Term)}}}$$

- While the first term is slightly less than 1, the second term is derived from the inverse of the squared step size.
- Since we assumed a sufficiently small step size ($\eta_t < c/L$), the term $\frac{1}{\eta_t^2 L^2}$ becomes **dominant**.

$\therefore$ The product remains **strictly greater than 1**, guaranteeing a valid $e > 0$.

# Step 3.6: Cross Term Bound without Parameter $e$

With the validated $e$, we can now get the following bound to the **Cross Term** from Eq (5):

$$-\eta_t(1-\eta_t L)\langle \nabla f(x_t), \Delta_g \rangle \le \frac{1}{2}(1-\eta_t L)\left[\begin{array}{l} \underbrace{-\frac{1-\eta_{t+1}L}{1-\eta_t L}\left\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\right\|^2}_{\text{Term for } \mathcal{V}_{t+1}} \\ + \underbrace{\left\|\nabla f(x_t) - \nabla f(y_t)\right\|^2}_{\text{Term for } \mathcal{V}_t} \\ + \underbrace{\eta_t^2(1+A_t)\left\|\Delta_g\right\|^2}_{\text{Error Part 1 + 2}} \end{array}\right]$$

where

$$A_t = \frac{(1-\eta_t)^2}{\eta_t^2 e} = \frac{L^2(1-\eta_t)^2}{\frac{1-\eta_t L}{1-\eta_{t+1}L} - \eta_t^2 L^2}$$

# Step 3.7: Finalizing the Bound

The error term coefficient can be bounded using update rule $\tilde{x}_t = x_t + \rho \frac{\nabla f(y_t)}{\|\nabla f(y_t)\|}$ and $L$-smoothness of $f$:

$$\|\nabla f(\tilde{x}_t) - \nabla f(x_t)\|^2 \le L^2 \|\tilde{x}_t - x_t\|^2 = L^2\rho^2 \quad \longrightarrow \quad \eta_t^2(1 + A_t)\|\Delta_g\|^2 \le \eta_t^2(1 + A_t)L^2\rho^2$$

Therefore, the upper-bound for the **Cross Term** from Eq (5),

$$-\eta_t(1 - \eta_t L)\langle \nabla f(x_t), \Delta_g\rangle \le \frac{1}{2}(1 - \eta_t L)\left[\begin{array}{l} \underbrace{-\frac{1 - \eta_{t+1}L}{1 - \eta_t L}\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\|^2}_{\text{Matches } \mathcal{V}_{t+1}} \\ + \underbrace{\|\nabla f(x_t) - \nabla f(y_t)\|^2}_{\text{Cancels in } \mathcal{V}_t} \\ + \underbrace{\eta_t^2(1 + A_t)L^2\rho^2}_{\text{Bounded Error}} \end{array}\right]$$

now perfectly aligns with the structure of $\mathcal{V}_t$ and $\mathcal{V}_{t+1}$.

# Step 4.1: Time-Step Separation for Potential Function

**Grouping terms by time step:** We move terms depending on $t+1$ to the LHS, keeping $t$ on the RHS.

$$f(x_{t+1}) \le f(x_t) - \eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2 + \frac{1}{2}\eta_t^2 L^3 \rho^2 \quad \text{(Eq (5))}$$

$$+ \frac{1}{2}(1-\eta_t L)\left[ -\frac{1-\eta_{t+1}L}{1-\eta_t L} \|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\|^2 + \|\nabla f(x_t) - \nabla f(y_t)\|^2 + \eta_t^2 (1+A_t) L^2 \rho^2 \right]$$

## Strategy for Final Form

1. Identify $\mathcal{V}_{t+1} = f(x_{t+1}) + \frac{1}{2}(1-\eta_{t+1}L)\|\nabla f(x_{t+1}) - \nabla f(y_{t+1})\|^2$.
2. Identify $\mathcal{V}_t = f(x_t) + \frac{1}{2}(1-\eta_t L)\|\nabla f(x_t) - \nabla f(y_t)\|^2$.
3. Collect all remaining "Error Terms" dependent on $\eta_t^2 \rho^2$.

# Step 4.2: Establishing the Recursive Descent Structure

By identifying the grouped terms as the potential function, we obtain:

$$\underbrace{f(x_{t+1}) + \frac{1}{2}(1 - \eta_{t+1} L) \left\| \nabla f(x_{t+1}) - \nabla f(y_{t+1}) \right\|^2}_{\mathcal{V}_{t+1}} \leq \underbrace{f(x_t) + \frac{1}{2}(1 - \eta_t L) \left\| \nabla f(x_t) - \nabla f(y_t) \right\|^2}_{\mathcal{V}_t}$$

$$\underbrace{-\eta_t \left(1 - \frac{\eta_t L}{2}\right) \left\| \nabla f(x_t) \right\|^2}_{\text{Descent Term}}$$

$$+ \quad \underbrace{\eta_t^2 \rho^2 C}_{\text{Controlled Error}}$$

This inequality guarantees that the potential energy decreases at every step, dominated by the descent term.

# Final Result and Interpretation

**Lemma 4.3 (The Descent Inequality)**

$$\mathcal{V}_{t+1} \le \mathcal{V}_t - \eta_t \left(1 - \frac{\eta_t L}{2}\right) \|\nabla f(x_t)\|^2 + \eta_t^2 \rho^2 C$$

**Interpretation:**

- The descent term $-\eta_t \|\nabla f\|^2$ drives the potential down continuously.
- The noise term $\eta_t^2 \rho^2 C$ resists convergence, but its influence decays faster than the descent term ($\eta_t^2 \ll \eta_t$).
- Since the total accumulated error is finite, the driving force ensures $\min_{t<T} \|\nabla f(x_t)\| \to 0$

# Conclusion and Significance

We have established the theoretical foundation of SAMPa.

1. **Foundation for Parallelism**
   - We proved that using the decoupled auxiliary sequence $y_{t+1}$ is **mathematically safe**.
   - **Impact:** This enables simultaneous computation of $\nabla f(\tilde{x}_t)$ and $\nabla f(y_{t+1})$, justifying the **2x speedup** in SAMPa.
2. **Mathematical Rigor & Correction**
   - **Step Size:** Corrected max $\to$ min condition prevents divergence.
   - **Telescoping:** Validated logic using magnitude analysis $(1/\eta_t^2 \gg 1)$.
3. **Road to Convergence Rate (Next Section)**
   - This Descent Lemma serves as the engine for **Theorem 4.4**.
   - Next, we will sum this inequality to derive the $\mathcal{O}(1/\sqrt{T})$ rate.

# SAM Parallelized (SAMPa)

- Convergence analysis

**Theorem 4.4.** SAMPa satisfies the following descent inequality for $\rho > 0$ and a decreasing sequence $(\eta_t)_{t\in\mathbb{N}}$ with $\eta_t \in (0, \min\{1, 1/2L\})$

$$\sum_{t=0}^{T-1} \frac{\eta_t(1-\eta_t L/2)}{\sum_{\tau=0}^{T-1}\eta_\tau(1-\eta_\tau L/2)} \|\nabla f(x_t)\|^2 \le \frac{\Delta_0 + C\rho^2 \sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t(1-\eta_t L/2)}$$

where $\Delta_0 = f(x_0) - \inf_{x\in\mathbb{R}^d} f(x)$ and $C = \frac{L^2+L^3}{2} + \frac{2L^4}{3}$

- Assumption 1: The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex
- Assumption 2: The operator $\nabla f : \mathbb{R}^d \to \mathbb{R}$ is $L$-Lipschitz with $L \in (0, \infty)$, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathbb{R}^n$$

# SAM Parallelized (SAMPa)

- Convergence analysis (**proof of Theorem 4.4**)
  - We start from the Lemma 4.3.

$$\mathcal{V}_{t+1} \;\le\; \mathcal{V}_t - \eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla f(x_t)\|^2 + \eta_t^2\rho^2 C$$

  - Summing over $t = 0, \dots, T-1$ gives

$$\mathcal{V}_T - \mathcal{V}_0 \;\le\; -\sum_{t=0}^{T-1}\eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla f(x_t)\|^2 + \rho^2 C\sum_{t=0}^{T-1}\eta_t^2 .$$

$$\mathcal{V}_T \ge f^\star \triangleq \inf_{x\in\mathbb{R}^d} f(x)$$

$$\sum_{t=0}^{T-1}\eta_t\left(1 - \frac{\eta_t L}{2}\right)\|\nabla f(x_t)\|^2 \;\le\; \mathcal{V}_0 - f^\star + \rho^2 C\sum_{t=0}^{T-1}\eta_t^2 . \quad (1)$$

# SAM Parallelized (SAMPa)

- Convergence analysis (**proof of Theorem 4.4**)
  - By using the definition of the potential function

$$\mathcal{V}_0 - f^\star = f(x_0) - f^\star + \frac{1}{2}(1-\eta_0 L)\|\nabla f(x_0) - \nabla f(y_0)\|^2$$

$$= \Delta_0 + \frac{1}{2}(1-\eta_0 L)\|\nabla f(x_0) - \nabla f(y_0)\|^2$$

$$\leq \Delta_0 + \frac{1}{2}\|\nabla f(x_0) - \nabla f(y_0)\|^2$$

$1 - \eta_0 L \leq 1$

  - Finally, dividing both sides of 1 by $\sum_{\tau=0}^{T-1} \eta_\tau(1-\eta_\tau L/2)$ yields the averaged bound:

$$\sum_{t=0}^{T-1} \frac{\eta_t(1-\eta_t L/2)}{\sum_{\tau=0}^{T-1}\eta_\tau(1-\eta_\tau L/2)}\|\nabla f(x_t)\|^2 \leq \frac{\Delta_0 + \frac{1}{2}\|\nabla f(x_0) - \nabla f(y_0)\|^2 + C\rho^2\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t(1-\eta_t L/2)}$$

# SAM Parallelized (SAMPa)

- Convergence analysis (**proof of Theorem 4.4**)
  - By using Lipschitz continuity from Assumption 4.2 we have that

$$\|\nabla f(x_0) - \nabla f(y_0)\|^2 \le L^2 \|x_0 - y_0\|^2 = 0$$

  - The last equality follows from picking the initialization $y_0 = x_0$
  - If we set $c = 0.5$, then $\eta_t < \min\{1, \frac{1}{2L}\}$

$$C = 0.5(L^2 + L^3 + \frac{1}{1-c^2} L^4) = \frac{L^2 + L^3}{2} + \frac{2L^4}{3}$$

$$\sum_{t=0}^{T-1} \frac{\eta_t(1-\eta_t L/2)}{\sum_{\tau=0}^{T-1} \eta_\tau (1-\eta_\tau L/2)} \|\nabla f(x_t)\|^2 \le \frac{\Delta_0 + C\rho^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t (1-\eta_t L/2)}$$

# SAM Parallelized (SAMPa)

- Convergence analysis (**proof of Theorem 4.4**)
  - Picking a fixed stepsize $\eta_t = \eta$, the convergence guarantee reduces to

$$\sum_{t=0}^{T-1} \frac{\eta_t(1-\eta_t L/2)}{\sum_{\tau=0}^{T-1}\eta_\tau(1-\eta_\tau L/2)}\|\nabla f(x_t)\|^2 \leq \frac{\Delta_0 + C\rho^2\sum_{t=0}^{T-1}\eta_t^2}{\sum_{t=0}^{T-1}\eta_t(1-\eta_t L/2)}$$

$$\eta_t = \eta, \forall t$$

$$\sum_{t=0}^{T-1} \frac{\eta(1-\eta L/2)}{T\eta(1-\eta L/2)}\|\nabla f(x_t)\|^2 \leq \frac{\Delta_0 + C\rho^2 T\eta^2}{T\eta(1-\eta L/2)}$$

$$\eta L \leq 0.5$$

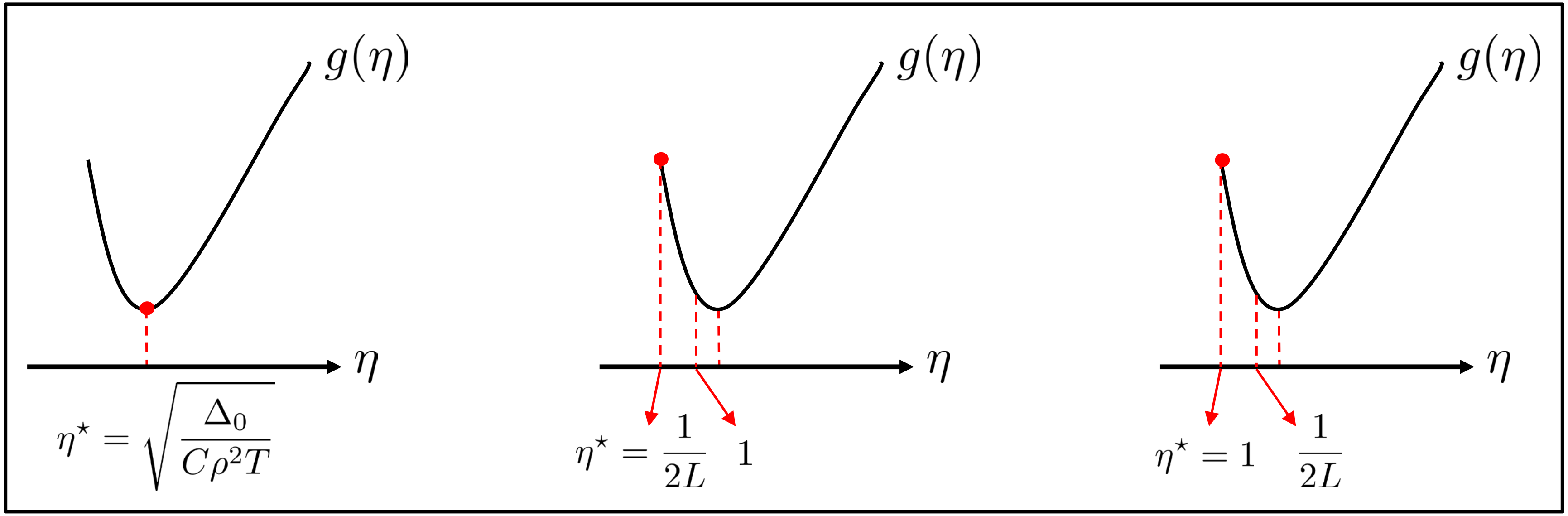
$$\sum_{t=0}^{T-1} \frac{1}{T}\|\nabla f(x_t)\|^2 \leq \frac{4}{3}\left(\frac{\Delta_0}{T\eta} + C\rho^2\eta\right) \triangleq g(\eta)$$

# SAM Parallelized (SAMPa)

Three possible values

$$\eta^\star = \min\left\{\sqrt{\frac{\Delta_0}{C\rho^2 T}}, \frac{1}{2L}, 1\right\}$$

- Convergence analysis (**proof of Theorem 4.4**)
  - Case study




$$\min_{t=0,\dots,T-1} \|\nabla f(x_t)\|^2 \leq \sum_{t=0}^{T-1} \frac{1}{T}\|\nabla f(x_t)\|^2 \leq g(\eta^\star) = \mathcal{O}\left(\frac{L\Delta_0}{T} + \frac{\rho\sqrt{\Delta_0 C}}{\sqrt{T}}\right)$$

# Contents

# Image classification

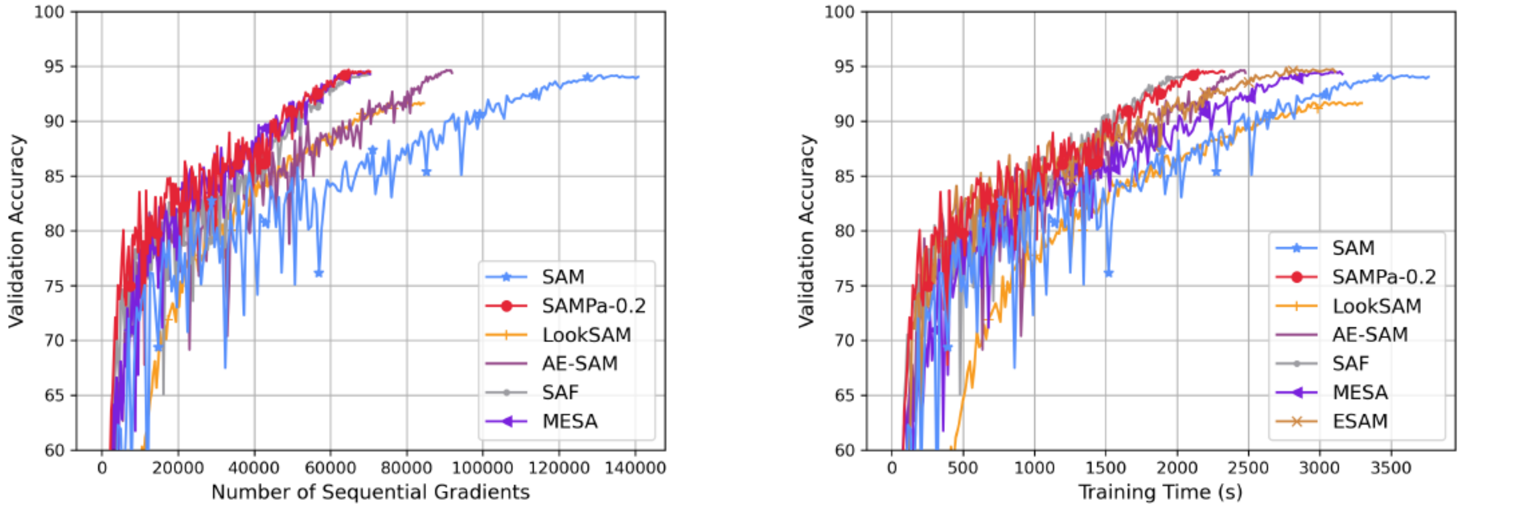
Table 1: **Test accuracies on CIFAR-10.** SAMPa-0.2 outperforms SAM across all models with halved total temporal cost."Temporal cost" represents the number of sequential gradient computations per update. SAMPa-0.2 with 400 epochs is included for comprehensive comparison with SGD and SAM.

| **Model**<br>Temporal cost/Epochs | **SGD**<br>$\times 1/400$ | **SAM**<br>$\times 2/200$ | **SAMPa-0**<br>$\times 1/200$ | **SAMPa-0.2**<br>$\times 1/200$ | **SAMPa-0.2**<br>$\times 1/400$ |
|---|---|---|---|---|---|
| DenseNet-121 | $96.14_{\pm 0.09}$ | $96.49_{\pm 0.14}$ | $96.53_{\pm 0.11}$ | $\mathbf{96.77}_{\pm 0.11}$ | $96.92_{\pm 0.09}$ |
| Resnet-56 | $94.20_{\pm 0.39}$ | $94.26_{\pm 0.70}$ | $94.31_{\pm 0.43}$ | $\mathbf{94.62}_{\pm 0.35}$ | $95.43_{\pm 0.25}$ |
| VGG19-BN | $94.76_{\pm 0.10}$ | $95.05_{\pm 0.17}$ | $95.06_{\pm 0.22}$ | $\mathbf{95.11}_{\pm 0.10}$ | $95.34_{\pm 0.07}$ |
| WRN-28-2 | $95.71_{\pm 0.19}$ | $95.98_{\pm 0.10}$ | $96.06_{\pm 0.10}$ | $\mathbf{96.13}_{\pm 0.14}$ | $96.31_{\pm 0.09}$ |
| WRN-28-10 | $96.77_{\pm 0.21}$ | $97.25_{\pm 0.09}$ | $97.24_{\pm 0.11}$ | $\mathbf{97.34}_{\pm 0.09}$ | $97.46_{\pm 0.07}$ |
| Average | $95.52_{\pm 0.10}$ | $95.81_{\pm 0.15}$ | $95.86_{\pm 0.10}$ | $\mathbf{95.99}_{\pm 0.08}$ | $96.29_{\pm 0.06}$ |

Table 2: **Test accuracies on CIFAR-100.** SAMPa-0.2 outperforms SAM across all models with halved total temporal cost. "Temporal cost" represents the number of sequential gradient computations per update. SAMPa-0.2 with 400 epochs is included for a comprehensive comparison.

| **Model**<br>Temporal cost/Epochs | **SGD**<br>$\times 1/400$ | **SAM**<br>$\times 2/200$ | **SAMPa-0**<br>$\times 1/200$ | **SAMPa-0.2**<br>$\times 1/200$ | **SAMPa-0.2**<br>$\times 1/400$ |
|---|---|---|---|---|---|
| DenseNet-121 | $81.08_{\pm 0.43}$ | $82.53_{\pm 0.22}$ | $82.50_{\pm 0.10}$ | $\mathbf{82.70}_{\pm 0.23}$ | $83.44_{\pm 0.21}$ |
| Resnet-56 | $74.09_{\pm 0.39}$ | $75.14_{\pm 0.15}$ | $75.22_{\pm 0.20}$ | $\mathbf{75.29}_{\pm 0.24}$ | $75.84_{\pm 0.27}$ |
| VGG19-BN | $74.85_{\pm 0.53}$ | $74.94_{\pm 0.12}$ | $74.94_{\pm 0.17}$ | $\mathbf{75.38}_{\pm 0.31}$ | $76.23_{\pm 0.16}$ |
| WRN-28-2 | $78.00_{\pm 0.17}$ | $78.50_{\pm 0.24}$ | $78.45_{\pm 0.29}$ | $\mathbf{78.82}_{\pm 0.22}$ | $79.46_{\pm 0.20}$ |
| WRN-28-10 | $81.56_{\pm 0.25}$ | $83.37_{\pm 0.30}$ | $83.46_{\pm 0.25}$ | $\mathbf{83.90}_{\pm 0.25}$ | $83.91_{\pm 0.13}$ |
| Average | $77.92_{\pm 0.17}$ | $78.90_{\pm 0.10}$ | $78.91_{\pm 0.09}$ | $\mathbf{79.22}_{\pm 0.11}$ | $79.78_{\pm 0.09}$ |

# Efficiency comparison with efficient SAM variants




(a) Number of sequential gradients (b) Actual running time

Figure 2: **Computational time comparison for efficient SAM variants.** SAMPa-0.2 requires near-minimal computational time in both ideal and practical scenarios.

Table 4: **Efficient SAM variants.** The best result is in bold and the second best is underlined.

| | SAM | SAMPa-0.2 | LookSAM | AE-SAM | SAF | MESA | ESAM |
|---|---|---|---|---|---|---|---|
| Accuracy | 94.26 | **94.62** | 91.42 | <u>94.46</u> | 93.89 | 94.23 | 94.21 |
| Time/Epoch (s) | 18.81 | <u>10.94</u> | 16.28 | 13.47 | **10.09** | 15.43 | 15.97 |

# Transfer learning: NLP fine-tuning

Table 6: **Test results of BERT-base fine-tuned on GLUE.**

| Method | GLUE | CoLA | SST-2 | MRPC | STS-B | QQP | MNLI | QNLI | RTE | WNLI |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Mcc.* | *Acc.* | *Acc./F1.* | *Pear./Spea.* | *Acc./F1.* | *Acc.* | *Acc.* | *Acc.* | *Acc.* |
| AdamW | 74.6 | 56.6 | 91.6 | 85.6/89.9 | 85.4/85.3 | 90.2/86.8 | 82.6 | 89.8 | 62.4 | 26.4 |
| -w SAM | 76.6 | 58.8 | 92.3 | 86.5/90.5 | 85.0/85.0 | 90.6/87.5 | 83.9 | 90.4 | 60.6 | 41.2 |
| -w SAMPa-0 | 76.9 | 58.9 | 92.5 | 86.4/90.4 | 85.0/85.0 | 90.6/87.6 | 83.8 | 90.4 | 60.4 | 43.2 |
| -w SAMPa-0.1 | **78.0** | 58.9 | 92.5 | 86.8/90.7 | 85.2/85.1 | 90.7/87.7 | 84.0 | 90.5 | 61.3 | 51.6 |

# Noisy Label task

Table 7: **Test accuracies of ResNet-32 models trained on CIFAR-10 with label noise.**

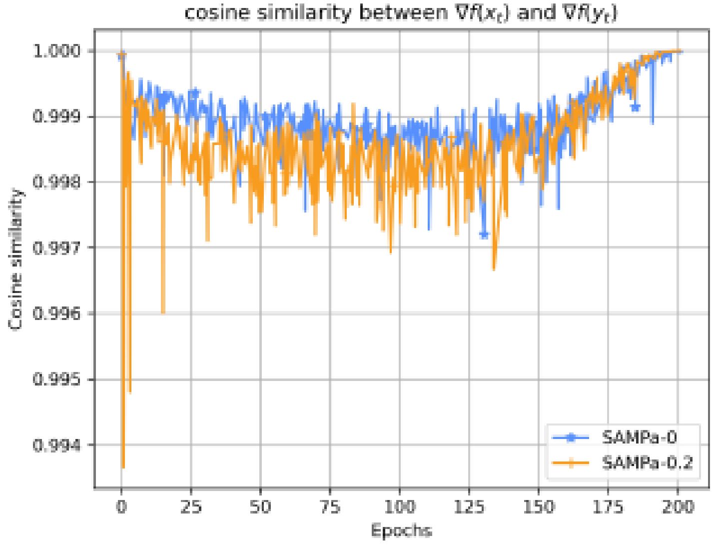
| Noise rate | SGD | SAM | SAMPa-0 | SAMPa-0.2 |
|---|---|---|---|---|
| 0% | $94.22_{\pm 0.14}$ | $94.36_{\pm 0.07}$ | $94.36_{\pm 0.12}$ | $\mathbf{94.41}_{\pm 0.08}$ |
| 20% | $88.65_{\pm 0.75}$ | $92.20_{\pm 0.06}$ | $92.22_{\pm 0.10}$ | $\mathbf{92.39}_{\pm 0.09}$ |
| 40% | $84.24_{\pm 0.25}$ | $89.78_{\pm 0.12}$ | $89.75_{\pm 0.15}$ | $\mathbf{90.01}_{\pm 0.18}$ |
| 60% | $76.29_{\pm 0.25}$ | $83.83_{\pm 0.51}$ | $83.81_{\pm 0.37}$ | $\mathbf{84.38}_{\pm 0.07}$ |
| 80% | $44.44_{\pm 1.20}$ | $48.01_{\pm 1.63}$ | $48.22_{\pm 1.71}$ | $\mathbf{49.92}_{\pm 1.12}$ |

# Incorporation with other SAM variants

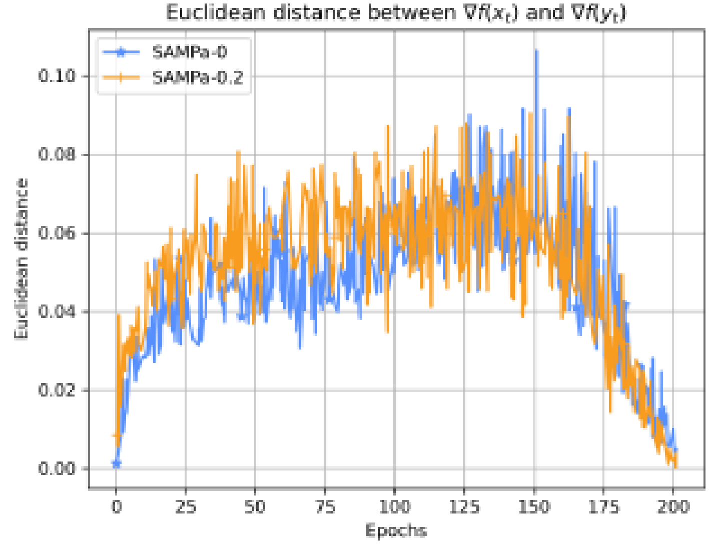
Table 8: **Incorporation with variants of SAM.** SAMPa in the table denotes SAMPa-0.2. The incorporation of SAMPa with SAM variants enhances both accuracy and efficiency.

| mSAM | +SAMPa | ASAM | +SAMPa | SAM-ON | +SAMPa | VaSSO | +SAMPa | BiSAM | +SAMPa |
|---|---|---|---|---|---|---|---|---|---|
| 94.28 | **94.71** | 94.84 | **94.95** | 94.44 | **94.51** | 94.80 | **94.97** | 94.49 | **95.13** |

# Appendix C. Choice of $y_{t+1}$




(a) Cosine similarity




(b) Euclidean distance

Figure 4: Difference between $\nabla f(x_t)$ and $\nabla f(y_t)$.

# Thank you for your attention