Sharpness-Aware Minimization Efficiently Selects Flatter Minima Late In Training

Zhanpeng Zhou[*1], Mingze Wang[*2], Yuchen Mao[1], Bingrui Li[3], Junchi Yan[1]

[1]Shanghai Jiao Tong University, [2]Peking University, [3]Tsinghua University

*Equal contribution.

Background: Sharpness-Aware Minimization

Sharpness-Aware Minimization (SAM) [1]

The central idea is to minimize the worst-case loss within a neighborhood of current weights, i.e.,

$$\min_{\boldsymbol{\theta}} \mathcal{L}_S^{\mathrm{SAM}}(\boldsymbol{\theta};\rho), \qquad \text{where } \mathcal{L}_S^{\mathrm{SAM}}(\boldsymbol{\theta};\rho) = \max_{||\epsilon||_2 \leq \rho} \mathcal{L}_S(\boldsymbol{\theta}+\boldsymbol{\epsilon}).^*$$

*$\mathcal{L}_S(\boldsymbol{\theta})$ denotes the total loss over the training set.

However, finding $\boldsymbol{\epsilon}$ can be computationally intractable in practice. Thus Foret et al. [1] used first-order approximation, i.e.,

$$\boldsymbol{\epsilon} \approx \arg\max_{||\boldsymbol{\epsilon}||_2 \leq \rho} \left(\mathcal{L}_S(\boldsymbol{\theta}) + \boldsymbol{\epsilon}^\top \nabla \mathcal{L}_S(\boldsymbol{\theta})\right) = \rho \nabla \mathcal{L}_S(\boldsymbol{\theta}) / \left|\left|\nabla \mathcal{L}_S(\boldsymbol{\theta})\right|\right|_{\mathbf{2}}.$$

Consequently, the update rule of SAM with stochastic gradient is,

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla \mathcal{L}_{\xi_t}\left(\boldsymbol{\theta} + \rho \nabla \mathcal{L}_{\xi_t}(\boldsymbol{\theta}) / \left|\left|\nabla \mathcal{L}_{\xi_t}(\boldsymbol{\theta})\right|\right|_{\mathbf{2}}\right).^*$$

*$\mathcal{L}_{\xi_t}(\boldsymbol{\theta})$ denotes the loss over a randomly sampled mini-batch ξ_t at iteration t. η denotes the learning rate.

[1] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization.

Background: Implicit Bias

*The effectiveness of gradient-based optimization methods can be attributed to their **implicit bias** toward solutions with favorable properties [2].*

Implicit Bias of SGD

- SGD and its variants tends to find flat minima, which often generalize well.

Implicit Bias of SAM

- SAM tends to find flatter minima over SGD, which represents a form of implicit bias*.

*Though SAM is inspired from sharpness regularization, its practical implementation, which minimizes a first-order approximation of the original objective, doesn't explicitly achieve this.

Understanding the mechanism behind the implicit bias of SAM towards flatter minima is crucial to explain its effectiveness.

[2] Gal Vardi. On the implicit bias in deep-learning algorithms.

SAM Selects Flatter Minima Late In Training

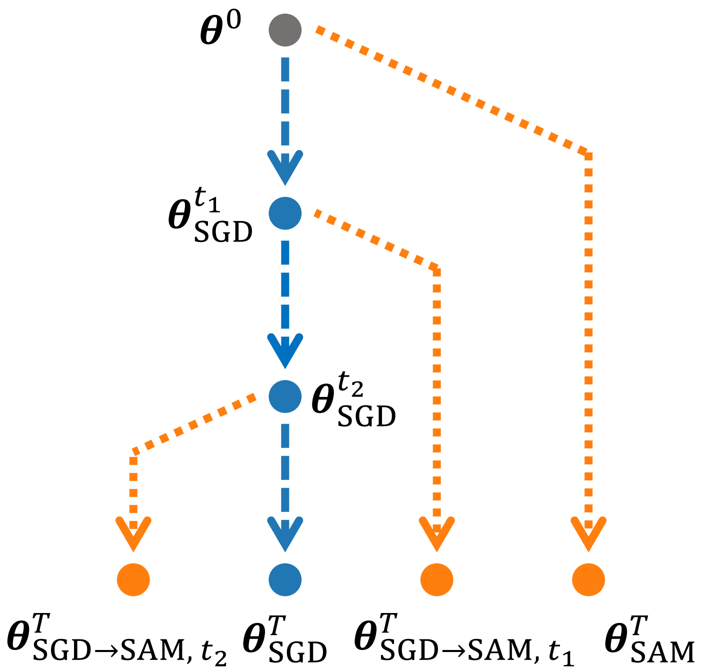

Fig. 1: **Illustration of the switching method.** Blue dashed lines represent SGD training, while orange dashed lines represent SAM training. t_1, t_2 denotes two switching points.

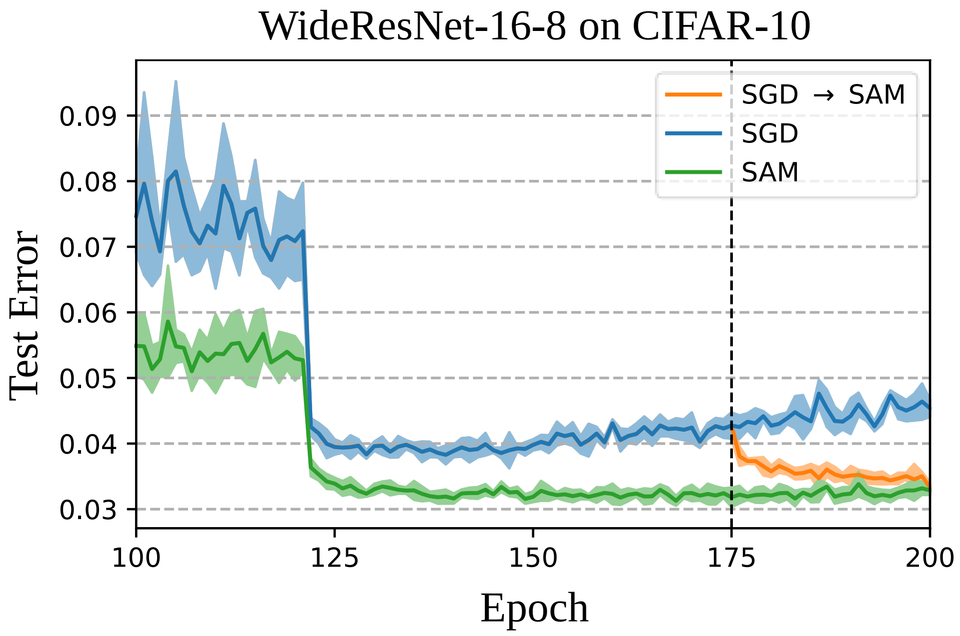

Fig. 2: **SAM operates efficiently late in training.** Blue line represents $\boldsymbol{\theta}^{T}_{\mathrm{SGD}}$, while orange lines represents $\boldsymbol{\theta}^{T}_{\mathrm{SAM}}$. Orange line represents $\boldsymbol{\theta}^{T}_{\mathrm{SGD}\to\mathrm{SAM},t}$, where $t = 175$.

SAM Selects Flatter Minima Late In Training

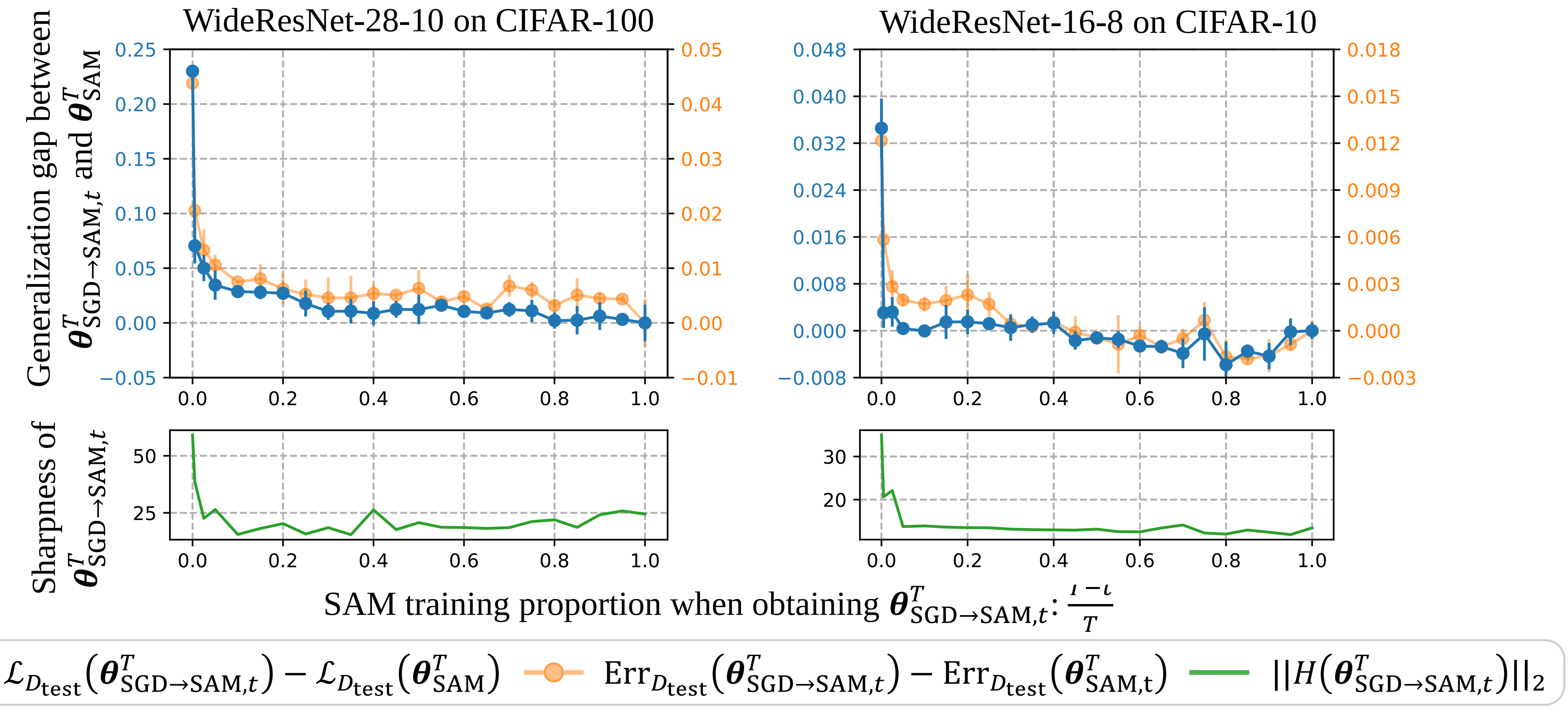

Fig. 3: **Few epochs of SAM substantially improves generalization/sharpness**.
We vary t while keep T fixed to adjust the SAM training proportion of $\boldsymbol{\theta}^T_{\mathrm{SGD}\to\mathrm{SAM},t}$.

How SAM Selects Flatter Minima late In Training

A Two-Phase Picture

We identify a two-phase picture in training dynamics after switching to SAM in the late training phase. This two-phase picture is characterized into four key claims **(P1-4)**, as outlined in Tab. 1.

Phase I. (Escape)	**(P1).** *SAM rapidly escapes from the minimum found by SGD;* **(P2).** *However, the iterator remains within the current valley.*	**Theorem 4.2** **Proposition 4.1**
Phase II. (Converge)	**(P3).** *SAM converges to a flatter minimum compared to SGD;* **(P4).** *The convergence rate of SAM is extremely fast.*	**Theorem 4.1** **Theorem 4.3**

Tab. 1: **Overview of the two-phase picture and corresponding theoretical results.**

How SAM Selects Flatter Minima late In Training

To understand the two-phase picture, let us first use a toy but representative example.

Example 4.1.

Consider using the shallow neural network $f(u, v; x) = tanh(v\, tanh(ux))$ *to fit a single data* $(x = 1, y = 0)$ *under the squared loss* $\ell(y; y') = (y - y')^2/2$, *then the loss landscape is* $\mathcal{L}(u, v) = \frac{1}{2} tanh^2(v\, tanh(u))$.

Note: If dynamics occurs around the set of global minima $\mathcal{M} = \{(u, v) | v = 0\}$, then small u implies flatter minima*.

Toy model on $\frac{1}{2}\tanh^2(v \tanh(ux))$

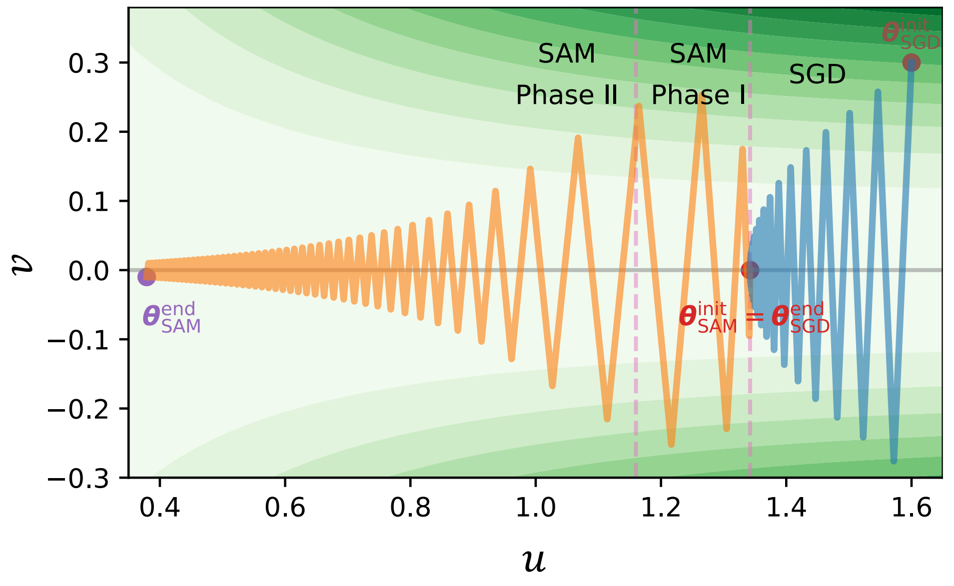

Fig. 4: **Visualization of the two-phase dynamics for Example 4.1.** The horizontal gray line represents $\mathcal{M}$. Blue lines trace SGD, while orange lines show SAM.

How SAM Selects Flatter Minima late In Training

Theoretical Support for P1 and P3: A Linear Stability Analysis [3].

Theorem 4.1 (P3)

Let $\boldsymbol{\theta}^*$ be a global minimum that is linearly stable for SAM and suppose Assumption 4.1 (see main paper) holds, then we have $\left\|H(\boldsymbol{\theta}^*)\right\|_F^2 \left(1 + \frac{\rho^2\gamma}{B}\left\|H(\boldsymbol{\theta}^*)\right\|_F^2\right) \le \frac{B}{\eta^2\gamma}$.

*B denotes the mini-batch size, and $\gamma \ge 0$ is a constant defined in Assumption 4.1.

Theorem 4.1 characterizes the sharpness of the global minima selected by SAM. In Tab. 2, SAM probably selects flatter minima than SGD **(P3)**.

	SAM (Theorem 4.1)	SGD [3]
Sharpness Bound	$\left\|H(\boldsymbol{\theta}^*)\right\|_F^2 \left(1 + \frac{\rho^2\gamma}{B}\left\|H(\boldsymbol{\theta}^*)\right\|_F^2\right) \le \frac{B}{\eta^2\gamma}$	$\left\|H(\boldsymbol{\theta}^*)\right\|_F^2 \le \frac{B}{\eta^2\gamma}$

Tab. 2: Comparison of the sharpness of global minima selected by SAM and SGD.

[3] Lei Wu, Mingze Wang, and Weijie Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis..

How SAM Selects Flatter Minima late In Training

Theorem 4.2 (P1)

Let $\boldsymbol{\theta}^*$ be a global minimum that is linearly stable for SAM and suppose Assumption 4.1 (see main paper) holds. If $\left|\left|H(\boldsymbol{\theta}^*)\right|\right|_F^2 \left(1 + \frac{\rho^2\gamma}{B}\left|\left|H(\boldsymbol{\theta}^*)\right|\right|_F^2\right) > \frac{B}{\eta^2\gamma}$, then $\boldsymbol{\theta}^*$ is linearly non-stable for SAM and $\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}^t\right)\right] \geq C^t\,\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}^0)]$ holds for all $t > 0$ with $C > 1$.

Theorem 4.2 characterizes the necessary condition of a linearly stable minimum for SAM. As SGD minimum cannot meet the stability condition of SAM, SAM will escape from the minimum found by SGD exponentially fast **(P1).**

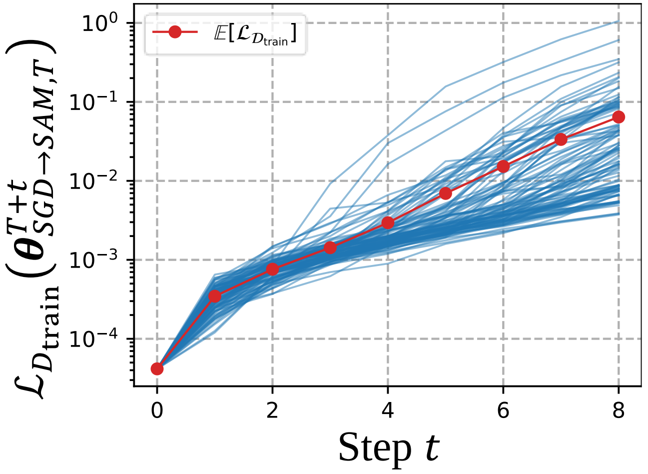

Fig. 5: **The exponentially fast escape from minima found by SGD.** Train loss $\mathcal{L}_{\mathcal{D}_{\text{train}}}\left(\boldsymbol{\theta}^{T+t}_{\text{SGD}\to\text{SAM},T}\right)$ vs. step t.

How SAM Selects Flatter Minima late In Training

Theoretical Support for P2: Beyond Local Analysis [4].

Proposition 4.1 (P2)

Under Definition 4.2 (see main paper), assume the landscape is sub-quadratic in the valley $V = [-2b, 2b]$*. Then,* $\forall\, \eta, \rho\ s.t.\ \eta < \min_{z \in V} b/|\mathcal{L}'(z)|, \rho \leq \min\{\frac{1}{a}, \eta \min_{0<|z|<b} |\mathcal{L}'(2z)/\mathcal{L}'(z)|\}$, and $\theta_0 \in (-b, b)$*, the full-batch SAM will remain within* V*, i.e.,* $\theta_t \in V, \forall t \in \mathbb{N}$*.*

Proposition 4.1 supports our key claim **P2** that SAM remains within the current valley during escape.

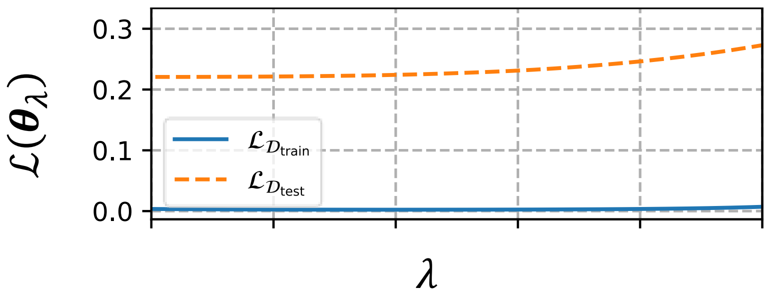

Fig. 6: **SAM converges to a flatter minimum within the same valley as the one found by SGD.** The loss of the interpolated model $\mathcal{L}_D(\boldsymbol{\theta}_\lambda)$ vs. interpolation coefficient λ. Here, $\boldsymbol{\theta}_\lambda = (1-\lambda)\boldsymbol{\theta}^{\text{end}}_{\text{SGD}\to\text{SAM}} + \lambda \boldsymbol{\theta}^{\text{end}}_{\text{SGD}}$.

[4] Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes..

How SAM Selects Flatter Minima late In Training

Theoretical Support for P4: Convergence Analysis.

Theorem 4.3 (P4)

Under Assumption 4.2 and 4.3 (main paper), let $\{\boldsymbol{\theta}^t\}_t$ *be the weights found by SAM. If* $\eta \leq \min\{\frac{1}{L}, \frac{\mu B}{2L\sigma^2}\}$ *and* $\rho \leq \min\left\{\frac{1}{L}, \frac{\mu B}{4L\sigma^2}, \frac{\eta\mu^2}{24L^2}\right\}$*, then we have* $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}^t)] \leq \left(1 - \frac{\eta\mu}{2}\right)^t \mathcal{L}(\boldsymbol{\theta}^0), \forall t \in \mathbb{N}$.

Theorem 4.3 supports our key claim **P4** that the convergence rate of stochastic SAM is significantly fast, and notably faster than the previous result on SAM's convergence rate [5].

5] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization..

Is SAM Still Necessary in Early Phase?

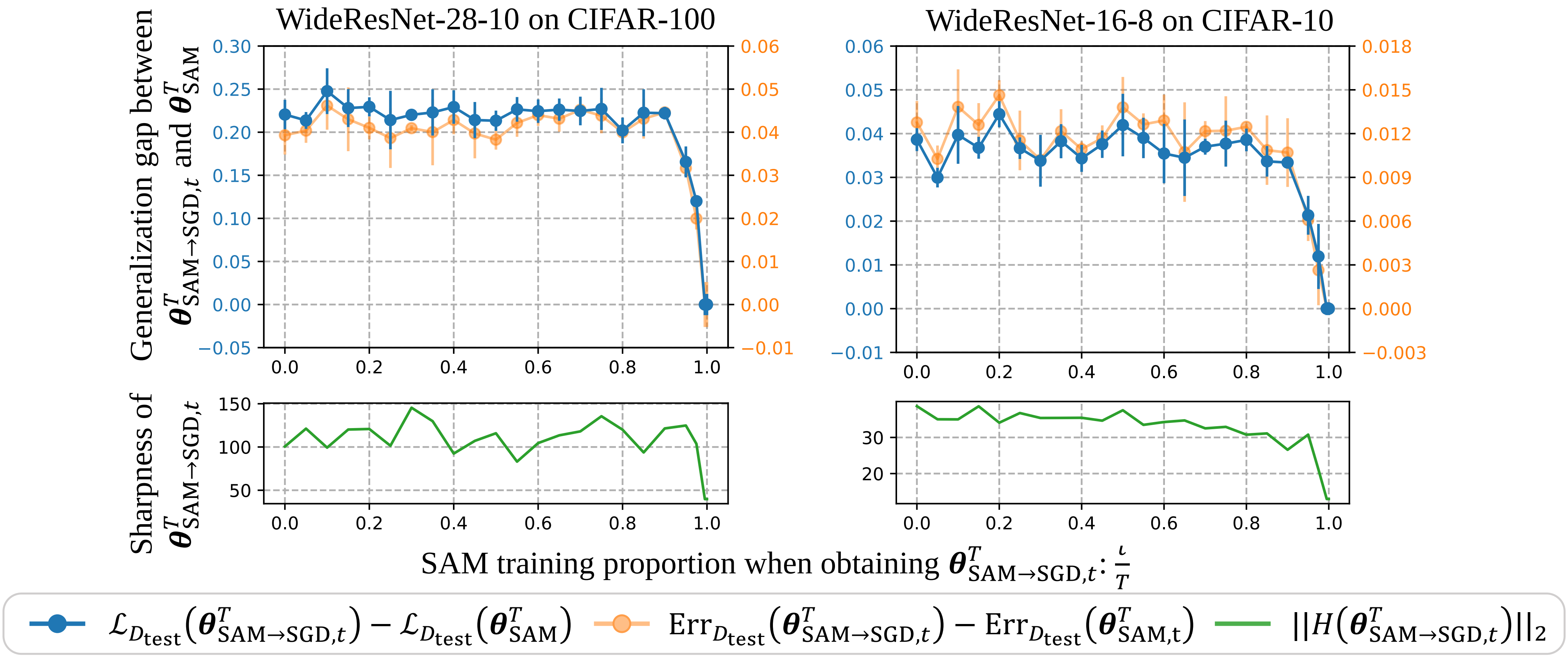

Fig. 7: **Early-phase SAM marginally improves generalization/sharpness**. We vary t while keep T fixed to adjust the SAM training proportion of $\boldsymbol{\theta}^T_{\text{SAM}\to\text{SGD},t}$.

Extend Findings to Adversarial Training

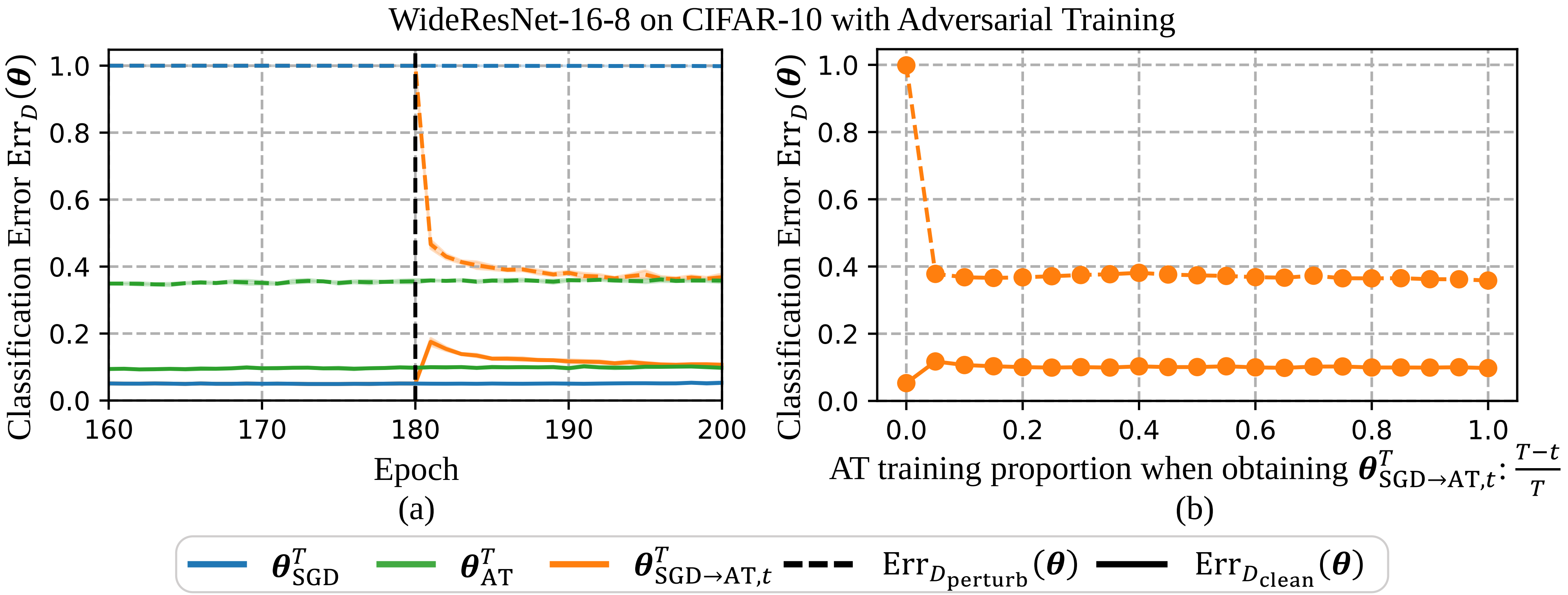

Fig. 8: **AT improves robustness efficiently even when applied only during the final few epochs of training**. **(a)** Robust/natural error vs. training epochs for model trained with different strategies. **(b)** Robust/natural error of $\boldsymbol{\theta}^T_{\mathrm{SGD}\to\mathrm{AT},t}$ vs. the proportion of AT epochs $\frac{T-t}{T}$.

Thank you!

Q&A