

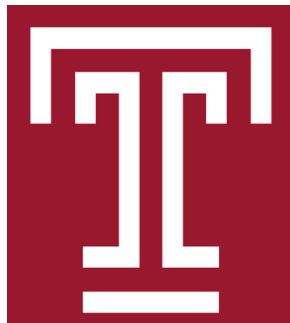
Sharpness-Aware Optimization Through Variance Suppression on Deep AUC Maximization

Xinwen Zhang, Hongchang Gao

PhD student, Assistant Professor

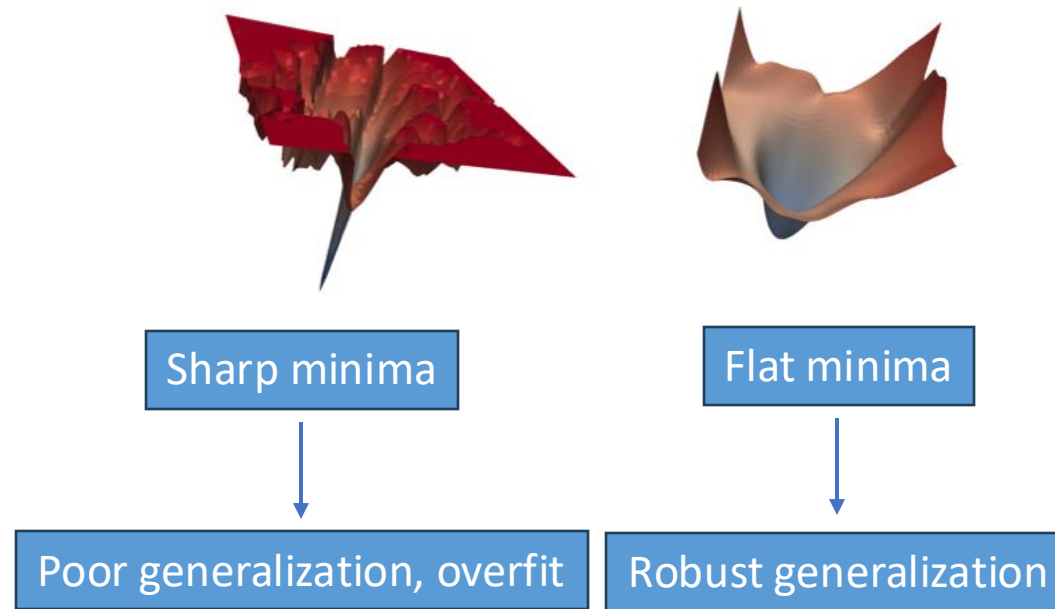
Department of Computer and Information Sciences

Temple University



Motivation

- Generalization in deep learning linked to loss landscape geometry.



- SAM (Sharpness-Aware Minimization) helps by smoothing the loss landscape. [Foret et al, 2021.]

From SAM to Minimax Objectives

- Instead of directly descending the gradient at w_t , SAM updates using the gradient evaluated at the adversarial perturbed point:

$$w_{t+1}^{SAM} = w_t - \eta \nabla L(w_{adv})$$

$$w_{adv} = w_t + \epsilon_t$$

Perturbation: $\epsilon_t = \rho \frac{\nabla L(w_t)}{\|\nabla L(w_t)\|_2}$

- SAM reformulates optimization as

$$\min_w \max_{\|\epsilon\| \leq \rho} L(w + \epsilon)$$

- Limitation: most sharpness-aware tasks target **minimization**, not **minimax** problems.
- This **motivates** us to explore SAM to minimax objectives.

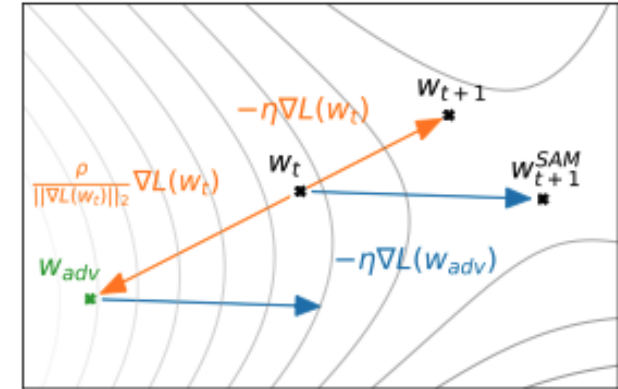


Figure 2: Schematic of the SAM parameter update.

Deep AUC Maximization

- AUC measures the probability that a randomly chosen positive sample is ranked higher than a negative one → crucial for **imbalanced classification**

$$\text{AUC} = P(f(x^+) > f(x^-))$$

- Optimizing AUC can be reformulated as a **minimax** problem: [Ying et al, 2016]

$$\min_{w,a,b} \max_{\alpha} \mathcal{L}_{\text{AUC}}(w, a, b, \alpha)$$

- In deep networks, this leads to a **non-convex-strongly-concave minimax** problem: [Liu et al, 2020]

$$\min_{x=(w,a,b)} \max_y f(x, y)$$

[1] Ying et al, 2016. Stochastic online AUC maximization, NeurIPS.

[2] Liu et al, 2020. Stochastic auc maximization with deep neural networks, ICLR.

Objective and Challenges: Generalization

- Objective: Sharpness-aware AUC maximization

$$\min_{x \in \mathbb{R}^{d_1}} \max_{\substack{\|\epsilon\| \leq \rho \\ y \in \mathbb{R}^{d_2}}} f(x + \epsilon, y)$$

- x : primal variable
- y : dual variable
- ϵ : perturbation bounded by ρ
- f : non-convex in x , strongly concave in y

Why no perturbation on y ?

Strong concavity ensures a unique optimal solution.

Difficult optimization landscape

Friendly adversary issue

- Key challenges:

- Algorithmic gap: how to design **perturbation updates** for **AUC loss** to mitigate **friendly adversary issue** caused by stochastic gradients.
- Theoretical gap: analyzing convergence under **minmax** + **perturbation** is hard.

Proposed Idea

- VaSSO-SGDAM: Variance-Suppressed Sharpness-aware Optimization + Stochastic Gradient Decent Ascent Momentum
 - Friendly Adversary: In stochastic SAM, adversary only reflects **local minibatch** sharpness $\rightarrow$ inconsistent **global** estimation $\rightarrow$ unstable training and weak generalization

- Solution: Introduce a **variance suppression variable** d_t :

$$d_t = (1 - \alpha\eta)d_{t-1} + \alpha\eta\nabla_x f(x_t, y_t; \xi_t) .$$

$$\epsilon_t = \rho\eta \frac{d_t}{\|d_t\|}$$

Use moving-average
to stabilize updates

- Suppress stochastic variance, mitigates friendly adversary, aligns more closely with global sharpness.

Algorithm

Algorithm 1 VaSSO-SGDAM

Input: $x_0, y_0, \rho, \alpha, \eta \in (0, 1), \gamma_x > 0, \gamma_y > 0, \beta_x > 0, \beta_y > 0$.

1: **for** $t = 0, \dots, T - 1$ **do**

2: Update variance suppression variable:

3: $d_t = (1 - \alpha\eta)d_{t-1} + \alpha\eta\nabla_x f(x_t, y_t; \xi_t)$,

4: $\epsilon_t = \rho\eta \frac{d_t}{\|d_t\|}$,

5: Update momentum:

6: **if** $t == 0$ **then**

7: $u_{\frac{1}{2}} = \nabla_x f(x_0 + \epsilon_0, y_0; \xi_0)$,

$v_{\frac{1}{2}} = \nabla_y f(x_0 + \epsilon_0, y_0; \xi_0)$,

8: **else**

9: $u_{t+\frac{1}{2}} = (1 - \beta_x\eta)u_{t-\frac{1}{2}} + \beta_x\eta\nabla_x f(x_t + \epsilon_t, y_t; \xi_t)$,

$v_{t+\frac{1}{2}} = (1 - \beta_y\eta)v_{t-\frac{1}{2}} + \beta_y\eta\nabla_y f(x_t + \epsilon_t, y_t; \xi_t)$,

10: **end if**

11: Update x and y :

12: $x_{t+1} = x_t - \gamma_x\eta u_{t+\frac{1}{2}}$,

13: $y_{t+1} = y_t + \gamma_y\eta v_{t+\frac{1}{2}}$.

14: **end for**

Variance suppression variable
for primal variable

Moving-average momentum
for both primal-dual variables

Primal-dual update

Theoretical Analysis

- Assumption:

- (1) L-smoothness (2) Bounded variance. (3) μ -strong concavity in y

- Auxiliary function:

$$y^*(x) = \arg \max_{y \in \mathbb{R}^{d_2}} f(x, y) ,$$

$$\Phi(x) = \max_{y \in \mathbb{R}^{d_2}} f(x, y(x)) .$$

- Establish convergence bound:

$$\frac{1}{T} \sum_t \mathbb{E} \|\nabla \Phi(x_t)\|^2 = O\left(\frac{1}{\sqrt{T}}\right)$$

- First $O\left(\frac{1}{\sqrt{T}}\right)$ rate for sharpness-aware **non-convex-strongly-concave** problems.

Experiments Setup

- Dataset: CatvsDog, CIFAR10, CIFAR100, STL10.
- Model: ResNet-20, for binary classification.
- Imbalance ratios: $\tau = \frac{N_{\text{pos}}}{N_{\text{pos}} + N_{\text{neg}}}$ (0.1, 0.05, 0.01).
- Baselines: SGDM, SAM, VaSSO,  Minimization Optimizers, standard CE loss
SGDAM, VaSSO-SGDAM  Minimax Optimizers, AUC loss
- Hyperparameter: same learning rate and momentum setup

Experiments Results: Benchmark Datasets

- VaSSO-SGDAM consistently outperforms the baseline methods.

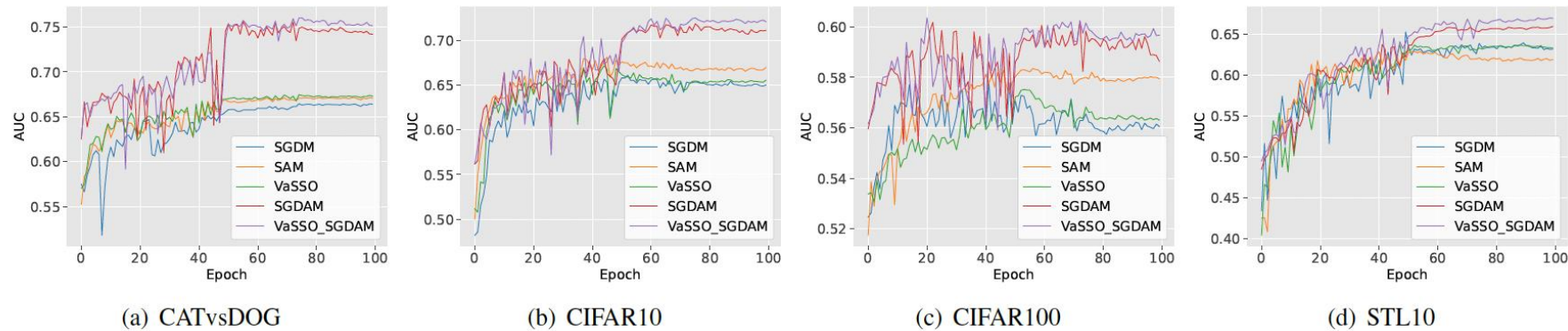


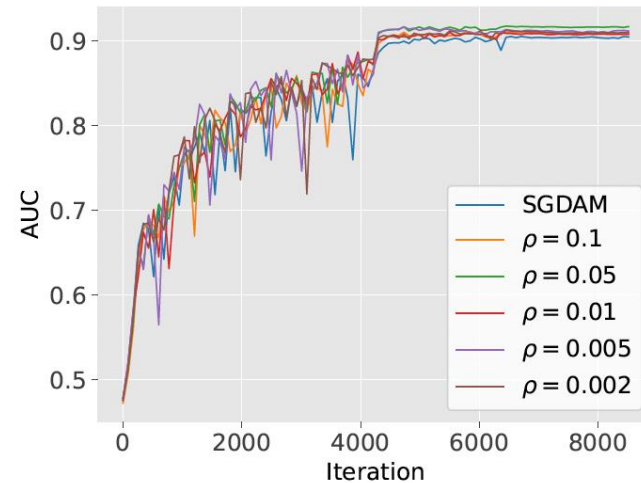
Fig. 3. Testing performance with AUC score versus the number of epochs when the imratio $\tau = 0.01$.

TABLE I
THE COMPARISON BETWEEN THE TEST AUC SCORE OF DIFFERENT ALGORITHMS ON ALL BENCHMARK DATASETS.

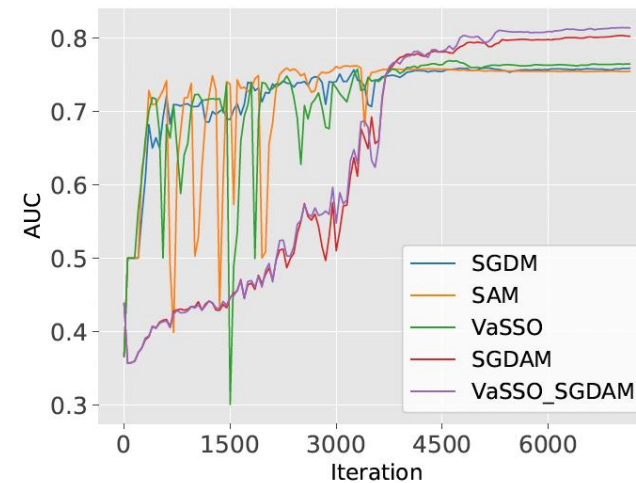
Imratio	Methods	AUC			
		CATvsDOG	CIFAR10	CIFAR100	STL10
$\tau = 0.1$	VaSSO-SGDAM	0.917	0.901	0.699	0.816
	SGDAM	0.905	0.890	0.690	0.798
	VaSSO	0.876	0.869	0.672	0.780
	SAM	0.882	0.872	0.678	0.764
	SGDM	0.874	0.873	0.677	0.771
$\tau = 0.05$	VASSO-SGDAM	0.872	0.855	0.666	0.769
	SGDAM	0.860	0.853	0.659	0.739
	VASSO	0.839	0.797	0.638	0.722
	SAM	0.827	0.779	0.641	0.725
	SGDM	0.846	0.782	0.637	0.707

Experiments Results: Medical Dataset

- Dataset: Melanoma ($\tau = 0.0178$)
- Model: DenseNet-121
- Demonstrates strong generalization on real-world imbalanced medical data.
- Tuning hyperparameter ρ : indicating its robustness



(a) Different ρ on CATvsDOG

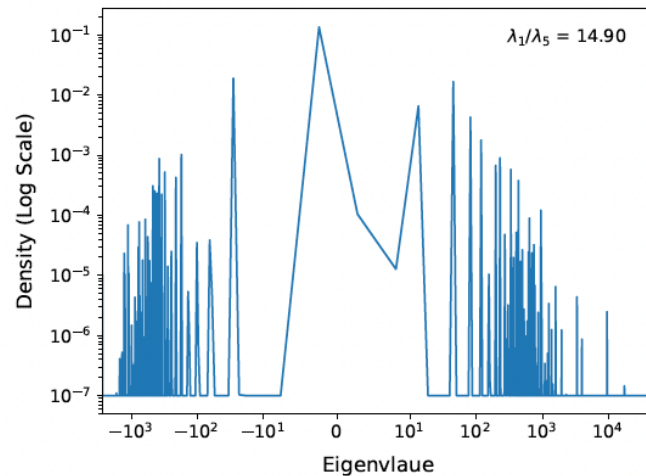


(b) Melanoma Dataset

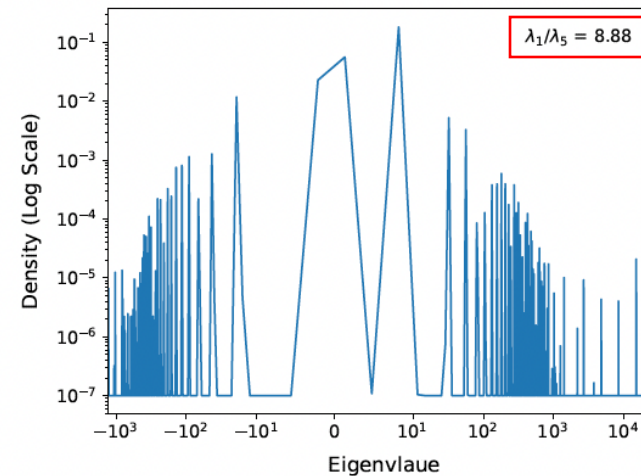
Fig. 4. Testing performance with AUC score versus the number of iteration

Hessian Spectrum Analysis

- Compare eigenvalue spectral densities on CIFAR10.
- VaSSO-SGDAM: lower $\lambda_1/\lambda_5 \longrightarrow$ Flatter minima
- Confirms theoretical link between sharpness suppression & generalization.



(a) SGDAM



(b) VaSSO-SGDAM

Fig. 5. Eigenvalue Spectral Density on CIFAR10.

Conclusion & Future Work

- Conclusion:
 - Bridge sharpness-aware optimization and AUC minimax framework.
 - Proposed VaSSO-SGDAM achieving $O\left(\frac{1}{\sqrt{T}}\right)$ convergence and strong empirical gains.
- Future work:
 - Extend to other minimax tasks (e.g., adversarial training, GANs).
 - Adaptive selection of perturbation ρ
 - Explore distributed(federated/decentralized) variants.