Flatness-aware Curriculum Learning via Adversarial Difficulty

Hiroaki Aizawa[1] Yoshikazu Hayashi[2]

[1] Graduate School of Advanced Science and Engineering, Hiroshima University, Japan [2] Intelligent Production Technology R&D Center for Aerospace, Gifu University, Japan

Summary

Neural networks trained by ERM tend to overfit. Curriculum Learning (CL) selects samples by difficulty to stabilize training. Flat minima improve generalization and robustness. SAM seeks flat minima but makes loss and input gradients uniformly small which weakens conventional difficulty measures. We propose the **Adversarial Difficulty Measure (ADM)** that uses a normalized loss gap between clean and FGSM adversarial examples. ADM keeps difficulty informative in flatter regions. We integrate ADM with CL and SAM and evaluate on image classification, corruption robustness, fine-grained recognition, and domain generalization. Our method preserves the benefits of CL and SAM and yields strong generalization.

Introduction

Background. Empirical Risk Minimization (ERM) trains models by minimizing the average loss on training data. At the same time ERM can overfit to spurious patterns in specific samples or domains. Such overfitting reduces generalization performance. Curriculum Learning (CL) addresses this issue by selecting training based on sample difficulty. Self-Paced Learning (SPL) [1] measures difficulty by the loss value and prioritizes low loss samples at early stages. Symmetric SPL (SSPL) [2] extends this idea with a symmetric scheduler and a difficulty measure based on the norm of the input gradient. From the optimization perspective, flat minima are associated with better robustness and generalization. Sharpness-Aware Minimization (SAM) [3] promotes convergence to flat minima by minimizing the worst case loss in a small neighborhood of the parameters. This optimization encourages solutions whose loss is insensitive to small parameter perturbations. As training progresses toward flatter regions both loss values and gradient norms tend to become small.

Problem. As training progresses toward flat minima under Sharpness-Aware Minimization (SAM), both the per-sample loss values and input-gradient norms become uniformly small, causing conventional difficulty measures based on loss or gradient to lose separability in the later stages of training and weakening curricula such as SPL and SSPL, as illustrated in *Figure 1 (right)*, where loss- and gradient-based difficulties collapse near flat minima and no longer distinguish samples.

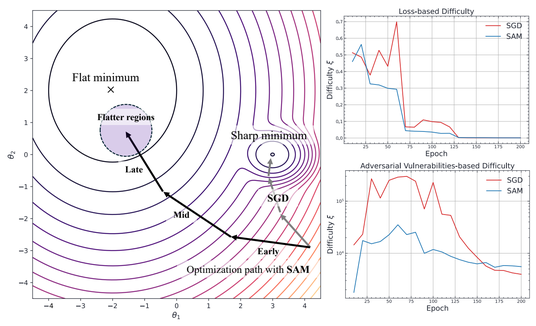

Figure: Illustration of the optimization trajectory toward a flat minimum (left) and visualizations of the loss-based difficulty (top-right) and our proposed measure (bottom-right). As the optimization trajectory moves toward the flat minimum in the late phase of training (approximately 125-200 epochs), the loss-based difficulty becomes uniformly small, making it hard to evaluate examples for CL. In contrast, our difficulty, measured as the normalized loss gap between clean and adversarial samples, remains informative signal for CL.

Idea. We define an *Adversarial Difficulty Measure (ADM)* that quantifies sample difficulty by the normalized loss gap between a clean input and its adversarial counterpart generated by a one-step FGSM variant [4]. The difficulty score $\xi_{\mathrm{adv}}(\boldsymbol{x}_i)$ is defined as the absolute loss difference divided by the clean loss plus a small constant. Since adversarial vulnerability remains informative even in flat regions, ADM preserves meaningful difficulty contrast when both loss and gradient magnitudes have diminished.

Contribution. We bridge Curriculum Learning (CL) and flatness-aware optimization by introducing the Adversarial Difficulty Measure (ADM) and integrating it with SPL and SSPL under Sharpness-Aware Minimization (SAM). The proposed mechanism is plug-and-play and requires almost no additional cost. Across standard image classification, corruption robustness, fine-grained recognition, and Domain Generalization, ADM-based curricula retain the benefits of SAM while sustaining informative sample selection throughout training. Empirically, models trained with ADM converge to flatter minima with smaller maximum Hessian eigenvalues and show improved generalization and robustness.

Method

Curriculum-Based Weighted Loss. We consider a K-class classification task with a labeled dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, \ldots, K\}$. A classification model $f(\cdot;\theta)$ with parameters θ is trained by minimizing the weighted loss

$$\mathcal{L}_{\mathrm{CL}}(\theta, \boldsymbol{v}) = \sum_{i=1}^{N} v_i \ell(f(\boldsymbol{x}_i;\theta), y_i),$$

where ℓ denotes the cross-entropy loss and $\boldsymbol{v} \in \mathbb{R}^N$ is the weight vector defining the curriculum. Setting all $v_i = 1$ corresponds to Empirical Risk Minimization (ERM). Each weight v_i is dynamically updated according to a sample difficulty score ξ_i, which can be based on the loss value (SPL) or the gradient norm (SSPL).

Seeking Flat Minima in Curriculum Learning. To seek flat minima from the weighted loss $\mathcal{L}_{\mathrm{CL}}$, we apply Sharpness-Aware Minimization (SAM) to promote robustness to parameter perturbations within an ℓ_2 ball of radius ρ :

$$\tilde{\mathcal{L}}_{\mathrm{CL}}(\theta, \boldsymbol{v}) = \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}_{\mathrm{CL}}(\theta + \epsilon, \boldsymbol{v})$$

where $\epsilon \in \mathbb{R}^{|\theta|}$ is a perturbation and $\rho > 0$ controls the neighborhood size. The perturbation that maximizes the loss is approximated as

$$\epsilon^* \approx \rho \frac{\nabla_\theta \mathcal{L}_{\mathrm{CL}}(\theta, \boldsymbol{v})}{\|\nabla_\theta \mathcal{L}_{\mathrm{CL}}(\theta, \boldsymbol{v})\|_2},$$

and the parameters are updated using the gradient at the perturbed point

$$\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{CL}}\left(\theta + \epsilon^*, \boldsymbol{v}\right),$$

where η is the learning rate. This procedure couples curriculum-based weighting with flatness-aware optimization.

Adversarial Difficulty Measure. Near flat minima, both loss values and input-gradient norms become uniformly small, which reduces the separability of conventional difficulty measures. To address this, we evaluate *adversarial vulnerability*. Using input gradients already computed in the SAM backward pass, we generate adversarial examples by a one-step FGSM variant

$$\boldsymbol{x}_i^{\mathrm{adv}} = \boldsymbol{x}_i + \alpha \operatorname{sign}\left(\nabla_{\boldsymbol{x}_i} \ell\left(f\left(\boldsymbol{x}_i;\theta\right), y_i\right)\right), \quad \alpha = \frac{8}{255}.$$

The difficulty score is defined as the normalized loss gap between the adversarial and clean inputs

$$\xi_{\mathrm{adv}}\left(\boldsymbol{x}_i\right) = \frac{\left|\ell\left(f\left(\boldsymbol{x}_i^{\mathrm{adv}};\theta\right), y_i\right) - \ell\left(f\left(\boldsymbol{x}_i;\theta\right), y_i\right)\right|}{\ell\left(f\left(\boldsymbol{x}_i;\theta\right), y_i\right) + \delta}, \quad \delta > 0.$$

Samples with small clean loss but large adversarial increase are considered *hard yet informative*. Normalization stabilizes the scale during late training and preserves difficulty contrast in flat regions.

Curriculum Update. Based on the difficulty scores ξ_i, the sample weights are updated following the SPL rule

$$v_i^{(t+1)} = \begin{cases} 1, & \xi_i \le \lambda^{(t)} \\ 0, & \text{otherwise} \end{cases}$$

Here, $\lambda^{(t)}$ is a threshold at epoch t. By gradually increasing $\lambda^{(t)}$, training starts with easy samples that have low ADM scores and progressively includes harder ones. Since ADM remains discriminative even in flat regions, it allows reliable sample selection and enables curriculum learning to stay effective under SAM.

Evaluation

Main Results

Fine-grained image classification accuracies (%)

Curr. v Diff. ξ		Cars	Aircraft	CUB	Food
Opt.: SGD					
ERM	-	$82.82_{\pm 0.24}$	$77.97_{\pm 0.57}$	$61.64_{\pm 0.20}$	$79.60_{\pm 0.01}$
SPL	Loss	$82.64_{\pm 0.06}$	$77.55_{\pm 0.43}$	$61.96_{\pm 0.22}$	$79.45_{\pm 0.10}$
	ADM	$82.14_{\pm 0.19}$	$78.44_{\pm 0.54}$	$61.95_{\pm 0.29}$	$79.66_{\pm 0.07}$
SSPL	Loss	$82.88_{\pm 0.27}$	$77.95_{\pm 0.19}$	$63.01_{\pm 0.13}$	$79.38_{\pm 0.14}$
	GDM	$82.64_{\pm 0.19}$	$77.87_{\pm 0.48}$	$63.03_{\pm 0.66}$	$79.61_{\pm 0.19}$
	ADM	$\mathbf{83.89}_{\pm 0.09}$	$\mathbf{78.80}_{\pm 0.35}$	$\mathbf{63.63}_{\pm 0.11}$	$\mathbf{80.20}_{\pm 0.08}$
Opt.: SAM					
ERM	-	$83.24_{\pm 0.32}$	$78.58_{\pm 0.40}$	$62.02_{\pm 0.40}$	$80.91_{\pm 0.13}$
SPL	Loss	$82.97_{\pm 0.55}$	$78.73_{\pm 0.61}$	$62.10_{\pm 0.73}$	$80.89_{\pm 0.05}$
	ADM	$83.26_{\pm 0.50}$	$78.44_{\pm 0.12}$	$62.65_{\pm 0.22}$	$80.93_{\pm 0.11}$
SSPL	Loss	$83.64_{\pm 0.56}$	$78.72_{\pm 0.42}$	$63.39_{\pm 0.56}$	$81.16_{\pm 0.11}$
	GDM	$83.33_{\pm 0.19}$	$78.36_{\pm 0.04}$	$63.28_{\pm 0.27}$	$80.95_{\pm 0.09}$
	ADM	$\mathbf{83.85}_{\pm 0.05}$	$\mathbf{79.32}_{\pm 0.41}$	$\mathbf{63.44}_{\pm 0.24}$	$\mathbf{82.01}_{\pm 0.09}$

Domain generalization (DG) accuracies (%)

Curr. v Diff. ξ		PACS	VLCS	OH	TI	Avg.
Opt.: Adam						
ERM	-	$\mathbf{82.16}_{\pm 0.77}$	$\mathbf{75.35}_{\pm 1.02}$	$55.32_{\pm 0.70}$	$42.07_{\pm 0.35}$	63.72
SPL	Loss	$80.36_{\pm 1.37}$	$74.95_{\pm 1.48}$	$55.18_{\pm 0.57}$	$39.64_{\pm 0.48}$	62.53
	ADM	$80.82_{\pm 0.04}$	$75.08_{\pm 0.11}$	$54.27_{\pm 0.32}$	$40.05_{\pm 1.90}$	62.55
SSPL	Loss	$80.08_{\pm 0.78}$	$73.86_{\pm 0.48}$	$46.43_{\pm 0.95}$	$38.92_{\pm 1.41}$	59.82
	GDM	$79.90_{\pm 0.41}$	$73.78_{\pm 0.37}$	$56.01_{\pm 0.60}$	$40.59_{\pm 1.68}$	62.57
	ADM	$81.11_{\pm 0.69}$	$75.03_{\pm 1.03}$	$\mathbf{57.10}_{\pm 0.37}$	$\mathbf{42.81}_{\pm 0.46}$	**64.01**
Opt.: SAM						
ERM	-	$84.66_{\pm 0.38}$	$76.84_{\pm 0.39}$	$60.76_{\pm 0.17}$	$44.55_{\pm 0.50}$	66.70
SPL	Loss	$84.54_{\pm 0.95}$	$76.88_{\pm 0.34}$	$\mathbf{61.19}_{\pm 0.31}$	$45.36_{\pm 0.57}$	66.99
	ADM	$\mathbf{85.43}_{\pm 0.40}$	$\mathbf{77.13}_{\pm 0.38}$	$59.32_{\pm 0.16}$	$43.19_{\pm 2.86}$	66.27
SSPL	Loss	$83.60_{\pm 0.31}$	$76.79_{\pm 0.53}$	$53.33_{\pm 0.87}$	$42.63_{\pm 0.63}$	64.09
	GDM	$82.47_{\pm 0.67}$	$75.06_{\pm 0.32}$	$58.36_{\pm 0.18}$	$40.39_{\pm 2.73}$	64.07
	ADM	$83.88_{\pm 1.32}$	$76.87_{\pm 0.36}$	$60.88_{\pm 0.26}$	$\mathbf{46.92}_{\pm 0.77}$	**67.14**
Flatness-aware DG approach: SAGM						
ERM	-	$83.72_{\pm 0.74}$	$76.31_{\pm 0.47}$	$60.34_{\pm 0.21}$	$46.00_{\pm 1.61}$	66.59

Visualization of Difficulties

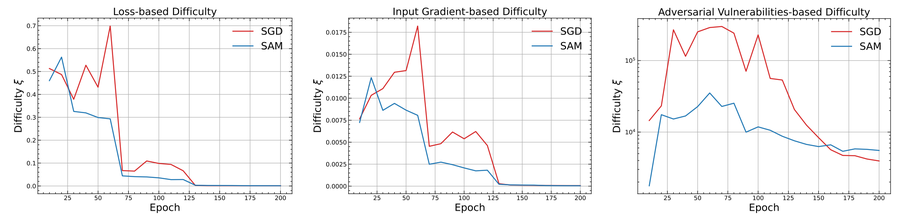

Figure: Visualization of the difficulties for curriculum learning during the training on CIFAR10 with SGD and SAM.

Flatness Analysis

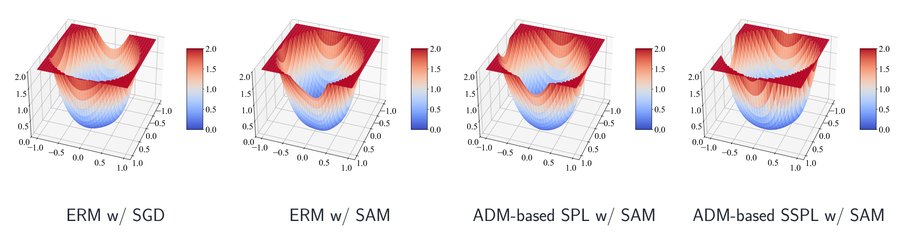

Figure: Visualization of loss landscapes on CIFAR-10.

Table: Maximum Hessian eigenvalue. Lower is flatter.

	CIFAR10		CIFAR100	
	SGD	SAM	SGD	SAM
ERM	8.25	3.03	**22.87**	9.47
SPL	**6.17**	**2.35**	33.97	**7.49**
SSPL	8.55	6.11	78.78	19.92

Experimental Summary

- ADM improves fine-grained classification, especially when combined with SSPL under SAM.
- ADM-based curricula enhance domain generalization by maintaining effective sample selection.
- Models trained with ADM exhibit flatter loss surfaces (smaller Hessian eigenvalues) and retain discriminative difficulty signals in flat regions.

References

1. Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. Self-paced curriculum learning. In AAAI, volume 29, 2015.
2. Di Zhao, Yun Sing Koh, Gillian Dobbie, Hongsheng Hu, and Philippe Fournier-Viger. Symmetric self-paced learning for domain generalization. In AAAI, volume 38, pages 1696116969, 2024.
3. Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In ICLR, 2021.
4. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.