SHARD-SAM: SIGNED HESSIAN-AMPLIFIED RADIUS AND DIRECTION FOR SHARPNESS-AWARE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconciling optimization and generalization remains a fundamental challenge in deep learning. Sharpness-Aware Minimization (SAM) advances this goal, yet its fixed perturbation radius and first-order geometric approximation limit its adaptability to heterogeneous loss landscapes. We propose SHARD-SAM (Signed Hessian-Amplified Radius and Direction), a sharpness-aware framework that jointly adapts perturbation magnitude and orientation through a shared curvature signal. SHARD-SAM constructs a signed Hessian-amplified spectral surrogate using zeroth-order curvature probes. A symmetric asymptotic envelope adjusts the perturbation radius relative to a running curvature baseline, while magnitude-adaptive gating combines the gradient with a spectral correction. A shared curvature-temperature parameter controls both modules, coupling radius and direction adaptation within a unified perturbation rule. Our theoretical analysis characterizes the surrogate's spectral properties, estimation error, and conditions for local directional benefits. Experiments on CIFAR-10, CIFAR-100, and ImageNet across CNN and ViT architectures demonstrate consistent generalization improvements over representative first-order and second-order geometry-adaptive SAM optimizers, with a favorable accuracy–efficiency trade-off among the evaluated curvature-aware methods. Ablations and direction-source controls support the contribution of the shared curvature signal and its two adaptive modules.

1 INTRODUCTION

Neural networks can fit even random labels (Zhang et al., 2021), yet their generalization depends on more than training loss. Flat minima are associated with robustness to parameter perturbations (Hochreiter & Schmidhuber, 1997; Keskar et al., 2017; Jiang et al., 2020), although their sensitivity to reparameterization motivates normalized measures (Dinh et al., 2017; Tsuzuku et al., 2020; Petzka et al., 2021; Andriushchenko et al., 2023b). SAM (Foret et al., 2021) and ASAM (Kwon et al., 2021) exploit this geometry through adversarial weight perturbations. Their behavior also involves implicit gradient regularization, gradient–Hessian alignment, and representation-rank reduction (Andriushchenko & Flammarion, 2022; Wen et al., 2023; Andriushchenko et al., 2023a). These findings motivate a closer look at how the perturbations are constructed. The perturbation radius determines how far the inner maximization explores, while its direction determines which local variations it probes. Standard SAM combines a fixed radius with a gradient-based direction, limiting its response to changes in local curvature. Methods such as LETS (Wang et al., 2024) and ARWP (Li et al., 2024a) adapt perturbations without explicitly using the Hessian spectrum. Eigen-SAM (Luo et al., 2024) and CR-SAM (Wu et al., 2024) incorporate Hessian information through leading eigenvectors and trace regularization, respectively, but require additional computation to estimate curvature. This suggests using a shared curvature signal to coordinate perturbation magnitude and orientation at a practical computational cost.

We propose SHARD-SAM (Signed Hessian-Amplified Radius and Direction), which uses a signed Hessian-amplified spectral surrogate to jointly adapt the perturbation radius and direction (Figure 1). We construct the surrogate from forward-only loss queries and extract its spectral direction through implicit matrix-vector products without forming a dense matrix. The resulting curvature score drives

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

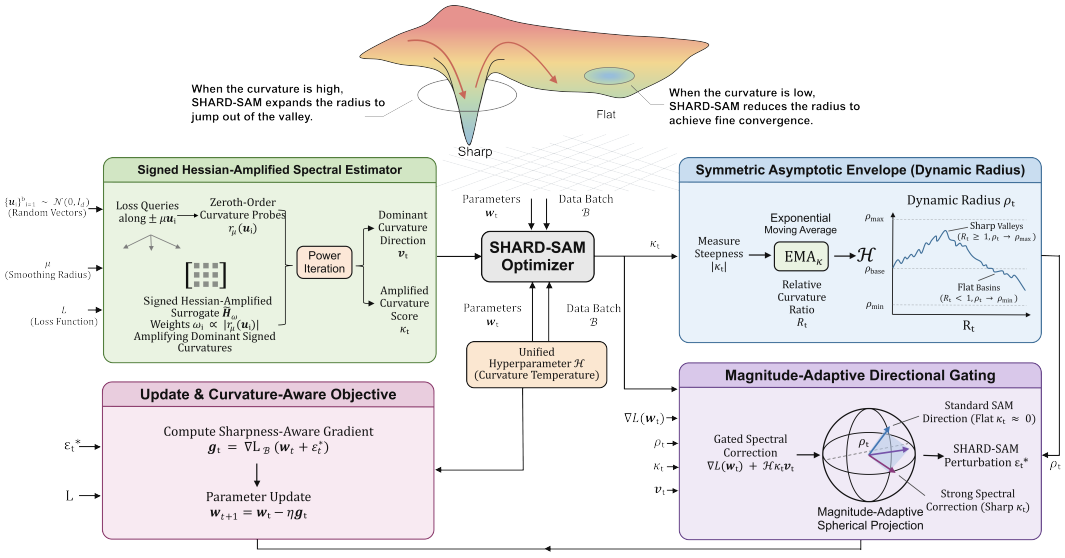


Figure 1: SHARD-SAM framework: a shared spectral signal controls the perturbation radius and direction.

two complementary modules. A bounded symmetric asymptotic envelope expands or contracts the radius according to the score magnitude relative to its moving-average baseline. A magnitude-adaptive directional gate combines the gradient with a score-weighted spectral correction and approaches the SAM direction as the correction vanishes. A shared curvature-temperature parameter controls both modules, linking the perturbation radius and direction through the same geometric signal. Experiments on CIFAR-10, CIFAR-100, and ImageNet show consistent generalization gains across the evaluated CNNs and ViTs, with a favorable accuracy–efficiency trade-off among the compared curvature-aware methods. Component ablations and matched direction-source controls examine how curvature information and the two adaptive modules contribute to these gains. Our analysis characterizes the population spectral structure, finite-probe estimation error, conditions for local directional benefit, and convergence. Detailed derivations are provided in the supplementary material.

In summary, we make the following contributions:

- We propose a unified curvature-aware perturbation rule that uses a signed zeroth-order spectral surrogate to jointly adapt SAM's radius and direction.
- We demonstrate generalization improvements across CNNs and ViTs, supported by component ablations, direction-source controls, and computational-efficiency evaluations.
- We provide theoretical support for the method through analyses of spectral structure, estimation error, local directional benefit conditions, and convergence, presented in the supplementary material.

2 Related Work

PAC-Bayesian analysis (Dziugaite & Roy, 2017) and empirical visualization (Li et al., 2018; Izmailov et al., 2018) have studied the relationship between loss-landscape flatness and generalization. We group sharpness-aware methods into first-order and second-order approaches.

First-order approaches include SAM, which minimizes the worst-case loss within a fixed perturbation neighborhood, and ASAM, which introduces scale-invariant perturbations. Subsequent methods improve computational efficiency or adapt perturbations using first-order signals and stochastic heuristics. LookSAM (Liu et al., 2022a) periodically reuses projected gradients, GSAM minimizes a surrogate sharpness gap (Zhuang et al., 2022), and other methods use Fisher information (Kim et al., 2022) or random weight perturbations (Du et al., 2022; Li et al., 2022). Sparse SAM restricts the set

of perturbed parameters (Mi et al., 2022), Random SAM smooths the inner maximization (Liu et al., 2022b), and VaSSO reduces perturbation variance (Li & Giannakis, 2023). Friendly-SAM removes the estimated full-gradient component (Li et al., 2024b), whereas SAMPa parallelizes the gradient computations of SAM (Xie et al., 2024). LETS learns the perturbation radius through bilevel optimization based on validation signals (Wang et al., 2024). These methods do not explicitly use the Hessian spectrum to jointly adapt the perturbation magnitude and direction.

Among second-order approaches, Eigen-SAM (Luo et al., 2024) incorporates the dominant Hessian eigenvector through intermittent Hessian-vector products with a fixed mixing coefficient. CR-SAM (Wu et al., 2024) regularizes the Hessian trace using finite differences and an additional backward pass. SASSHA (Shin et al., 2025) combines sharpness-aware optimization with a stabilized Hessian approximation and lazy updates. SHARD-SAM uses a shared forward-only spectral signal to control both perturbation-radius adaptation and directional gating.

3 Preliminaries

3.1 Sharpness-Aware Minimization

We consider a neural network parameterized by $\boldsymbol{w} \in \mathbb{R}^d$ and an empirical loss $L_{\mathcal{S}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \boldsymbol{w}))$. The standard SAM framework (Foret et al., 2021) minimizes the worst-case loss within a Euclidean ρ-ball:

$$\min_{\boldsymbol{w}} \max_{\|\boldsymbol{\varepsilon}\|_2 \le \rho} L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\varepsilon}) + \frac{\lambda^{(1)}}{2}\|\boldsymbol{w}\|_2^2. \tag{1}$$

A first-order Taylor expansion gives the approximate adversarial perturbation $\boldsymbol{\varepsilon}_t = \rho \frac{\nabla L_{\mathcal{S}}(\boldsymbol{w}_t)}{\|\nabla L_{\mathcal{S}}(\boldsymbol{w}_t)\|_2}$ and the weight update:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \alpha_t \left(\nabla L_{\mathcal{S}}(\boldsymbol{w}_t + \boldsymbol{\varepsilon}_t) + \lambda^{(1)} \boldsymbol{w}_t \right). \tag{2}$$

This formulation uses a fixed radius ρ and does not explicitly account for curvature.

3.2 Gaussian Smoothing for Curvature Estimation

We use Gaussian smoothing (Ye et al., 2025) to estimate curvature without forming the full Hessian $\nabla^2 L(\boldsymbol{w}) \in \mathbb{R}^{d \times d}$. Gaussian smoothing and random directional finite differences are standard tools in derivative-free and zeroth-order optimization (Ghadimi & Lan, 2013; Nesterov & Spokoiny, 2017). Higher-order Stein identities also allow Hessian estimation from function-value queries (Balasubramanian & Ghadimi, 2022). Define the smoothed loss as $L_\mu(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{u}\sim\mathcal{N}(\mathbf{0}, I_d)}[L(\boldsymbol{w} + \mu\boldsymbol{u})]$. For b random directions $\{\boldsymbol{u}_i\}_{i=1}^{b} \sim \mathcal{N}(\mathbf{0}, I_d)$, the finite-difference estimator is

$$\widetilde{H} = \frac{1}{b}\sum_{i=1}^{b} \frac{L(\boldsymbol{w} + \mu\boldsymbol{u}_i) + L(\boldsymbol{w} - \mu\boldsymbol{u}_i) - 2L(\boldsymbol{w})}{2\mu^2} \boldsymbol{u}_i \boldsymbol{u}_i^\top + \lambda^{(2)} I_d. \tag{3}$$

The estimator satisfies $\mathbb{E}[\widetilde{H}] = \nabla^2 L_\mu(\boldsymbol{w}) + c \cdot I_d$ (Ye et al., 2025), where c is a scalar identity shift. This shift leaves the population smoothed Hessian's eigenspaces unchanged. Finite-sample eigendirections require a separate error analysis. The next section introduces curvature weighting without explicit Hessian-vector products.

4 Methodology: SHARD-SAM

4.1 Signed Hessian-Amplified Spectral Estimation

For a fixed mini-batch and b Gaussian probes $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, I_d)$, two loss queries per direction at scale $\mu > 0$ give the zeroth-order curvature response

$$r_\mu(\boldsymbol{u}_i) \triangleq \frac{L(\boldsymbol{w} + \mu\boldsymbol{u}_i) + L(\boldsymbol{w} - \mu\boldsymbol{u}_i) - 2L(\boldsymbol{w})}{2\mu^2} \approx \frac{1}{2}\boldsymbol{u}_i^\top \nabla^2 L(\boldsymbol{w}) \boldsymbol{u}_i. \tag{4}$$

This scalar estimates directional quadratic curvature. SHARD-SAM gives larger-magnitude responses more weight through

$$w_i = \frac{|r_\mu(\boldsymbol{u}_i)|}{\sum_{j=1}^{b} |r_\mu(\boldsymbol{u}_j)| + \epsilon}, \tag{5}$$

where ϵ is a small stability constant. The resulting signed Hessian-amplified surrogate is constructed implicitly as

$$\widetilde{\boldsymbol{H}}_w \approx \sum_{i=1}^{b} w_i r_\mu(\boldsymbol{u}_i) \boldsymbol{u}_i \boldsymbol{u}_i^\top = \sum_{i=1}^{b} \frac{\operatorname{sign}(r_\mu(\boldsymbol{u}_i)) r_\mu(\boldsymbol{u}_i)^2}{\sum_{j=1}^{b} |r_\mu(\boldsymbol{u}_j)| + \epsilon} \boldsymbol{u}_i \boldsymbol{u}_i^\top. \tag{6}$$

The factor $r_\mu|r_\mu|$ preserves probe-response signs, not necessarily every Hessian eigenvalue's sign. This signed surrogate is neither the raw Hessian nor an unsigned Hessian square.

We use power iteration to estimate a dominant-magnitude eigendirection of the signed surrogate. The matrix-vector product required at iteration k can be computed without materializing $\widetilde{\boldsymbol{H}}_w$:

$$\widetilde{\boldsymbol{H}}_w \boldsymbol{v}_k = \sum_{i=1}^{b} \left(w_i r_\mu(\boldsymbol{u}_i) \langle \boldsymbol{u}_i, \boldsymbol{v}_k \rangle \right) \boldsymbol{u}_i. \tag{7}$$

The post-query algebraic cost is $\mathcal{O}(d^2)$ for a dense matrix product, compared with $\mathcal{O}(b \cdot d)$ for this implicit calculation. After power iteration, we compute the Hessian-amplified spectral score

$$\kappa_t \triangleq \boldsymbol{v}_t^\top \widetilde{\boldsymbol{H}}_w \boldsymbol{v}_t. \tag{8}$$

Here, κ_t is a spectral control signal, not an unbiased estimate of a raw Hessian eigenvalue. Appendix A characterizes the population target; Sections A.2–A.5 quantify finite-probe, smoothing, and iteration errors and the limits of high-dimensional recovery.

4.2 Symmetric Asymptotic Envelope for Dynamic Radius

To adapt the radius to changes in the local spectral signal, we track the magnitude of κ_t with an exponential moving average:

$$\bar{\kappa}_t \triangleq |\kappa_t|, \quad \mathrm{EMA}_{\kappa,t} = \beta \cdot \mathrm{EMA}_{\kappa,t-1} + (1-\beta) \cdot \bar{\kappa}_t. \tag{9}$$

We then define the relative curvature ratio as

$$R_t = \frac{\bar{\kappa}_t}{\mathrm{EMA}_{\kappa,t} + \epsilon}. \tag{10}$$

We map this ratio to the perturbation radius through a bounded, C^0-continuous symmetric asymptotic envelope governed by curvature temperature $\mathcal{H} \in (0, 1]$:

$$\rho_t = \begin{cases} \rho_{\max} - (\rho_{\max} - \rho_{\text{base}}) \left(\frac{1}{R_t} \right)^{\mathcal{H}}, & R_t \geq 1, \\ \rho_{\min} + (\rho_{\text{base}} - \rho_{\min}) (R_t)^{\mathcal{H}}, & R_t < 1. \end{cases} \tag{11}$$

For $R_t \geq 1$, the radius increases toward $\rho_{\max}$; for $R_t < 1$, it decreases toward $\rho_{\min}$ as the score falls relative to its historical baseline. It equals ρ_{base} at $R_t = 1$. This is relative-curvature control, not a prescribed decay schedule: a constant nonzero score gives $R_t \approx 1$ after the EMA settles. Sections B.2–B.3 establish envelope stability and propagate estimation errors.

4.3 Magnitude-Adaptive Directional Gating

The radius determines perturbation magnitude; the gate determines its orientation. SHARD-SAM combines the gradient with a score-weighted spectral direction and normalizes the result to the prescribed radius:

$$\boldsymbol{\varepsilon}_t^* = \rho_t \cdot \frac{\nabla L(\boldsymbol{w}_t) + \mathcal{H} \kappa_t \boldsymbol{v}_t}{\| \nabla L(\boldsymbol{w}_t) + \mathcal{H} \kappa_t \boldsymbol{v}_t \|_2}. \tag{12}$$

We orient the unit vector so that $\boldsymbol{v}_t^\top \nabla L(\boldsymbol{w}_t) \geq 0$ before gating. For a nonzero gradient, the direction approaches SAM as $\kappa_t \to 0$; if the projection denominator is zero, we set the perturbation to zero. Appendix B.1 derives local benefit conditions for a small positive gate with partially gradient-aligned positive curvature; Section B.3 propagates finite-probe error.

4.4 Unified Algorithmic Framework

Combining the radius and gate gives the following stepwise perturbed-loss formulation:

$$\min_{\boldsymbol{w}} \mathcal{L}^{\text{SHARD}}(\boldsymbol{w}) \triangleq \min_{\boldsymbol{w}} L\left(\boldsymbol{w} + \boldsymbol{\varepsilon}_t^*(\boldsymbol{w}, \kappa_t)\right), \tag{13}$$

where $\boldsymbol{\varepsilon}_t^*$ is defined in Eq. 12. This formulation is history-dependent: the update differentiates the loss while holding the current perturbation fixed, without differentiating through the spectral estimate or EMA. Algorithm 1 summarizes the resulting plain-gradient update. The experiments use the momentum and AdamW variants specified in Appendix E.4; the convergence analysis in Appendix C concerns the plain update. Appendix D complements this analysis with a PAC-Bayesian risk bound for an analytical curvature-oriented posterior, exposing the roles of radius and orientation.

Algorithm 1 SHARD-SAM Optimization Framework

Input: Data $\mathcal{S}$, Loss L, L.R. η, Batch B, Sample size b, Smoothing scale μ, Temp $\mathcal{H} \in (0, 1]$, Structural bounds $\rho_{\text{base}}, \rho_{\min}, \rho_{\max}, \beta$.
Output: Optimized weights $\boldsymbol{w}^*$.

1: Initialize $\boldsymbol{w}_0$, $t = 0$, $\text{EMA}_{\kappa,-1} = 0$;
2: **while** not converged **do**
3: Sample batch $\mathcal{B} \subset \mathcal{S}$ and random vectors $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_b\} \sim \mathcal{N}(\mathbf{0}, I_d)$;
4: Phase 1 (Spectral Est.): Compute $r_\mu(\boldsymbol{u}_i)$ via Eq. 4 and weights $w_i \propto |r_\mu(\boldsymbol{u}_i)|$; run Power Iterations via Eq. 7 to extract $\boldsymbol{v}_t$ and compute $\kappa_t = \boldsymbol{v}_t^\top \widetilde{H}_w \boldsymbol{v}_t$. If a matrix-vector product is zero, stop, retain the current unit vector, and set $\kappa_t = 0$;
5: Phase 2 (Radius Envelope): Update $\bar{\kappa}_t = |\kappa_t|$ and $\text{EMA}_{\kappa,t} = \beta\text{EMA}_{\kappa,t-1} + (1-\beta)\bar{\kappa}_t$; compute $R_t = \bar{\kappa}_t / (\text{EMA}_{\kappa,t} + \epsilon)$ and dynamic ρ_t via Eq. 11;
6: Phase 3 (Direction Gate): Orient $\boldsymbol{v}_t$ toward $\nabla L_{\mathcal{B}}(\boldsymbol{w}_t)$; compute gated perturbation $\boldsymbol{\varepsilon}_t^*$ via Eq. 12;
7: Phase 4 (Update): $\boldsymbol{g}_t = \nabla L_{\mathcal{B}}(\boldsymbol{w}_t + \boldsymbol{\varepsilon}_t^*)$; update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}_t$; $t = t + 1$;
8: **end while**
9: return $\boldsymbol{w}^* = \boldsymbol{w}_t$.

5 Experiments

5.1 Experimental Setup

We conduct experiments on CIFAR-10/100 (Krizhevsky, 2009) and ImageNet-1K (Deng et al., 2009) using PyTorch (Paszke et al., 2019) on RTX A6000 GPUs. Models include ResNet-18/50/101 (He et al., 2016), WRN-28-10 (Zagoruyko & Komodakis, 2016), DenseNet-121 (Huang et al., 2022), PyramidNet-110/272 (Han et al., 2017), and ViT-B/16 for CIFAR and ViT-B/32 for ImageNet (Dosovitskiy et al., 2021). We compare SHARD-SAM with five first-order sharpness-aware methods, SAM (Foret et al., 2021), ASAM (Kwon et al., 2021), LookSAM (Liu et al., 2022a), F-SAM (Li et al., 2024b), and LETS (Wang et al., 2024), and two second-order methods, Eigen-SAM (Luo et al., 2024) and CR-SAM (Wu et al., 2024). All methods use the same random cropping and horizontal flipping, without MixUp (Zhang et al., 2018), CutMix (Yun et al., 2019), or RandAugment (Cubuk et al., 2020). In the epoch-matched benchmarks, CIFAR CNNs and ViT-B/16 use a global batch size of 256 and train for 200 and 300 epochs, respectively. On ImageNet, ResNet-50 uses a global batch size of 256 and trains for 90 epochs, while ViT-B/32 uses a global batch size of 512 and trains for 300 epochs. Learning-rate schedules are matched within each setting, but wall-clock and FLOP budgets are not. We fix $\rho_{\text{base}} = 0.1$, $\rho_{\max} = 5.0\rho_{\text{base}}$, $\rho_{\min} = 0.05\rho_{\text{base}}$, and $\beta = 0.9$ across architectures. With the estimator settings fixed, we select the curvature temperature from $\mathcal{H} \in \{0.3, 0.5, 0.7\}$. Appendices E.1 and E.2 evaluate sensitivity to the radius bounds and EMA decay on CIFAR-100/WRN-28-10. Appendix E.4 provides the training recipes, including initialization, smoothing scale, and power-iteration count. The additional comparison in Table 3 includes SASSHA. Method-specific configurations are selected using held-out training data, independently of the five final seeds. Mechanism interventions keep the training recipe fixed. Appendix E.4 specifies the splits, search budgets, and paired-run protocol. These additional runs are separate from the original comparisons reported below.

5.2 Main Results on Generalization

Table 1 reports top-1 test accuracy over five independent runs, with Welch's t-test comparisons at $p < 0.05$. On CIFAR-100/WRN-28-10, SHARD-SAM reaches 84.93%, exceeding SAM and CR-SAM by 1.78 and 0.48 percentage points. On ImageNet-1K/ViT-B/32, it reaches 75.95%, compared with 75.28% for CR-SAM. These gains extend across the CNN and ViT settings in the table.

Table 2 broadens the CNN comparison to ResNet-18, ResNet-101, DenseNet-121, and PyramidNet-110/272. SHARD-SAM improves over SAM by 1.98 percentage points on CIFAR-100/DenseNet-121 and reaches 90.75% on PyramidNet-272, compared with 90.12% for CR-SAM. Together, the tables show consistent gains across the evaluated model capacities.

Figure 2(a) shows radius-only and direction-only accuracies of 83.92% and 84.18%, versus 83.15% for SAM and 84.93% jointly. Panel (b) replaces the direction source while retaining the training recipe, hyperparameters, and adaptive-control rules; random, unweighted, SHARD, and HVP replacements gain approximately 0.13, 0.33, 1.78, and 0.83 points. HVP is a direction control, not Eigen-SAM. Appendix E.3 contrasts this adaptive-framework test with fixed-control training in Figure 7(c).

Table 3 complements the original benchmark in Table 1 with method-specific development-set selection and separate evaluation runs, adding SASSHA to the comparison. SHARD's advantage persists: it exceeds SASSHA by 0.30, 0.22, and 0.43 percentage points across the three settings. The WRN paired difference is 0.302 points (95% interval [0.203, 0.401]; Appendix E.4).

Table 1: Top-1 test accuracy (%) under basic augmentation (mean ± SD, five runs). †: all eight SHARD-SAM results differ from the second-best results at $p < 0.05$ (Welch's t-test). Bold: highest mean per setting. The base-optimizer row uses SGD for CNNs and AdamW for ViTs (Appendix E.4).

Method	CIFAR				ImageNet-1K	
	Dataset	ResNet-50	WRN-28-10	ViT-B/16	ResNet-50	ViT-B/32
Base optimizer	CIFAR-10	95.10±0.15	95.89±0.21	98.03±0.10	75.94±0.10	71.25±0.21
	CIFAR-100	78.42±0.20	81.84±0.13	88.61±0.10		
SAM	CIFAR-10	96.35±0.14	97.05±0.15	98.42±0.07	76.48±0.13	73.51±0.10
	CIFAR-100	80.85±0.16	83.15±0.14	89.53±0.16		
ASAM	CIFAR-10	96.45±0.12	97.15±0.10	98.47±0.12	76.63±0.16	74.05±0.17
	CIFAR-100	81.10±0.14	83.42±0.10	89.67±0.11		
LookSAM	CIFAR-10	96.40±0.15	97.10±0.12	98.44±0.05	76.65±0.08	73.85±0.14
	CIFAR-100	80.95±0.15	83.21±0.11	89.55±0.08		
F-SAM	CIFAR-10	96.52±0.10	97.26±0.11	98.56±0.08	76.69±0.12	73.79±0.12
	CIFAR-100	81.26±0.13	83.43±0.10	89.73±0.11		
LETS	CIFAR-10	96.64±0.09	97.38±0.08	98.60±0.09	76.68±0.11	74.22±0.10
	CIFAR-100	81.39±0.11	84.36±0.13	89.80±0.12		
Eigen-SAM	CIFAR-10	96.75±0.10	97.40±0.10	98.65±0.10	77.51±0.10	75.04±0.06
	CIFAR-100	81.67±0.12	84.04±0.10	89.82±0.12		
CR-SAM	CIFAR-10	96.72±0.05	97.60±0.09	98.60±0.13	77.64±0.12	75.28±0.08
	CIFAR-100	81.63±0.15	84.45±0.09	89.78±0.16		
SHARD-SAM†	CIFAR-10	**97.18±0.08**	**98.02±0.06**	**98.82±0.05**	**78.08±0.10**	**75.95±0.12**
	CIFAR-100	**82.40±0.09**	**84.93±0.05**	**90.34±0.08**		

5.3 Loss Landscape Visualization

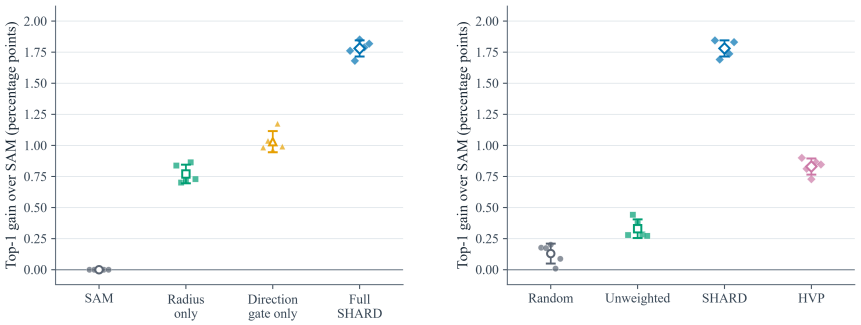
Figure 3 compares CIFAR-100/WRN-28-10 training-loss slices around the final solutions. Following the filter-normalized protocol of Li et al. (2018), we evaluate the loss in the plane spanned by two random orthogonal directions. SHARD-SAM has the broadest low-loss region in the displayed slices, consistent with reduced sensitivity to these perturbations.

Table 2: Top-1 test accuracy (%) comparison across extended CNN architectures on CIFAR-10 and CIFAR-100 under basic data augmentation. (Notation conventions are identical to those in Table 1.)

	CIFAR-10			CIFAR-100		
Architecture	SAM	CR-SAM	SHARD-SAM	SAM	CR-SAM	SHARD-SAM
ResNet-18	95.85±0.16	96.42±0.13	**96.88±0.09**†	79.81±0.18	80.76±0.21	**81.35±0.14**†
ResNet-101	96.68±0.14	97.15±0.11	**97.56±0.08**†	82.11±0.12	83.03±0.17	**83.65±0.11**†
DenseNet-121	92.00±0.17	93.15±0.12	**93.85±0.08**†	69.84±0.12	71.05±0.15	**71.82±0.10**†
PyramidNet-110	97.45±0.12	97.92±0.08	**98.35±0.06**†	84.89±0.09	85.68±0.14	**86.25±0.08**†
PyramidNet-272	98.30±0.08	98.65±0.06	**98.98±0.04**†	89.36±0.20	90.12±0.15	**90.75±0.09**†

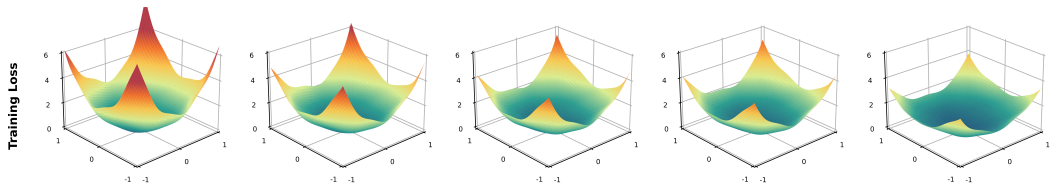
Table 3: Additional comparison with method-specific development-set selection, including SASSHA. These are separate evaluation runs from Table 1 (Appendix E.4). Top-1 accuracy (%, mean ± SD, five seeds); bold: highest mean per column.

Method	CIFAR-100	ImageNet-1K	
	WRN-28-10	ResNet-50	ViT-B/32
SAM	83.31±0.11	76.61±0.13	73.68±0.16
Eigen-SAM	84.29±0.13	77.61±0.12	75.15±0.10
CR-SAM	84.57±0.09	77.74±0.11	75.37±0.10
SASSHA	84.61±0.10	77.85±0.12	75.51±0.14
SHARD-SAM	**84.91±0.07**	**78.07±0.09**	**75.94±0.11**



(a) (b)

Figure 2: CIFAR-100/WRN-28-10 ablations with $\rho_{\text{base}} = 0.1$: (a) radius, direction-gate, and joint variants; (b) direction-source replacements with unchanged training and adaptive-control rules. Gains over SAM are in percentage points. Small points: five runs; large markers/bars: mean ± SD. Protocols are in Appendix E.3.



(a) (b) (c) (d) (e)

Figure 3: CIFAR-100/WRN-28-10 training-loss landscapes along two filter-normalized random directions. Panels (a)–(e): SGD, SAM, Eigen-SAM, CR-SAM, and SHARD-SAM.

5.4 Probe Budget and Wall-Clock Efficiency

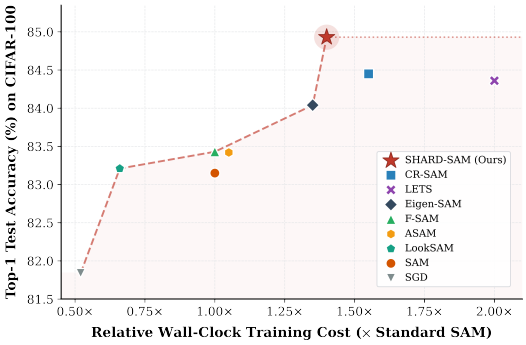


Figure 4: Accuracy versus relative training time on CIFAR-100/WRN-28-10.

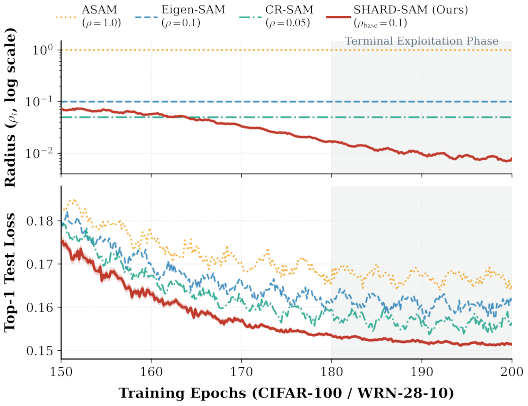


Figure 5: Terminal micro-convergence dynamics on CIFAR-100/WRN-28-10.

Each update uses $2b$ extra forward-only loss queries, three implicit matrix-vector iterations costing $\mathcal{O}(3bd)$ after these queries, and $\mathcal{O}(bd)$ probe storage. Figure 4 reports approximately $1.4\times$ SAM runtime at $b = 10$ on CIFAR-100/WRN-28-10. Appendix E.5 separates end-to-end timing, resource configuration, and the additional fixed-budget comparison.

In the CIFAR-100 sweep (Figure 6), WRN-28-10 and ViT-B/16 gain 1.78/0.81 points over SAM at $b = 10$, 1.79/0.82 at $b = 20$, and 1.77/0.80 at $b = 30$. From 10 to 30 probes, relative time rises from 1.40 to 1.94 for WRN and from 1.32 to 1.96 for ViT: about 39% and 48% extra time without higher mean accuracy. The plateau supports $b = 10$ within these evaluated settings.

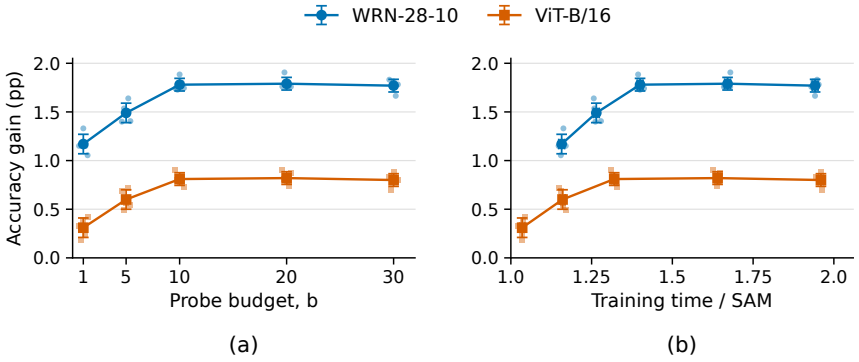


Figure 6: CIFAR-100 probe-budget sweep (CNN: WRN-28-10; ViT: ViT-B/16): (a) top-1 accuracy gain over SAM versus b; (b) gain versus training time normalized to SAM within each architecture. Small points: five runs; large markers/bars: mean $\pm$ SD. Gains are in percentage points.

On ImageNet-1K, SHARD-SAM with $b = 10$ reaches 78.08% on ResNet-50 and 75.95% on ViT-B/32, exceeding SAM by 1.60 and 2.44 percentage points at $1.24\times$ and $1.32\times$ its training time, respectively. Increasing b to 30 raises training time to $1.72\times$ and $1.96\times$ that of SAM without improving final accuracy. These runs use matched epoch counts; Appendix E.5 provides the trajectories and normalization details.

In the additional WRN comparison, independent runs with a budget of 10 GPU hours reach 84.54% with SHARD, 84.30% with SASSHA, and 84.22% with Eigen-SAM. SHARD reaches 84% in 8.6 GPU hours, compared with 9.1 for Eigen-SAM and 9.3 for SASSHA. Eigen-SAM reaches 83% earlier (6.7 versus 7.0 hours). The benefit of additional computation therefore depends on the target accuracy (Appendix E.5).

5.5 Micro-Convergence Dynamics and Radius Lifecycle

Over the final 50 WRN epochs, Figure 5 shows a contracting SHARD-SAM radius and a lower, less oscillatory test loss than ASAM and Eigen-SAM. Contraction reflects $R_t < 1$, a curvature-score magnitude below its EMA rather than low absolute curvature. The smaller radius narrows the neighborhood explored by the inner maximization, reducing the scale of adversarial weight perturbations during late training. The accompanying test-loss trajectory suggests that this contraction is compatible with maintaining predictive performance while reducing fluctuations across epochs.

5.6 Full-Network Direction and Control Diagnostics

Figure 7 compares direction sources. Unweighted uses SHARD's probes and responses $r_\mu(u_i)$ with uniform weights $w_i = 1/b$, removing extra magnitude weighting. HVP uses restarted Lanczos and autodifferentiated Hessian-vector products to estimate a reference eigendirection. Panels (a,c) fix ρ, z in $\epsilon = \rho(u+zv)/\|u+zv\|$, where $u = g/\|g\|$; panel (b) fixes the deflection angle. SAM remains unrotated. Frozen-state tests share checkpoints, batches, and stochastic states (Appendix E.6).

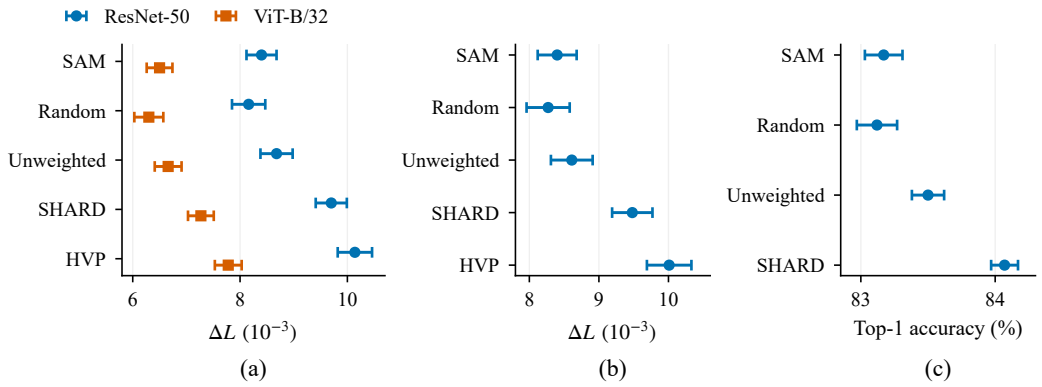


Figure 7: Controlled direction tests (mean ± SD, five seeds). (a) ResNet-50/ViT-B/32 raw loss increments at $\rho = 0.1, z = 0.2$. (b) ResNet-50 at $\rho = 0.1$ and 10° deflection; SAM remains unrotated. (c) CIFAR-100/WRN-28-10 training at fixed $\rho = 0.1, z = 0.2$. HVP is a direction reference, not an optimizer.

At fixed z, SHARD increases the ResNet-50 batch loss by 9.70×10^{-3}, compared with 8.16×10^{-3} for random directions and 8.68×10^{-3} without curvature weights. Its advantage persists at a fixed deflection angle, separating directional geometry from rotation magnitude. Fixed-control WRN training reaches 84.07%, compared with 83.17% for SAM, showing that the direction signal benefits training when the radius and gate amplitude are fixed. Fixing the gate yields 84.50%, while fixing the radius yields 84.19%, compared with 84.91% under joint adaptation (Appendix E.3). Appendix E.6 reports alignment and resampling diagnostics. Appendix E.7 checks WRN finite-difference responses against FP64 HVP references, separately from spectral recovery.

Negative scores can reverse the first-order perturbation (Appendix B.4). Across the five WRN runs, 2,215 of 195,000 updates have negative scores, and 14 of 195,000 (0.00718%) reverse the perturbation. The mean loss increment across these 14 events is negative. Signed and negative-off training differ by 0.006 percentage points in accuracy (paired 95% interval $[-0.025, 0.037]$; Appendix E.8).

6 Conclusion

In this work, we introduced SHARD-SAM, which uses a signed zeroth-order spectral surrogate to jointly adapt SAM's perturbation radius and direction. Experiments on CIFAR and ImageNet show consistent generalization gains across the evaluated CNNs and ViTs, with a favorable accuracy–efficiency trade-off among the compared curvature-aware methods. Ablations support the contributions of curvature information and both adaptive modules, while theoretical analyses characterize spectral structure, estimation error, perturbation stability, and convergence of the plain update.

AI USE STATEMENT

We used generative AI tools solely to assist with the formatting and layout of the manuscript text, figures, and tables. These tools were not used to generate scientific content, develop methods or proofs, conduct experiments, or analyze or interpret results. The authors reviewed all AI-assisted formatting changes to ensure that the original content and data were preserved, and take full responsibility for the accuracy and integrity of the final manuscript.

REPRODUCIBILITY STATEMENT

The supplementary material provides the population and finite-probe analysis (Appendix A), local directional-gating analysis and its negative-score boundary (Appendix B), convergence proof (Appendix C), and PAC-Bayesian motivation (Appendix D). Appendix E contains training and selection recipes, ablations, sensitivity analyses, runtime and fixed-budget comparisons, full-network direction diagnostics, finite-difference precision checks, and negative-branch measurements. Algorithm 1 in the main paper provides the pseudocode.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 639–668. PMLR, 2022. URL https://proceedings.mlr.press/v162/andriushchenko22a.html.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47032–47051, 2023a. doi: 10.52202/075280-2037. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/92dd1adab39f362046f99dfe3c39d90f-Abstract.html.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 840–902. PMLR, 2023b. URL https://proceedings.mlr.press/v202/andriushchenko23a.html.

Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022. doi: 10.1007/s10208-021-09499-8.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009. doi: 10.1109/CVPR.2009.5206848.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Y. F. Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=n0OeTdNRG0Q.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017. URL https://auai.org/uai2017/proceedings/papers/173.pdf.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5927–5935, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1.

Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens van der Maaten, and Kilian Q. Weinberger. Convolutional networks with dense connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8704–8716, 2022. doi: 10.1109/TPAMI.2019.2918284. URL https://doi.org/10.1109/TPAMI.2019.2918284.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 876–885. AUAI Press, 2018. URL https://auai.org/uai2018/proceedings/papers/313.pdf.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11148–11161. PMLR, 2022. URL https://proceedings.mlr.press/v162/kim22f.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5905–5914. PMLR, 2021. URL https://proceedings.mlr.press/v139/kwon21b.html.

Bingcong Li and Georgios Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70861–70879, 2023. doi: 10.52202/075280-3104. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e095c0a3717629aa5497601985bfcf0e-Abstract-Conference.html.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html.

Tao Li, Weihao Yan, Zehao Lei, Yingwen Wu, Kun Fang, Ming Yang, and Xiaolin Huang. Efficient generalization improvement guided by random weight perturbation. arXiv preprint arXiv:2211.11489, 2022. URL https://arxiv.org/abs/2211.11489.

Tao Li, Qinghua Tao, Weihao Yan, Yingwen Wu, Zehao Lei, Kun Fang, Mingzhen He, and Xiaolin Huang. Revisiting random weight perturbation for efficiently improving generalization. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=WbbgOHpoPX.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5631–5640, 2024b.

Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022a.

Yong Liu, Siqi Mai, Minhao Cheng, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Random sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24543–24556, 2022b. doi: 10.52202/068431-1782. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9b79416c0dc4b09feaa169ed5cdd63d4-Abstract-Conference.html.

Haocheng Luo, Tuan Truong, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. Explicit eigenvalue regularization improves sharpness-aware minimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4424–4453, 2024. doi: 10.52202/079017-0144. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0845e3f7447ff9e36a26033bb7334674-Abstract-Conference.html.

Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Xiaoshuai Sun, Rongrong Ji, and Dacheng Tao. Make sharpness-aware minimization stronger: A sparsified perturbation approach. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30950–30962, 2022. doi: 10.52202/068431-2244. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c859b99b5d717c9035e79d43dfd69435-Abstract-Conference.html.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18420–18432, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/995f5e03890b029865f402e83a81c29d-Abstract.html.

Dahun Shin, Dongyeop Lee, Jinseok Chung, and Namhoon Lee. SASSHA: Sharpness-aware adaptive second-order optimization with stable Hessian approximation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55145–55171. PMLR, 2025. URL https://proceedings.mlr.press/v267/shin25d.html.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9636–9647. PMLR, 2020. URL https://proceedings.mlr.press/v119/tsuzuku20a.html.

Xuehao Wang, Weisen Jiang, Shuai Fu, and Yu Zhang. Enhancing sharpness-aware minimization by learning perturbation radius. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, volume 14942 of *Lecture Notes in Computer Science*, pp. 375–391. Springer, 2024. doi: 10.1007/978-3-031-70344-7_22.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5spDgWmpY6x.

Tao Wu, Tie Luo, and Donald C. Wunsch, II. CR-SAM: Curvature regularized sharpness-aware minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):6144–6152, 2024. doi: 10.1609/aaai.v38i6.28431.

Wanyun Xie, Thomas Pethick, and Volkan Cevher. SAMPa: Sharpness-aware minimization parallelized. In *Advances in Neural Information Processing Systems*, volume 37, pp. 51333–51357, 2024. doi: 10.52202/079017-1625. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5bf2b802e24106064dc547ae9283bb0c-Abstract-Conference.html.

Haishan Ye, Zhichao Huang, Cong Fang, Chris Junchi Li, and Tong Zhang. Hessian-aware zeroth-order optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4869–4877, 2025. doi: 10.1109/TPAMI.2025.3548810.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, 2016. doi: 10.5244/C.30.87. URL https://bmva-archive.org.uk/bmvc/2016/papers/paper087/index.html.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C. Dvornek, Sekhar Tatikonda, James S. Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=edONMAnhLu-.