REVISITING SHARPNESS AND MODEL GENERALIZATION THROUGH LOCAL RADEMACHER COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-aware minimization (SAM) and its variants can improve test accuracy by minimizing the maximum increase in empirical loss under local model-parameter perturbations. This improvement is commonly attributed to the view that higher loss sharpness is strongly associated with poorer generalization. However, several studies challenge this view and report weak or even opposite correlations. Regardless of the correlation, one more fundamental question is whether the sharpness alone can explain model generalization. Using local Rademacher complexity, we develop a new theorem on excess risk, which shows that, in comparison with reducing the sharpness alone, jointly reducing sharpness and complexity will lead to a compacter excess-risk bound. Focusing on Rademacher complexity within the theorem, we derive a bound which increases as the sample-wise classification margins decrease. Furthermore, we build a Rademacher-guided plug-in mechanism, termed Rad, that adjusts the direction of parameter perturbations according to sample-wise classification margins. Extensive experiments on multiple datasets and model architectures demonstrate that Rad combined with SAM or its variants improves the test accuracy and reduces the magnitude of the train–test accuracy gap. Grounded in local Rademacher complexity, our work offers a new interpretation of the relationship between loss sharpness and model generalization.

1 INTRODUCTION

Generalization performance measures the ability of neural networks to apply knowledge acquired from finite training samples to unseen data. Many studies relate model generalization to the sharpness of loss landscape around learned solutions (Tsuzuku et al., 2020; Petzka et al., 2021; Haddouche et al., 2025; Peng et al., 2025).

Existing studies propose various measures to quantify the sharpness of the loss landscape (Foret et al., 2021; Sankar et al., 2021; Luo et al., 2024; Zhang et al., 2026). Among them, Foret et al. (2021) quantify sharpness through the loss increase under constrained parameter perturbations and propose Sharpness-Aware Minimization (SAM). Following this, subsequent studies extend SAM from different aspects, such as model parameter rescaling (Kwon et al., 2021), key gradient components in parameter perturbations (Li et al., 2024), perturbation normalization (Tahmasebi et al., 2024), perturbation direction (Chen & Zhou, 2026), and training efficiency (Zhou et al., 2025). More recently, Liu et al. (2026) apply SAM to parameter-efficient fine-tuning of large models. These studies report higher test accuracy across different models and tasks, which reflects better model generalization. Meanwhile, many studies provide theoretical evidences that lower sharpness can lead to better model generalization (Tsuzuku et al., 2020; Petzka et al., 2021; Haddouche et al., 2025; Peng et al., 2025). However, some studies report that models with high sharpness can still generalize well and existing sharpness measures do not consistently predict generalization across different settings (Dinh et al., 2017; Andriushchenko et al., 2023; Kaur et al., 2023; Liang et al., 2025). Overall, these two views do not establish a consistent relationship between sharpness and model generalization. This motivates the question of whether sharpness alone can explain model generalization.

To investigate this question, we focus on the perturbation-based sharpness measure used in SAM, which we refer to as sharpness hereafter. The sharpness formulation is the maximum of the empirical loss at a perturbed parameter minus that at the current parameter. However, empirical loss averages the losses over all training samples. Large changes in individual sample losses may still occur even

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.

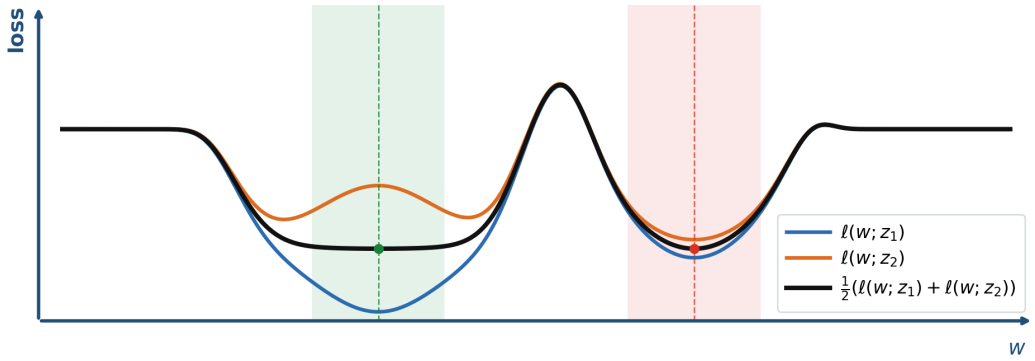


Figure 1: A schematic comparison between individual sample losses and their average under local parameter perturbations

when the average increases only slightly. As shown in Figure 1, the blue and orange curves represent the losses of two samples as the model parameters vary, and the black curve represents their average loss. Compared with the red region, the two sample losses vary more in the green region, but their average loss remains relatively flat. Therefore, this sharpness may not fully capture changes in individual sample losses.

The local Rademacher complexity characterizes the ability of a local function class to fit random signs assigned to individual samples and is widely used to bound excess risk (Cheng et al., 2025). The excess risk represents the population-risk gap between a learned predictor and the best predictor. The local function class consists of functions that satisfy a given second-moment constraint. In this work, we derive an excess-risk bound that contains the empirical sharpness of the population-risk minimizer and the local Rademacher complexity of the corresponding loss-difference function class. The former measures the maximum increase in the average loss, and the latter evaluates the corresponding function class through its values on individual samples. Specifically, under the stated conditions, empirical sharpness provides an upper bound for the second-moment radius of the corresponding loss-difference function class. Therefore, controlling both empirical sharpness and local Rademacher complexity can benefit model generalization.

However, the local Rademacher complexity is generally impossible to compute directly. We derive an upper bound on the local Rademacher complexity which depends on the sample-wise classification margins. The upper bound increases as the sample-wise classification margins decrease. This result motivates a Rademacher-guided mechanism (Rad) that uses sample margin information during perturbation construction and gives larger weights to low-margin samples. Experiments show that combining Rad with SAM or its variants improves the test accuracy and reduces the train-test accuracy gap. Our main contributions are summarized as follows:

- We establish an excess-risk bound that contains both empirical sharpness and local Rademacher complexity, indicating that sharpness alone is insufficient for the analysis of model generalization.
- To estimate local Rademacher complexity without direct computation, we derive an upper bound depends on sample-wise classification margins.
- We propose a Rademacher-guided plug-in mechanism (Rad) that incorporates margin information into perturbation construction and provides a way to refine the perturbation direction.
- Compared with their corresponding baselines, Rad-enhanced SAM-based methods achieve higher test accuracy and generally train–test accuracy gaps closer to zero across multiple datasets and architectures.

1.1 RELATED WORK

Sharpness and Generalization: The relationship between sharpness and generalization is studied from various perspectives. Keskar et al. (2017) empirically associate the generalization gap observed in large-batch SGD with the sharpness of the loss landscape. From a theoretical perspective, Tsuzuku et al. (2020) derive PAC-Bayesian generalization error bounds based on normalized sharpness measures. Building on this line of research, Petzka et al. (2021) examine the interpolation from representative data to investigate when sharpness can explain generalization. More recently, Haddouche et al. (2025) establish a PAC-Bayesian generalization bound that mainly contains a sharpness term and a KL-based complexity term. From a different perspective, Peng et al. (2025) establish an information-theoretic generalization bound for SGD based on the local curvature of the loss landscape. However, existing studies later show that lower sharpness does not correspond to better generalization. Dinh et al. (2017) show that functionally equivalent networks with the same generalization performance can have different sharpness values under different parameterizations. Extending this concern to modern training settings, Andriushchenko et al. (2023) report a weak correlation between sharpness and generalization and observe better generalization of sharper minima under distribution shifts. Similarly, Kaur et al. (2023) provide empirical evidence that models with smaller maximum Hessian eigenvalues do not always generalize better. Theoretically, Liang et al. (2025) show that smaller sharpness can benefit model generalization, but this benefit decreases as the input dimension increases. These findings imply that sharpness alone is insufficient to explain generalization, and additional factors need to be considered.

Local Rademacher Complexity and Excess-Risk Bounds: Global Rademacher complexity characterizes the complexity of an entire function class (Bartlett & Mendelson, 2002), whereas the local Rademacher complexity focuses on localized subclasses (Bartlett et al., 2005). Local Rademacher complexity has been further used to derive excess-risk bounds, where the excess risk measures the population-risk gap between the learned predictor and the best predictor in the function class (Koltchinskii, 2006). In multi-label learning, Xu et al. (2016) show that the local Rademacher complexity is upper-bounded by the tail singular values of the predictor. For neural networks, Lu et al. (2022) derive an upper bound on local Rademacher complexity in terms of layer-wise dropout rates and network weights, and use it to construct a regularizer. Moreover, Shultzman et al. (2023) derive upper bounds on the generalization and estimation errors of model-based neural networks using global and local Rademacher complexities. Kanade et al. (2024) extend the classical excess-risk analysis based on local Rademacher complexity to nonconvex and improper learning settings through offset Rademacher complexity. Yang (2025) extends local Rademacher complexity to transductive learning and derives a new excess-risk bound. Overall, local Rademacher complexity is commonly used to derive excess-risk bounds.

2 NOTATION AND PRELIMINARIES

Notations. Let $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathcal{Y}$, and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ denote the input, label, and sample spaces, respectively. Consider a training dataset $S = \{z_i\}_{i=1}^n$, where the samples $z_i = (x_i, y_i)$ are drawn i.i.d. from distribution $\mathcal{D}$. Let $\mathcal{W}$ be the parameter space. For each $w \in \mathcal{W}$, $f_w : \mathcal{X} \to \mathbb{R}^K$ denotes the model with parameter w. Let $\ell(w; z)$ denote the loss of model f_w on sample z. The empirical risk is defined as $L_S(w) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(w; z_i)$, and the population risk as $L_{\mathcal{D}}(w) \triangleq \mathbb{E}_{z \sim \mathcal{D}}[\ell(w; z)]$.

Preliminaries. SAM (Foret et al., 2021) defines sharpness as the maximum increase in empirical risk under constrained weight perturbations. Given a perturbation radius $\rho > 0$, we denote the local parameter neighborhood of w by $\mathcal{W}_\rho(w) = \{\widetilde{w} = w + \epsilon \in \mathcal{W} : \|\epsilon\|_2 \leq \rho\}$, where ϵ denotes a weight perturbation. The sharpness at w is defined as:

$$\mathrm{Sharp}_{S,\rho}(w) = \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \left[L_S(\widetilde{w}) - L_S(w) \right]. \tag{1}$$

A smaller sharpness value corresponds to a flatter loss landscape. We now review the Rademacher complexity which quantifies the capacity of the function class to fit random noise. A larger Rademacher complexity generally leads to a looser generalization bound. Let $\mathcal{L} = \{z \mapsto \ell(w; z) \mid w \in \mathcal{W}\}$ represent the loss function class induced by the whole parameter space

$\mathcal{W}$. The empirical Rademacher complexity of $\mathcal{L}$ is defined as:

$$R_S(\mathcal{L}) = \mathbb{E}_{\sigma_i}\left[\sup_{w\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(w; z_i)\right]. \tag{2}$$

where $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables that take values ± 1 each with equal probability. The expected Rademacher complexity is obtained by taking the expectation of the empirical Rademacher over the training set S, denoted by $R_{\mathcal{D}}(\mathcal{L})$. Unlike global Rademacher complexity, local Rademacher complexity focuses on a constrained subset within the entire function class. Xu et al. (2016) present the following excess-risk bound based on local Rademacher complexity.

Theorem 2.1. *Let $\mathcal{W}' \subseteq \mathcal{W}$, $\mathcal{L}' = \{z \mapsto \ell(w; z) \mid w \in \mathcal{W}'\}$ and $\Delta(\mathcal{W}', \mathcal{W}) = \inf_{w\in\mathcal{W}'} L_{\mathcal{D}}(w) - \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w)$ denote the gap between the minimum population risks over $\mathcal{W}'$ and $\mathcal{W}$. Given $\delta > 0$, suppose that $\widehat{w}$ is learned from n training points. Assume that there is some $r > 0$ such that for all $w \in \mathcal{W}'$, $\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(w; z)^2\right] \le r$. Then with probability at least $1 - \delta$, we have*

$$L_{\mathcal{D}}(\widehat{w}) \le \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w) + 8R_S(\mathcal{L}') + \sqrt{\frac{8r\log(2/\delta)}{n}} + \frac{32\log(2/\delta)}{3n} + \Delta(\mathcal{W}', \mathcal{W}). \tag{3}$$

The local parameter set $\mathcal{W}'$ should be specified independently of the training set S and contain the learned parameter $\widehat{w}$. In particular, $\Delta(\mathcal{W}', \mathcal{W}) = 0$ if $\mathcal{W}'$ contains a population-risk minimizer over $\mathcal{W}$.

3 Excess-Risk Analysis via Sharpness and Local Rademacher Complexity

In this section, we will derive a new excess-risk upper bound that is dominated jointly by empirical sharpness and local Rademacher complexity. First, we define population sharpness.

Definition 3.1 (Population Sharpness)**.** Given a perturbation radius $\rho > 0$ and a parameter $w \in \mathcal{W}$, the population sharpness of w is defined as

$$\mathrm{Sharp}_{\mathcal{D},\rho}(w) \triangleq \max_{\widetilde{w}\in\mathcal{W}_\rho(w)} \left[L_{\mathcal{D}}(\widetilde{w}) - L_{\mathcal{D}}(w)\right]. \tag{4}$$

The population sharpness measures the maximum increase in population risk under constrained weight perturbations. Unlike empirical sharpness, it depends on the unknown distribution $\mathcal{D}$ and cannot be evaluated directly from the training set. Next, we review the η-approximate realizability condition, which assumes that the minimum population risk over the parameter space is at most η.

Definition 3.2 (η-approximate realizability (Shalev-Shwartz & Ben-David, 2014))**.** Given a constant $\eta \ge 0$, the learning problem satisfies the η-approximate realizability condition if there exists a population-risk minimizer $w^\star$ over $\mathcal{W}$ such that

$$L_{\mathcal{D}}(w^\star) \le \eta. \tag{5}$$

In the subsequent analysis, we consider the case where the learned parameter $\widehat{w}$ lies within a local parameter neighborhood of a population-risk minimizer $w^\star$, and analyze the local sharpness and local Rademacher complexity within this region. In the following, we define the local loss-difference function class induced by constrained parameter perturbations.

Definition 3.3. Given $w \in \mathcal{W}$ and a perturbation radius $\rho > 0$, for each $\widetilde{w} \in \mathcal{W}_\rho(w)$, denote the corresponding local loss-difference function $g_{\widetilde{w}} : \mathcal{Z} \to \mathbb{R}$ by $g_{\widetilde{w}}(z) = \ell(\widetilde{w}; z) - \ell(w; z)$. The local loss-difference function class induced by $\mathcal{W}_\rho(w)$ is defined as:

$$\mathcal{G}_\rho(w) = \{g_{\widetilde{w}} \mid \widetilde{w} \in \mathcal{W}_\rho(w)\}.$$

The class $\mathcal{G}_\rho(w)$ consists of loss-difference functions induced by local parameter perturbations. The subsequent analysis of local Rademacher complexity is conducted over this class $\mathcal{G}_\rho(w)$. Since population sharpness cannot be directly computed, the following lemma upper-bounds it by empirical sharpness, so that the excess-risk bound can be stated in terms of empirical quantities.

Lemma 3.4. *Let $S = \{z_i\}_{i=1}^n$ be a training set. Given a perturbation radius $\rho > 0$ and $w \in \mathcal{W}$. Suppose that $\mathcal{W}_\rho(w) = \{w + \epsilon : \|\epsilon\|_2 \le \rho\} \subseteq \mathcal{W}$, for any $\widetilde{w} \in \mathcal{W}_\rho(w)$ and $z \in \mathcal{Z}$, $0 \le \ell(\widetilde{w}; z) \le a$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathrm{Sharp}_{\mathcal{D},\rho}(w) \le \mathrm{Sharp}_{S,\rho}(w) + a\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{6}$$

The proof of Lemma 3.4 can be found in Appendix A. Unlike global Rademacher complexity, local Rademacher complexity restricts the analysis to a localized subclass. We impose a Bernstein-type condition to relate the second moment of each loss-difference function to its expectation (Bartlett & Mendelson, 2006). According to Lemma 3.4 and Theorem 2.1, we derive an excess-risk upper bound that contains the empirical sharpness, the local Rademacher complexity of the loss-difference function class, and the confidence term. The main theorem is stated as follows:

Theorem 3.5. *Given a perturbation radius $\rho > 0$, let $S = \{z_i\}_{i=1}^n$ be a training set and $\widehat{w}$ be the parameter learned from S. Assume that there exists a population-risk minimizer $w^\star \in \mathcal{W}$ such that $L_{\mathcal{D}}(w^\star) \le \eta$. Further assume that $\widehat{w} \in \mathcal{W}_\rho(w^\star)$ and $\mathcal{G}_\rho(w^\star)$ satisfies the Bernstein-type condition with a constant $\nu > 0$. Under the assumptions of Lemma 3.4, with probability at least $1 - \delta$, we have*

$$L_{\mathcal{D}}(\widehat{w}) \le \inf_{w \in \mathcal{W}} L_{\mathcal{D}}(w) + 8R_S(\mathcal{G}_\rho(w^\star)) + \sqrt{\frac{8\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)\log(4/\delta)}{n}} + \Omega_n(\delta, \eta), \tag{7}$$

where $\Omega_n(\delta, \eta) = \sqrt{\frac{8a\eta\log(4/\delta)}{n}} + 2\sqrt{\nu a}\left(\frac{2\log(4/\delta)}{n}\right)^{3/4} + \frac{32\log(4/\delta)}{3n}$.

The proof of theorem 3.5 is placed in Appendix B. Empirical sharpness characterizes the maximum empirical loss increase over the local loss-difference function class, whereas local Rademacher complexity measures the complexity of the same function class. Our new bound shows that the excess risk depends not only on empirical sharpness but also on the Rademacher complexity of the local loss-difference function class. This result provides a theoretical basis for the joint consideration of these two quantities in the following section. The theoretical analysis considers the case where the learned solution lies within a local region around a population-risk minimizer. During training, we use the parameter at each optimization step as the reference parameter and account jointly for flatness and complexity within its local parameter region.

4 Rademacher-Guided Sharpness-Aware Optimization

We derive an upper bound on local Rademacher complexity and propose Rad to emphasize small-margin samples during perturbation construction.

4.1 Margin-Dependent Rademacher Complexity Bound

Compared with reducing empirical sharpness alone, theorem 3.5 suggests that jointly reducing empirical sharpness and local Rademacher complexity can lead to better model generalization. Since local Rademacher complexity is difficult to compute directly, existing studies often derive tractable upper bounds and use them to guide model optimization. Along this line, we consider a K-class classifier trained with the softmax cross-entropy loss, which is commonly used in classification tasks with SAM and its variants. As commonly assumed in theoretical analyses of SAM (Andriushchenko & Flammarion, 2022), each sample-wise loss satisfies β-smoothness over $\mathcal{W}_\rho(w)$. This condition ensures that the norm of the gradient variation does not exceed β times the parameter variation and controls the local loss change under parameter perturbations. Under this setting, we have the following upper bound for the Rademacher complexity of the local loss-difference function class.

Theorem 4.1. *Given a perturbation radius $\rho > 0$, a training set $S = \{z_i\}_{i=1}^n$, and $w \in \mathcal{W}$, consider a K-class classifier $f_w : \mathcal{X} \to \mathbb{R}^K$. Suppose that ℓ is the softmax cross-entropy loss. For each $i \in \{1, \dots, n\}$, assume that $\ell(\cdot; z_i)$ is β-smooth over $\mathcal{W}_\rho(w)$ and that the operator norm of the Jacobian of $f_w(x_i)$ with respect to w is bounded by κ. Then,*

$$R_S\big(\mathcal{G}_\rho(w)\big) \le \frac{\sqrt{2}\rho\kappa}{n}\sqrt{\sum_{i=1}^n q_i^2(w)} + \frac{\beta\rho^2}{2}, \tag{8}$$

where $q_i(w) = \frac{(K-1)e^{-m_i(w)}}{1+(K-1)e^{-m_i(w)}}$, $m_i(w) = f_{w,y_i}(x_i) - \max_{c \neq y_i} f_{w,c}(x_i)$ *is the classification margin of* z_i*, and* $f_{w,c}(x)$ *denotes the logit of class* c*.*

The proof of Theorem 4.1 is given in Appendix C. The classification margin $m_i(w)$ measures the difference between the logit of the ground-truth class and the largest logit among the incorrect classes. When ρ, β,κ, and K are fixed, the upper bound decreases as the sample-wise classification margins $m_i(w)$ increase. Here, $q_i(w)$ is a monotonically decreasing function of the classification margin and characterizes the contribution of each sample to the margin-dependent complexity term. Therefore, small-margin samples have larger $q_i(w)$ values and contribute more to the complexity upper bound.

4.2 Optimization Procedure

In order to obtain a flatter solution, SAM and its variants search for the perturbation that maximizes the empirical loss within a local parameter neighborhood then update the model at the perturbed parameters. Theorem 3.5 shows that empirical sharpness and the Rademacher complexity should be controlled simultaneously. Theorem 4.1 further provides an upper bound on this complexity, where smaller sample margins lead to larger contributions to the bound. Motivated by these results, we retain the perturbation construction used in SAM or its variants, and additionally introduce a margin-based per-sample weighting term. Specifically, at iteration t, let w_t denote the current model parameters and $B = \{z_i\}_{i=1}^m$ denote the current mini-batch. For each sample z_i, we compute its classification margin as $m_i(w_t) = f_{w_t,y_i}(x_i) - \max_{c \neq y_i} f_{w_t,c}(x_i)$, where $f_{w_t,c}(x_i)$ denotes the logit corresponding to class c. Let $\bar{m}_t = \frac{1}{m}\sum_{j=1}^m m_j(w_t)$ denote the average classification margin over the mini-batch. We define a margin-based coefficient according to the deviation of each sample margin:

$$\gamma_i = \mathrm{clip}\left(\frac{\bar{m}_t - m_i(w_t)}{\frac{1}{m}\sum_{j=1}^m |m_j(w_t) - \bar{m}_t| + \varepsilon}, -1, 1\right), \tag{9}$$

where $\mathrm{clip}(\cdot, -1, 1)$ truncates the margin deviation scaled by the mini-batch mean absolute deviation to $[-1, 1]$, preventing samples with extreme margin deviations from dominating the perturbation direction, and $\varepsilon > 0$ ensures numerical stability. This design assigns larger weights to small-margin samples and smaller weights to large-margin samples for perturbation construction. We now give the Rademacher-guided adjustment (Rad) as:

$$\Delta g_t^{\mathrm{Rad}} = \frac{\lambda}{m}\sum_{i=1}^m \gamma_i \nabla_w \ell(w_t; z_i), \tag{10}$$

where λ is a hyperparameter. Let g_t^{base} denote the perturbation direction used by SAM or its variants before normalization. Rad adds the margin-based adjustment $\Delta g_t^{\mathrm{Rad}}$ to this direction, and the perturbation is constructed as:

$$\epsilon_t^{\mathrm{Rad}} = \rho \frac{g_t^{\mathrm{base}} + \Delta g_t^{\mathrm{Rad}}}{\left\| g_t^{\mathrm{base}} + \Delta g_t^{\mathrm{Rad}} \right\|_2}. \tag{11}$$

We then add $\epsilon_t^{\mathrm{Rad}}$ to w_t to obtain the perturbed parameters $w_t + \epsilon_t^{\mathrm{Rad}}$, and compute the gradient of the empirical loss at the perturbed parameters to update w_t:

$$w_{t+1} = w_t - \eta \nabla_w L_B\left(w_t + \epsilon_t^{\mathrm{Rad}}\right), \tag{12}$$

where η denotes the learning rate. The complete algorithm is summarized in Appendix D.

5 Experiments

To experimentally evaluate the effectiveness of the proposed Rad mechanism, we consider multiple image classification benchmark datasets and various network architectures.

5.1 Experimental Setup

Datasets and Models. First, we choose CIFAR-10 and CIFAR-100 datasets (Krizhevsky & Hinton, 2009) with VGG16 (Simonyan & Zisserman, 2014), ResNet18 and ResNet50 (He et al., 2016),

and Wide ResNet-28-10 (Zagoruyko & Komodakis, 2016) architectures to experimentally verify the performance. Second, to further evaluate the effectiveness on large-scale visual recognition test, we train ResNet50 from scratch on ImageNet-1K. Finally, we extend our experiments to the CUB-200-2011 (Wah et al., 2011), Stanford Dogs (Khosla et al., 2011), and Stanford Cars (Krause et al., 2013) datasets and adopt ImageNet-pretrained ViT as backbone architectures. To ensure a fair comparison, all methods are evaluated using identical data-processing and training protocols. We report the best test accuracy across all training epochs and the corresponding accuracy gap. Complete implementation details are provided in Appendix E.

Comparison Baselines and Evaluation Metrics. To verify the effectiveness of the proposed Rad mechanism, we select stochastic gradient descent (SGD) (Bottou, 2010), SAM (Foret et al., 2021), adaptive sharpness-aware minimization (ASAM) (Kwon et al., 2021) and eXplicit Sharpness-Aware Minimization (XSAM) (Chen & Zhou, 2026) as comparison baselines. We use SGD as the standard optimizer and adopt SAM, ASAM, and XSAM as sharpness-aware baselines, where XSAM is the most recent method to the best of our knowledge. Furthermore, we integrate Rad into SAM and ASAM, denoted as SAM+Rad and ASAM+Rad, respectively. We use test accuracy and the absolute version of the generalization gap in (Liu et al., 2022), termed the train–test accuracy gap, as evaluation metrics. Test accuracy measures the classification performance of the model, and accuracy gap measures the difference between train accuracy and test accuracy. A smaller accuracy gap means that the model maintains more similar performance on the train and test sets, which indicates better generalization. For ImageNet-1K, we report Top-1 and Top-5 accuracies together with their corresponding accuracy gaps.

5.2 Experimental Results

We first conduct experiments on CIFAR-10 and CIFAR-100, two commonly used image classification datasets, as shown in Tables 1 and 2. Results that improve upon the corresponding method without Rad are highlighted in bold. Across four network architectures, SAM and ASAM with Rad consistently achieve higher test accuracy than their original versions. In most experimental settings, SAM and ASAM with Rad also achieve smaller accuracy gaps, which indicates better model generalization. Since the value of the accuracy gap depends on the optimization method, SAM and ASAM have different baseline gaps, and the effect of Rad should be evaluated relative to the corresponding original method. These improvements across different architectures and perturbation strategies suggest that Rad does not depend on a specific model architecture or perturbation strategy. Compared with XSAM (Chen & Zhou, 2026), our method achieves competitive performance across different network architectures. XSAM requires an additional explicit search for the maximum-loss direction, whereas Rad does not introduce such a search.

Table 1: Comparison of test accuracy and accuracy gap on CIFAR-10.

Model	Metric	SGD	SAM	ASAM	XSAM	SAM+Rad	ASAM+Rad
VGG16	Acc.↑	$94.71_{\pm 0.04}$	$95.36_{\pm 0.10}$	$95.46_{\pm 0.06}$	$95.39_{\pm 0.13}$	$\mathbf{95.46_{\pm 0.10}}$	$\mathbf{95.52_{\pm 0.07}}$
	Gap↓	$3.75_{\pm 0.04}$	$2.85_{\pm 0.12}$	$2.96_{\pm 0.14}$	$2.87_{\pm 0.19}$	$\mathbf{2.81_{\pm 0.13}}$	$\mathbf{2.91_{\pm 0.12}}$
ResNet18	Acc.↑	$96.19_{\pm 0.15}$	$96.60_{\pm 0.09}$	$96.67_{\pm 0.08}$	$96.60_{\pm 0.12}$	$\mathbf{96.74_{\pm 0.07}}$	$\mathbf{96.78_{\pm 0.04}}$
	Gap↓	$3.06_{\pm 0.21}$	$2.66_{\pm 0.05}$	$2.75_{\pm 0.13}$	$2.67_{\pm 0.08}$	$\mathbf{2.42_{\pm 0.09}}$	$\mathbf{2.64_{\pm 0.10}}$
ResNet50	Acc.↑	$95.92_{\pm 0.15}$	$96.65_{\pm 0.06}$	$96.82_{\pm 0.05}$	$96.56_{\pm 0.11}$	$\mathbf{96.82_{\pm 0.06}}$	$\mathbf{97.11_{\pm 0.06}}$
	Gap↓	$3.13_{\pm 0.15}$	$2.51_{\pm 0.07}$	$2.62_{\pm 0.07}$	$2.50_{\pm 0.10}$	$\mathbf{2.26_{\pm 0.07}}$	$\mathbf{2.29_{\pm 0.08}}$
WRN-28-10	Acc.↑	$96.87_{\pm 0.07}$	$97.17_{\pm 0.08}$	$97.51_{\pm 0.05}$	$97.42_{\pm 0.10}$	$\mathbf{97.32_{\pm 0.06}}$	$\mathbf{97.55_{\pm 0.02}}$
	Gap↓	$2.67_{\pm 0.08}$	$2.32_{\pm 0.11}$	$2.10_{\pm 0.07}$	$2.16_{\pm 0.12}$	$\mathbf{2.16_{\pm 0.07}}$	$\mathbf{2.05_{\pm 0.02}}$

To further evaluate Rad on a large-scale visual recognition task, we conduct experiments on ImageNet-1K with ResNet50, as shown in Table 3. Compared with SAM and ASAM, their Rad-enhanced versions achieve higher Top-1 and Top-5 accuracies, while most accuracy gaps are reduced. Compared with XSAM, SAM+Rad achieves higher Top-1 and Top-5 accuracies.

Table 4 compares the test accuracy and accuracy gap on CUB-200-2011, Stanford Dogs, and Stanford Cars. These datasets are widely used for fine-grained visual recognition tasks. Their categories

Table 2: Comparison of test accuracy and accuracy gap on CIFAR-100.

Model	Metric	SGD	SAM	ASAM	XSAM	SAM+Rad	ASAM+Rad
VGG16	Acc.↑	$75.07_{\pm 0.33}$	$76.73_{\pm 0.14}$	$76.59_{\pm 0.17}$	$76.74_{\pm 0.08}$	$\mathbf{76.82}_{\pm \mathbf{0.24}}$	$\mathbf{76.89}_{\pm \mathbf{0.12}}$
	Gap↓	$20.75_{\pm 0.29}$	$18.88_{\pm 0.31}$	$19.54_{\pm 0.49}$	$19.24_{\pm 0.04}$	$\mathbf{18.88}_{\pm \mathbf{0.23}}$	$\mathbf{19.41}_{\pm \mathbf{0.29}}$
ResNet18	Acc.↑	$79.75_{\pm 0.10}$	$80.29_{\pm 0.16}$	$81.52_{\pm 0.21}$	$81.16_{\pm 0.10}$	$\mathbf{80.69}_{\pm \mathbf{0.08}}$	$\mathbf{81.65}_{\pm \mathbf{0.07}}$
	Gap↓	$19.31_{\pm 0.10}$	$18.86_{\pm 0.13}$	$17.78_{\pm 0.27}$	$17.93_{\pm 0.14}$	$\mathbf{18.43}_{\pm \mathbf{0.09}}$	$\mathbf{17.65}_{\pm \mathbf{0.07}}$
ResNet50	Acc.↑	$79.94_{\pm 0.33}$	$81.55_{\pm 0.54}$	$82.50_{\pm 0.53}$	$82.33_{\pm 0.39}$	$\mathbf{82.36}_{\pm \mathbf{0.09}}$	$\mathbf{82.76}_{\pm \mathbf{0.44}}$
	Gap↓	$18.85_{\pm 0.37}$	$17.29_{\pm 0.54}$	$16.86_{\pm 0.54}$	$16.20_{\pm 0.26}$	$\mathbf{16.60}_{\pm \mathbf{0.08}}$	$\mathbf{16.54}_{\pm \mathbf{0.39}}$
WRN-28-10	Acc.↑	$82.49_{\pm 0.09}$	$83.60_{\pm 0.16}$	$84.14_{\pm 0.20}$	$84.33_{\pm 0.16}$	$\mathbf{83.78}_{\pm \mathbf{0.17}}$	$\mathbf{84.38}_{\pm \mathbf{0.14}}$
	Gap↓	$16.82_{\pm 0.09}$	$15.78_{\pm 0.14}$	$15.31_{\pm 0.20}$	$15.08_{\pm 0.19}$	$\mathbf{15.57}_{\pm \mathbf{0.19}}$	$\mathbf{15.00}_{\pm \mathbf{0.17}}$

Table 3: Comparison of test accuracy and accuracy gap on ImageNet-1K with ResNet50.

Metric	SGD	SAM	ASAM	XSAM	SAM+Rad	ASAM+Rad
Top-1 Acc.↑	$76.76_{\pm 0.10}$	$77.26_{\pm 0.11}$	$76.70_{\pm 0.01}$	$77.11_{\pm 0.09}$	$\mathbf{77.39}_{\pm \mathbf{0.03}}$	$\mathbf{76.93}_{\pm \mathbf{0.05}}$
Gap@1↓	$2.27_{\pm 0.05}$	$\mathbf{0.83}_{\pm \mathbf{0.01}}$	$2.64_{\pm 0.06}$	$1.35_{\pm 0.08}$	$0.91_{\pm 0.06}$	$\mathbf{2.47}_{\pm \mathbf{0.03}}$
Top-5 Acc.↑	$93.21_{\pm 0.07}$	$93.62_{\pm 0.08}$	$93.19_{\pm 0.08}$	$93.42_{\pm 0.09}$	$\mathbf{93.66}_{\pm \mathbf{0.06}}$	$\mathbf{93.43}_{\pm \mathbf{0.01}}$
Gap@5↓	$0.48_{\pm 0.05}$	$1.43_{\pm 0.03}$	$0.38_{\pm 0.06}$	$0.68_{\pm 0.07}$	$\mathbf{1.25}_{\pm \mathbf{0.04}}$	$\mathbf{0.11}_{\pm \mathbf{0.05}}$

exhibit small inter-class differences and large intra-class variations. Rad improves the test accuracy of both SAM and ASAM across all benchmarks and reduces the accuracy gap in most cases. Compared with XSAM, the Rad-enhanced methods achieve competitive accuracy and consistently smaller accuracy gaps. Overall, Rad remains effective on these challenging tasks.

Table 4: Comparison of test accuracy and accuracy gap on fine-grained visual classification datasets with an ImageNet-pretrained ViT-B/16.

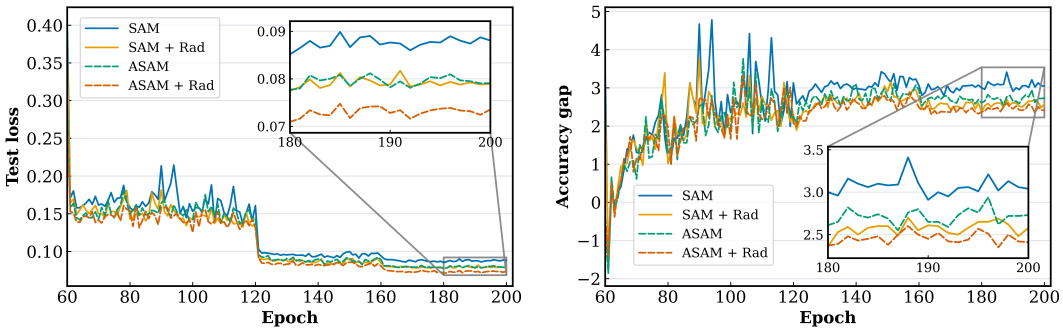
Dataset	Metric	SGD	SAM	ASAM	XSAM	SAM+Rad	ASAM+Rad
CUB-200-2011	Acc.↑	$91.09_{\pm 0.09}$	$91.16_{\pm 0.02}$	$91.40_{\pm 0.16}$	$91.17_{\pm 0.13}$	$\mathbf{91.20}_{\pm \mathbf{0.06}}$	$\mathbf{91.45}_{\pm \mathbf{0.02}}$
	Gap↓	$7.84_{\pm 0.21}$	$8.01_{\pm 0.07}$	$7.37_{\pm 0.11}$	$7.97_{\pm 0.30}$	$\mathbf{7.90}_{\pm \mathbf{0.20}}$	$\mathbf{7.15}_{\pm \mathbf{0.16}}$
Stanford Dogs	Acc.↑	$90.50_{\pm 0.14}$	$90.61_{\pm 0.18}$	$90.79_{\pm 0.16}$	$90.79_{\pm 0.09}$	$\mathbf{90.73}_{\pm \mathbf{0.12}}$	$\mathbf{90.94}_{\pm \mathbf{0.07}}$
	Gap↓	$7.82_{\pm 0.13}$	$8.03_{\pm 0.08}$	$7.51_{\pm 0.18}$	$8.24_{\pm 0.06}$	$\mathbf{7.57}_{\pm \mathbf{0.14}}$	$\mathbf{7.50}_{\pm \mathbf{0.07}}$
Stanford Cars	Acc.↑	$90.85_{\pm 0.15}$	$90.71_{\pm 0.13}$	$90.98_{\pm 0.09}$	$91.29_{\pm 0.06}$	$\mathbf{91.43}_{\pm \mathbf{0.31}}$	$\mathbf{91.13}_{\pm \mathbf{0.12}}$
	Gap↓	$8.15_{\pm 0.13}$	$8.59_{\pm 0.01}$	$8.20_{\pm 0.13}$	$8.17_{\pm 0.08}$	$\mathbf{8.07}_{\pm \mathbf{0.12}}$	$\mathbf{8.01}_{\pm \mathbf{0.05}}$

5.3 Further Analysis

Figure 2 shows (a) the test loss and (b) the train–test accuracy gap for SAM, SAM+Rad, ASAM, and ASAM+Rad across training epochs on CIFAR-10 with ResNet50. The test loss is computed on the test set at the end of each training epoch. It can be observed that SAM+Rad and ASAM+Rad achieve lower test losses than SAM and ASAM, respectively, in the later training stage. Meanwhile, SAM+Rad and ASAM+Rad generally maintain smaller train-test accuracy gaps than their corresponding baselines. These results again demonstrate that Rad can effectively improve the generalization performance.

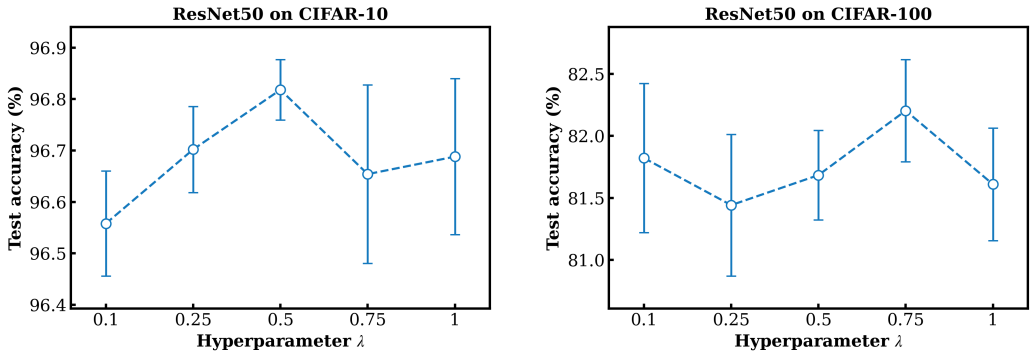
Figure 3 presents the test accuracy under different λ settings on CIFAR-10 and CIFAR-100. We evaluate the influence of λ using ResNet50, where each result is averaged over five independent runs. It can be observed that Rad maintains relatively stable performance across different λ values and achieves the best results at $\lambda = 0.5$ on CIFAR-10 and $\lambda = 0.75$ on CIFAR-100. Additional hyperparameter sensitivity results are provided in Appendix F.

Visualization Analysis: First, we analyze the Hessian eigenvalue distributions of solutions obtained by different optimization methods on CIFAR-10 and CIFAR-100. The maximum Hessian



(a) Test loss. (b) Train-test accuracy gap

Figure 2: Test performance of different methods on CIFAR-10 using ResNet50. The inset plots magnify epochs 180–200.



(a) (b)

Figure 3: Test accuracy under different λ settings on CIFAR-10 and CIFAR-100.

eigenvalue is commonly used to characterize the local curvature of the loss landscape. The results show that introducing Rad does not lead to a sharper local loss landscape. Detailed results are provided in Appendix G. Furthermore, we analyze the effect of Rad on sample margins. The results show that Rad yields more pronounced margin improvements for hard samples, which is consistent with our theoretical analysis. Detailed results are provided in Appendix H.

Computational Cost: Since Rad is a plug-and-play perturbation enhancement mechanism, it only introduces an additional sample weighting operation during perturbation construction and does not require extra forward or backward passes. Therefore, Rad can be directly integrated into existing SAM or its variants with comparable computational cost.

6 CONCLUSION

In this paper, we revisit the relationship between the sharpness of loss and model generalization, and establish a new generalization bound from the perspective of excess risk. The bound shows that better generalization requires simultaneous control of both empirical sharpness and local Rademacher complexity. To avoid the direct computation of local Rademacher complexity, we derive an upper bound characterized by sample margins. Motivated by this bound, we develop a Rademacher-guided mechanism (Rad) that uses sample margin information to guide parameter perturbation construction. Extensive experiments demonstrate the effectiveness of the proposed method on SAM and its variants. Future work will explore more accurate local complexity measures and their applications to broader models.

AI USE STATEMENT

In this work, we use generative AI tools to assist with the design of and feedback on research methodology and experiments, as well as translation. We do not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, implement methods, or interpret results. Generating synthetic datasets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we use generative AI tools for sourcing and searching for information, editing the manuscript to improve readability, and identifying relevant literature. We review all AI-assisted work. Information and literature suggested by generative AI tools are manually verified against the original sources, and AI-assisted edits are reviewed by the authors to ensure that they do not alter the intended technical meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on theoretical analysis and optimization methods for deep neural networks. The experiments use publicly available benchmark datasets and do not involve human subjects, sensitive personal data, or other specific ethical concerns.

REPRODUCIBILITY STATEMENT

The assumptions and theoretical results are presented in Sections 3 and 4.1, and the complete proofs are provided in Appendices A–C. The optimization procedure of Rad is described in Section 4.2, and the complete algorithm is provided in Appendix D. Dataset preprocessing, model architectures, training configurations, implementation details, and hyperparameter settings for the experiments are provided in Appendix E. In addition, anonymous source code is provided in the supplementary material to further support the reproducibility of our results.

REFERENCES

Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 639–668, 2022.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning*, pp. 840–902, 2023.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, 2006.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Amir Beck. *First-Order Methods in Optimization*, volume 10 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2017.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186, 2010.

Jianlong Chen and Zhiming Zhou. Revisiting sharpness-aware minimization: A more faithful and effective implementation. In *International Conference on Learning Representations*, pp. 150036–150062, 2026.

Xuewei Cheng, Ke Huang, and Shujie Ma. Generalization and risk bounds for recurrent neural networks. *Neurocomputing*, 616:128825, 2025.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A pac-bayesian link between generalisation and flat minima. In *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, pp. 1–31, 2025.

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.

Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. Exponential tail local rademacher complexity risk bounds without the bernstein condition. *Journal of Machine Learning Research*, 25 (388):1–43, 2024.

Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum hessian eigenvalue and generalization. In *Proceedings on "I Can't Believe It's Not Better! – Understanding Deep Learning Through Empirical Falsification" at NeurIPS 2022 Workshops*, pp. 51–65, 2023.

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.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1–2, 2011.

Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive Sharpness-Aware Minimization for Scale-Invariant Learning of Deep Neural Networks. In *International Conference on Machine Learning*, volume 139, pp. 5905–5914, 2021.

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, 2024.

Tongtong Liang, Dan Qiao, Yu-Xiang Wang, and Rahul Parhi. Stable minima of relu neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24686–24750, 2025.

Yuhan Helena Liu, Arna Ghosh, Blake Richards, Eric Shea-Brown, and Guillaume Lajoie. Beyond accuracy: generalization properties of bio-plausible temporal credit assignment rules. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23077–23097, 2022.

Yuhang Liu, Tao Li, Zhehao Huang, Zuopeng Yang, and Xiaolin Huang. Bi-lora: Efficient sharpness-aware minimization for fine-tuning large-scale models. In *International Conference on Learning Representations*, volume 2026, pp. 96782–96813, 2026.

Ziqing Lu, Chang Xu, Bo Du, Takashi Ishida, Lefei Zhang, and Masashi Sugiyama. Localdrop: A hybrid regularization for deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3590–3601, 2022.

Haocheng Luo, Tuan Truong, Tung Pham, Mehrtash Harandi, Dinh Phung, and Trung Le. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4424–4453, 2024.

Ze Peng, Jian Zhang, Yisen Wang, Lei Qi, Yinghuan Shi, and Yang Gao. Leveraging flatness to improve information-theoretic generalization bounds for sgd. In *International Conference on Learning Representations*, volume 2025, pp. 43012–43060, 2025.

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.

Adepu Ravi Sankar, Yash Khasbage, Rahul Vigneswaran, and Vineeth N. Balasubramanian. A deeper look at the hessian eigenspectrum of deep neural networks and its applications to regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9481–9488, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Avner Shultzman, Eyar Azar, Miguel R. D. Rodrigues, and Yonina C. Eldar. Generalization and estimation error bounds for model-based neural networks. In *International Conference on Learning Representations*, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Behrooz Tahmasebi, Ashkan Soleymani, Dara Bahri, Stefanie Jegelka, and Patrick Jaillet. A universal class of sharpness-aware minimization algorithms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 47418–47440, 2024.

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*, pp. 9636–9647, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Chang Xu, Tongliang Liu, Dacheng Tao, and Chao Xu. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016.

Yingzhen Yang. A new concentration inequality for sampling without replacement and its application for transductive learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70472–70497, 2025.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE International Conference on Big Data*, pp. 581–590, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference*, 2016.

Qiaozhe Zhang, Jun Sun, and Yingzhuang Liu. Rényi sharpness: A novel sharpness that strongly correlates with generalization. In *International Conference on Learning Representations*, 2026.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training. In *International Conference on Learning Representations*, volume 2025, pp. 20949–20980, 2025.

A Proof of Lemma 3.4

Given a parameter $w \in \mathcal{W}$ and a perturbation radius $\rho > 0$, consider any $\widetilde{w} \in \mathcal{W}_\rho(w)$. According to Definition 3.3, the perturbation from w to $\widetilde{w}$ induces the local loss-difference function $g_{\widetilde{w}}(z) = \ell(\widetilde{w}; z) - \ell(w; z)$. Since the loss function is bounded in $[0, a]$, we have $g_{\widetilde{w}}(z) \in [-a, a]$. By the definitions of empirical sharpness and population sharpness, we have:

$$\begin{aligned}
\mathrm{Sharp}_{S,\rho}(w) &= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \left[L_S(\widetilde{w}) - L_S(w) \right] \\
&= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(\widetilde{w}; z_i) - \ell(w; z_i) \right] \\
&= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z_i),
\end{aligned} \tag{13}$$

and

$$\begin{aligned}
\mathrm{Sharp}_{\mathcal{D},\rho}(w) &= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \left[L_{\mathcal{D}}(\widetilde{w}) - L_{\mathcal{D}}(w) \right] \\
&= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \mathbb{E}_{z \sim \mathcal{D}} \left[\ell(\widetilde{w}; z) - \ell(w; z) \right] \\
&= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \mathbb{E}_{z \sim \mathcal{D}} \left[g_{\widetilde{w}}(z) \right].
\end{aligned} \tag{14}$$

Moreover, using the inequality that the maximum of expectations is no larger than the expectation of the maximum, we obtain the relation between population sharpness and the expectation of empirical sharpness:

$$\begin{aligned}
\mathrm{Sharp}_{\mathcal{D},\rho}(w) &= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \mathbb{E}_{z \sim \mathcal{D}} \left[g_{\widetilde{w}}(z) \right] \\
&= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z_i) \right] \\
&\leq \mathbb{E}_S \left[\max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z_i) \right] \\
&= \mathbb{E}_S \left[\mathrm{Sharp}_{S,\rho}(w) \right].
\end{aligned} \tag{15}$$

Next, we establish the bounded-difference property of empirical sharpness, which allows us to apply McDiarmid's inequality. Specifically, let S' be another training set obtained by replacing the j-th sample z_j in S with z'_j. For any $\widetilde{w} \in \mathcal{W}_\rho(w)$, we have:

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z_i) - \frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z'_i) \right| \\
&= \left| \frac{1}{n} \left(\sum_{i \neq j} g_{\widetilde{w}}(z_i) + g_{\widetilde{w}}(z_j) \right) - \frac{1}{n} \left(\sum_{i \neq j} g_{\widetilde{w}}(z_i) + g_{\widetilde{w}}(z'_j) \right) \right| \\
&= \frac{1}{n} \left| g_{\widetilde{w}}(z_j) - g_{\widetilde{w}}(z'_j) \right| \leq \frac{2a}{n}.
\end{aligned} \tag{16}$$

Then, the empirical sharpness on S and S' satisfies:

$$\begin{aligned}
\mathrm{Sharp}_{S,\rho}(w) &= \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z_i) \\
&\leq \max_{\widetilde{w} \in \mathcal{W}_\rho(w)} \left[\frac{1}{n} \sum_{i=1}^{n} g_{\widetilde{w}}(z'_i) + \frac{2a}{n} \right] \\
&= \mathrm{Sharp}_{S',\rho}(w) + \frac{2a}{n}.
\end{aligned} \tag{17}$$

By exchanging the roles of S and S', we obtain the same bound in the reverse direction. Therefore, replacing a single training sample changes the empirical sharpness by at most $2a/n$, and $\mathrm{Sharp}_{S,\rho}(w)$ satisfies the bounded difference condition. According to McDiarmid's inequality, for any $t > 0$,

$$\Pr_S\left(\mathbb{E}_S\left[\mathrm{Sharp}_{S,\rho}(w)\right] - \mathrm{Sharp}_{S,\rho}(w) \geq t\right) \leq \exp\left(-\frac{nt^2}{2a^2}\right). \tag{18}$$

Taking $t = a\sqrt{\frac{2\log(1/\delta)}{n}}$, we obtain that, with probability at least $1-\delta$,

$$\mathbb{E}_S\left[\mathrm{Sharp}_{S,\rho}(w)\right] \leq \mathrm{Sharp}_{S,\rho}(w) + a\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{19}$$

Substituting equation 19 into equation 15, we have:

$$\mathrm{Sharp}_{\mathcal{D},\rho}(w) \leq \mathrm{Sharp}_{S,\rho}(w) + a\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{20}$$

B Proof of Theorem 3.5

Given the population-risk minimizer $w^\star$ and a perturbation radius $\rho > 0$, we have the local parameter class $\mathcal{W}_\rho(w^\star)$ and the corresponding local loss-difference function class $\mathcal{G}_\rho(w^\star)$. The Bernstein-type condition relates the second moment of a function to its expectation. It is commonly used in local Rademacher complexity analyses to control localized function classes Cheng et al. (2025). For any $g_{\widetilde{w}} \in \mathcal{G}_\rho(w^\star)$, the Bernstein-type condition and Definition 3.3 give:

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{D}}\left[g_{\widetilde{w}}^2(z)\right] &\leq \nu\mathbb{E}_{z\sim\mathcal{D}}\left[g_{\widetilde{w}}(z)\right] \\
&= \nu\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(\widetilde{w};z) - \ell(w^\star;z)\right] \\
&= \nu\left[L_{\mathcal{D}}(\widetilde{w}) - L_{\mathcal{D}}(w^\star)\right] \\
&\leq \nu\,\mathrm{Sharp}_{\mathcal{D},\rho}(w^\star).
\end{aligned} \tag{21}$$

Since Lemma 3.4 holds for any fixed $w \in \mathcal{W}$, setting $w = w^\star$ gives, with probability at least $1-\delta/2$,

$$\mathrm{Sharp}_{\mathcal{D},\rho}(w^\star) \leq \mathrm{Sharp}_{S,\rho}(w^\star) + a\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{22}$$

Substituting the empirical upper bound on $\mathrm{Sharp}_{\mathcal{D},\rho}(w^\star)$ into the equation 21, with probability at least $1-\delta$, for every $g_{\widetilde{w}} \in \mathcal{G}_\rho(w^\star)$,

$$\mathbb{E}_{z\sim\mathcal{D}}\left[g_{\widetilde{w}}^2(z)\right] \leq \nu\left[\mathrm{Sharp}_{S,\rho}(w^\star) + a\sqrt{\frac{2\log(2/\delta)}{n}}\right]. \tag{23}$$

The equation 23 shows that the second moment of each function in the local loss-difference function class is bounded by the empirical sharpness together with a confidence term, which is required in the subsequent local Rademacher analysis. For notational simplicity, denote the right-hand side of equation 23 by r_{flat}. Moreover, since $0 \leq \ell(w^\star;z) \leq a$, we have $\ell^2(w^\star;z) \leq a\ell(w^\star;z)$. Taking expectation over $z \sim \mathcal{D}$ and applying the η-approximate realizability condition gives:

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(w^\star;z)\right] &\leq a\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(w^\star;z)\right] \\
&= aL_{\mathcal{D}}(w^\star) \leq a\eta.
\end{aligned} \tag{24}$$

For any $\widetilde{w} \in \mathcal{W}_\rho(w^\star)$, we decompose $\ell(\widetilde{w};z)$ as $\ell(w^\star;z)$ plus $g_{\widetilde{w}}(z)$. Since the second moments of both terms are bounded, the Cauchy–Schwarz inequality can be used to derive an upper bound on

the second moment of $\ell(\widetilde{w}; z)$:

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(\widetilde{w}; z)\right] &= \mathbb{E}_{z\sim\mathcal{D}}\left[\left(\ell(w^\star; z) + g_{\widetilde{w}}(z)\right)^2\right] \\
&= \mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(w^\star; z)\right] + 2\mathbb{E}_{z\sim\mathcal{D}}\left[\ell(w^\star; z) g_{\widetilde{w}}(z)\right] + \mathbb{E}_{z\sim\mathcal{D}}\left[g^2_{\widetilde{w}}(z)\right] \\
&\leq \mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(w^\star; z)\right] + 2\sqrt{\mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(w^\star; z)\right]}\sqrt{\mathbb{E}_{z\sim\mathcal{D}}\left[g^2_{\widetilde{w}}(z)\right]} + \mathbb{E}_{z\sim\mathcal{D}}\left[g^2_{\widetilde{w}}(z)\right] \\
&= \left(\sqrt{\mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(w^\star; z)\right]} + \sqrt{\mathbb{E}_{z\sim\mathcal{D}}\left[g^2_{\widetilde{w}}(z)\right]}\right)^2 \\
&\leq \left(\sqrt{a\eta} + \sqrt{r_{\mathrm{flat}}}\right)^2 . \qquad (25)
\end{aligned}$$

Equation 23 provides a second-moment bound for every function in the local loss-difference function class $\mathcal{G}_\rho(w^\star)$. This bound is then substituted into equation 25, and the triangle inequality gives the following second-moment bound for every function in $\mathcal{L}_\rho(w^\star)$:

$$\begin{aligned}
\sqrt{\mathbb{E}_{z\sim\mathcal{D}}\left[\ell^2(\widetilde{w}; z)\right]} &\leq \sqrt{a\eta} + \sqrt{r_{\mathrm{flat}}} \\
&\leq \sqrt{a\eta} + \sqrt{\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)} + \sqrt{\nu a}\left(\frac{2\log(2/\delta)}{n}\right)^{1/4} . \qquad (26)
\end{aligned}$$

Let $\sqrt{r_{\eta,\mathrm{flat}}}$ be the right-hand side of equation 26. Each $\widetilde{w} \in \mathcal{W}_\rho(w^\star)$ corresponds to a loss function $\ell_{\widetilde{w}} \in \mathcal{L}_\rho(w^\star)$. Equation 25 shows that every function in $\mathcal{L}_\rho(w^\star)$ satisfies the uniform second-moment bound $r_{\eta,\mathrm{flat}}$. We apply Theorem 2.1 to the local parameter set $\mathcal{W}_\rho(w^\star)$ and its corresponding loss class $\mathcal{L}_\rho(w^\star)$, for which the second-moment condition holds with $r = r_{\eta,\mathrm{flat}}$. Setting the confidence parameter to $\delta/2$, the following bound holds with probability at least $1 - \delta/2$:

$$\begin{aligned}
L_{\mathcal{D}}(\widehat{w}) \leq \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w) + 8R_S(\mathcal{L}_\rho(w^\star)) + \sqrt{\frac{8r_{\eta,\mathrm{flat}}\log(4/\delta)}{n}} \\
+ \frac{32\log(4/\delta)}{3n} + \Delta(\mathcal{W}_\rho(w^\star), \mathcal{W}) . \qquad (27)
\end{aligned}$$

Since $w^\star$ is a population-risk minimizer over $\mathcal{W}$ and belongs to the local parameter set $\mathcal{W}_\rho(w^\star)$, the local parameter set attains the same optimal population risk as $\mathcal{W}$. Therefore, $\Delta(\mathcal{W}_\rho(w^\star), \mathcal{W}) = 0$. We substitute the expression for $\sqrt{r_{\eta,\mathrm{flat}}}$ into the square-root term in equation 27 and obtain:

$$\begin{aligned}
\sqrt{\frac{8r_{\eta,\mathrm{flat}}\log(4/\delta)}{n}} &\leq \sqrt{\frac{8\log(4/\delta)}{n}}\left[\sqrt{a\eta} + \sqrt{\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)} + \sqrt{\nu a}\left(\frac{2\log(4/\delta)}{n}\right)^{1/4}\right] \\
&= \sqrt{\frac{8\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)\log(4/\delta)}{n}} + \sqrt{\frac{8a\eta\log(4/\delta)}{n}} + 2\sqrt{\nu a}\left(\frac{2\log(4/\delta)}{n}\right)^{3/4} . \qquad (28)
\end{aligned}$$

Substituting this into the equation 27, with probability at least $1 - \delta$, we obtain:

$$L_{\mathcal{D}}(\widehat{w}) \leq \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w) + 8R_S(\mathcal{L}_\rho(w^\star)) + \sqrt{\frac{8\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)\log(4/\delta)}{n}} + \Omega_n(\delta, \eta), \qquad (29)$$

where $\Omega_n(\delta, \eta) = \sqrt{\frac{8a\eta\log(4/\delta)}{n}} + 2\sqrt{\nu a}\left(\frac{2\log(4/\delta)}{n}\right)^{3/4} + \frac{32\log(4/\delta)}{3n}$.

To make the sharpness and complexity terms characterize the same class of local loss variations, we examine the relationship between the Rademacher complexities of the localized loss class $\mathcal{L}_\rho(w^\star)$ and the local loss-difference function class $\mathcal{G}_\rho(w^\star)$. The localized loss class is obtained by adding the same fixed function $\ell(w^\star; \cdot)$ to every function in the local loss-difference function class. Since Rademacher complexity is invariant under translation by a fixed function, the two classes have the

same Rademacher complexity:

$$\begin{aligned} R_S(\mathcal{L}_\rho(w^\star)) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\widetilde{w}\in\mathcal{W}_\rho(w^\star)} \frac{1}{n}\sum_{i=1}^n \sigma_i\left(\ell(w^\star;z_i)+g_{\widetilde{w}}(z_i)\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\ell(w^\star;z_i)+\sup_{\widetilde{w}\in\mathcal{W}_\rho(w^\star)}\frac{1}{n}\sum_{i=1}^n\sigma_i g_{\widetilde{w}}(z_i)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i\ell(w^\star;z_i)\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\widetilde{w}\in\mathcal{W}_\rho(w^\star)}\frac{1}{n}\sum_{i=1}^n\sigma_i g_{\widetilde{w}}(z_i)\right] \\ &= R_S(\mathcal{G}_\rho(w^\star)), \end{aligned} \tag{30}$$

Substituting equation 30 into the equation 29, we obtain, with probability at least $1-\delta$,

$$L_{\mathcal{D}}(\widehat{w}) \le \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w) + 8R_S(\mathcal{G}_\rho(w^\star)) + \sqrt{\frac{8\nu\,\mathrm{Sharp}_{S,\rho}(w^\star)\log(4/\delta)}{n}} + \Omega_n(\delta,\eta), \tag{31}$$

where $\Omega_n(\delta,\eta) = \sqrt{\frac{8a\eta\log(4/\delta)}{n}} + 2\sqrt{\nu a}\left(\frac{2\log(4/\delta)}{n}\right)^{3/4} + \frac{32\log(4/\delta)}{3n}$.

C Proof of Theorem 4.1

Given $w\in\mathcal{W}$ and a perturbation radius $\rho>0$, consider the local parameter neighborhood $\mathcal{W}_\rho(w) = \{\widetilde{w} = w+\epsilon\in\mathcal{W} : \|\epsilon\|_2\le\rho\}$, where ϵ denotes a weight perturbation. The corresponding local loss-difference function class is $\mathcal{G}_\rho(w) = \{g_{\widetilde{w}} \mid \widetilde{w}\in\mathcal{W}_\rho(w)\}$, where $g_{\widetilde{w}}(z) = \ell(\widetilde{w};z)-\ell(w;z)$. Since $\ell(\cdot;z_i)$ is β-smooth over the local parameter neighborhood $\mathcal{W}_\rho(w)$, the descent lemma (Beck, 2017) implies that, for any $\widetilde{w}\in\mathcal{W}_\rho(w)$:

$$\begin{aligned} \ell(\widetilde{w};z_i)-\ell(w;z_i) &\le \nabla_w\ell(w;z_i)^\top(\widetilde{w}-w)+\frac{\beta}{2}\|\widetilde{w}-w\|_2^2 \\ &= \nabla_w\ell(w;z_i)^\top\epsilon+\frac{\beta}{2}\|\epsilon\|_2^2 \\ &\le \nabla_w\ell(w;z_i)^\top\epsilon+\frac{\beta\rho^2}{2}. \end{aligned} \tag{32}$$

We define $r_i(\epsilon) = \ell(\widetilde{w};z_i)-\ell(w;z_i)-\nabla_w\ell(w;z_i)^\top\epsilon$, which measures the difference between the exact loss change and its first-order approximation. Equation 32 gives $r_i(\epsilon)\le\frac{\beta\rho^2}{2}$. Clearly, the negated loss $-\ell(\cdot;z_i)$ only reverses the gradient direction and does not change the norm of gradient differences. Therefore, $-\ell(\cdot;z_i)$ is also β-smooth over $\mathcal{W}_\rho(w)$. Applying the descent lemma (Beck, 2017) to $-\ell(\cdot;z_i)$ gives $r_i(\epsilon)\ge-\frac{\beta\rho^2}{2}$. Together with equation 32, we can have:

$$\left|\ell(\widetilde{w};z_i)-\ell(w;z_i)-\nabla_w\ell(w;z_i)^\top\epsilon\right| \le \frac{\beta\rho^2}{2}. \tag{33}$$

We focus on the empirical Rademacher complexity of the local loss-difference function class $\mathcal{G}_\rho(w)$. For any $\widetilde{w} = w+\epsilon\in\mathcal{W}_\rho(w)$, the local loss-difference function can be rewritten as:

$$g_{\widetilde{w}}(z_i) = \ell(\widetilde{w};z_i)-\ell(w;z_i) = \nabla_w\ell(w;z_i)^\top\epsilon + r_i(\epsilon). \tag{34}$$

Therefore, substituting equation 34 into the definition of empirical Rademacher complexity gives:

$$\begin{aligned} R_S\big(\mathcal{G}_\rho(w)\big) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\epsilon\|_2\le\rho}\frac{1}{n}\sum_{i=1}^n\sigma_i\left(\nabla_w\ell(w;z_i)^\top\epsilon+r_i(\epsilon)\right)\right] \\ &\le \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\epsilon\|_2\le\rho}\epsilon^\top\sum_{i=1}^n\sigma_i\nabla_w\ell(w;z_i)\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\epsilon\|_2\le\rho}\frac{1}{n}\sum_{i=1}^n\sigma_i r_i(\epsilon)\right] \end{aligned} \tag{35}$$

We first consider the first term in equation 35. For any vector v, the duality of the Euclidean norm gives $\sup_{\|\epsilon\|_2 \le \rho} \epsilon^\top v = \rho \|v\|_2$. Setting $v = \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i)$, we obtain

$$\frac{1}{n}\mathbb{E}_\sigma \left[\sup_{\|\epsilon\|_2 \le \rho} \epsilon^\top \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right] = \frac{\rho}{n} \mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2 . \tag{36}$$

For notational simplicity, let $X = \|\sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i)\|_2$. Since $X \ge 0$, applying Jensen's inequality to the convex function $f(x) = x^2$ and then taking square roots gives $\mathbb{E}_{\boldsymbol{\sigma}}[X] \le \sqrt{\mathbb{E}_{\boldsymbol{\sigma}}[X^2]}$. Hence,

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2 \le \sqrt{\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2^2}. \tag{37}$$

Expanding the squared norm in the above expression, we obtain:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2^2 &= \mathbb{E}_{\boldsymbol{\sigma}} \left[\left\langle \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i), \sum_{j=1}^n \sigma_j \nabla_w \ell(w; z_j) \right\rangle \right] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_{\boldsymbol{\sigma}} [\sigma_i \sigma_j \langle \nabla_w \ell(w; z_i), \nabla_w \ell(w; z_j) \rangle] \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_{\boldsymbol{\sigma}} [\sigma_i \sigma_j] \langle \nabla_w \ell(w; z_i), \nabla_w \ell(w; z_j) \rangle . \end{aligned} \tag{38}$$

Since the Rademacher variables are independent and satisfy $\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i] = 0$ and $\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i^2] = 1$, we have:

$$\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i \sigma_j] = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases} \tag{39}$$

Thus, all cross terms vanish, and

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2^2 = \sum_{i=1}^n \|\nabla_w \ell(w; z_i)\|_2^2 . \tag{40}$$

Together with equation 40 and equation 36, the first-order term is bounded as:

$$\mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^n \sigma_i \nabla_w \ell(w; z_i) \right\|_2 \le \frac{\rho}{n} \sqrt{\sum_{i=1}^n \|\nabla_w \ell(w; z_i)\|_2^2}. \tag{41}$$

Next, we consider the second term in equation 35. Since $|\sigma_i| = 1$ and $|r_i(\epsilon)| \le \beta\rho^2/2$ for all $\|\epsilon\|_2 \le \rho$, we have:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\epsilon\|_2 \le \rho} \frac{1}{n} \sum_{i=1}^n \sigma_i r_i(\epsilon) \right] &\le \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{\|\epsilon\|_2 \le \rho} \frac{1}{n} \sum_{i=1}^n |r_i(\epsilon)| \right] \\ &\le \frac{1}{n} \sum_{i=1}^n \frac{\beta\rho^2}{2} = \frac{\beta\rho^2}{2}. \end{aligned} \tag{42}$$

Substituting the equation 41 and equation 42 into equation 35 gives:

$$R_S\big(\mathcal{G}_\rho(w)\big) \le \frac{\rho}{n} \sqrt{\sum_{i=1}^n \|\nabla_w \ell(w; z_i)\|_2^2} + \frac{\beta\rho^2}{2}. \tag{43}$$

Therefore, the empirical Rademacher complexity is bounded by a first-order term that depends on the sample-wise gradient norms and a constant term determined by the smoothness parameter β and the perturbation radius ρ.

For a given sample $z_i = (x_i, y_i)$, let $f_w(x_i) = (f_{w,1}(x_i), \ldots, f_{w,K}(x_i))^\top$ denote the logit vector produced by the model, where K denotes the number of classes. The corresponding softmax probability vector is denoted by $p_i(w) = (p_{i,1}(w), \ldots, p_{i,K}(w))^\top$ where $p_{i,c}(w)$ is given by:

$$p_{i,c}(w) = \frac{\exp(f_{w,c}(x_i))}{\sum_{k=1}^{K} \exp(f_{w,k}(x_i))}, \qquad c = 1, \ldots, K. \tag{44}$$

The cross-entropy loss is $\ell(w; z_i) = -\log p_{i,y_i}(w)$. For each class $c \in \{1, \ldots, K\}$, we compute the derivative of the cross-entropy loss with respect to the corresponding logit $f_{w,c}(x_i)$. According to the chain rule,

$$\begin{aligned}
\frac{\partial \ell(w; z_i)}{\partial f_{w,c}(x_i)} &= -\frac{1}{p_{i,y_i}(w)} \frac{\partial}{\partial f_{w,c}(x_i)} \left[\frac{\exp\left(f_{w,y_i}(x_i)\right)}{\sum_{k=1}^{K} \exp\left(f_{w,k}(x_i)\right)} \right] \\
&= -\frac{1}{p_{i,y_i}(w)} p_{i,y_i}(w) \left(\mathbf{1}\{c = y_i\} - p_{i,c}(w) \right) \\
&= p_{i,c}(w) - \mathbf{1}\{c = y_i\},
\end{aligned} \tag{45}$$

where $\mathbf{1}\{c = y_i\}$ is the indicator function, which equals 1 when $c = y_i$ and 0 otherwise. The derivatives over all classes form the vector

$$\nabla_{f_w(x_i)} \ell(w; z_i) = p_i(w) - e_{y_i}. \tag{46}$$

Let $J_i(w) = \frac{\partial f_w(x_i)}{\partial w} \in \mathbb{R}^{K \times d}$ denote the Jacobian matrix of the logit vector with respect to the model parameters, where d is the dimension of w. The chain rule gives:

$$\nabla_w \ell(w; z_i) = J_i(w)^\top \nabla_{f_w(x_i)} \ell(w; z_i) = J_i(w)^\top \left(p_i(w) - e_{y_i} \right). \tag{47}$$

Under the assumed Jacobian bound $\|J_i(w)\|_{\mathrm{op}} \le \kappa$, we have:

$$\begin{aligned}
\|\nabla_w \ell(w; z_i)\|_2 &= \left\| J_i(w)^\top \left(p_i(w) - e_{y_i} \right) \right\|_2 \\
&\le \|J_i(w)\|_{\mathrm{op}} \left\| p_i(w) - e_{y_i} \right\|_2, \\
&= \|J_i(w)\|_{\mathrm{op}} \sqrt{\left(1 - p_{i,y_i}(w)\right)^2 + \sum_{c \neq y_i} p_{i,c}^2(w)} \\
&\le \sqrt{2}\, \|J_i(w)\|_{\mathrm{op}} \left(1 - p_{i,y_i}(w)\right) \\
&\le \sqrt{2}\kappa \left(1 - p_{i,y_i}(w)\right).
\end{aligned} \tag{48}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm induced by the Euclidean norm. Define the classification margin of sample z_i as $m_i(w) = f_{w,y_i}(x_i) - \max_{c \neq y_i} f_{w,c}(x_i)$. For every $c \neq y_i$, the margin definition implies $f_{w,c}(x_i) - f_{w,y_i}(x_i) \le -m_i(w)$. Therefore,

$$\begin{aligned}
1 - p_{i,y_i}(w) &= \frac{\sum_{c \neq y_i} \exp\left(f_{w,c}(x_i) - f_{w,y_i}(x_i)\right)}{1 + \sum_{c \neq y_i} \exp\left(f_{w,c}(x_i) - f_{w,y_i}(x_i)\right)} \\
&\le \frac{(K-1)\exp\left(-m_i(w)\right)}{1 + (K-1)\exp\left(-m_i(w)\right)}.
\end{aligned} \tag{49}$$

For notational simplicity, denote the margin-dependent upper bound by

$$q_i(w) = \frac{(K-1)\exp\left(-m_i(w)\right)}{1 + (K-1)\exp\left(-m_i(w)\right)}. \tag{50}$$

We substitute equation 48 and equation 50 into equation 43 and obtain

$$R_S\big(\mathcal{G}_\rho(w)\big) \leq \frac{\rho}{n}\sqrt{\sum_{i=1}^{n}\Big(\sqrt{2}\kappa q_i(w)\Big)^2} + \frac{\beta\rho^2}{2}$$
$$= \frac{\sqrt{2}\rho\kappa}{n}\sqrt{\sum_{i=1}^{n} q_i^2(w)} + \frac{\beta\rho^2}{2}. \tag{51}$$

D ALGORITHM

Theorem 3.5 shows that empirical sharpness and local Rademacher complexity should be considered simultaneously, while Theorem 4.1 further shows that the upper bound on local Rademacher complexity depends on sample-wise classification margins. Based on this result, we retain the perturbation construction used in SAM or its variants, and introduce an adjustment based on sample-wise classification margins. Specifically, at iteration t, we first compute the classification margin of each sample in the current mini-batch and obtain the coefficient γ_i from its deviation from the mini-batch average margin. These coefficients are then used to weight the sample gradients and construct the Rademacher-guided adjustment $\Delta g_t^{\mathrm{Rad}}$. We add this adjustment to the original perturbation direction g_t^{base} to obtain the Rad perturbation $\epsilon_t^{\mathrm{Rad}}$. Finally, we compute the mini-batch loss gradient at the perturbed parameters and update the model. The complete optimization procedure is summarized in Algorithm 1.

Algorithm 1 Optimization procedure of Rad

Require: Training set S, iterations T, perturbation radius ρ, learning rate η, and hyperparameter λ
1: Initialize model parameters w_0
2: **for** $t = 0, \ldots, T-1$ **do**
3: Sample a mini-batch $B = \{z_i\}_{i=1}^{m}$
4: Compute the classification margin $m_i(w_t)$ for each sample and γ_i according to equation 9
5: Compute the Rademacher-guided adjustment $\Delta g_t^{\mathrm{Rad}}$ according to equation 10
6: Obtain the perturbation direction g_t^{base} and construct $\epsilon_t^{\mathrm{Rad}}$ according to equation 11
7: Compute $g_t = \nabla_w L_B(w_t + \epsilon_t^{\mathrm{Rad}})$
8: Update $w_{t+1} = w_t - \eta g_t$
9: **end for**
10: **return** w_T

E EXPERIMENT SETTING

CIFAR-10 and CIFAR-100 We evaluate VGG16, ResNet18, ResNet50, and WideResNet on CIFAR-10 and CIFAR-100. Each model is trained for 200 epochs with batch size of 64. Training images are randomly cropped to 32×32 after four-pixel padding, horizontally flipped, and augmented with Cutout, which masks a 16×16 region with probability 0.5. For each dataset, we compute the per-channel mean and standard deviation from its training images and use these values to normalize both train and test images. All methods use momentum SGD as the base optimizer, with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 5×10^{-4}. The learning rate is multiplied by 0.2 at epochs 60, 120, and 160, and the label-smoothing coefficient is set to 0.1. The perturbation radius is $\rho = 0.05$ for SAM and $\rho = 2$ for ASAM. We set $\lambda = 0.75$ for VGG16 and $\lambda = 0.5$ for ResNet18, ResNet50, and WideResNet on CIFAR-10. On CIFAR-100, we set $\lambda = 0.75$ for VGG16, ResNet50, and WideResNet, and $\lambda = 0.5$ for ResNet18. The XSAM baseline uses an initial perturbation radius of 0.15, a maximum radius of 0.3, one ascent step, and dynamic radius adjustment. We repeat all experiments five times. We implement all methods in PyTorch and run the experiments on two NVIDIA Tesla V100 GPUs.

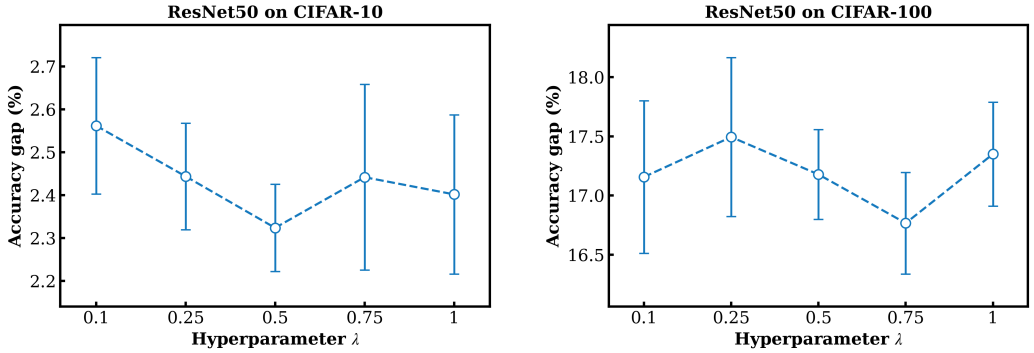
ImageNet-1K We train ResNet50 from scratch on ImageNet-1K for 90 epochs with batch size of 128. Training images undergo random resized cropping to 224×224 and random horizontal flipping.

Validation images are resized to 256 pixels and center-cropped to 224×224. Both splits are normalized using RGB means of $(0.485, 0.456, 0.406)$ and standard deviations of $(0.229, 0.224, 0.225)$. All methods use momentum SGD as the base optimizer, with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 10^{-4}. The learning rate follows a cosine schedule over 90 epochs, and training uses cross-entropy loss without label smoothing. We set the perturbation radius to $\rho = 0.07$ for SAM and ASAM, including their regularized variants. The value of λ is set to 0.5. We evaluate top-1 and top-5 accuracy on the validation split and repeat each experiment three times. The implementation is based on PyTorch and runs on two NVIDIA Tesla V100 GPUs.

Fine-grained classification datasets We evaluate ViT-B/16 on CUB-200-2011, Stanford Dogs, and Stanford Cars. The model is initialized from a pretrained ViT-B/16 checkpoint and fine-tuned for 50 epochs with batch size of 32. Images are resized to 512×512 using bicubic interpolation and cropped to 384×384. Training uses random cropping and horizontal flipping, while evaluation uses center cropping. All images are normalized using RGB means of $(0.485, 0.456, 0.406)$ and standard deviations of $(0.229, 0.224, 0.225)$. All methods use momentum SGD as the base optimizer, with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 5×10^{-4}. The learning rate is multiplied by 0.2 at epochs 15, 30, and 40. We use a label-smoothing coefficient of 0.1 and a drop-path rate of 0.2. The perturbation radius is $\rho = 0.05$ for SAM and $\rho = 2$ for ASAM. We set $\lambda = 0.5$ on CUB-200-2011 and $\lambda = 0.75$ on Stanford Dogs. On Stanford Cars, we use $\lambda = 0.5$ for SAM+Rad and $\lambda = 0.75$ for ASAM+Rad. The XSAM baseline uses an initial perturbation radius of 0.15, a maximum radius of 0.3, one ascent step, and dynamic radius adjustment. We repeat each experiment three times. We implement all methods in PyTorch and run the experiments on two NVIDIA Tesla V100 GPUs.

F Hyperparameter Sensitivity Analysis

Figure 4 presents the accuracy gap under different λ settings on CIFAR-10 and CIFAR-100. We evaluate the influence of λ using ResNet50, where each result is averaged over five independent runs. It can be observed that Rad maintains relatively consistent accuracy gaps across different λ values and achieves lower gaps at $\lambda = 0.5$ on CIFAR-10 and $\lambda = 0.75$ on CIFAR-100. These results provide additional analysis of the effect of the weighting coefficient λ.



(a) (b)

Figure 4: Accuracy gap under different λ settings on CIFAR-10 and CIFAR-100.

G Hessian Eigenvalue Visualization

Figure 5 presents the Hessian eigenvalue distributions on CIFAR-10. Each row reports the results for one network architecture: VGG16, ResNet18, ResNet50, and WRN-28-10 from top to bottom. Within each row, the five panels from left to right correspond to SGD, SAM, SAM+Rad, ASAM, and ASAM+Rad. We use PyHessian (Yao et al., 2020) to compute the Hessian eigenvalue spectral density, the maximum eigenvalue $\lambda_{\max}$, and the ratio λ_1/λ_5 between the largest and fifth-largest

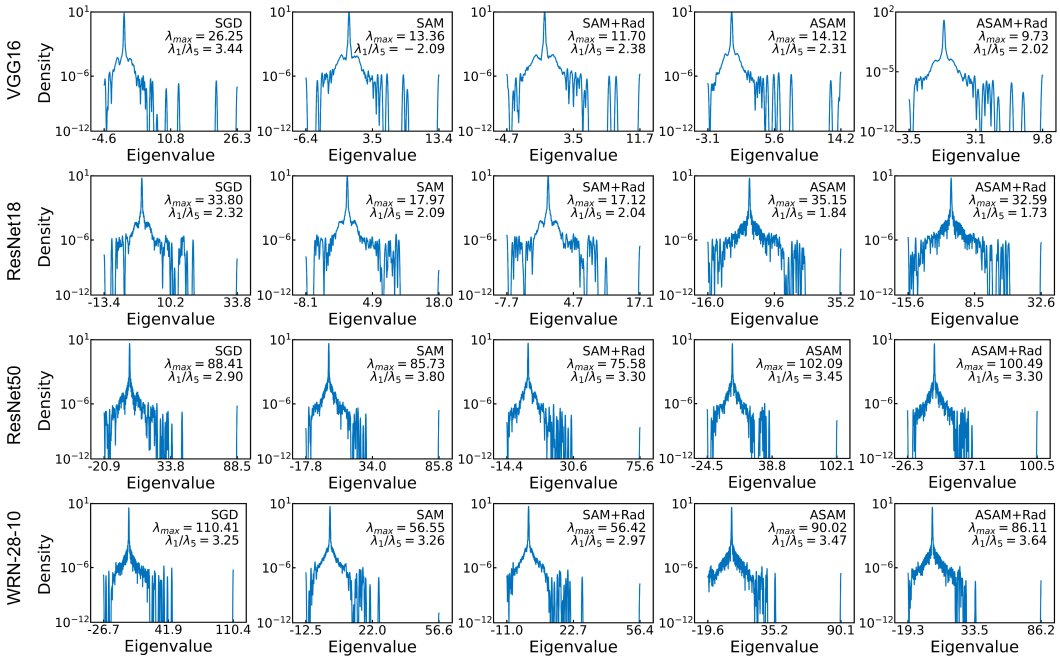


Figure 5: Hessian eigenvalue distributions of different optimization methods on CIFAR-10 across four network architectures.

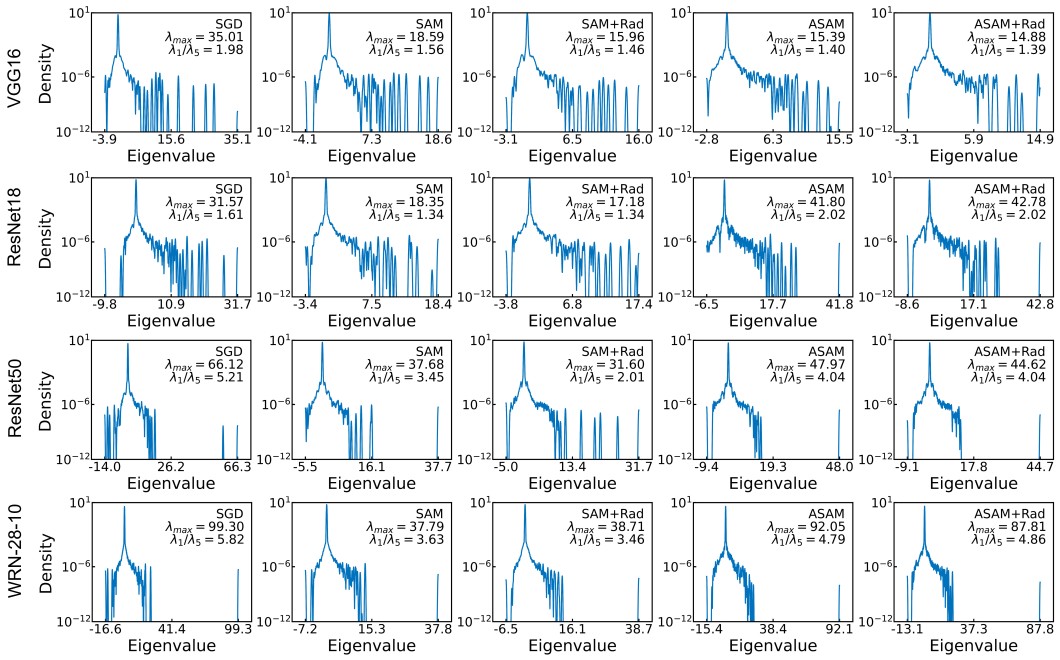


Figure 6: Hessian eigenvalue distributions of different optimization methods on CIFAR-100 across four network architectures.

eigenvalues. The computation uses 2,000 training samples with a batch size of 128. For the spectral density, we use 100 Lanczos iterations and one random probe vector, followed by Gaussian smoothing with a bandwidth of 10^{-4}. The spectral density reflects the overall distribution of local curvature, while $\lambda_{\max}$ characterizes the strongest curvature direction. A smaller $\lambda_{\max}$ indicates a flatter loss landscape along this direction. The ratio λ_1/λ_5 measures the relative spread of the leading Hessian

eigenvalues. Compared with SAM, SAM+Rad reduces $\lambda_{\max}$ from 13.36 to 11.70, 17.97 to 17.12, 85.73 to 75.58, and 56.55 to 56.42 on VGG16, ResNet18, ResNet50, and WRN-28-10, respectively. ASAM+Rad also yields smaller $\lambda_{\max}$ than ASAM across all four architectures, with reductions from 14.12 to 9.73, 35.15 to 32.59, 102.09 to 100.49, and 90.02 to 86.11. These results show that the introduction of Rad does not increase the local curvature around the solutions obtained by SAM or ASAM and, in most cases, leads to a smaller maximum Hessian eigenvalue.

Figure 6 presents the corresponding Hessian eigenvalue distributions on CIFAR-100. The four rows correspond to VGG16, ResNet18, ResNet50, and WRN-28-10 from top to bottom, and the five panels in each row correspond to SGD, SAM, SAM+Rad, ASAM, and ASAM+Rad from left to right. We follow the same Hessian computation and visualization settings as those used for CIFAR-10. Compared with SAM, SAM+Rad reduces $\lambda_{\max}$ from 18.59 to 15.96, 18.35 to 17.18, and 37.68 to 31.60 on VGG16, ResNet18, and ResNet50, respectively, while a slight increase from 37.79 to 38.71 is observed on WRN-28-10. For ASAM, the introduction of Rad decreases $\lambda_{\max}$ from 15.39 to 14.88, 47.97 to 44.62, and 92.05 to 87.81 on VGG16, ResNet50, and WRN-28-10, respectively, with a small increase from 41.80 to 42.78 on ResNet18. Overall, Rad preserves or reduces the maximum local curvature in most settings on CIFAR-100, which shows that introducing Rad does not generally lead to a sharper local loss landscape.

H Sample-Level Margin Change Visualization

Figure 7 presents the sample-level margin changes after introducing Rad into SAM and ASAM on CIFAR-10 and CIFAR-100. We conduct this analysis on VGG16, ResNet18, ResNet50, and WRN-28-10, which are presented from the top to the bottom rows, respectively. For each architecture, the first two columns show the margin changes from SAM to SAM+Rad and from ASAM to ASAM+Rad on CIFAR-10, while the last two columns show the corresponding results on CIFAR-100. Each panel visualizes the distribution of $\Delta m = m_{\mathrm{Rad}} - m_{\mathrm{base}}$ rather than the margins of an individual method. Specifically, for each baseline–Rad pair, we first rank the test samples according to the margins produced by the baseline model and divide them into the hardest 20% and the remaining 80%. We then compute $\Delta m = m_{\mathrm{Rad}} - m_{\mathrm{base}}$ for each sample. The violin plots show the distributions of Δm, while the overlaid box plots summarize their central tendency and dispersion. The percentage of samples with $\Delta m > 0$ and the median value of Δm are also reported in each panel.

For the hardest 20% of samples, Rad yields positive median margin changes across almost all settings, and more than half of the samples obtain larger margins in most cases. This trend is particularly clear on CIFAR-100, where the median improvements are generally larger than those on CIFAR-10. These results indicate that Rad places greater emphasis on samples with relatively small baseline margins. For the remaining 80% of samples, the median margin changes stay close to zero and are slightly negative in many settings. The proportion of samples with increased margins is also generally lower than that of the hardest group. This suggests that the margin improvement introduced by Rad is mainly concentrated on difficult samples rather than uniformly distributed across all samples.

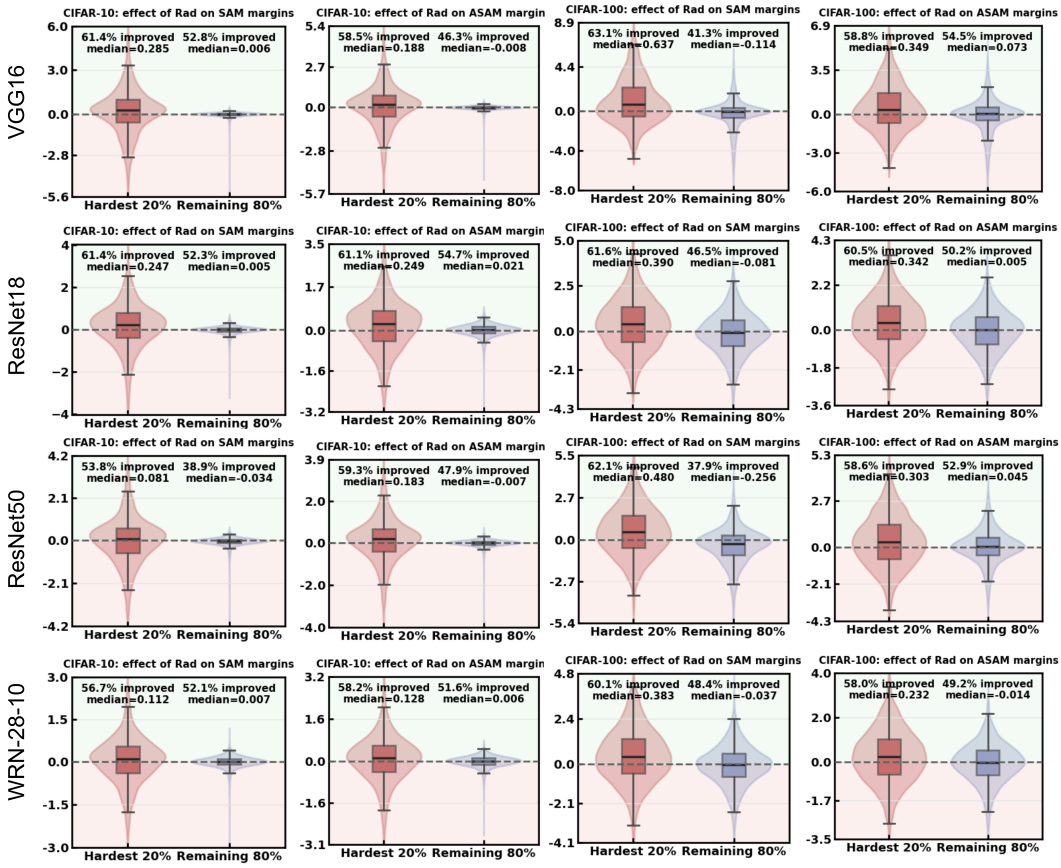


Figure 7: Visualization of sample-level classification margin changes after introducing Rad into SAM and ASAM on CIFAR-10 and CIFAR-100.