Sharpness-diversity tradeoff: improving flat ensembles with SharpBalance

Haiquan Lu[1]*, **Xiaotian Liu**[2]*, **Yefan Zhou**[2]*, **Qunli Li**[3]*,
Kurt Keutzer[4], **Michael W. Mahoney**[4,5,6], **Yujun Yan**[2], **Huanrui Yang**[4], **Yaoqing Yang**[2]
[1] Nankai University
[2] Dartmouth College
[3] University of California San Diego
[4] University of California at Berkeley
[5] International Computer Science Institute
[6] Lawrence Berkeley National Laboratory

Abstract

Recent studies on deep ensembles have identified the sharpness of the local minima of individual learners and the diversity of the ensemble members as key factors in improving test-time performance. Building on this, our study investigates the interplay between sharpness and diversity within deep ensembles, illustrating their crucial role in robust generalization to both in-distribution (ID) and out-of-distribution (OOD) data. We discover a trade-off between sharpness and diversity: minimizing the sharpness in the loss landscape tends to diminish the diversity of individual members within the ensemble, adversely affecting the ensemble's improvement. The trade-off is justified through our theoretical analysis and verified empirically through extensive experiments. To address the issue of reduced diversity, we introduce `SharpBalance`, a novel training approach that balances sharpness and diversity within ensembles. Theoretically, we show that our training strategy achieves a better sharpness-diversity trade-off. Empirically, we conducted comprehensive evaluations in various data sets (CIFAR-10, CIFAR-100, TinyImageNet) and showed that `SharpBalance` not only effectively improves the sharpness-diversity trade-off, but also significantly improves ensemble performance in ID and OOD scenarios. Our code has been made open-source.[†]

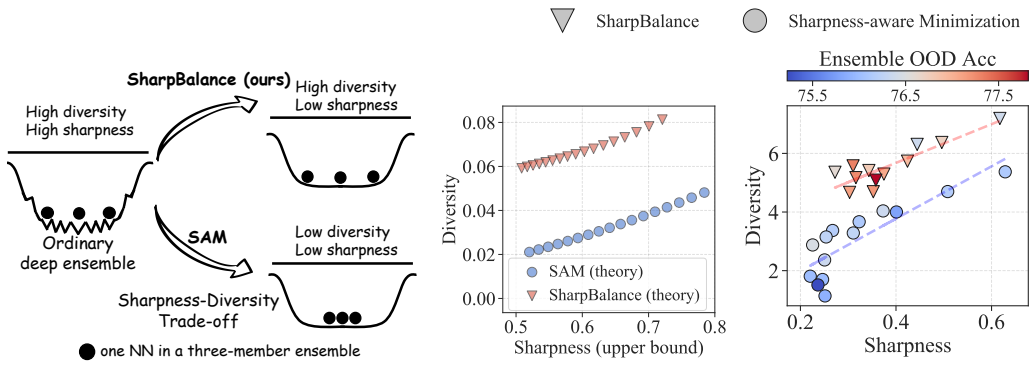
1 Introduction

There has been interest in understanding the properties of neural networks (NNs) and their implications for robust generalization to both in-distribution (ID) and out-of-distribution (OOD) data [Hendrycks and Dietterich, 2019a]. Two properties of particular importance, sharpness (or flatness) [Granziol, 2020, Andriushchenko et al., 2023, Yang et al., 2021, Dinh et al., 2017, Yao et al., 2020] and diversity [Laviolette et al., 2017, Fort et al., 2019, Yao et al., 2020, Dietterich, 2000, Ortega et al., 2022, Theisen et al., 2023], have been shown to have a significant influence on performance. In the context of *deep ensembles* [Ovadia et al., 2019, Lakshminarayanan et al., 2017, Fort et al., 2019, Mehrtash et al., 2020, Ganaie et al., 2022], diversity (which measures the variance in output between independently-trained models) is shown to be critical in enhancing ensemble accuracy. Sharpness, on the other hand, quantifies the curvature of local minima and is believed to be empirically correlated with an individual model's generalization ability.

*First four authors contributed equally.
[†]https://github.com/haiquanlu/SharpBalance

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-4344



(a) Overview (b) Theoretical results (c) Empirical results

Figure 1: **(Sharpness-diversity trade-off and `SharpBalance`). (a)** Caricature illustrating the sharpness-diversity trade-off that emerges in an ensemble's loss landscape induced by the Sharpness-aware Minimization (`SAM`) optimizer. We propose `SharpBalance` to address this trade-off. Each black circle represents an individual NN in a three-member ensemble. The distance between circles represents the diversity between NNs and the ruggedness of the basin represents the sharpness of each NN. **(b)** Theoretically proving the existence of the sharpness-diversity trade-off and improvement from `SharpBalance`, plotting the analytic representation of sharpness and diversity from Theorem 1 and Theorem 2 by changing the perturbation radius ρ of `SAM`. `SharpBalance` achieves a larger diversity for the same level of sharpness. **(c)** Empirical results of verifying sharpness-diversity trade-off improvement from `SharpBalance`. Each marker represents a three-member ResNet18 ensemble trained on CIFAR-10. Diversity is measured by the variance of individual models' predictions, and sharpness is measured by the adaptive worst-case sharpness, both defined in Section 2.

Recent research on loss landscapes [Yang et al., 2021] highlights that a single structural property of the loss landscape is insufficient to fully capture a model's generalizability, and it underscores the importance of a joint analysis of sharpness and diversity. Despite significant efforts in studying sharpness and diversity individually, a gap persists in understanding their relationship, particularly in the context of ensemble learning. Our work seeks to bridge this gap by investigating ensemble learning through the lens of loss landscapes, with a specific focus on the interplay between sharpness and diversity.

Sharpness-diversity trade-off. Our examination of loss landscape structure for ensembling revealed a "trade-off" between the diversity of individual NNs and the sharpness of the local minima to which they converge. This trade-off introduces a potential limitation to the achievable performance of the deep ensemble: the test accuracy of individual NN can be improved as the sharpness is reduced, but it simultaneously reduces diversity, thereby compromising the ensembling improvement (evidence in Section 4.2 and 4.4). This trade-off is visually summarized in the lower transition branch in Figure 1a. We also developed theories (in Section 3) to verify the trade-off. The theoretical results characterizing this phenomenon are visualized in Figure 1b, and the experimental observation is presented in Figure 1c. In Section 4.2, we also verified the existence of the trade-off by varying the experimental setting to include different datasets and different levels of overparameterization (e.g., changing model width).

`SharpBalance` **mitigates the trade-off and improves ensembling performance.** To address the challenge presented by the sharpness-diversity tradeoff, we propose a novel ensemble training method called `SharpBalance`. This method aims to simultaneously reduce the sharpness of individual NNs and prevent diversity reduction among them, as demonstrated in the upper transition branch of Figure 1a. This method is designed based on our theoretical results, which suggest that training different ensemble members using a loss function that aims to reduce sharpness on different subsets of the training data can improve the trade-off between sharpness and diversity. Our theoretical results are summarized in Figure 1b. Aligned with theoretical insights, our `SharpBalance` method lets each ensemble member minimize the sharpness objective exclusively on a subset of training data, termed the *sharpness-aware set*. The sharpness-aware set of each ensemble member is diversified by an adaptive strategy based on data-dependent sharpness measures. As shown in Figure 1c, we verify

that SharpBalance improves the sharpness-diversity tradeoff in training the ResNet18 ensemble on CIFAR10. We conducted experiments on three classification datasets to show that SharpBalance boosts ensembling performance in ID and OOD data.

Our contributions are summarized as follows:

- **Comprehensive identification of the sharpness-diversity trade-off:** This work provides a thorough examination of the phenomenon sharpness-diversity trade-off where reducing the sharpness of individual models can decrease diversity between models within an ensemble. We demonstrate this effect through extensive experiments across various settings,using different sharpness and diversity measures, as well as different model capacities. Our findings show that this trade-off can negatively affect the ensemble improvements.
- **Novel theory:** We prove the existence of the trade-off under a novel theoretical framework based on rigorous analysis of sharpness-aware training objectives [Foret et al., 2021, Behdin and Mazumder, 2023]. Our analysis borrows tools from analyzing Wishart moments [Bishop et al., 2018], and characterizes the exact dynamics of training, bias-variance tradeoff, and the upper and lower bounds of sharpness. Notably, our novel theoretical analysis generalizes existing analysis to ensemble members trained with different data, which is the key to analyzing our own training method SharpBalance.
- **Effective approach:** To mitigate the sharpness-diversity trade-off, we introduce SharpBalance, an ensemble training approach. Our theoretical framework demonstrates that SharpBalance provably achieves improvements on the sharpness-diversity trade-off by reducing sharpness while mitigating the decrease in diversity. Empirically, we confirm this improvement and demonstrate that SharpBalance enhances overall ensemble performance, outperforming baseline methods in CIFAR-10, CIFAR-100 [Krizhevsky, 2009], TinyImageNet [Le and Yang, 2015], and their corrupted versions to assess OOD performance.

We provide a more detailed discussion on related work in Appendix B.

2 Background

Preliminaries. We use a NN denoted as $f_{\boldsymbol{\theta}} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$, where $\boldsymbol{\theta} \in \mathbb{R}^p$ denotes the trainable parameters. The training dataset comprises n data-label pairs $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$. The training loss of NN $f_{\boldsymbol{\theta}}$ over a dataset $\mathcal{D}$ can be defined as $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right), \boldsymbol{y}_i\right)$. Here $\ell(\cdot)$ is a loss function, which, for instance, can be the cross entropy loss or ℓ_2 loss. We construct a deep ensemble consisting of m distinct NNs $f_{\boldsymbol{\theta}_1}, \ldots, f_{\boldsymbol{\theta}_m}$. For classification tasks, the ensemble's output is derived by averaging the predicted logits of these individual networks. We use *flat ensemble* to mean the deep ensemble in which each ensemble member is trained using a sharpness-aware optimization method [Foret et al., 2021], differentiating it from other ensemble approaches.

Diversity metrics. Distinct measures of diversity have been proposed in the literature [Laviolette et al., 2017, Fort et al., 2019, Dietterich, 2000, Baek et al., 2022, Ortega et al., 2022, Theisen et al., 2023], and they are primarily calculated using the predictions made by individual models. Ortega et al. [2022] define diversity $\mathbb{D}(\boldsymbol{\theta})$ to be the variance of model outputs averaged over the data-generating distribution, which we adopt in the theoretical analysis:

$$\mathbb{D}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}[\text{Var}(f_{\boldsymbol{\theta}}(\mathcal{D}))]. \tag{1}$$

In our experiments, diversity is measured using variance defined above, as well as two other widely used metrics in ensemble learning, namely Disagreement Error Ratio (DER) [Theisen et al., 2023] defined in equation (2), and KL divergence [Kullback and Leibler, 1951] defined in equation (11) in the appendices. We show in Section 4.2 that our main claim generalizes to these three metrics in characterizing the diversity between members within an ensemble. Specifically, denote $\mathcal{P}$ as the distribution of model weights $\boldsymbol{\theta}$ after training. Then, the DER is defined as

$$\text{DER} = \frac{E_{\boldsymbol{\theta},\boldsymbol{\theta}' \sim \mathcal{P}}[\text{Dis}(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})]}{E_{\boldsymbol{\theta} \sim \mathcal{P}}[\mathcal{E}(f_{\boldsymbol{\theta}})]}, \tag{2}$$

where $\text{Dis}(f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'})$ is the prediction disagreement [Masegosa, 2020, Mukhoti et al., 2021, Jiang et al., 2022] between two classifier $f_{\boldsymbol{\theta}}, f_{\boldsymbol{\theta}'}$, and $\mathcal{E}(f_{\boldsymbol{\theta}})$ is the prediction error.

 https://doi.org/10.52202/079017-4344

Sharpness Metric. In accordance with the definition proposed by Foret et al. [2021], we characterize the *first-order sharpness* of a model as the worst-case perturbation within a radius of ρ_0. Mathematically, the sharpness κ of a model $\boldsymbol{\theta}$ is expressed as follows:

$$\kappa(\boldsymbol{\theta}; \rho_0) = \max_{\|\boldsymbol{\varepsilon}\|_2 \leq \rho_0} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}).$$

Empirically, we measure the sharpness of the NN via the adaptive worst-case sharpness [Kwon et al., 2021, Andriushchenko et al., 2023]. The adaptive worst-case sharpness captures how much the loss can increase within the perturbation radius ρ_0 of $\boldsymbol{\theta}$:

$$\max_{\|T_{\boldsymbol{\theta}}^{-1}\boldsymbol{\varepsilon}\|_2 \leq \rho_0} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_l]$, and $T_{\boldsymbol{\theta}} = \operatorname{diag}(|\theta_1|, \ldots, |\theta_l|)$. $T_{\boldsymbol{\theta}}^{-1}$ is a normalization operator to make sharpness "scale-free", that is, such that scaling operations on $\boldsymbol{\theta}$ that do not alter NN predictions will not impact the sharpness measure.

Ensembling. We characterize the effectiveness of ensembling by the metric called ensemble improvement rate (EIR) [Theisen et al., 2023], which is defined as the ensembling improvement over the average performance of single models. Let $\mathcal{E}_{\text{ens}}$ denote the test error of an ensemble; the EIR is then defined as follows:

$$\text{EIR} = \frac{E_{\boldsymbol{\theta}\sim\mathcal{P}}[\mathcal{E}(f_{\boldsymbol{\theta}})] - \mathcal{E}_{\text{ens}}}{E_{\boldsymbol{\theta}\sim\mathcal{P}}[\mathcal{E}(f_{\boldsymbol{\theta}})]}. \tag{4}$$

Sharpness Aware Minimization (`SAM`). `SAM` [Foret et al., 2021] has been shown to be an effective method for improving the generalization of NNs by reducing the sharpness of local minima. It essentially functions by penalizing the maximum loss within a specified radius ρ of the current parameter θ. The training objective of `SAM` is to minimize the following loss function:

$$\mathcal{L}_{\mathcal{D}}^{\text{SAM}}(\boldsymbol{\theta}) := \max_{\|\boldsymbol{\varepsilon}\|_2 \leq \rho} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) + \lambda\|\boldsymbol{\theta}\|_2^2, \tag{5}$$

where λ is the hyperparameter of a standard ℓ_2 regularization term.

3 Theoretical Analysis of Sharpness-diversity Trade-off

This section theoretically analyzes the sharpness-diversity trade-off. The diversity among individual models is quantified using equation (1). The first theorem establishes the existence of a trade-off between sharpness and diversity. The second theorem demonstrates that training models with only a subset of data samples leads to a more favorable trade-off between these two metrics.

Sharpness and Diversity of `SAM`. Assume the training data matrix $\mathbf{A} \in \mathbb{R}^{n_{\text{tr}} \times d_{\text{in}}}$ and test data matrix $\mathbf{T} \in \mathbb{R}^{n_{\text{te}} \times d_{\text{in}}}$ are random with entries drawn from Gaussian $\mathcal{N}(0, \frac{1}{d_{\text{in}}}\mathbf{I})$. Suppose the model weight at the 0-th time step, $\boldsymbol{\theta}_0$, is initialized randomly such that $\mathbb{E}[\boldsymbol{\theta}_0] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T] = \sigma^2\mathbf{I}$ and updated with a quadratic optimization objective through `SAM`. The learned weight matrix after k time steps is denoted as $\boldsymbol{\theta}_k$. Let $\boldsymbol{\theta}^*$ be the teacher model (i.e., ground-truth model) such that $\mathbf{A}\boldsymbol{\theta}^* = \mathbf{y}^{(\mathbf{A})}$ and $\mathbf{T}\boldsymbol{\theta}^* = \mathbf{y}^{(\mathbf{T})}$, where $\mathbf{y}^{(\mathbf{D})}$ is the label vector for data matrix $\mathbf{D}$. Given a perturbation radius ρ_0, the sharpness of a model after k iteration under the random matrix assumption is defined as

$$\kappa(\boldsymbol{\theta}_k; \rho_0) = \mathbb{E}_{\mathbf{A}}[\max_{\|\boldsymbol{\varepsilon}\|_2 \leq \rho_0} f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k\right] + \boldsymbol{\varepsilon}; \mathbf{A}\right) - f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k\right]; \mathbf{A}\right)],$$

which is the expected fluctuation of the model output after perturbation over the data distribution. For simplicity, we denote $\kappa(\boldsymbol{\theta}_k^{SAM}; \rho_0) = \kappa_k^{SAM}$ for the rest of the paper. We derive an explicit formulation of diversity and upper and lower bounds of sharpness for models optimized with `SAM` in Theorem 1. Detailed proof can be found in Appendix C.1.

Theorem 1 (Diversity and Sharpness of `SAM`). *Let $\boldsymbol{\theta}_0$ be initialized randomly such that $\mathbb{E}[\boldsymbol{\theta}_0] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T] = \sigma^2\mathbf{I}$. Suppose $\boldsymbol{\theta}_k^{SAM}$ is the model weight after k iterations of training with `SAM` on $\mathbf{A} \in \mathbb{R}^{n_{\text{tr}} \times d_{\text{in}}}$ and evaluated on $\mathbf{T} \in \mathbb{R}^{n_{\text{te}} \times d_{\text{in}}}$. Let η be the step size, ρ be the perturbation radius in `SAM` and ρ_0 be the radius for measuring sharpness κ_k^{SAM}. Then*

$$\mathbb{D}(\boldsymbol{\theta}_k^{SAM}) = \phi(2k, 0)\sigma^2,$$

$$\frac{\rho_0^2}{2}\left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} - 1\right)^2 + \rho_0\sqrt{\phi(2k,2)}\|\boldsymbol{\theta}^*\|_2 - G \leq \kappa_k^{SAM} \leq \frac{\rho_0^2}{2}\left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} + 1\right)^2 + \rho_0\sqrt{\phi(2k,2)}\|\boldsymbol{\theta}^*\|_2,$$

where

$$\phi(i,j) := \mathbb{1}_{j=0} + \sum_{k_1+k_2+k_3=i} \frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\left(\frac{n_{\text{tr}}}{d_{\text{in}}}\right)^m \sum_{l=1}^{m}\left(\frac{d_{\text{in}}}{n_{\text{tr}}}\right)^{m-l}\mathcal{O}(1+1/d_{\text{in}})N_{m,l},$$

$G = \frac{\phi(4k,4)-\phi(2k,2)^2}{2\phi(2k,2)^{3/2}\|\boldsymbol{\theta}^*\|_2}$, *and* $m = k_2 + 2k_3 + j$. $N_{m,l} = \frac{1}{l}\binom{m-1}{l-1}\binom{m}{l-1}$ *is the Narayana number.*

To provide a clearer understanding of the relationship between sharpness and diversity, Figure 2 presents a trade-off curve between these two metrics. The estimated sharpness and diversity are displayed on the x and y axes, respectively. Each point in the plot corresponds to a model trained using SAM with a different ρ value, showcasing the outcome of varying perturbation radius. In these experiments, we evaluated the sharpness and diversity of the models empirically and compared them to the estimates obtained using Theorem 1. The soundness of Theorem 1 and tightness of the derived bounds are further supported by empirical evidence, as depicted in Figure 2. Further verification results supporting our theoretical analysis are provided in Appendix C.3

Training with Data Subsets. Assume $\mathbf{A}$ is partitioned into S horizontal submatrices, such that $\mathbf{A} = [\mathbf{A}_1^T, \mathbf{A}_2^T, \ldots, \mathbf{A}_S^T]^T$. We show in Theorem 2 a similar analysis of the sharpness and diversity of ensembles for which each model is trained with a submatrix. Under this setting, we first selected a subset of data $\mathbf{A}_s$ uniformly at random and then train the model with the selected subset with SAM.

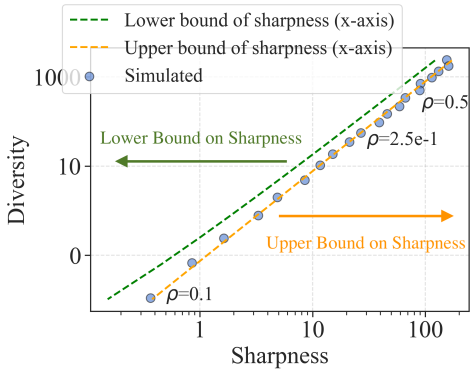


Figure 2: **(Theoretical vs. Simulated sharpness-diversity trade-off)**. This figure illustrates the relationship between sharpness (upper and lower bounds) and diversity as predicted by Thereom 1 and as observed in simulations. Note that the upper and lower bounds correspond to the sharpness values plotted along the x-axis, with the upper bound positioned to the right and the lower bound to the left. Also, note that the bounds provided are for the expected sharpness, which means that random fluctuations can cause the simulation results to move beyond these bounds.

Theorem 2 (Diversity and Sharpness when Models are Trained on Subsets)**.** *Suppose the training data matrix* $\mathbf{A}$ *is partitioned into* S *horizontal submatrices. Let* $\boldsymbol{\theta}_0$ *be initialized randomly such that* $\mathbb{E}[\boldsymbol{\theta}_0] = \mathbf{0}$ *and* $\mathbb{E}[\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T] = \sigma^2\mathbf{I}$. *Let* $\boldsymbol{\theta}_k^{SharpBal}$ *be the model weight trained with* SAM *for* k *iterations on the submatrix* $\mathbf{A}_s \in \mathbb{R}^{\frac{n_{\text{tr}}}{S}\times d_{\text{in}}}$, *selected uniformly at random, and evaluated on test data* $\mathbf{T} \in \mathbb{R}^{n_{\text{te}}\times d_{\text{in}}}$. *Let* η *be the step size,* ρ *be the perturbation radius in* SAM, ρ_0 *be the radius for measuring sharpness* κ_k^{SAM}, *and* $r = \frac{n_{\text{tr}}}{Sd_{\text{in}}}$. *Then*

$$\mathbb{D}(\boldsymbol{\theta}_k^{SharpBal}) = \phi'(2k,0)\sigma^2 + \frac{S-1}{d_{\text{in}}S}\left(\phi'(2k,0) - \phi'(k,0)^2\right)\|\boldsymbol{\theta}^*\|_2^2,$$

and

$$\kappa_k^{SharpBal}(\rho_0) \le \frac{\rho_0^2}{2}\left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}}+1\right)^2 + \frac{\rho_0}{S}\sqrt{C}\|\boldsymbol{\theta}^*\|_2,$$

where

$$\begin{aligned} C =& S\phi'(2k,2) + 2rS(S-1)\phi'(2k,1) + 2S(S-1)\phi'(k,2)\phi'(k,0) \\ &+ r(1+r)S(S-1)\phi'(2k,0) + 2S(S-1)\phi'(k,1)\phi'(k,1) \\ &+ \frac{3}{2}r(1+r)S(S-1)(S-2)\phi'(k,0)^2 + \frac{3}{2}r^2S(S-1)(S-2)\phi'(2k,0) \\ &+ 3rS(S-1)(S-2)\phi'(k,0)\phi'(k,1) + r^2S(S-1)(S-2)(S-3)\phi'(k,0)^2, \end{aligned}$$

$$\phi'(i,j) := \mathbb{1}_{j=0} + \sum_{k_1+k_2+k_3=i} \frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\left(\frac{n_{\text{tr}}}{Sd_{\text{in}}}\right)^m \sum_{l=1}^{m}\left(\frac{Sd_{\text{in}}}{n_{\text{tr}}}\right)^{m-l}\mathcal{O}(1+\frac{1}{d_{\text{in}}})N_{m,l},$$

where $m = k_2 + 2k_3 + j$. $N_{m,l} = \frac{1}{l}\binom{m-1}{l-1}\binom{m}{l-1}$ *is the Narayana number.*

The proof of Theorem 2 is provided in Appendix C.2. Similar experimental validations are conducted to verify Theorem 2, with results also presented in Appendix C. The main insight from Theorem 2 is

https://doi.org/10.52202/079017-4344

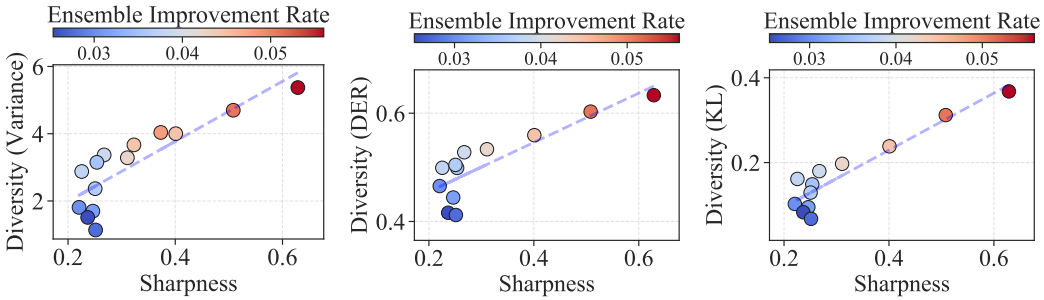
that training models on a randomly selected data subset offers a better trade-off between sharpness and diversity compared to training on the complete dataset. This idea is further illustrated in Figure 1b, where we compare the sharpness upper bound and diversity of models trained on the full dataset (labeled as `SAM`) and those trained on subsets (labeled as `SharpBalance`). The results demonstrate that `SharpBalance` achieves a more favorable trade-off. For a given level of sharpness, deep ensembles with models trained on subsets of the data exhibit higher diversity compared to those trained on the entire dataset. This indicates that minimizing sharpness on randomly sampled data subsets for each model within the ensemble promotes the diversity among the models, thereby enhancing the sharpness-diversity trade-off.

4 Experiments

In this section, we describe our experiments. In particular, following Section 4.1 where we describe our experimental setup, in Section 4.2, we provide an empirical evaluation across various datasets to explore the trade-off between sharpness and diversity. We also examine how this trade-off changes with different levels of overparameterization. Then, in Section 4.3 and 4.4, we elaborate the `SharpBalance` algorithm and compare its performance with baseline methods.

4.1 Experimental setup

Here, we describe the experiment setup for Section 4.2. Each ensemble member is trained individually using `SAM` with a consistent perturbation radius ρ, as defined in equation (5). We adjust ρ across different ensembles to achieve varying levels of minimized sharpness. Sharpness for each NN was measured using the adaptive worst-case sharpness metric, defined in equation (3). The sharpness measurement was done on the training set, using 100 batches of size 5. The diversity between NNs is measured using DER defined in equation (2). The diversity between ensemble members is tested on OOD data. We evaluated this trade-off using a variety of image classification datasets, including CIFAR-10, CIFAR-100 [Krizhevsky, 2009], TinyImageNet [Le and Yang, 2015], and their corrupted versions [Hendrycks and Dietterich, 2019b]. For the setup of Section 4.4, we used the same datasets and architecture. The hyperparameters of the baseline methods has been carefully tuned. The hyperparameters for conducting the experiments are detailed in Appendix D.



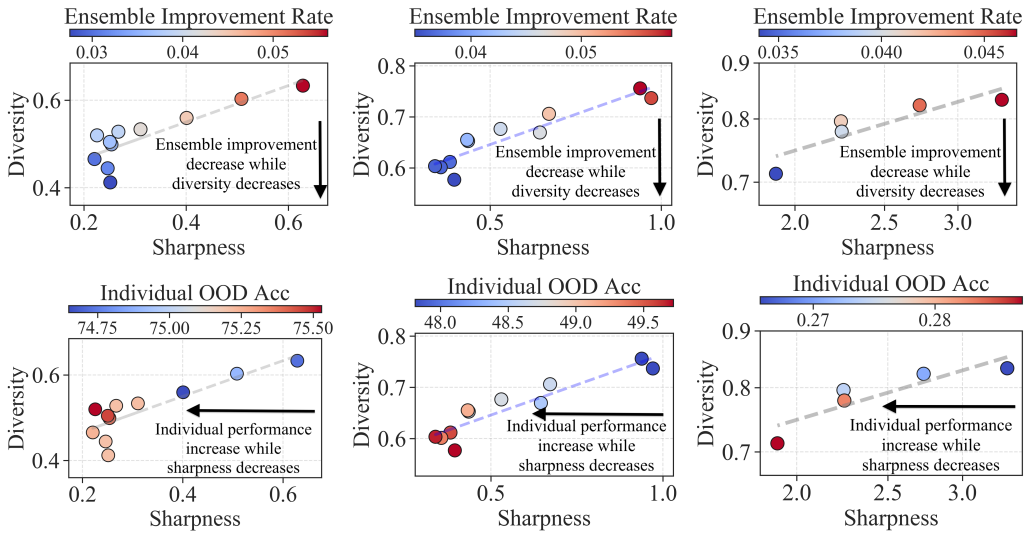
(a) Measuring diversity via Variance (b) Measuring diversity via DER (c) Measuring diversity via KL

Figure 3: **(Varying diversity measure in empirical study).** Three different metrics are employed to measure the diversity of individual models within an ensemble, i.e., Variance in equation (1), DER in equation (2), and KL divergence in equation (11). The results of the three metrics show consistent trends, demonstrating the sharpness-diversity trade-off: lower sharpness is correlated with lower diversity. The experiment is conducted by training a three-member ResNet18 ensemble on CIFAR10.

4.2 Empirical validation of Sharpness-diversity trade-off

We provide empirical observation to validate and explore the sharpness-diversity trade-off. Figure 3 presents the validation of observing the trade-off phenomenon on training ResNet18 ensembles on CIFAR10 applying three different metrics to measure the diversity. The results demonstrate that this trade-off phenomenon generalizes to the three diversity metrics defined in Section 2. Figure 4 presents the validation on three different datasets. In the following empirical study, DER will be the primary metric for measuring diversity of models.

https://doi.org/10.52202/079017-4344



(a) CIFAR-10 (b) CIFAR-100 (c) TinyImageNet

Figure 4: **(Empirical observations of sharpness-diversity trade-off).** The identified trade-off shows that while reducing sharpness enhances individual model performance, it concurrently lowers diversity and thus diminishes the ensemble improvement rate. *First row*: the color encoding represents the ensemble improvement rate (EIR) defined in equation (4), from red to blue means ensembling improvement decreases. *Second row*: the color encoding represents the individual ensemble member's OOD accuracy, from blue to red means individual performance becomes better. Each marker represents a three-member ResNet18 ensemble trained with `SAM` with a different perturbation radius.

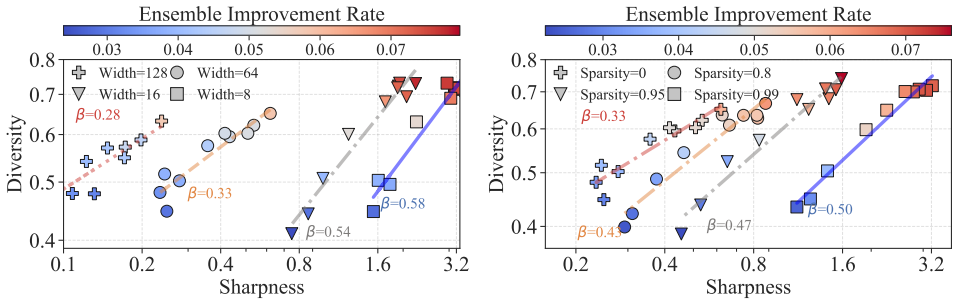
Experimental results obtained with the other two metrics are available in Appendix E. The three sets of results first verify that minimizing individual member's sharpness indeed reduces diversity. This is confirmed by the consistent trends of markers moving from upper right to lower left. Second, the first row of Figure 4 shows that an ensemble with decreased diversity (lower in y-axis) shows a lower ensemble improvement rate (from red to blue), highlighting the negative impact of this trade-off. Lastly, the second row shows when the sharpness of the individual model is reduced (lower in x-axis), the individual model's OOD accuracy is improved (from blue to red), demonstrating the benefits of minimizing sharpness. We verify the robustness of the phenomenon by measuring the sharpness and diversity using different metrics in Appendix E.

Figure 5 illustrates the trade-off curves as the overparameterization level of the model is adjusted by changing width or sparsity (introduced using model pruning). This visualization confirms that the trade-off is a consistent phenomenon across models of different sizes, and the ensemble provides less improvement (blue color) at the lower left end of each trade-off curve. It also highlights that models with smaller or sparser configurations show a more significant trade-off effect, as evidenced by the steeper slopes and higher coefficient values of the linear fitting curves. As sparse ensembles are now being used to demonstrate the benefits of ensembling for efficient models [Liu et al., 2022, Diffenderfer et al., 2021, Whitaker and Whitley, 2022, Kobayashi et al., 2022, Zimmer et al., 2024], addressing the conflict between sharpness and diversity becomes particularly crucial.

4.3 Our `SharpBalance` method

Here, we describe the design and implementation of our main method, `SharpBalance`. Figure 6 provides an overview. Our approach is motivated by the theoretical analysis in Section 3, which suggests that having each ensemble member minimize sharpness on diverse subsets of the data can lead to a better trade-off between sharpness and diversity. `SharpBalance` aims to achieve the optimal balance by applying `SAM` to a carefully selected subset of the data, while performing standard optimization on the remaining samples. More specifically, for each ensemble member NN f_{θ_i}, our method divides the entire training dataset $\mathcal{D}$ into two distinct subsets: sharpness-aware set $\mathcal{D}^i_{\text{SAM}}$ and normal set $\mathcal{D}^i_{\text{Normal}}$. The model is trained to optimize the sharpness reduction objective on $\mathcal{D}^i_{\text{SAM}}$,

https://doi.org/10.52202/079017-4344



(a) Varying model width

(b) Varying model sparsity

Figure 5: **(Sharpness-diversity trade-off in models varying overparameterization levels).** Different types of markers represent models with varying degrees of overparameterization, determined by changing the model width (a) or sparsity (b). Each marker represents a three-member ensemble trained with `SAM` with a different perturbation radius. The β reflects the rate of decline in the trade-off curve, calculated via applying linear fitting over the ensembles at each level of overparameterization. A higher β points to a steeper decline in the trade-off. Ensembles with narrower widths or increased sparsity display more pronounced trade-off effects. The model used in ResNet18 and the dataset is CIFAR-10.

while it optimizes the normal training objective on $\mathcal{D}^i_{\text{Normal}}$. These training objectives are denoted as $\mathcal{L}^{\text{SAM}}_{\mathcal{D}^i_{\text{SAM}}}(\boldsymbol{\theta}_i)$ and $\mathcal{L}_{\mathcal{D}^i_{\text{Normal}}}(\boldsymbol{\theta}_i)$, respectively. The $\mathcal{D}^i_{\text{SAM}}$ is selected by an adaptive strategy from the whole dataset $\mathcal{D}$: it is composed of the union of samples that are deemed "sharp" by all other members of the ensemble except the i-th. Specifically, for each model, we pick the subset of data samples with the top-k% highest "per-data-sample sharpness." Then, we take the union of all such subsets expect the i-th for creating the subset $\mathcal{D}^i_{\text{SAM}}$. This partition of data samples can be efficiently computed in parallel as there is no sequential dependency on the training of the ensemble members. However, `SharpBalance` can be easily adapted for sequential training if memory constraints permit training only one model at a time.

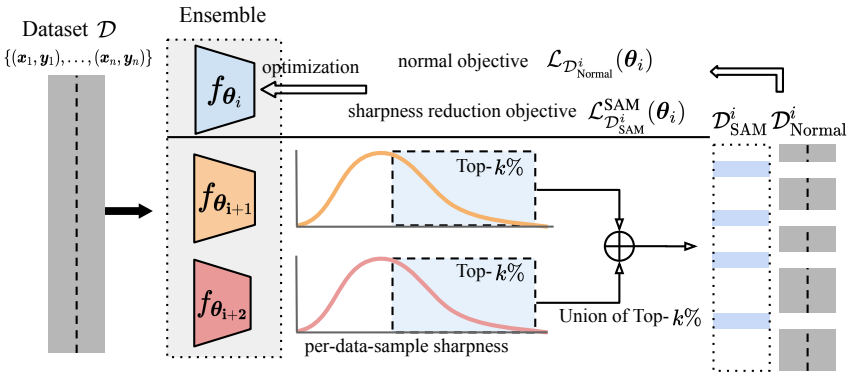


Figure 6: **(System diagram of `SharpBalance`).** Each ensemble member $f_{\boldsymbol{\theta}_i}$ optimizes the sharpness reduction objective on subset $\mathcal{D}^i_{\text{SAM}}$ and the normal training objective on $\mathcal{D}^i_{\text{Normal}}$. $\mathcal{D}^i_{\text{SAM}}$ is formed by selecting data samples from $\mathcal{D}$ that significantly affect the loss landscape sharpness of other ensemble members.

Per-data-sample sharpness. This metric is designed to efficiently assess the sharpness of a model for individual data samples. For each data point $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, sharpness is quantified using the Fisher Information Matrix (FIM), which is expressed as $\nabla_{\boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \boldsymbol{y}_j) \nabla_{\boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \boldsymbol{y}_j)^T$. Following a well-established approach [Bottou et al., 2018], we approximate the trace of the FIM by computing the squared ℓ_2 norm of the gradient: $\|\nabla_{\boldsymbol{\theta}} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_j), \boldsymbol{y}_j)\|_2^2$. Other common sharpness metrics, such as worst-case sharpness, trace of the Hessian, or Hessian eigenvalues, are computationally slightly more expensive to approximate [Yao et al., 2020, 2021], but are expected to lead to similar results.

4.4 Empirical evaluation of `SharpBalance`

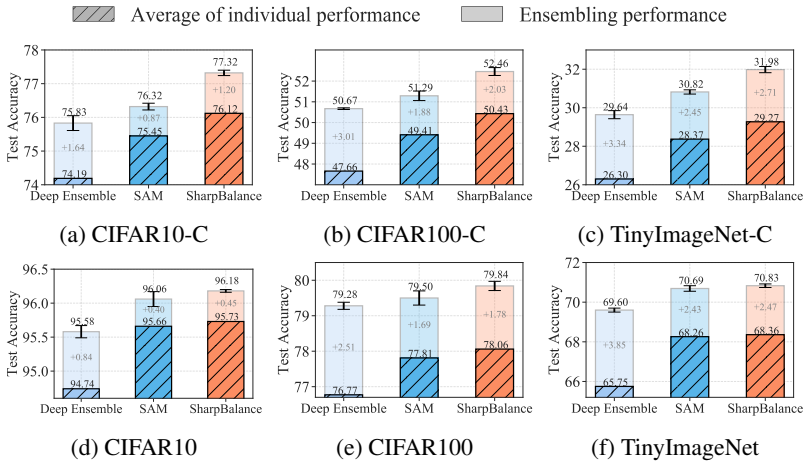


Figure 7: **(Main results: `SharpBalance` improves the overall ensembling performance and mitigates the reduced ensembling improvement caused by sharpness-diversity trade-off).** The three-member ResNet18 ensemble is trained with different methods on three datasets. The first row reports the OOD accuracy and the second row reports the ID accuracy. The lower part of each bar with the diagonal lines represents the individual model performance. The upper part of each bar represents the ensembling improvement. The results are reported by averaging three ensembles, and each ensemble is comprised of three models.

We evaluate `SharpBalance` by benchmarking it against both a standard Deep Ensemble, trained using SGD, and a Deep Ensemble enhanced with `SAM`. The results are presented in Figure 7 for CIFAR-10, CIFAR-100, and TinyImageNet. The comparison between the middle and left bars shows that `SAM` enhances individual model performance by reducing sharpness. However, this reduction in sharpness also diminishes the overall ensemble effectiveness by lowering diversity, exemplifying the sharpness-diversity trade-off discussed in Section 4.2. Further comparison between the right and middle bars shows that `SharpBalance` maintains or improves individual performance while improving ensemble effectiveness.

We also evaluate `SharpBalance` on different ensemble sizes. As shown in Figure 8, `SharpBalance` demonstrates more pronounced empirical improvements as the number of ensemble models increases. The accuracy difference between `SharpBalance` and the baseline methods becomes more significant, especially on corrupted data. Specifically, `SharpBalance` outperforms the baselines by up to 1.30% when ensembling 5 models on CIFAR100-C dataset.

To further evaluate `SharpBalance`, we provide corroborating results in Appendix F, which includes:

- We evaluate `SharpBalance` on different severity of the corruption on CIFAR10-C, CIFAR100-C and Tiny-ImageNet-C. `SharpBalance` increasingly outperforms the baselines as the severity of the corruption increases. We also evaluate the proposed method using uncertainty metrics such as negative log-likelihood and expected calibration error.
- We further evaluate `SharpBalance` on other model architectures and tasks, such as WideResNet, ViT, and ALBERT [Lan et al., 2020] on language tasks.
- We compare our method of measuring sharpness with another method of measuring the curvature of the loss around a data point [Garg and Roy, 2023] and show the strong correlation between these two methods.
- We further compare `SharpBalance` with ensemble baseline EoA [Arpit et al., 2022], an improved version of `SAM` (for which individual models in an ensemble are trained with different ρ values) and GSAM [Zhuang et al., 2022]. Results show that `SharpBalance` can significantly outperform the baselines.

https://doi.org/10.52202/079017-4344

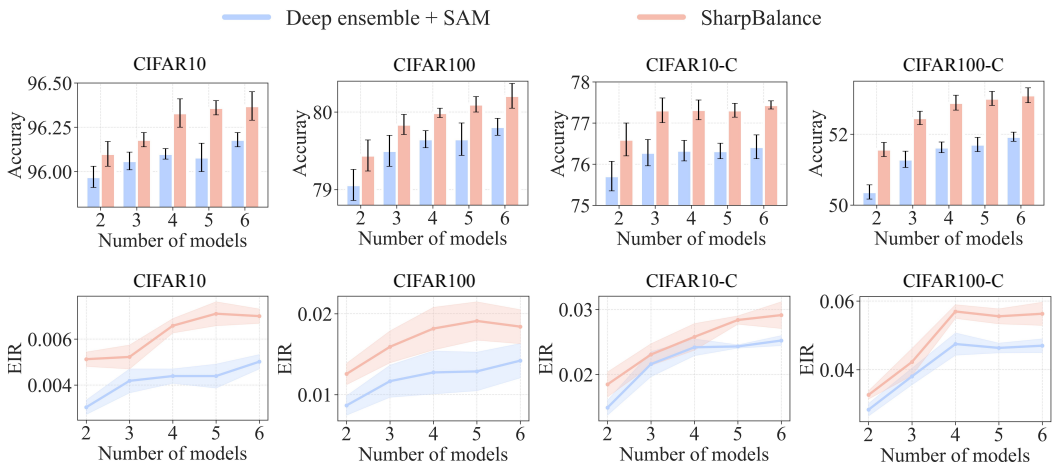


Figure 8: `SharpBalance` **achieves more pronounced improvement when increasing the number of ensembling models.** "EIR" represents the ensemble inprovement rate, which is defined in Section 2, the larger the better. x-axis represents the number of individual models in one ensemble.

- We demonstrate that, compared to training a deep ensemble with `SAM`, our method adds only minimal computational cost. The extra time complexity is dominated by the computation of Fisher trace for evaluating per-sample sharpness, which empirically increases the training time by 1%.

5 Conclusion

Our theoretical and empirical analyses demonstrate the existence of a sharpness-diversity trade-off when sharpness-minimization training methods are applied to deep ensembles. This leads to two main insights that are relevant for improving model performance. First, reducing the sharpness in individual models proves to be beneficial in enhancing the performance of the ensemble as a whole. Second, the accompanying reduction in diversity suggests that popular ensembling methods have limitations, and also highlights the potential for more sophisticated designs that promote diversity among models with lower sharpness. These results are particularly timely, given recent theoretical work on characterizing ensemble improvement [Theisen et al., 2023]. In response to these findings, we have proposed `SharpBalance`, which "diagnoses" the training data by evaluating the sharpness of each sample and then fine-tunes the training of individual models to focus on a diverse subset of the sharpest training data samples. This targeted approach helps maintain diversity among models while also reducing their individual sharpness. Extensive evaluations indicate that `SharpBalance` not only improves the sharpness-diversity trade-off but also delivers superior OOD performance for both dense and sparse models across various datasets and architectures when compared to other ensembling approaches.

Limitations. One limitation of the study is that our theoretical analysis in Section 3 relies on the assumption that the data matrices $\mathbf{A}, \mathbf{T}$ follow a Gaussian distribution and assumed the optimization objective to be quadratic, which may not always hold in practice. Despite the potentially strong assumptions, our empirical findings in Section 4 show that the conclusions remain robust in real-world datasets with various model architectures. This suggests the insights discovered in our study are applicable to a wider range of real-world scenarios, beyond just those strictly adhering to the Gaussian assumption. Nevertheless, future research could explore how such assumptions can be relaxed and extend the theoretical analysis to a weaker condition.

Acknowledgements. Michael W. Mahoney would like to acknowledge the UC Berkeley CLTC, ARO, IARPA (contract W911NF20C0035), NSF, and ONR for providing partial support of this work. Kurt Keutzer would like to acknowledge support from Berkeley Deep Drive. Yaoqing Yang would like to acknowledge support from DOE under Award Number DE-SC0025584, DARPA under Agreement

number HR00112490441, and Dartmouth College. Our conclusions do not necessarily reflect the position or the policy of our sponsors, and no official endorsement should be inferred.

 https://doi.org/10.52202/079017-4344

References

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

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 2022.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

Kayhan Behdin and Rahul Mazumder. On statistical properties of sharpness-aware minimization: Provable guarantees. *arXiv preprint arXiv:2302.11836*, 2023.

Adrian N Bishop, Pierre Del Moral, Angèle Niclas, et al. An introduction to wishart matrix moments. *Foundations and Trends® in Machine Learning*, 11(2):97–218, 2018.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Anh Bui, Vy Vo, Tung Pham, Dinh Phung, and Trung Le. Diversity-aware agnostic ensemble of sharpness minimizers. *arXiv preprint arXiv:2403.13204*, 2024.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

James Diffenderfer, Brian Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. *Advances in neural information processing systems*, 34:664–676, 2021.

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

Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2024.

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

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yoav Freund. Boosting a weak learning algorithm by majority. *Information and computation*, 121(2): 256–285, 1995.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

Xiang Gao, Meera Sitharam, and Adrian E. Roitberg. Bounds on the jensen gap, and implications for mean-concentrated distributions. *The Australian Journal of Mathematical Analysis and Applications*, 2019.

Isha Garg and Kaushik Roy. Samples with low loss curvature improve data efficiency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20290–20300, 2023.

Diego Granziol. Flatness is a false friend. *arXiv preprint arXiv:2006.09091*, 2020.

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*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019a.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019b.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, 1993.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, 1997.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), 2018.

Weisen Jiang, Hansi Yang, Yu Zhang, and James Kwok. An adaptive policy to employ sharpness-aware minimization. In *International Conference on Learning Representations*, 2023.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

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*, 2016.

Sosuke Kobayashi, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Diverse lottery tickets boost ensemble from a single pretrained model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694.

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*, pages 5905–5914. PMLR, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

 https://doi.org/10.52202/079017-4344

François Laviolette, Emilie Morvant, Liva Ralaivola, and Jean-Francis Roy. Risk upper bounds for general ensemble methods with an application to multiclass classification. *Neurocomputing*, 219: 15–25, 2017.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Yoonho Lee, Huaxiu Yao, and Chelsea Finn. Diversify and disambiguate: Learning from underspecified data. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022.

Shiwei Liu, Tianlong Chen, Zahra Atashgahi, Xiaohan Chen, Ghada Sokar, Elena Mocanu, Mykola Pechenizkiy, Zhangyang Wang, and Decebal Constantin Mocanu. Deep ensembling with no overhead for either training or testing: The all-round blessings of dynamic sparsity. In *International Conference on Learning Representations*, 2022.

Andres Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems*, 33:5479–5491, 2020.

Alireza Mehrtash, Purang Abolmaesumi, Polina Golland, Tina Kapur, Demian Wassermann, and William Wells. Pep: Parameter ensembling by perturbation. *Advances in neural information processing systems*, 33:8895–8906, 2020.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

Luis A Ortega, Rafael Cabañas, and Andres Masegosa. Diversity and generalization in neural network ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 11720–11743. PMLR, 2022.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019.

Jack Parker-Holder, Luke Metz, Cinjon Resnick, Hengyuan Hu, Adam Lerer, Alistair Letcher, Alexander Peysakhovich, Aldo Pacchiano, and Jakob Foerster. Ridge rider: Finding diverse solutions by following eigenvectors of the hessian. *Advances in Neural Information Processing Systems*, 33:753–765, 2020.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:10821–10836, 2022.

Ryan Theisen, Hyunsuk Kim, Yaoqing Yang, Liam Hodgkinson, and Michael W Mahoney. When are ensembles really effective? *Advances in neural information processing systems*, 2023.

Giorgio Valentini and Thomas G Dietterich. Low bias bagged support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 752–759, 2003.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Tim Whitaker and Darrell Whitley. Prune and tune ensembles: low-cost ensemble learning with sparse independent subnetworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8638–8646, 2022.

Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733, 2021.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

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 (Big data)*, pages 581–590. IEEE, 2020.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10665–10673, 2021.

Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha C Dvornek, James s Duncan, Ting Liu, et al. Surrogate gap minimization improves sharpness-aware training. In *International Conference on Learning Representations*, 2022.

Max Zimmer, Christoph Spiegel, and Sebastian Pokutta. Sparse model soups: A recipe for improved pruning via model averaging. In *International Conference on Learning Representations*, 2024.

 https://doi.org/10.52202/079017-4344

Appendix

A Impact Statement

This paper uncovers a trade-off between sharpness and diversity in deep ensembles and introduces a novel training strategy to achieve an optimal balance between these two crucial metrics. While the proposed method could potentially be misused for malicious purposes, we believe that the study itself does not pose any direct negative societal impact. More importantly, this research advances the field of ensemble learning and contribute to the development of more reliable deep ensemble models. These advancements consequently result in enhanced robustness when dealing with OOD data and enable the quantification of uncertainty, thereby strengthening the reliability and applicability of deep learning systems in real-world scenarios.

B Related work

Ensembling. Diversity is one of the major factors that contribute to the success of the ensembling method. Popular ensemble techniques have been developed for tree-type individual learners, which are known to have a high variance. This is evident such as in [Breiman, 2001, Chen and Guestrin, 2016, Freund, 1995, Freund and Schapire, 1997]. In contrast, more stable algorithms, such as support vector machines (SVM) type learners, are less commonly used for ensembles, unless they are tuned to a low-bias, high-variance regime, as explored in [Valentini and Dietterich, 2003]. When it comes to diversity and ensembling, NNs are known to exhibit properties different than traditional models, e.g., as described in recent theoretical and empirical work on loss landscapes and emsemble improvement [Theisen et al., 2023, Yang et al., 2021]. Therefore, ensembling techniques that work well for traditional models (e.g., tree-type models) often underperform the simple yet efficient deep ensembles method [Fort et al., 2019, Ortega et al., 2022] that uses the independent initialization and optimization. Previous literature has explored various new methods to learn diverse NNs [Lee et al., 2022, Rame et al., 2022, Pang et al., 2019, Parker-Holder et al., 2020]. Our work is different from previous work in that we study flat ensembles obtained from sharpness-aware training methods, especially focusing on diversifying flat ensembles by reducing the overlap between sharpness-aware data subsets. While our work demonstrates significant improvements in OOD generalization, it is known that (in some cases, see also Theisen et al. [2023]) deep ensembling is a simple, yet effective method to improve OOD performance [Diffenderfer et al., 2021]. Therefore, we compare the OOD performance of `SharpBalance` to deep ensembles.

Sharpness and generalization. A large body of work has studied the relationship between the sharpness (or flatness) of minima and the generalizability of models [Hochreiter and Schmidhuber, 1997, Hinton and van Camp, 1993, Keskar et al., 2016, Neyshabur et al., 2018, Yang et al., 2021, Kaddour et al., 2022, Yao et al., 2021, 2020]. Works such as those by [Hochreiter and Schmidhuber, 1997] and [Hinton and van Camp, 1993] use Bayesian learning and minimum description length to explain why we should train models to flat minima. [Keskar et al., 2016] introduces a sharpness-based metric, demonstrating how large-batch training can skew NNs towards sharp local minima, adversely affecting generalization. In addition, [Neyshabur et al., 2018] uses a PAC-Bayesian framework to prove bounds on generalization, which can be interpreted as the relationship between sharpness and test accuracy. Furthermore, [Cha et al., 2021] presents a theoretical exploration of the link between the sharpness of minima and OOD generalization.

Motivated by the good generalization property of flat minima, variants of sharpness-guided optimization techniques have been proposed [Yao et al., 2018, 2021, Du et al., 2024, Jiang et al., 2023], including sharpness-aware minimization [Foret et al., 2021]. The DiWA method [Rame et al., 2022] observed that `SAM` can decrease the diversity of models in the context of weight averaging (WA) [Izmailov et al., 2018]. However, WA imposes constraints on different models, requiring them to share the same initialization and stay close to each other in the parameter space. In contrast, our work focuses on deep ensembles that do not pose additional constraints on the training trajectories of individual ensemble members. Previous work by [Behdin and Mazumder, 2023] provided a theoretical characterization of important statistical properties for kernel regression models and single-layer ReLU networks, optimized using `SAM` on noisy datasets. Our theoretical analysis borrows ideas from [Behdin and Mazumder, 2023] and extends the analysis using random matrix theory. DASH was proposed in [Bui et al., 2024] to minimize the generalization loss by adding KL divergence constraint

on the output logits of ensemble members. The authors believe that the decrease in diversity is a result of models being initialized closely and updated with the same direction. In contrast, `SharpBalance` observed that the sharpness-diversity trade-off is ubiquitous across various settings and provides a rigorous theoretical quantification that characterizes the interplay of the two metrics. Compare to DASH, `SharpBalance` provably achieves improved performance and is simple, effective, and computationally cheap to implement.

C Proof of Theorems in Section 3

Recall that `SAM` updates the model weights, ignoring the normalization constant and regularization, through the following recursive rule

$$\boldsymbol{\theta}_{k+1}^{SAM} = \boldsymbol{\theta}_k^{SAM} - \eta \nabla f\left(\boldsymbol{\theta}_k^{SAM} + \rho \nabla f(\boldsymbol{\theta}_k^{SAM})\right).$$

We first show an unrolling of the iterative optimization on a quadratic objective.

Theorem 3 (Unrolling `SAM`)**.** *Let $\boldsymbol{\theta}^*$ be the teacher model. Let $\boldsymbol{\theta}_0$ be randomly initialized and updated with `SAM` to solve a quadratic objective $\mathcal{L}_{\mathbf{A}}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{A}^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Then,*

$$\boldsymbol{\theta}_{k+1}^{SAM} = \eta \sum_{i=0}^{k} \mathbf{B}^i \left(\mathbf{A}^T\mathbf{A} + \rho(\mathbf{A}^T\mathbf{A})^2\right)\boldsymbol{\theta}^* + \mathbf{B}^{k+1}\boldsymbol{\theta}_0,$$

where $\mathbf{B} = \mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2$.

Proof. The gradient of the objective f is given by $\nabla f(\boldsymbol{\theta}) = \mathbf{A}^T\mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. Therefore,

$$\boldsymbol{\theta}_k^{SAM} + \rho\nabla f(\boldsymbol{\theta}_k^{SAM}) = (\mathbf{I} + \rho\mathbf{A}^T\mathbf{A})\boldsymbol{\theta}_k^{SAM} - \rho\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*.$$

With SAM update,

$$\begin{aligned}
\boldsymbol{\theta}_{k+1}^{SAM} &= \boldsymbol{\theta}_k^{SAM} - \eta\nabla f\left(\boldsymbol{\theta}_k^{SAM} + \rho\nabla f(\boldsymbol{\theta}_k^{SAM})\right) \\
&= \boldsymbol{\theta}_k^{SAM} - \eta\mathbf{A}^T\mathbf{A}\left(\boldsymbol{\theta}_k^{SAM} + \rho\nabla f(\boldsymbol{\theta}_k^{SAM}) - \boldsymbol{\theta}^*\right) \\
&= \boldsymbol{\theta}_k^{SAM} - \eta\mathbf{A}^T\mathbf{A}\left((\mathbf{I} + \rho\mathbf{A}^T\mathbf{A})\boldsymbol{\theta}_k^{SAM} - \rho\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^* - \boldsymbol{\theta}^*\right) \\
&= \left(\mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2\right)\boldsymbol{\theta}_k^{SAM} + \eta\left(\mathbf{A}^T\mathbf{A} + \rho(\mathbf{A}^T\mathbf{A})^2\right)\boldsymbol{\theta}^* \\
&= \eta\sum_{i=0}^{k}\mathbf{B}^i\left(\mathbf{A}^T\mathbf{A} + \rho(\mathbf{A}^T\mathbf{A})^2\right)\boldsymbol{\theta}^* + \mathbf{B}^{k+1}\boldsymbol{\theta}_0,
\end{aligned}$$

where the last equation is obtained by recursively unrolling the weight by previous updates. □

Theorem 3 offers a valuable tool to analyze the statistical behavior of the models optimized by `SAM`. However, one more ingredient is required to arrive at the interesting conclusions claimed in Section 3, the random matrix theory. Recall that the data matrix $\mathbf{A} \in \mathbb{R}^{n_{\text{tr}} \times d_{\text{in}}}$ is random with entries drawn from Gaussian $\mathcal{N}(0, \mathbf{I}/d_{\text{in}})$. As a result, entries in $\mathbf{A}^T\mathbf{A}$ follows the Wishart distribution and according to Corollary 3.3 in Bishop et al. [2018], for $k \geq 1$,

$$\mathbb{E}[(\mathbf{A}^T\mathbf{A})^k] = \left(\frac{n_{\text{tr}}}{d_{\text{in}}}\right)^k \sum_{i=1}^{k}\left(\frac{d_{\text{in}}}{n_{\text{tr}}}\right)^{k-i} \mathcal{O}\left(1 + 1/d_{\text{in}}\right) N_{k,i}\mathbf{I}, \tag{6}$$

where $N_{k,i} = \frac{1}{i}\binom{k-1}{i-1}\binom{k}{i-1}$ is the Narayana number. With the help of this Corollary, we now prove a proposition on the expectation of $\mathbf{B}^k$.

Proposition 1 (Expectation of Wishart Moments)**.** *Let i, j be non-negative integers, then*

$$\mathbb{E}_{\mathbf{A}}[\mathbf{B}^i(\mathbf{A}^T\mathbf{A})^j] = \phi(i,j)\mathbf{I},$$

where

$$\phi(i,j) := \mathbb{1}_{j=0} + \sum_{k_1+k_2+k_3=i} \frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\left(\frac{n_{tr}}{d_{in}}\right)^m \sum_{l=1}^{m}\left(\frac{d_{in}}{n_{tr}}\right)^{m-l}\mathcal{O}(1 + 1/d_{in})N_{m,l},$$

and $m = k_2 + 2k_3 + j$.

https://doi.org/10.52202/079017-4344

Proof. By Multinomial Theorem,

$$\begin{aligned}\mathbf{B}^i(\mathbf{A}^T\mathbf{A})^j &= \left(\sum_{k_1+k_2+k_3=i}\frac{i!}{k_1!k_2!k_3!}\mathbf{I}^{k_1}(-\eta\mathbf{A}^T\mathbf{A})^{k_2}(-\eta\rho(\mathbf{A}^T\mathbf{A})^2)^{k_3}\right)(\mathbf{A}^T\mathbf{A})^j \\ &= \sum_{k_1+k_2+k_3=i}\frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}(\mathbf{A}^T\mathbf{A})^{k_2+2k_3+j}.\end{aligned}$$

Let $m = k_2 + 2k_3 + j$ and taking the expectation with equation (6) gives

$$\begin{aligned}\mathbb{E}_{\mathbf{A}}[\mathbf{B}^i(\mathbf{A}^T\mathbf{A})^j] &= \sum_{k_1+k_2+k_3=i}\frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\mathbb{E}_{\mathbf{A}}[(\mathbf{A}^T\mathbf{A})^{k_2+2k_3+j}] \\ &= \sum_{k_1+k_2+k_3=i}\frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\left(\frac{n_{\text{tr}}}{d_{\text{in}}}\right)^m\sum_{l=1}^{m}\left(\frac{d_{\text{in}}}{n_{\text{tr}}}\right)^{m-l}\mathcal{O}(1+1/d_{\text{in}})N_{m,l}\mathbf{I}.\end{aligned}$$

If $j = 0$, then there is a case when $k_2 = k_3 = 0$, and the expectation of $(\mathbf{A}^T\mathbf{A})^0$ simply becomes $\mathbf{I}$. Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{A}}[\mathbf{B}^i(\mathbf{A}^T\mathbf{A})^j] &= \mathbb{1}_{j=0}\mathbf{I} + \sum_{k_1+k_2+k_3=i}\frac{i!}{k_1!k_2!k_3!}(-\eta)^{k_2+k_3}\rho^{k_3}\left(\frac{n_{\text{tr}}}{d_{\text{in}}}\right)^m\sum_{l=1}^{m}\left(\frac{d_{\text{in}}}{n_{\text{tr}}}\right)^{m-l}\mathcal{O}(1+1/d_{\text{in}})N_{m,l}\mathbf{I} \\ &= \phi(i,j)\mathbf{I}.\end{aligned}$$

□

C.1 Proof of Theorem 1

In this subsection, we show a proof for Theorem 1.

Proof. Apply Singular Value Decomposition (SVD) to obtain $\mathbf{A} = \mathbf{V}\Sigma\mathbf{U}^T$ and $\mathbf{A}^T\mathbf{A} = \mathbf{U}\Sigma^2\mathbf{U}^T$. Let $\mathbf{D} = \Sigma^2$. By Theorem 3,

$$\begin{aligned}\boldsymbol{\theta}_k^{SAM} &= \eta\sum_{i=0}^{k-1}\mathbf{B}^i(\mathbf{A}^T\mathbf{A} + \rho(\mathbf{A}^T\mathbf{A})^2)\boldsymbol{\theta}^* + \mathbf{B}^k\boldsymbol{\theta}_0 \\ &= \eta\sum_{i=0}^{k}\left(\mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2\right)^i\left(\mathbf{A}^T\mathbf{A} + \rho(\mathbf{A}^T\mathbf{A})^2\right)\boldsymbol{\theta}^* + \mathbf{B}^k\boldsymbol{\theta}_0 \\ &= \eta\sum_{i=0}^{k-1}\mathbf{U}(\mathbf{I} - \eta\mathbf{D} - \eta\rho\mathbf{D}^2)^i\mathbf{U}^T\mathbf{U}(\mathbf{D} + \rho\mathbf{D}^2)\mathbf{U}^T\boldsymbol{\theta}^* + \mathbf{B}^k\boldsymbol{\theta}_0 \\ &= \eta\mathbf{U}\cdot\mathbf{diag}\left(\left\{\sum_{i=0}^{k-1}(1 - \eta d_j - \eta\rho d_j^2)^i(d_j + \rho d_j^2)\right\}_{j=1}^{d_{\text{in}}}\right)\mathbf{U}^T\boldsymbol{\theta}^* + \mathbf{B}^k\boldsymbol{\theta}_0 \\ &= \eta\mathbf{U}\cdot\mathbf{diag}\left(\left\{\frac{1 - (1 - \eta d_j - \eta\rho d_j^2)^k}{\eta d_j + \eta\rho d_j^2}(d_j + \rho d_j^2)\right\}_{j=1}^{d_{\text{in}}}\right)\mathbf{U}^T\boldsymbol{\theta}^* + \mathbf{B}^k\boldsymbol{\theta}_0 \\ &= \mathbf{U}\left(\mathbf{I} - (\mathbf{I} - \eta\mathbf{D} - \eta\rho\mathbf{D}^2)^k\right)\mathbf{U}^T\boldsymbol{\theta}^* + \left(\mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2\right)^k\boldsymbol{\theta}_0 \\ &= \boldsymbol{\theta}^* + \left(\mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2\right)^k(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*).\end{aligned}$$

As a result, $\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_k^{SAM}] = \boldsymbol{\theta}^* - \left(\mathbf{I} - \eta\mathbf{A}^T\mathbf{A} - \eta\rho(\mathbf{A}^T\mathbf{A})^2\right)^k \boldsymbol{\theta}^* = \boldsymbol{\theta}^* - \mathbf{B}^k\boldsymbol{\theta}^*$. By definition,

$$\begin{aligned}
n_{\text{te}}\text{Bias}^2(\boldsymbol{\theta}_k^{SAM}) &= \mathbb{E}_{\mathbf{A},\mathbf{T}}[\sum_{i=1}^{p}(\mathbb{E}_{\boldsymbol{\theta}_0}[f(\boldsymbol{\theta}_k^{SAM};\mathbf{T}_i)] - y_i^{(\mathbf{T})})^2] \\
&= \mathbb{E}_{\mathbf{A},\mathbf{T}}[(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_k^{SAM}] - \boldsymbol{\theta}^*)^T\mathbf{T}^T\mathbf{T}(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_k^{SAM}] - \boldsymbol{\theta}^*)] \\
&= \mathbb{E}_{\mathbf{A}}[(\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_k^{SAM}] - \boldsymbol{\theta}^*)^T\mathbb{E}_{\mathbf{T}}[\mathbf{T}^T\mathbf{T}](\mathbb{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\theta}_k^{SAM}] - \boldsymbol{\theta}^*)] \\
&= \frac{n_{\text{te}}}{d_{\text{in}}}\mathbb{E}_{\mathbf{A}}[(\boldsymbol{\theta}^*)^T\mathbf{B}^{2k}\boldsymbol{\theta}^*] \\
&= \frac{n_{\text{te}}}{d_{\text{in}}}\phi(2k,0)\|\boldsymbol{\theta}^*\|_2^2, \\
n_{\text{te}}\text{Error}(\boldsymbol{\theta}_k^{SAM}) &= \mathbb{E}_{\mathbf{A},\mathbf{T},\theta_0}[\sum_{i=1}^{p}(y_i^{(\mathbf{T})} - f(\boldsymbol{\theta}_k^{SAM};\mathbf{T}_i))^2] \\
&= \mathbb{E}_{\mathbf{A},\mathbf{T},\theta_0}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^{SAM})^T\mathbf{T}^T\mathbf{T}(\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^{SAM})] \\
&= \mathbb{E}_{\mathbf{A},\mathbf{T},\theta_0}[(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)^T\mathbf{B}^k\mathbf{T}^T\mathbf{T}\mathbf{B}^k(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)] \\
&= \mathbb{E}_{\mathbf{A},\mathbf{T},\theta_0}[(\boldsymbol{\theta}^*)^T\mathbf{B}^k\mathbf{T}^T\mathbf{T}\mathbf{B}^k\boldsymbol{\theta}^*] + \mathbb{E}_{\mathbf{A},\mathbf{T},\theta_0}[\boldsymbol{\theta}_0^T\mathbf{B}^k\mathbf{T}^T\mathbf{T}\mathbf{B}^k\boldsymbol{\theta}_0] \\
&= \frac{n_{\text{te}}}{d_{\text{in}}}\phi(2k,0)\|\boldsymbol{\theta}^*\|_2^2 + n_{\text{te}}\phi(2k,0)\sigma^2,
\end{aligned}$$

Since $\mathbb{E}_{\mathbf{T}}[\mathbf{T}^T\mathbf{T}] = \frac{n_{\text{te}}}{d_{\text{in}}}\mathbf{I}$ and $\mathbb{E}[\boldsymbol{\theta}_0\boldsymbol{\theta}_0^T] = \sigma^2\mathbf{I}$. Hence,

$$\mathbb{D}(\boldsymbol{\theta}_k^{SAM}) = \text{Var}\left(f(\boldsymbol{\theta}_k^{SAM};\mathbf{T})\right) = \frac{1}{n_{\text{te}}}\left(n_{\text{te}}\text{Error}(\boldsymbol{\theta}_k^{SAM}) - n_{\text{te}}\text{Bias}^2(\boldsymbol{\theta}_k^{SAM})\right) = \phi(2k,0)\sigma^2.$$

Recall that given a perturbation radius ρ_0, the sharpness is defined as

$$\kappa(\boldsymbol{\theta}_k) = \mathbb{E}_A[\max_{\|\boldsymbol{\varepsilon}\|_2\le\rho_0} f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k\right] + \boldsymbol{\varepsilon}\right) - f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k\right]\right)].$$

We first compute

$$\begin{aligned}
f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k^{SAM}\right] + \boldsymbol{\varepsilon};\mathbf{A}\right) &= \frac{1}{2}(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k^{SAM}\right] + \boldsymbol{\varepsilon} - \boldsymbol{\theta}^*)^T A^T A(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k^{SAM}\right] + \boldsymbol{\varepsilon} - \boldsymbol{\theta}^*) \\
&= \frac{1}{2}(\boldsymbol{\varepsilon} - \mathbf{B}^k\boldsymbol{\theta}^*)^T\mathbf{A}^T\mathbf{A}(\boldsymbol{\varepsilon} - \mathbf{B}^k\boldsymbol{\theta}^*) \\
&= \frac{1}{2}\boldsymbol{\varepsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^* + \frac{1}{2}(\boldsymbol{\theta}^*)^T\mathbf{B}^{2k}\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*. \qquad (7)
\end{aligned}$$

Similarly,

$$f\left(\mathbb{E}_{\boldsymbol{\theta}_0}\left[\boldsymbol{\theta}_k^{SAM}\right];\mathbf{A}\right) = \frac{1}{2}(\boldsymbol{\theta}^*)^T\mathbf{B}^{2k}\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*. \qquad (8)$$

Let λ_{min} be the least eigenvalue of $\mathbf{A}^T\mathbf{A}$. By subtracting equation (7) with equation (8), we have

$$\begin{aligned}
\kappa_k^{SAM} &= \mathbb{E}_{\mathbf{A}}[\max_{\|\boldsymbol{\varepsilon}\|_2\le\rho_0} \frac{1}{2}\boldsymbol{\varepsilon}^T\mathbf{A}^T\mathbf{A}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*] \\
&\ge \mathbb{E}_{\mathbf{A}}[\max_{\|\boldsymbol{\varepsilon}\|_2=\rho_0} \frac{1}{2}\lambda_{min}\|\mathbf{U}^T\boldsymbol{\varepsilon}\|_2^2 - \boldsymbol{\varepsilon}^T\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*] \\
&\ge \mathbb{E}_{\mathbf{A}}[\max_{\substack{\|\boldsymbol{\varepsilon}\|_2=\rho_0 \\ \boldsymbol{\varepsilon}=\mathbf{U}\mathbf{v}}} \frac{1}{2}\lambda_{min}\|\mathbf{U}^T\boldsymbol{\varepsilon}\|_2^2 - \boldsymbol{\varepsilon}^T\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*] \\
&= \mathbb{E}_{\mathbf{A}}[\max_{\|\mathbf{v}\|_2=\rho_0} \frac{1}{2}\lambda_{min}\|\mathbf{v}\|_2^2 - \min_{\|\boldsymbol{\varepsilon}\|_2=\rho_0} \boldsymbol{\varepsilon}^T\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*] \\
&= \mathbb{E}_{\mathbf{A}}[\frac{1}{2}\lambda_{min}\rho_0^2 + \rho_0\|\mathbf{B}^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*\|_2].
\end{aligned}$$

The smallest singular value λ_{min} of a random $n \times d_{\text{in}}$ matrix $\mathbf{A}$ can be bounded by the following inequality on the smallest singular value $\sigma_{min}(A)$ by Vershynin [2018], assuming $n_{\text{tr}} \ge d_{\text{in}}$, then almost surely

https://doi.org/10.52202/079017-4344

$$\mathbb{E}_{\mathbf{A}}[\sigma_{min}(\mathbf{A})] \geq \sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} - 1.$$

Therefore, $\mathbb{E}_A[\lambda_{min}] \geq \mathbb{E}_A[\sigma_{min}(A)]^2 \geq \left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} - 1\right)^2$. Now we show a lower bound on $\mathbb{E}_{\mathbf{A}}[\rho_0 \|\mathbf{B}^k \mathbf{A}^T \mathbf{A} \boldsymbol{\theta}^*\|_2]$. By Gao et al. [2019], the Jensen gap $(\mathbb{E}[Z])^{1/2} - \mathbb{E}[(Z)^{1/2}]$ is upper bounded by $\frac{\text{Var}(Z)}{2}$ when Z is non-negative and $\mathbb{E}[Z] = 1$. Notice that

$$\mathbb{E}_{\mathbf{A}}[\rho_0 \|\mathbf{B}^k \mathbf{A}^T \mathbf{A} \boldsymbol{\theta}^*\|_2] = \rho_0 \mathbb{E}_{\mathbf{A}}[\left((\boldsymbol{\theta}^*)^T \mathbf{B}^{2k} (\mathbf{A}^T \mathbf{A})^2 \boldsymbol{\theta}^*\right)^{1/2}],$$

and we let $Z = (\boldsymbol{\theta}^*)^T \mathbf{B}^{2k} (\mathbf{A}^T \mathbf{A})^2 \boldsymbol{\theta}^*$. Then $\mathbb{E}_{\mathbf{A}}[Z] = \phi(2k, 2) \|\boldsymbol{\theta}^*\|_2^2$ and

$$\text{Var}[Z] = \left(\phi(4k, 4) - \phi(2k, 2)^2\right) \|\boldsymbol{\theta}^*\|_2^2.$$

By normalizing Z and applying the Jensen gap upperbound, we have

$$\mathbb{E}_{\mathbf{A}}[\rho_0 \|\mathbf{B}^k \mathbf{A}^T \mathbf{A} \boldsymbol{\theta}^*\|_2] \geq \rho_0 \sqrt{\phi(2k, 2)} \|\boldsymbol{\theta}^*\|_2^2 - \frac{\phi(4k, 4) - \phi(2k, 2)^2}{2\phi(2k, 2)^{3/2} \|\boldsymbol{\theta}^*\|_2}.$$

As a result,

$$\kappa_k^{SAM} \geq \frac{\rho_0^2}{2} \left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} - 1\right)^2 + \rho_0 \sqrt{\phi(2k, 2)} \|\boldsymbol{\theta}^*\|_2 - \frac{\phi(4k, 4) - \phi(2k, 2)^2}{2\phi(2k, 2)^{3/2} \|\boldsymbol{\theta}^*\|_2}.$$

The derivation of the upper bound follows from a similar proof, ignoring the Jensen gap.

□

C.2 Proof of Theorem 2

Below we show a proof of Theorem 2.

Proof. We apply SVD to $\mathbf{A}_s$ to obtain $\mathbf{A}_s = \mathbf{V}_s \Sigma_s \mathbf{U}_s^T$ and $\mathbf{A}_s^T \mathbf{A} = \mathbf{U}_s \Sigma_s^2 \mathbf{U}_s^T$. Let $\mathbf{D}_s = \Sigma_s^2$ and $\mathbf{B}_s = \mathbf{I} - \eta \mathbf{A}_s^T \mathbf{A}_s - \eta\rho(\mathbf{A}_s^T \mathbf{A}_s)^2$. By Theorem 3 and a similar derivation in the proof of Theorem 1,

$$\begin{aligned}
\boldsymbol{\theta}_k^{SharpBal} &= \eta \sum_{j=0}^{k-1} \mathbf{B}_s^j \left(\mathbf{A}_s^T \mathbf{A}_s + \rho(\mathbf{A}_s^T \mathbf{A}_s)^2\right) \boldsymbol{\theta}^* + \mathbf{B}_s^k \boldsymbol{\theta}_0 \\
&= \boldsymbol{\theta}^* + \left(\mathbf{I} - \eta \mathbf{A}_s^T \mathbf{A}_s - \eta\rho(\mathbf{A}_s^T \mathbf{A}_s)^2\right)^k (\boldsymbol{\theta}_0 - \boldsymbol{\theta}^*).
\end{aligned}$$

As a result, $\mathbb{E}_{\boldsymbol{\theta}_0, s}[\boldsymbol{\theta}_k^{Sharpbal}] = \mathbb{E}_s[\boldsymbol{\theta}^* - \mathbf{B}_s^k \boldsymbol{\theta}^*] = \boldsymbol{\theta}^* - \frac{1}{S} \sum_{s=1}^{S} \mathbf{B}_s^k \boldsymbol{\theta}^*$.

Applying Proposition 1, we have

$$\mathbb{E}_{\mathbf{A}}[\mathbf{B}_s^i (\mathbf{A}_s^T \mathbf{A}_s)^j] = \phi'(i, j),$$

where

$$\phi'(i, j) = \mathbb{1}_{j=0} + \sum_{k_1+k_2+k_3=i} \frac{i!}{k_1! k_2! k_3!} (-\eta)^{k_2+k_3} \rho^{k_3} \left(\frac{n_{\text{tr}}}{S d_{\text{in}}}\right)^m \sum_{l=1}^{m} \left(\frac{S d_{\text{in}}}{n_{\text{tr}}}\right)^{m-l} N_{m,l}.$$

Then,

$$\begin{aligned}
n_{\text{te}} \text{Bias}^2(\boldsymbol{\theta}_k^{SAM}) &= \mathbb{E}_{\mathbf{A}, \mathbf{T}}[\left(\mathbb{E}_{\boldsymbol{\theta}_0, s}[\boldsymbol{\theta}_k^{Sharpbal}] - \boldsymbol{\theta}^*\right)^T \mathbf{T}^T \mathbf{T} \left(\mathbb{E}_{\boldsymbol{\theta}_0, s}[\boldsymbol{\theta}_k^{Sharpbal}] - \boldsymbol{\theta}^*\right)] \\
&= \frac{n_{\text{te}}}{d_{\text{in}}} \mathbb{E}_{\mathbf{A}}[(-\frac{1}{S} \sum_{s=1}^{S} \mathbf{B}_s^k \boldsymbol{\theta}^*)^T (-\frac{1}{S} \sum_{s'=1}^{S} \mathbf{B}_{s'}^k \boldsymbol{\theta}^*)] \\
&= \frac{n_{\text{te}}}{d_{\text{in}} S^2} \mathbb{E}_{\mathbf{A}}[\sum_{s=1}^{S} \mathbf{B}_s^k \boldsymbol{\theta}^* \sum_{s'=1}^{S} \mathbf{B}_{s'}^k] \|\boldsymbol{\theta}^*\|_2^2 \\
&= \frac{n_{\text{te}}}{d_{\text{in}} S} \left(\phi'(2k, 0) + (s-1)\phi'(k, 0)^2\right) \|\boldsymbol{\theta}^*\|_2^2.
\end{aligned}$$

The last equality is the result of applying $\mathbb{E}_{\mathbf{A}}[B_s^i] = \phi'(i,0)$ with different combinations of $\mathbf{B}_s$, $\mathbf{B}_{s'}$, counting multiplicity. Similarly,

$$\begin{aligned} n_{\text{te}}\text{Error}(\boldsymbol{\theta}_k^{Sharpbal}) &= \mathbb{E}_{\mathbf{A},\mathbf{T},\boldsymbol{\theta}_0,s}[(\boldsymbol{\theta}^*)^T\mathbf{B}_s^k\mathbf{T}^T\mathbf{T}\mathbf{B}_s^k\boldsymbol{\theta}^*] + \mathbb{E}_{\mathbf{A},\mathbf{T},\boldsymbol{\theta}_0,s}[\boldsymbol{\theta}_0^T\mathbf{B}_s^k\mathbf{T}^T\mathbf{T}\mathbf{B}_s^k\boldsymbol{\theta}_0] \\ &= \frac{n_{\text{te}}}{d_{\text{in}}}\phi'(2k,0)\|\boldsymbol{\theta}^*\|_2^2 + n_{\text{te}}\phi'(2k,0)\sigma^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Var}\left(f(\boldsymbol{\theta}_k^{SharpBal};\mathbf{T})\right) =& \frac{1}{n_{\text{te}}}\left(n_{\text{te}}\text{Error}(\boldsymbol{\theta}_k^{SharpBal}) - n_{\text{te}}\text{Bias}^2(\boldsymbol{\theta}_k^{SharpBal})\right) \\ =& \phi'(2k,0)\sigma^2 + \frac{S-1}{d_{\text{in}}S}\left(\phi'(2k,0) - \phi'(k,0)^2\right)\|\boldsymbol{\theta}^*\|_2^2. \end{aligned}$$

When the model is trained on the submatrix, the sharpness of model $\boldsymbol{\theta}_k^{SharpBal}$ is defined as

$$\kappa_k^{SharpBal} = \mathbb{E}_{\mathbf{A}}[\max_{\|\boldsymbol{\varepsilon}\|_2\leq\rho_0} f\left(\mathbb{E}_{\boldsymbol{\theta}_0,s}\left[\boldsymbol{\theta}_k^{SharpBal}\right] + \boldsymbol{\varepsilon};\mathbf{A}\right) - f\left(\mathbb{E}_{\boldsymbol{\theta}_0,s}\left[\boldsymbol{\theta}_k^{SharpBal}\right];\mathbf{A}\right)].$$

From a similar analysis of the proof for Theorem 1,

$$\kappa_k^{SharpBal} \leq \frac{\rho_0^2}{2}\left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} + 1\right)^2 + \frac{\rho_0}{S}\mathbb{E}_{\mathbf{A}}[\|\sum_{s=1}^{S}\mathbf{B}_s^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*\|_2],$$

and with $r = \frac{n_{\text{tr}}}{Sd_{\text{in}}}$,

$$\begin{aligned} \mathbb{E}_{\mathbf{A}}[\|\sum_{s=1}^{S}\mathbf{B}_s^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*\|_2] =& \mathbb{E}_{\mathbf{A}}[((\boldsymbol{\theta}^*)^T\mathbf{A}^T\mathbf{A}\sum_{s=1}^{S}\mathbf{B}_s^k\sum_{s'=1}^{S}\mathbf{B}_s'^k\mathbf{A}^T\mathbf{A}\boldsymbol{\theta}^*)^{1/2}] \\ \leq& \left((\boldsymbol{\theta}^*)^T\mathbb{E}_{\mathbf{A}}[\sum_{j=1}^{S}\mathbf{A}_j^T\mathbf{A}_j\sum_{s=1}^{S}\mathbf{B}_s^k\sum_{s'=1}^{S}\mathbf{B}_s'^k\sum_{l=1}^{S}\mathbf{A}_l^T\mathbf{A}_l]\boldsymbol{\theta}^*\right)^{1/2} \\ =& (S\phi'(2k,2) + 2rS(S-1)\phi'(2k,1) + 2S(S-1)\phi'(k,2)\phi'(k,0) \\ &+ r(1+r)S(S-1)\phi'(2k,0) + 2S(S-1)\phi'(k,1)\phi'(k,1) \\ &+ \frac{3}{2}r(1+r)S(S-1)(S-2)\phi'(k,0)^2 \\ &+ \frac{3}{2}r^2S(S-1)(S-2)\phi'(2k,0) \\ &+ 3rS(S-1)(S-2)\phi'(k,1)\phi'(k,0) \\ &+ r^2S(S-1)(S-2)(S-3)\phi'(k,0)^2)^{1/2}\|\boldsymbol{\theta}^*\|_2. \end{aligned}$$

The last equality is the result of applying $\mathbb{E}_{\mathbf{A}}[B_s^i(\mathbf{A}_s^T\mathbf{A}_s)^j] = \phi'(i,j)$ with different combinations of $\mathbf{B}_s$, $\mathbf{B}_{s'}$, $\mathbf{A}_j^T\mathbf{A}_j$, and $\mathbf{A}_l^T\mathbf{A}_l$, counting multiplicity and the fact that $\mathbb{E}_{\mathbf{A}}[(\mathbf{A}_s^T\mathbf{A}_s)^2] = r(1+r)\mathbf{I}$. In conclusion,

$$\kappa_k^{SharpBal} \leq \frac{\rho_0^2}{2}\left(\sqrt{\frac{n_{\text{tr}}}{d_{\text{in}}}} + 1\right)^2 + \frac{\rho_0}{S}\sqrt{C}\|\boldsymbol{\theta}^*\|_2,$$

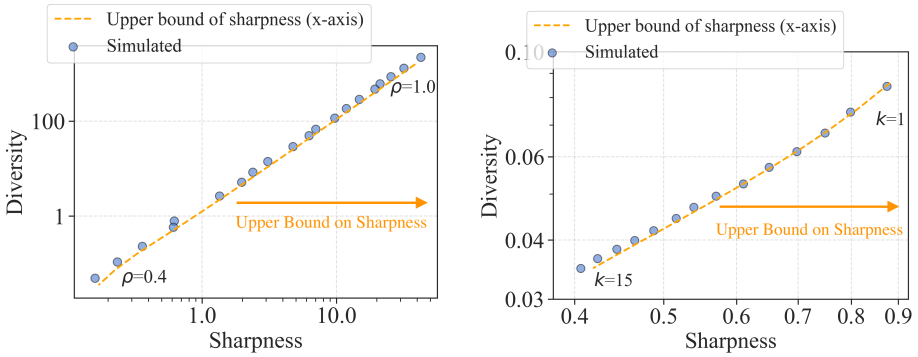
where

$$\begin{aligned} C =& S\phi'(2k,2) + 2rS(S-1)\phi'(2k,1) + 2S(S-1)\phi'(k,2)\phi'(k,0) \\ &+ r(1+r)S(S-1)\phi'(2k,0) + 2S(S-1)\phi'(k,1)\phi'(k,1) \\ &+ \frac{3}{2}r(1+r)S(S-1)(S-2)\phi'(k,0)^2 + \frac{3}{2}r^2S(S-1)(S-2)\phi'(2k,0) \\ &+ 3rS(S-1)(S-2)\phi'(k,0)\phi'(k,1) + r^2S(S-1)(S-2)(S-3)\phi'(k,0)^2. \end{aligned}$$

□

The claims in Theorem 2 is further supported by the experimental validations with results presented in Figure 9.

 https://doi.org/10.52202/079017-4344



(a) Varying perturbation radius ρ (b) Varying number of training iterations k

Figure 9: **(Theoretical vs. Simulated sharpness-diversity trade-off in `SharpBalance`)** This figure illustrates the relationship between sharpness(upper bound) and diversity as predicted by Theorem 2 and as observed in simulations under different configurations. **(a)** validates our theoretical results by varying the perturbation radius ρ from 1.0 to 0.4. **(b)** validates the derivation by varying number of iterations k from 1 to 15. These results demonstrate the soundness of our derivation across a range of parameters.

C.3 Empirical Verification of Theorem 1 and 2

To demonstrate the robustness and tightness of the bounds presented in Theorem 1, we provide verification results across a range of parameter configurations. Interestingly, the observed model behaviors closely align with the upper bound derived in Theorem 1, highlighting the effectiveness of our theoretical analysis in capturing the underlying dynamics of the ensemble. Figure 10 illustrates these results, with each sub-figure corresponding to a specific combination of k and η with ρ from range 0.5 to 0.3. In these experiment, we generated 50 random data matrices $\mathbf{A}$ of size 3000×150 and test data $\mathbf{T}$ of size 1000×150. For each random dataset, we initialized 50 random model weights $\boldsymbol{\theta}_0$ and collected the expected statistics of interest after training. To measure the sharpness κ_k^{SAM}, we employed projected gradient ascent to find the optimal perturbation, using a step size of 0.01 and a maximum of 50 steps. Similar experiments are performed to verify the derivations in Theorem 2 with results presented in Figure 11, with the number of partitions $S = 10$.

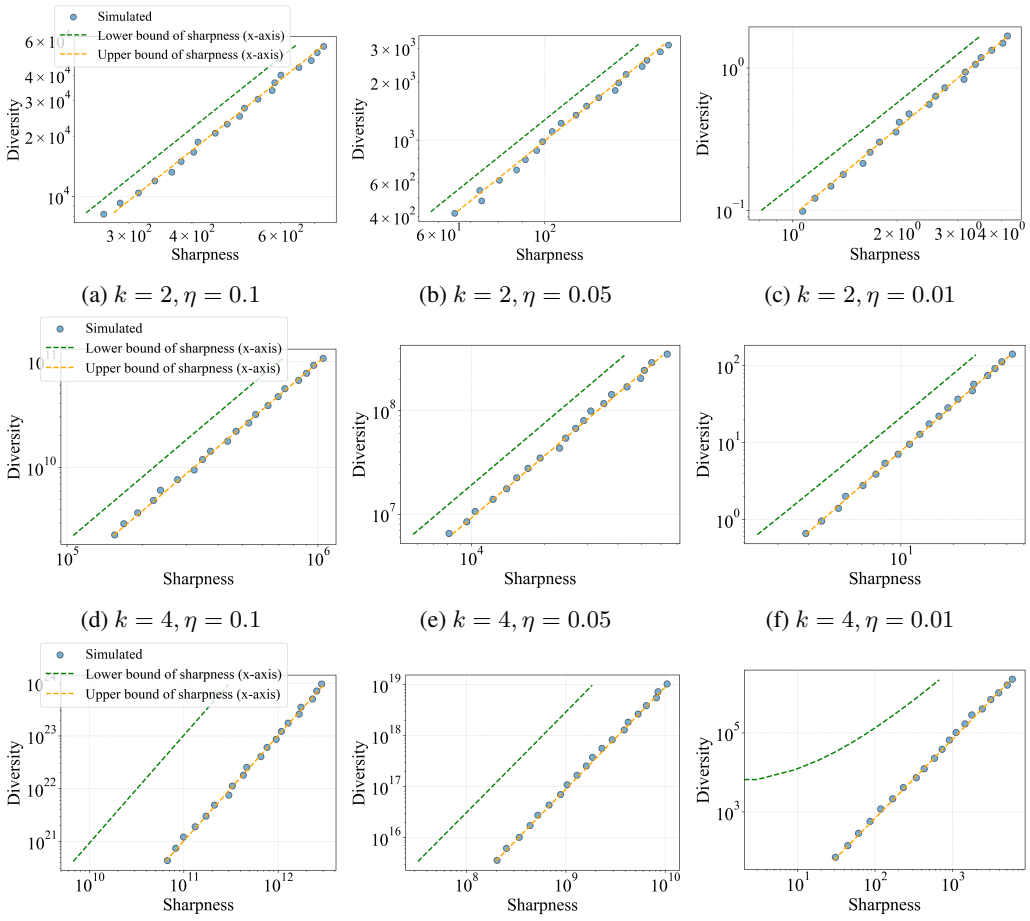
D Hyperparamter setting

D.1 Datasets

We first evaluate on image classification datasets CIFAR-10 and CIFAR-100. The corresponding OOD robustness is evaluated on CIFAR-10C and CIFAR-100C [Hendrycks and Dietterich, 2019b]. The experiments are carried out on ResNet18 [He et al., 2016]. We use a batch size of 128, a momentum of 0.9, and a weight decay of 0.0005 for model training. TinyImageNet is an image classification dataset consisting of 100K images for training and 10K images for in-distribution testing. We evaluate ensemble's OOD robustness on TinyImageNetC [Hendrycks and Dietterich, 2019b].

D.2 Hyperparamter setting for empirical sharpness-diversity trade-off

Here, we provide the hyperparameter for the experiments in Section 4.2. When using adaptive worst-case sharpness for sharpness measurement, the size of neighborhood γ defined in equation (3) needed to be specified, we use a γ of 0.5 for all the results in Figure 1 and Figure 5. Additionally, when training NNs in the ensemble, we change the perturbation radius ρ of `SAM` so that we can study the trade-off. The range of ρ for the results in Figure 1 is $\{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2, 0.3\}$, the range of ρ for the results in Figure 5 is $\{0.01, 0.015, 0.02, 0.025, 0.03, 0.05, 0.1, 0.2, 0.3, 0.4\}$.



(a) $k = 2, \eta = 0.1$ (b) $k = 2, \eta = 0.05$ (c) $k = 2, \eta = 0.01$

(d) $k = 4, \eta = 0.1$ (e) $k = 4, \eta = 0.05$ (f) $k = 4, \eta = 0.01$

(g) $k = 8, \eta = 0.1$ (h) $k = 8, \eta = 0.05$ (i) $k = 8, \eta = 0.01$

Figure 10: **(Theoretical vs. Simulated sharpness-diversity trade-off in `SAM`).** This figure compares the sharpness and diversity as predicted by Theorem 1 and as observed in simulations under various parameter configurations. Results demonstrates the robustness of our theoretical analysis and tightness of the derived sharpness upper bound.

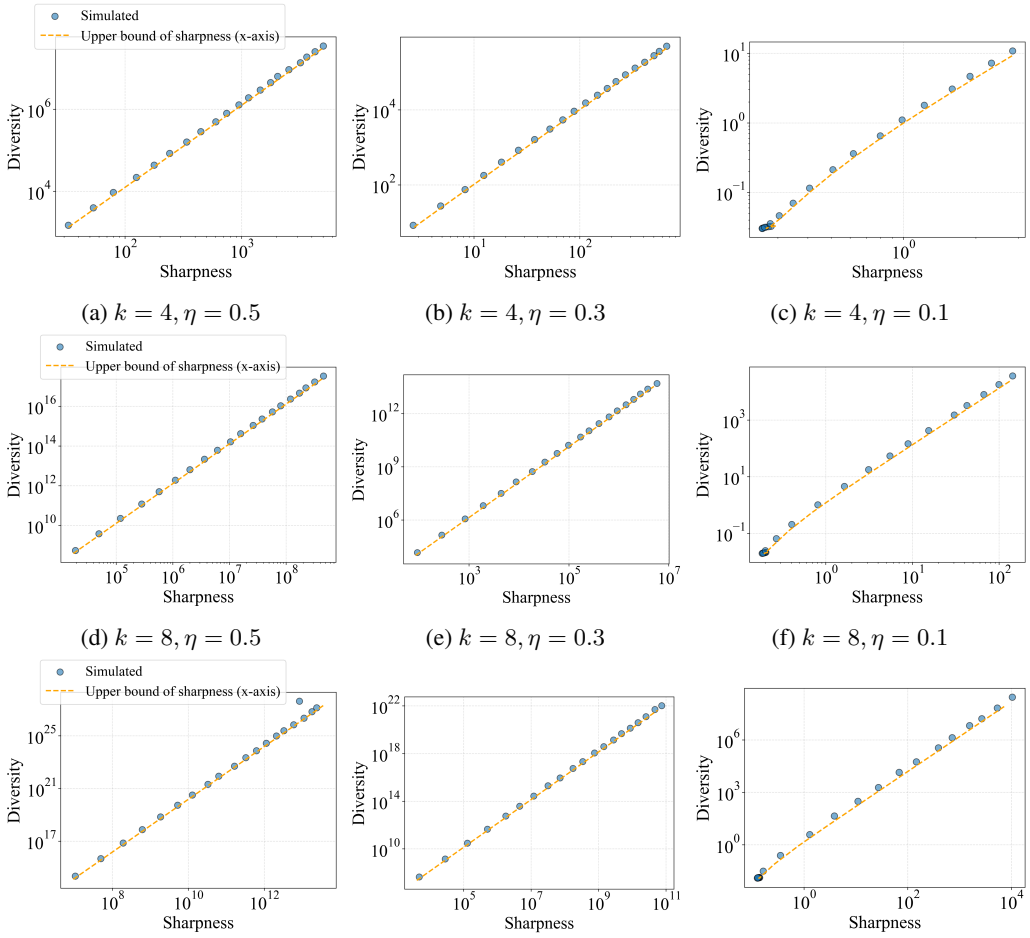
D.3 Hyperparamter setting for `SharpBalance`

Hyperparameter setting on CIFAR10/100. For experiments on CIFAR10/100, we train an NN from scratch with basic data augmentations, including random cropping, padding by four pixels, and random horizontal flipping. We use a batch size of 128, a momentum of 0.9, and a weight decay of 0.0005. For deep ensemble, we train each model for 200 epochs.

In addition, we use 10% of the training set as the validation set for selecting ρ and k based on the ensemble's performance. We make a grid search for ρ over $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. For `SharpBalance`, we use the same ρ as `SAM` and search k over $\{0.2, 0.3, 0.4, 0.5, 0.6\}$. T_d is another hyperparameter introduced by `SharpBalance`, we use a T_d of 10 for all experiments on CIFAR10, a T_d of 100 and 150 respectively when training dense and sparse models on CIFAR100. See Table 1 for the optimal ρ and k after grid search.

Hyperparameter setting on TinyImageNet. For experiments on TinyImageNet, we adopt basic data augmentations, including random cropping, padding by four pixels, and random horizontal flipping. We train each model for 200 epochs. We use a batch size of 128, a momentum of 0.9, a weight decay of 5e-4, a T_d of 100, an initial learning rate of 0.1, and decay it with a factor of 10 at Epoch 100 and 150. We search ρ and k in the same range as what we do on CIFAR10/100. See Table 1 for the optimal ρ and k after grid search.

https://doi.org/10.52202/079017-4344



(a) $k = 4, \eta = 0.5$ (b) $k = 4, \eta = 0.3$ (c) $k = 4, \eta = 0.1$

(d) $k = 8, \eta = 0.5$ (e) $k = 8, \eta = 0.3$ (f) $k = 8, \eta = 0.1$

(g) $k = 12, \eta = 0.5$ (h) $k = 12, \eta = 0.3$ (i) $k = 12, \eta = 0.1$

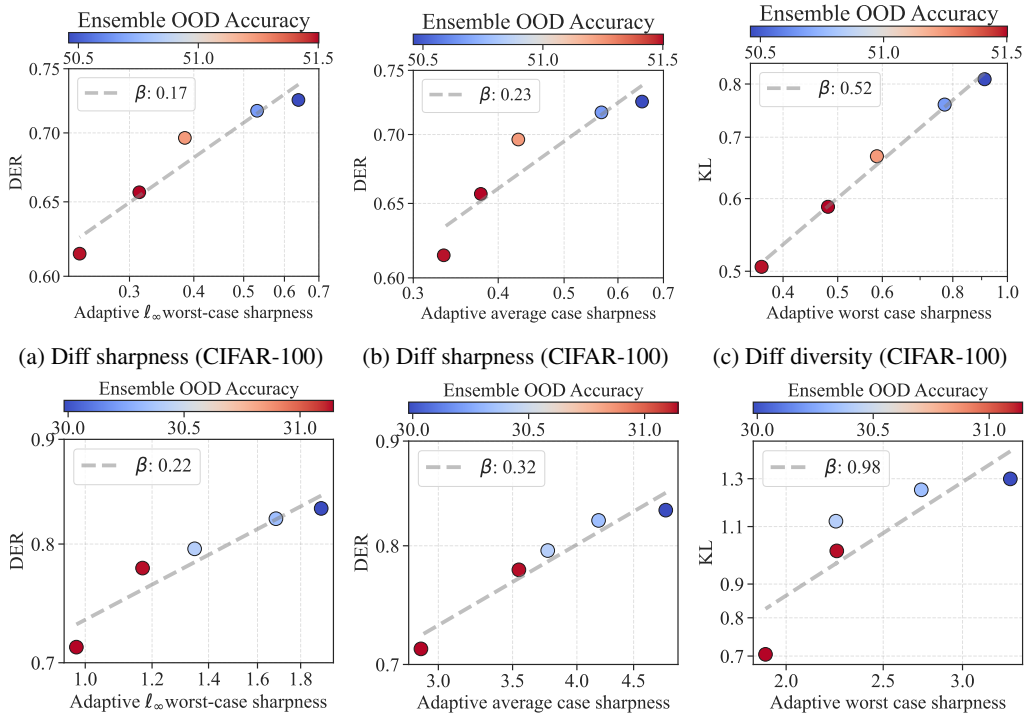
Figure 11: **(Theoretical vs. Simulated sharpness-diversity trade-off in `SharpBalance`).** This figure compares the sharpness and diversity as predicted by Theorem 2 and as observed in simulations under various parameter configurations. The observed model behaviors align closely with our derived upper bounds.

Dataset	Model	Method	ρ	k	T_d
CIFAR10	ResNet18	Deep Ensemble	-	-	-
	ResNet18	Deep Ensemble+SAM	0.2	-	-
	ResNet18	SharpBalance	0.2	0.4	100
CIFAR100	ResNet18	Deep Ensemble	-	-	-
	ResNet18	Deep Ensemble+SAM	0.2	-	-
	ResNet18	SharpBalance	0.2	0.5	100
TinyImageNet	ResNet18	Deep Ensemble	-	-	-
	ResNet18	Deep Ensemble+SAM	0.2	-	-
	ResNet18	SharpBalance	0.2	0.3	100

Table 1: Hyperparamter setting for results in Section 4.4, we report the optimal ρ and k after grid search. Each result in Figure 7 is averaged over three ensembles, which corresponds to 9 random seeds, the random seeds we use are $\{13, 17, 27, 113, 117, 127, 43, 59, 223\}$.

E Ablation studies on loss landscape metrics

In this section, we show that the sharpness-diversity trade-off generalizes to different measurements of sharpness and diversity. The results are presented in Figure 12.



(a) Diff sharpness (CIFAR-100) (b) Diff sharpness (CIFAR-100) (c) Diff diversity (CIFAR-100)

(d) Diff sharpness (TIN) (e) Diff sharpness (TIN) (f) Diff diversity (TIN)

Figure 12: **(Ablation study of varying sharpness and diversity metrics to corroborate existence of sharpness-diversity trade-off).** (a)(d) Varying sharpness metric by using the adaptive ℓ_∞ worst-case sharpness. (b)(e) Varying sharpness metric by using the adaptive ℓ_2 average case sharpness. (c)(f) Varying diversity metric by using the KL divergence. The sharpness-diversity trade-off is still observed in all the settings. The x-axis and y-axis are in log scale. The notation β stands for the slope of the linear regression function fitted on all the ensembles trained by `SAM`.

Sharpness metric. In the main paper, we use adaptive worst-case sharpness defined in equation (3), the parameter neighborhood is bounded by ℓ_2 norm. In this section, we consider two more sharpness metrics [Kwon et al., 2021, Andriushchenko et al., 2023]: adaptive worst-case sharpness with the parameter neighborhood bounded by ℓ_∞ norm (referred to as adaptive ℓ_∞ worst-case sharpness); and adaptive average case sharpness bounded by ℓ_2 norm (termed average case sharpness).
The adaptive ℓ_∞ worst-case sharpness is defined as:

$$\max_{\|T_{\boldsymbol{\theta}}^{-1}\boldsymbol{\varepsilon}\|_\infty \leq \rho_0} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}+\boldsymbol{\varepsilon}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}). \tag{9}$$

The average case sharpness is defined as:

$$\mathbb{E}_{\boldsymbol{\varepsilon}\sim\mathcal{N}\left(0,\rho_0^2\,\mathrm{diag}(T_{\boldsymbol{\theta}}{}^2)\right)} \quad \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}+\boldsymbol{\varepsilon}) - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}), \tag{10}$$

where ρ_0 is the neighborhood size of current parameter $\boldsymbol{\theta}$. $T_{\boldsymbol{\theta}}$ is a normalization operator that ensures the sharpness measure is invariant with respect to the re-scaling operation of the parameter. The results, illustrated in Figures 12, corroborate our observation of a trade-off between sharpness and diversity.

Diversity metric. We consider Kullback–Leibler (KL) Divergence [Kullback and Leibler, 1951] as an alternative diversity metric, which is also widely used in previous literature to gauge the diversity of two ensemble members [Fort et al., 2019, Liu et al., 2022]. Specifically, the KL-divergence between the outputs of two ensemble members given a data sample $(\boldsymbol{x}, \boldsymbol{y})$ is defined as:

$$\mathrm{KL}\left(f_{\boldsymbol{\theta}_1}(\boldsymbol{x}), f_{\boldsymbol{\theta}_2}(\boldsymbol{x})\right) = \mathbb{E}_{f_{\boldsymbol{\theta}_1}(\boldsymbol{x})}\left[\log f_{\boldsymbol{\theta}_1}(\boldsymbol{x}) - \log f_{\boldsymbol{\theta}_2}(\boldsymbol{x})\right]. \tag{11}$$

We measure the KL divergence on each data sample in the test data and then average the measured KL divergence. The results for KL-divergence are shown in Figure 12, which demonstrate the trade-off remains consistent for different diversity metrics.

https://doi.org/10.52202/079017-4344

F More results

F.1 Evaluation on different corruption severity

`SharpBalance` 's main advantage lies in OOD scenarios. As shown in Table 2-4, `SharpBalance` consistently outperforms the baselines on different levels of corruption.

Table 2: Results of different severity levels on CIFAR10-C.

Corruption Severity	1	2	3	4	5
Deep ensemble	88.90	83.67	77.56	70.37	58.63
Deep ensemble+SAM	89.44	84.24	78.16	71.04	58.77
`SharpBalance`	89.75 (+0.31)	84.80 (+0.56)	78.98 (+0.82)	72.25 (+1.21)	60.78 (+2.01)

Table 3: Results of different severity levels on CIFAR100-C.

Corruption Severity	1	2	3	4	5
Deep ensemble	65.78	57.77	51.30	44.33	34.16
Deep ensemble+SAM	66.39	58.47	51.89	44.90	34.81
`SharpBalance`	67.23 (+0.84)	59.53 (+1.06)	53.14 (+1.25)	46.19 (+1.29)	36.20 (+1.39)

Table 4: Results of different severity levels on Tiny-ImageNet-C.

Corruption Severity	1	2	3	4	5
Deep ensemble	43.62	36.65	28.96	22.08	16.86
Deep ensemble+SAM	45.20	38.04	30.19	22.98	17.71
`SharpBalance`	46.48 (+1.28)	39.53 (+1.49)	31.70 (+1.51)	24.27 (+1.29)	18.69 (+0.98)

F.2 Evaluation on different model architectures

We extend the evaluations on more architectures such as WideResNet (WRN), ViT, and ALBERT. Here we describe the experimental setup. For vision tasks with WRN, we trained the ensemble members from scratch on CIFAR-10 and CIFAR-100. For vision tasks with transformers, we constructed the three-member ensemble by fine-tuning the pre-trained ViT-T/16 model on the CIFAR-100 dataset, evaluated on in-distribution and CIFAR100-C test sets. For language tasks, we constructed the three-member ensemble by fine-tuning the pre-trained ALBERT-Base model on Microsoft Research Paraphrase Corpus (MRPC) dataset and evaluated the performance on its validation set. The hyperparameter search and setup are the same as in Appendix D.3.

These results in Figure 13 and Table 5 confirm that `SharpBalance` consistently boosts both ID and OOD performance across the models and datasets studied.

Method	ViT-T/16		ALBERT-B
	CIFAR100	CIFAR100-C	MRPC
Deep ensemble	88.34	66.64	89.50
Deep ensemble + SAM	88.48	66.89	89.89
`SharpBalance`	**88.68**	**67.21**	**90.11**

Table 5: **(Additional experiments on Transformer-architecture)**. The ensemble test accuracy is reported and each ensemble comprises three members. The observation is consistent with the residual network results in the main paper: SAM improves the Deep Ensemble, and `SharpBalance` outperforms both two baselines.

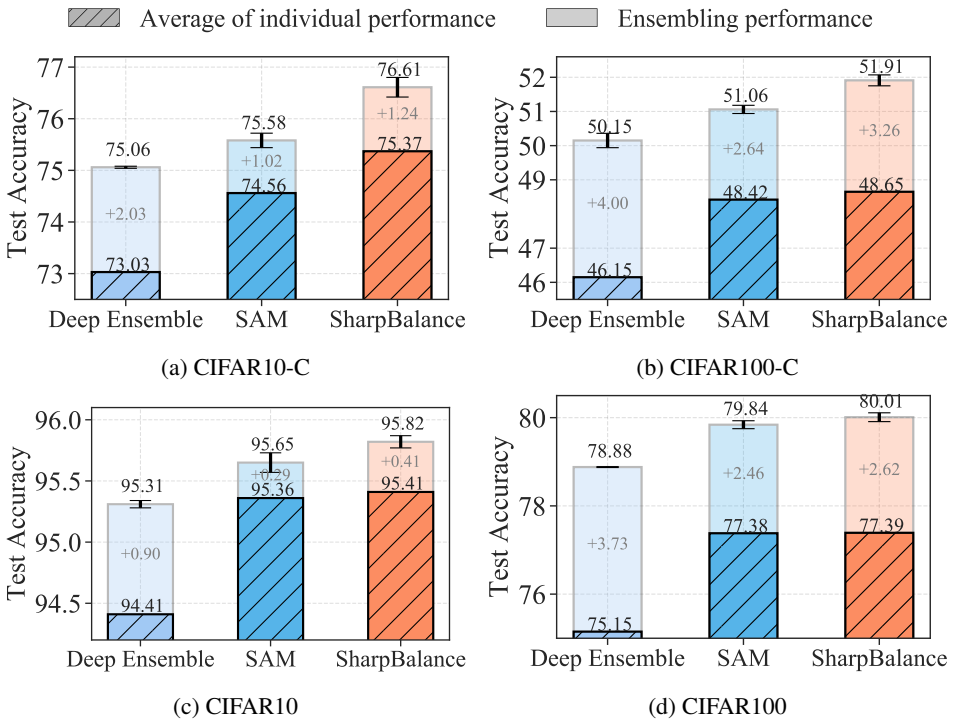


Figure 13: The three-member WRN-40-2 ensemble is trained with different methods on two datasets. The first row reports the OOD accuracy and the second row reports the ID accuracy. The lower part of each bar with the diagonal lines represents the individual model performance. The upper part of each bar represents the ensembling improvement. The results are reported by averaging three ensembles, and each ensemble is comprised of three models.

F.3 Sharpness-aware set: hard vs easy examples

`SharpBalance` aims to achieve the optimal balance by applying `SAM` to a carefully selected subset of the data while performing standard optimization on the remaining samples. In our work, sharpness is determined by the curvature of the loss around the model's weights, whereas [Garg and Roy, 2023] determines it based on the curvature of the loss around a data point. In Figure 14, we rank 1000 samples using both metrics and found a strong correlation between these two.

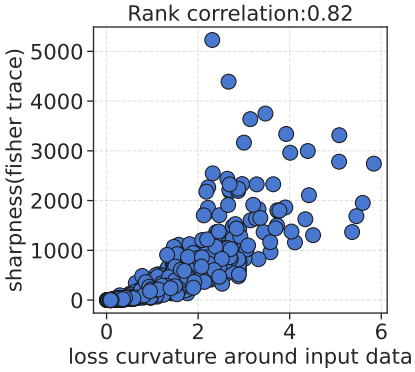


Figure 14: Rank correlation between fisher trace and loss curvature around input data

https://doi.org/10.52202/079017-4344

F.4 Comparison with more baselines

We compare SharpBalance with stronger ensemble method EoA [Arpit et al., 2022] and stronger SAM methods. We carefully tuned the hyperparameters for EoA. EoA fine-tuned a pre-trained model; and in our paper, all models are trained from scratch. We compare SharpBalance with another SAM baseline: SAM +, where three individual models are trained with different ρ values, e.g., 0.05, 0.1, and 0.2, respectively. From Table 6, SharpBalance outperforms these two baselines both in-distribution and OOD generalization.

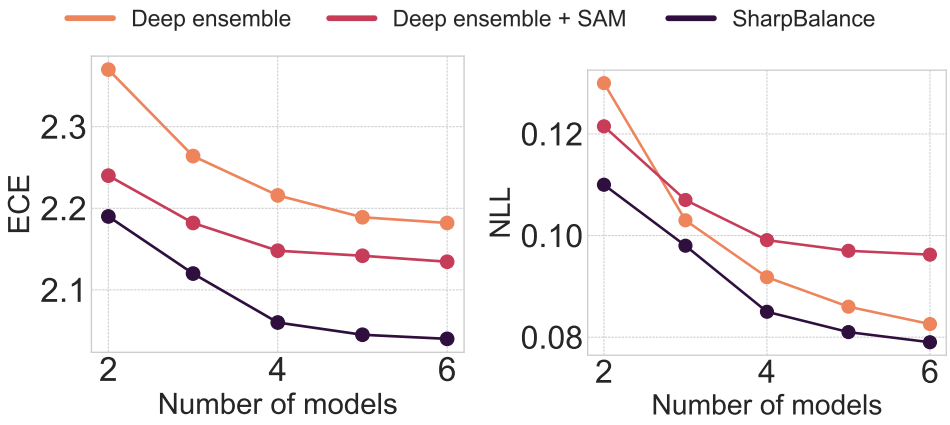
In Table 7, we combine GSAM [Zhuang et al., 2022] with Deep Ensemble as a new baseline method "Deep Ensemble + GSAM", and incorporate the GSAM into our method SharpBalance. The results show that the new baseline with GSAM outperforms the original baseline in ID and OOD performance but still underperforms SharpBalance (w/ SAM). Furthermore, we enhance SharpBalance by replacing the SAM with GSAM, which leads to better ID performance.

Dataset	Method	ACC	cACC
CIFAR10	SAM +	96.03	76.29
	EoA	95.55	75.57
	SharpBalance	96.18 (+0.15)	77.32 (+1.03)
CIFAR100	SAM +	79.67	51.28
	EoA	79.53	51.45
	SharpBalance	79.84 (+0.17)	52.46 (+1.01)

Table 6: SharpBalance outperforms EoA and SAM + both in-distribution and OOD generalization on CIFAR10 and CIFAR100.

Method	CIFAR100 Acc	CIFAR100-C Acc
Deep Ensemble	79.28	50.67
Deep Ensemble + SAM	79.50	51.28
SharpBalance (SAM)	79.84	**52.46**
Deep Ensemble + GSAM	79.74	51.37
SharpBalance (GSAM)	**80.01**	51.92

Table 7: **(Comparing our method SharpBalance with stronger SAM baseline).** The ensemble test accuracy is reported and each ensemble comprises three members. GSAM improves the original baseline method with SAM and SharpBalance. The model is ResNet18.



(a) Expected Calibration Error (ECE) (b) Negative Log Likelihood (NLL)

Figure 15: **(Uncertainty metrics on CIFAR100-C).** "ECE" represents expected calibration error, and "NLL" represents negative log-likelihood. Both metrics are lower the better. The model architecture is ResNet-18. The uncertainty metrics demonstrate the superior performance of SharpBalance. x-axis represents the number of individual models in one ensemble.

F.5 Evaluation on uncertainty metrics

In Figure 15, we present the results of uncertainty metrics, i.e., negative log-likelihood and expected calibration error. These uncertainty metrics exhibit trends similar to the accuracy metrics: "Deep Ensemble + SAM" outperforms "Deep Ensemble", and our method outperforms both baselines. The experiments were conducted using ResNet-18 on CIFAR100, with metrics reported on corrupted datasets. Additionally, we observe that both metrics improve as the number of ensemble members increases for all three methods.

G Experiments Compute Resources

All codes are implemented in PyTorch, and the experiments are conducted on 3 Nvidia Quadro RTX 6000 GPUs for training an ensemble of 3 models. Compared to `SAM`, our method adds a minimal computational cost. The extra time comes from using Fisher trace to compute the per-sample sharpness. Therefore, computing the per-sample sharpness requires one single forward pass and one backward pass. We report the additional training cost in Table 8. `SharpBalance` only increases the training time by 1%: $0.83\ (84.48 - 0.83) \times 100\% \approx 1\%$.

Additional training cost	Total training cost
0.83 min	84.48 min

Table 8: Additional training cost introduced by `SharpBalance`. We train a ResNet18 on CIFAR10 for 200 epochs.

https://doi.org/10.52202/079017-4344

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main claims made in the abstract and introduction accurately reflect the scope and contribution of the paper, supported by our theoretical and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Proofs of all theoretical results are provided in Appendix C and the assumptions are clearly stated in the theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all relevant information including detailed description of the algorithms, datasets, experimental set up in Section 4 and hyperparameters in Appendix D to reproduce the main experimental results claimed in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-4344

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The implementation can be found through the anonymous github repository and the zip file uploaded as the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings with hyperparameters are provided in both Section 4 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The results are accompanied by error bars, for the experiments in Section 4.4 that support the main claims of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources for experiments can be found in Appendix G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The impact statement of the study can be found in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-4344

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original papers that produced the code or dataset are appropriately cited in this work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-4344

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.