AGNOSTIC SHARPNESS-AWARE MINIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sharpness-aware minimization (SAM) has been instrumental in improving deep neural network training by minimizing both the training loss and the sharpness of the loss landscape, leading the model into flatter minima that are associated with better generalization properties. In another aspect, Model-Agnostic Meta-Learning (MAML) is a framework designed to improve the adaptability of models. MAML optimizes a set of meta-models that are specifically tailored for quick adaptation to multiple tasks with minimal fine-tuning steps and can generalize well with limited data. In this work, we explore the connection between SAM and MAML in enhancing model generalization. We introduce Agnostic-SAM, a novel approach that combines the principles of both SAM and MAML. Agnostic-SAM adapts the core idea of SAM by optimizing the model toward wider local minima using training data, while concurrently maintaining low loss values on validation data. By doing so, it seeks flatter minima that are not only robust to small perturbations but also less vulnerable to data distributional shift problems. Our experimental results demonstrate that Agnostic-SAM significantly improves generalization over baselines across a range of datasets and under challenging conditions such as noisy labels or data limitation.

1 INTRODUCTION

Deep neural networks have become the preferred method for analyzing data, surpassing traditional machine learning models in complex tasks such as classification. These networks process input through numerous parameters and operations to predict classes. The learning process involves finding parameters within a model space that minimize errors or maximize performance for a given task. Typically, training data, denoted as S, is finite and drawn from an unknown true data distribution $\mathcal{D}$. Larger or more aligned training sets lead to more efficient models.

Despite their ability to learn complex patterns, deep learning models can also capture noise or random fluctuations in training data, leading to overfitting. This results in excellent performance on training data but poor predictions on new, unseen data, especially with domain shifts. Generalization, measured by comparing prediction errors on S and $\mathcal{D}$, becomes crucial. Balancing a model's ability to fit training data with its risk of overfitting is a key challenge in machine learning.

Several studies have been done on this problem, both theoretically and practically. Statistical learning theory has proposed different complexity measures that are capable of controlling generalization errors (Vapnik, 1998; Bartlett & Mendelson, 2003; Mukherjee et al., 2002; Bousquet & Elisseeff, 2002; Poggio et al., 2004). In general, they develop a bound for general error on $\mathcal{D}$. Theory suggests that minimizing the intractable general error on $\mathcal{D}$ is equivalent to minimizing the empirical loss on S with some constraints to the complexity of models and training size (Alquier et al., 2016b). An alternative strategy for mitigating generalization errors involves the utilization of an optimizer to learn optimal parameters for models with a specific local geometry. This approach enables models to locate wider local minima, known as flat minima, which makes them more robust against data shift between training and testing sets (Jiang et al., 2020; Petzka et al., 2021; Dziugaite & Roy, 2017).

The connection between a model's generalization and the width of minima has been investigated theoretically and empirically in many studies, notably (Hochreiter & Schmidhuber, 1994; Neyshabur et al., 2017; Dinh et al., 2017; Fort & Ganguli, 2019). A specific method within this paradigm is Sharpness-aware Minimisation (SAM) (Foret et al., 2021), which has emerged as an effective technique for enhancing the generalization ability of deep learning models. SAM seeks a perturbed model within the vicinity of a current model that maximizes the loss over a training set. Eventually, SAM leads the model to the region where both the current model and its perturbation model have low loss values, which ensure flatness. The success of SAM and its variants (Kwon et al., 2021; Kim et al., 2022; Truong et al., 2023) has inspired further investigation into its formulation and behavior, as evidenced by recent works such as (Kaddour et al., 2022; Möllenhoff & Khan, 2022; Andriushchenko & Flammarion, 2022).

SAM significantly enhances robustness against shifts between training and testing datasets, thereby reducing overfitting and improving overall performance across different datasets and domains. This robust optimization approach aligns particularly well with the principles of Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017). MAML aims to find a set of meta-model parameters that not only generalize well on current tasks but can also be quickly adapted to a wide range of new tasks. Furthermore, the agnostic perspective of MAML is particularly enticing for enhancing generalization ability because it endeavors to learn the optimal meta-model from meta-training sets capable of achieving minimal losses on independent meta-testing sets, thus harmonizing with the goal of generalization.

In this paper, inspired by MAML and leveraging SAM, we initially approach the problem of learning the best model over a training set from an agnostic viewpoint. Subsequently, we harness this perspective with sharpness-aware minimization to formulate an agnostic optimization problem. However, a naive solution akin to MAML does not suit our objectives. We propose a novel solution for this agnostic optimization problem, resulting in an approach called *AgnosticSAM*. In summary, our contributions to this work are as follows:

- We proposed a framework inspired by SAM and MAML works, called Agnostic-SAM to improve model flatness and robustness against noise. Agnostic-SAM updates a model to a region that minimizes the sharpness on the training set while also implicitly performing well on the validation set by using a combination of gradients on both training and validation sets.
- We demonstrate the effectiveness of Agnostic-SAM in improving generalization performance. Our initial examination focuses on image classification tasks, including training from scratch and transfer learning across a range of datasets, from small to large scale. We also extend this experiment under noisy label conditions with varying levels of noise. Additionally, we apply Agnostic-SAM in MAML settings to validate the effectiveness of our method in generalizing beyond the meta-training tasks and its adaptability across different domains. The consistent improvement in performance across experiments indicates that Agnostic-SAM not only enhances robustness against label noise and improves the model's generalization across diverse tasks, but also contributes to more stable and reliable predictions in different settings.

2 Related works

Sharpness-Aware Minimization. The correlation between the wider minima and the generalization capacity has been extensively explored both theoretically and empirically in various studies (Jiang et al., 2020; Petzka et al., 2021; Dziugaite & Roy, 2017). Many works suggested that finding flat minimizers might help to reduce generlization error and increase robustness to data distributional shift problems in various settings (Jiang et al., 2020; Petzka et al., 2021; Dziugaite & Roy, 2017). There are multiple works have explored the impact of different training parameters, including batch size, learning rate, gradient covariance, and dropout, on the flatness of discovered minima such as (Keskar et al., 2017; Jastrzebski et al., 2017; Wei et al., 2020).

Sharpness-aware minimization (SAM) (Foret et al., 2021) is a recent optimization technique designed to improve the generalization error of neural networks by considering the sharpness of the loss landscape during training. SAM minimizes the worst-case loss around the current model and effectively updates models towards flatter minima to achieve low training loss and maximize generalization performance on new and unseen data. SAM has been successfully applied to various tasks and domains, such as vision models (Chen et al., 2021), language models (Bahri et al., 2022), federated learning (Qu et al., 2022), Bayesian Neural Networks (Nguyen et al., 2023), domain generalization (Cha et al., 2021), multi-task learning (Phan et al., 2022) and meta-learning bi-level optimization (Abbas et al., 2022). In Abbas et al. (2022), authors discussed SAM's effectiveness in enhancing meta-learning bi-level optimization, while SAM's superior convergence rates in federated learning compared to existing approaches in Qu et al. (2022) along with proposing a generalization bound for the global model. Additionally, multiple varieties of SAM have been proposed (Kwon et al., 2021), (Li et al., 2024), (Du et al., 2022) to tackle the different problems of the original method.

Model-Agnostic Machine Learning. Model-agnostic machine learning techniques have significant advances and offer flexible solutions applicable across various models and tasks. In which, MAML (Finn et al., 2017) stands out as the most compelling model-agnostic technique that formulates meta-learning as an optimization problem, enabling models to improve the model ability to quickly adapt to new tasks with minimal task-specific modifications or limited additional data. Subsequent research has largely focused on addressing the computational challenges of MAML (Chen et al., 2023; Wang et al., 2023) or proposing novel approaches that exploit the concept of model agnostic from MAML across a wide range of tasks, including non-stationary environments (Al-Shedivat et al., 2018), alternative optimization strategies (Rajeswaran et al., 2019), and uncertainty estimation for robust adaptation (Finn et al., 2018). Recently, Abbas et al. (2022) analyzed the loss-landscape of MAML models and proposed the integration of SAM in training to improve the generalization performance of a meta-model.

3 PROPOSED FRAMEWORK

3.1 NOTIONS

We start by introducing the notions used throughout our paper. We denote $\mathcal{D}$ as the data/label distribution to generate pairs of data/label (x, y). Given a model with the model parameter θ, we denote the per sample loss induced by (x, y) as $\ell(x, y; \theta)$. Let S be the training set drawn from the distribution $\mathcal{D}$. We denote the empirical and general losses as $\mathcal{L}_S(\theta) = \mathbb{E}_S[\ell(x, y; \theta)]$ and $\mathcal{L}_\mathcal{D}(\theta) = \mathbb{E}_\mathcal{D}[\ell(x, y; \theta)]$ respectively. We define $\mathcal{L}_\mathcal{D}(\theta \mid S)$ as an *upper bound defined over* S of the general loss $\mathcal{L}_\mathcal{D}(\theta)$. Note that inspired by SAM (Foret et al., 2021), we use the sharpness over S to define $\mathcal{L}_\mathcal{D}(\theta \mid S)$. Finally, we use $|A|$ to denote the cardinality of a set A.

3.2 PROBLEM FORMULATION

Given a training set S^t whose examples are sampled from $\mathcal{D}$ (i.e., $S^t \sim \mathcal{D}^{N_t}$ with $N_t = |S^t|$), we use $\mathcal{L}_\mathcal{D}(\theta \mid S^t)$ to train models. Among the models that minimize this loss, we select the one that minimizes the general loss as follows:

$$\min_{\theta^*} \mathcal{L}_\mathcal{D}(\theta^*) \text{ s.t. } \theta^* \in \mathcal{A}_\mathcal{D}\left(S^t\right) = \text{argmin}_\theta \mathcal{L}_\mathcal{D}\left(\theta \mid S^t\right). \tag{1}$$

The reason for the formulation in (1) is that although $\mathcal{L}_\mathcal{D}(\theta \mid S^t)$ is an upper bound of the general loss $\mathcal{L}_\mathcal{D}(\theta)$, there always exists a gap between them. Therefore, the additional outer minimization helps to refine the solutions. We now denote S^v (i.e., $S^v \sim \mathcal{D}^{N_v}$ with $N_v = |S^v|$) as a valid set sampled from $\mathcal{D}$. With respect to this valid set, we have the following theorem.

Theorem 1. *Denote $\mathcal{L}_\mathcal{D}(\theta \mid S) := \max_{\theta' : \|\theta' - \theta\|_2 \le \rho} \mathcal{L}_S(\theta')$. Under some mild condition similar to SAM (Foret et al., 2021), with a probability greater than $1 - \delta$ (i.e., $\delta \in [0, 1]$) over the choice of $S^v \sim \mathcal{D}^{N_v}$, we then have for any optimal models $\theta^* \in \mathcal{A}_\mathcal{D}(S^t)$:*

$$\mathcal{L}_\mathcal{D}(\theta^*) \le \mathcal{L}_\mathcal{D}(\theta^* \mid S^v) + \frac{4L}{\sqrt{N_v}} \left[k \log\left(1 + \frac{\|\theta^*\|^2}{\rho}\left(1 + \sqrt{\log N_v / k}\right)\right) + 2\sqrt{\log \frac{N_v + k}{\delta} + O(1)} \right], \tag{2}$$

where L is the upper-bound of the loss function (i.e., $\ell(x, y; \theta) \le L, \forall x, y, \theta$), k is the model size, and $\rho > 0$ is the perturbation radius.

Our theorem 1 (proof can be found in Appendix A.1) can be viewed as an extension of Theorem 1 in Foret et al. (2021), where we apply the Bayes-PAC theorem from Alquier et al. (2016a) to prove an upper bound for the general loss of any bounded loss, instead of the 0-1 loss in Foret et al. (2021).

We can generalize this proof for S^t to explain why we use $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^t\right) := \max_{\theta':\|\theta'-\theta\|_2\le\rho} \mathcal{L}_{S^t}\left(\theta'\right)$ as an objective to minimize, as in (1). Based on Theorem 1, we can rewrite the objectives in (1) as:

$$\min_{\theta^*} \mathcal{L}_{\mathcal{D}}\left(\theta^* \mid S^v\right) \text{ s.t. } \theta^* \in \mathcal{A}_{\mathcal{D}}\left(S^t\right) = \operatorname{argmin}_\theta \mathcal{L}_{\mathcal{D}}\big(\theta \mid S^t\big), \tag{3}$$

where $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S\right) := \max_{\theta':\|\theta'-\theta\|_2\le\rho} \mathcal{L}_{S}\left(\theta'\right)$.

3.3 Our Solution

Our motivation here is to primarily optimize the loss over the training set S^t, while using S^v to further enhance the generalization ability. Our agnostic formulation has the same form as MAML (Finn et al., 2017), developed for meta-learning. Inspired by MAML, a naive approach would be to consider $\theta^* = \theta^*(\theta)$ and then minimize $\mathcal{L}_{\mathcal{D}}\left(\theta^*\left(\theta\right) \mid S^v\right)$ with respect to θ. However, this naive approach does not align with our objective, as it mainly focuses on optimizing the loss $\mathcal{L}_{\mathcal{D}}\left(\theta^*\left(\theta\right) \mid S^v\right)$ over the validation set S^v.

We interpret the bi-level optimization problem in (3) as follows: at each iteration, our primary objective is to optimize $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^t\right)$, primarily based on its gradients, in such a way that future models are able to implicitly perform well on S^v. To achieve this, similar to SAM (Foret et al., 2021), we approximate $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^t\right) = \max_{\|\theta'-\theta\|\le\rho} \mathcal{L}_{S^t}\left(\theta'\right) \approx \mathcal{L}_{S^t}\left(\theta + \eta_1 \nabla \mathcal{L}_{S^t}\left(\theta\right)\right)$ for a sufficient small learning rate $\eta_1 > 0$ (i.e., $\eta_1 \|\nabla \mathcal{L}_{S^t}\left(\theta\right)\| \le \rho$) and $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^v\right) = \max_{\|\theta'-\theta\|\le\rho} \mathcal{L}_{S^v}\left(\theta'\right) \approx \mathcal{L}_{S^v}\left(\theta + \eta_2 \nabla \mathcal{L}_{S^v}\left(\theta\right)\right)$ for a sufficient small learning rate $\eta_2 > 0$ (i.e., $\eta_2 \|\nabla \mathcal{L}_{S^v}\left(\theta\right)\| \le \rho$).

At each iteration, while primarily using the gradients of $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^t\right)$ for optimization, we also utilize the gradient of $\mathcal{L}_{\mathcal{D}}\left(\theta \mid S^v\right)$ in an auxiliary manner to ensure congruent behavior between these two gradients. Specifically, at the l-th iteration, we update as follows:

$$\tilde{\theta}_l^v = \theta_l + \eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\theta_l\right), \tag{4}$$

$$\tilde{\theta}_l^t = \theta_l + \eta_1 \nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) - \eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right), \tag{5}$$

$$\theta_{l+1} = \theta_l - \eta \nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right), \tag{6}$$

where $\eta_1 > 0, \eta_2 > 0$, and $\eta > 0$ are the learning rates, while $\mathcal{L}_{B^t}\left(\theta_l\right)$ and $\mathcal{L}_{B^v}\left(\theta_l\right)$ represent the empirical losses over the mini-batches $B^t \sim S^t$ and $B^v \sim S^v$ respectively.

According to the update in (6) (i.e., $\theta_{l+1} = \theta_l - \eta \nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$), θ_{l+1} is updated to minimize $\mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$. We now do first-order Taylor expansion for $\mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$ as

$$\mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) = \mathcal{L}_{B_t}\left(\theta_l\right) + \eta_1 \|\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right)\|_2^2 - \eta_2 \nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right), \tag{7}$$

where $\cdot$ specifies the dot product.

From (7), we reach the conclusion that the update in (6) (i.e., $\theta_{l+1} = \theta_l - \eta \nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$) aims to *minimize* simultaneously (i) $\mathcal{L}_{B_t}\left(\theta_l\right)$, (ii) $\|\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right)\|_2^2$, and *maximize* (iii) $\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$. While the effects in (i) and (ii) are similar to SAM (Foret et al., 2021), maximizing $\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$) encourages two gradients of the losses over B^t and B^v to become more congruent.

Theorem 2. *For sufficiently small learning rates* $\eta_1 \le \frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{12\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}(\theta_l) \nabla_\theta \mathcal{L}_{B^t}(\theta_l)\right|}$ *and* $\eta_2 \le \min\left\{\frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}(\theta_l) \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}, \frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^v}\left(\tilde{\theta}_l^v\right) \nabla_\theta \mathcal{L}_{B^t}(\theta_l)\right|}\right\}$*, we have*

$$\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) \ge \begin{cases} \frac{1}{2} \nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) & \textit{if}\, \nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) \ge 0 \\ \frac{3}{2} \nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) & \textit{otherwise} \end{cases} \tag{8}$$

Theorem 2 (proof can be found in Appendix A.1) indicates that two gradients $\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$ and $\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ are encouraged to be more congruent since our update aims to maximize its lower bound $c \times \nabla_\theta \mathcal{L}_{B^t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ (i.e., $c = 0.5$ or $c = 1.5$). Notice that the negative gradient $-\eta \nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$ is used to update θ_l to θ_{l+1}, hence this update can have an implicit impact on minimizing $\mathcal{L}_{\mathcal{D}}(\theta \mid S^v)$ since the negative gradient $-\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ targets to minimize $\mathcal{L}_{\mathcal{D}}(\theta \mid S^v) = \max_{\|\theta' - \theta\| \leq \rho} \mathcal{L}_{S^v}(\theta') \approx \mathcal{L}_{S^v}(\theta + \eta_2 \nabla \mathcal{L}_{S^v}(\theta))$.

Practical Algorithm. Inspired by SAM Foret et al. (2021), we set $\eta_1 = \rho_1 \frac{\nabla_\theta \mathcal{L}_{B^t}(\theta_l)}{\|\nabla_\theta \mathcal{L}_{B^t}(\theta_l)\|_2}$ and $\eta_2 = \rho_2 \frac{\nabla_\theta \mathcal{L}_{B^v}(\theta_l)}{\|\nabla_\theta \mathcal{L}_{B^v}(\theta_l)\|_2}$, where $\rho_1 > 0$ and $\rho_2 > 0$ are perturbation radius. Furthermore, instead of splitting the training set S into two fixed subsets, S^t and S^v, which reduces the number of training samples, we set $S^t = S^v = S$, allowing the entire training set to be used for updating the model. This approach is especially beneficial for training on small datasets. Optionally, we apply momentum with a factor β to approximate the gradient of the full validation set using gradients from mini-batches. The effectiveness of this term will be discussed in section 5.

The pseudo-code of Agnostic-SAM is summarized in Algorithm 1.

Algorithm 1 Pseudo-code of Agnostic-SAM

Input: $\rho_1, \rho_2, \eta, \beta$, the number of iterations L_{iter}, and the training set S. Initialize gradient on the validation set $g_v \leftarrow 0$
Output: the optimal model θ_L.
for $l = 1$ to L_{iter} **do**
  Sample mini-batch $B^t \sim S^t$, $B^v \sim S^v$.
  Compute $\tilde{\theta}_l^v = \theta_l + \rho_2 \frac{\nabla_\theta \mathcal{L}_{B^v}(\theta_l)}{\|\nabla_\theta \mathcal{L}_{B^v}(\theta_l)\|_2}$
  $g_v \leftarrow \beta g_v + (1-\beta) \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$
  Compute $\tilde{\theta}_l^t \leftarrow \theta_l + \rho_1 \frac{\nabla_\theta \mathcal{L}_{B^t}(\theta_l)}{\|\nabla_\theta \mathcal{L}_{B^t}(\theta_l)\|_2} - \rho_2 \frac{g_v}{\|g_v\|_2}$.
  Compute $\theta_{l+1} \leftarrow \theta_l - \eta \nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$.
end for

4 Experiments

In this section, we present the results of various experiments to evaluate the effectiveness of our Agnostic-SAM, including training from scratch, transfer learning on different dataset sizes, learning with noisy labels, and MAML setting. For all experiments of Agnostic-SAM, we consistently use a fixed value of momentum factor $\beta = 0.9$ and mini-batch size of validation set $4|B^v| = |B^t|$. The effectiveness of these hyper-parameters on performance and training complexity will be discussed in Section 5.

4.1 Image Classification From Scratch

We first conduct experiments on ImageNet, Food101, and CIFAR datasets with standard image classification settings trained from scratch. The performance is compared with baseline models trained with the SGD, SAM, ASAM, and the integration of ASAM and Agnostic-SAM.

ImageNet dataset We use ResNet18 and ResNet34 models for experiments on the ImageNet dataset, with an input size of 224×224. For all experiments of Agnostic-SAM and its variations, we consistently set $\rho_1 = 2\rho_2 = 2\rho$, where ρ represents the perturbation radius for the respective SAM method. Specifically, in this experiment, we set $\rho = 0.1$, $\rho_1 = 0.2$, and $\rho_2 = 0.1$. The models are trained for 200 epochs with basic data augmentations (random cropping, horizontal flipping, and normalization). We use an initial learning rate of 0.1, a batch size of 2048 for the training set, and 512 for the validation set, following a cosine learning schedule across all experiments in this paper. We extend this experiment to the mid-sized Food101 dataset using the same settings, except for a batch size of 128 for the training set and 32 for the validation set. Performance results are detailed in Table 1.

Table 1: Classification accuracy on the ImageNet and Food101 datasets. All models are trained from scratch with 200 epochs.

Dataset	Method	Resnet18		Resnet34	
		Top-1	Top-5	Top-1	Top-5
ImageNet	SAM	62.46	84.19	63.73	84.95
	Agnostic-SAM	**63.64**	**85.22**	**65.89**	**86.84**
Food101	SAM	73.15	89.85	73.87	90.84
	Agnostic-SAM	**73.45**	**90.35**	**74.47**	**91.27**

CIFAR dataset We used three architectures: WideResnet28x10, Pyramid101, and Densenet121 with an input size of 32×32 for CIFAR datasets. To replicate the baseline experiments, we followed the hyperparameters provided in the original papers. Specifically, for CIFAR-100, we set $\rho = 0.1$, $\rho_1 = 0.2$, and $\rho_2 = 0.1$, and for CIFAR-10, we used $\rho = 0.05$, $\rho_1 = 0.1$, and $\rho_2 = 0.05$. The same procedure and settings were applied to ASAM and Agnostic-ASAM, with the perturbation radius ρ for ASAM being 10 times larger than that of the SAM method. Other training configurations are consistent with those used in the ImageNet experiments, except for data augmentations (horizontal flipping, four-pixel padding, and random cropping). Results are reported in Tables 2, while the SGD results are referenced from Foret et al. (2021).

Our proposed method consistently outperforms the baselines across various settings. On both ImageNet and Food101, it significantly surpasses the baselines, with a notable improvement in both Top-1 and Top-5 accuracy. For CIFAR-10, performance is close to the saturation point, making further improvements challenging. Nevertheless, Agnostic-SAM achieves slight enhancements across all cases. On CIFAR-100, where models are more prone to overfitting compared to CIFAR-10, Agnostic-SAM still delivers competitive results.

Table 2: Classification accuracy on the CIFAR datasets. All models are trained from scratch three times with different random seeds and we report the mean and standard deviation of accuracies.

Dataset	Method	WideResnet28x10	Pyramid101	Densenet121
CIFAR-100	SGD	81.20 ± 0.200	80.30 ± 0.300	-
	SAM	83.00 ± 0.035	81.99 ± 0.636	68.72 ± 0.409
	Agnostic-SAM	$\mathbf{83.49 \pm 0.049}$	$\mathbf{82.38 \pm 0.282}$	$\mathbf{69.10 \pm 0.311}$
	ASAM	83.16 ± 0.296	82.02 ± 0.134	69.62 ± 0.120
	Agnostic-ASAM	$\mathbf{83.68 \pm 0.042}$	$\mathbf{82.29 \pm 0.183}$	$\mathbf{69.79 \pm 0.339}$
CIFAR-10	SGD	96.50 ± 0.100	96.00 ± 0.100	-
	SAM	96.87 ± 0.027	96.17 ± 0.174	91.28 ± 0.241
	Agnostic-SAM	$\mathbf{96.88 \pm 0.007}$	$\mathbf{96.47 \pm 0.219}$	$\mathbf{91.31 \pm 0.707}$
	ASAM	96.91 ± 0.063	96.45 ± 0.042	$\mathbf{92.04 \pm 0.240}$
	Agnostic-ASAM	$\mathbf{97.15 \pm 0.063}$	$\mathbf{96.73 \pm 0.261}$	92.02 ± 0.000

4.2 TRANSFER LEARNING

In this subsection, we further evaluate Agnostic-SAM in the transfer learning setting using the ImageNet pre-trained models to fine-tune both small-size, mid-size, and large-size datasets. All initialized weights are available on the Pytorch library.

Table 3: Transfer learning on ImageNet with Resnet models.

Model	Top-1 Acc		Top-5 Acc	
	SAM	Agnostic-SAM	SAM	Agnostic-SAM
Resnet18	70.52	**70.88**	89.60	**89.94**
Resnet34	73.06	**73.84**	91.29	**91.81**
Resnet50	75.17	**75.91**	92.58	**92.83**

First, we conduct experiments on ImageNet by using three models from the Resnet family. These base models are both pre-trained on ImageNet by SGD and then fine-tuned for 50 epochs by SAM or Agnostic-SAM with a learning rate of 0.01. We $\rho = 0.05$ for SAM, and $\rho_1 = 2\rho_2 = 0.1$ for Agnostic-SAM and basic augmentation techniques, which are the same as training from the scratch setting. Results reported in Table 3 show that our methods outperform baselines with a significant gap in both top-1 and top-5 accuracies.

Next, we examine this setting on small and mid-sized datasets on three models of the EfficientNet family. We fine-tune with a learning rate of 0.05 in 50 epochs and use $\rho = 0.1$ for all experiments of SAM (as accuracies tend to decrease when reducing ρ), $\rho_1 = 2\rho_2 = 0.1$ for all experiments of Agnostic-SAM. In Table 4, Agnostic-SAM achieves a noticeable improvement compared to most of the baselines on all small-size, mid-size, and large-size datasets, demonstrating its robustness and stability across various experiment settings.

4.3 Train With Noisy Label

In addition to mitigating data shifts between training and testing datasets, we evaluate the robustness of Agnostic-SAM against noisy labels on standard training procedure. Specifically, we adopt a classical noisy-label setting for CIFAR-10 and CIFAR-100, in which a portion of the training set's labels are symmetrically flipped with noise fractions {0.2, 0.4, 0.6, 0.8}, while the testing set's labels remain unchanged.

Table 4: Transfer learning accuracy of small and medium datasets. All models are fine-tuned from pre-trained weights on ImageNet.

Dataset	Top-1 Acc			Top-5 Acc		
	SGD	SAM	Agnostic-SAM	SGD	SAM	Agnostic-SAM
EfficientNet-B2						
Stanford Cars	89.14 ± 0.11	89.68 ± 0.17	**90.34 ± 0.07**	97.60 ± 0.20	98.04 ± 0.07	**98.24 ± 0.09**
FGVC-Aircraft	85.83 ± 0.23	86.25 ± 0.36	**87.27 ± 0.27**	95.72 ± 0.02	95.87 ± 0.06	**96.05 ± 0.03**
Oxford IIIT Pets	92.17 ± 0.19	92.34 ± 0.11	**92.58 ± 0.17**	99.23 ± 0.02	99.35 ± 0.02	**99.35 ± 0.07**
Flower102	95.06 ± 0.01	95.22 ± 0.14	**95.56 ± 0.10**	99.08 ± 0.18	99.11 ± 0.19	**99.27 ± 0.02**
Food101	83.50 ± 0.01	85.12 ± 0.07	**85.51 ± 0.02**	96.10 ± 0.32	96.83 ± 0.08	**97.14 ± 0.00**
Country211	11.94 ± 0.14	12.48 ± 0.03	**13.28 ± 0.00**	23.70 ± 0.13	25.49 ± 0.07	**26.95 ± 0.16**
EfficientNet-B3						
Stanford Cars	89.01 ± 0.19	89.40 ± 0.09	**90.09 ± 0.14**	97.73 ± 0.21	98.03 ± 0.07	**98.13 ± 0.01**
FGVC-Aircraft	84.88 ± 0.08	85.19 ± 0.11	**85.99 ± 0.25**	95.53 ± 0.12	95.67 ± 0.00	**96.08 ± 0.10**
Oxford IIIT Pets	92.68 ± 0.25	92.58 ± 0.02	**92.75 ± 0.19**	99.00 ± 0.01	99.19 ± 0.05	**99.20 ± 0.11**
Flower102	94.59 ± 0.10	94.73 ± 0.14	**95.16 ± 0.26**	98.95 ± 0.08	99.12 ± 0.16	**99.18 ± 0.07**
Food101	83.75 ± 0.12	85.79 ± 0.13	**86.17 ± 0.13**	96.22 ± 0.02	97.12 ± 0.00	**97.38 ± 0.00**
Country211	12.96 ± 0.01	13.38 ± 0.09	**13.63 ± 0.05**	26.11 ± 0.56	25.78 ± 0.08	**26.71 ± 0.26**
EfficientNet-B4						
Stanford Cars	84.72 ± 0.04	85.08 ± 0.16	**85.79 ± 0.32**	96.41 ± 0.07	96.45 ± 0.01	**96.77 ± 0.00**
FGVC-Aircraft	79.95 ± 0.61	79.96 ± 0.04	**80.80 ± 0.51**	94.87 ± 0.08	94.65 ± 0.08	**94.95 ± 0.01**
Oxford IIIT Pets	91.89 ± 0.13	92.02 ± 0.23	**92.02 ± 0.00**	99.28 ± 0.10	99.43 ± 0.07	**99.44 ± 0.02**
Flower102	92.73 ± 0.04	93.02 ± 0.14	**93.03 ± 0.16**	98.49 ± 0.07	**98.68 ± 0.02**	98.59 ± 0.05
Food101	84.55 ± 0.14	86.13 ± 0.06	**86.15 ± 0.44**	96.31 ± 0.03	97.07 ± 0.01	**97.22 ± 0.02**
Country211	14.63 ± 0.09	14.80 ± 0.13	**15.10 ± 0.16**	27.60 ± 0.00	29.09 ± 1.77	**28.38 ± 0.14**

Table 5: Results under label noise on CIFAR dataset with ResNet32. Each experiment is conducted three times using different random seeds, and we report the average and standard deviation of the results.

Method	Noise rate (%)			
	0.2	0.4	0.6	0.8
Dataset CIFAR-100				
SGD	66.22 ± 0.355	59.26 ± 0.045	46.77 ± 0.020	26.49 ± 0.640
SAM	66.16 ± 0.721	59.95 ± 0.622	50.81 ± 0.353	24.26 ± 1.209
FSAM	65.73 ± 0.219	58.96 ± 0.381	49.36 ± 1.103	25.92 ± 1.173
Agnostic-SAM	$\mathbf{66.64 \pm 0.657}$	$\mathbf{61.13 \pm 0.636}$	$\mathbf{52.26 \pm 0.502}$	$\mathbf{27.66 \pm 1.265}$
ASAM	66.88 ± 0.593	61.53 ± 0.487	52.77 ± 0.561	30.33 ± 1.788
Agnostic-ASAM	$\mathbf{67.38 \pm 0.106}$	$\mathbf{62.72 \pm 0.304}$	$\mathbf{54.58 \pm 0.572}$	$\mathbf{32.77 \pm 0.388}$
Dataset CIFAR-10				
SGD	89.98 ± 0.070	84.83 ± 0.085	75.06 ± 0.385	54.47 ± 1.265
SAM	91.26 ± 0.007	88.19 ± 1.060	83.43 ± 0.622	61.69 ± 0.289
FSAM	91.35 ± 0.318	87.58 ± 0.353	82.78 ± 2.057	58.09 ± 2.276
Agnostic-SAM	$\mathbf{92.38 \pm 0.007}$	$\mathbf{90.20 \pm 0.318}$	$\mathbf{85.33 \pm 0.268}$	$\mathbf{70.02 \pm 0.403}$
ASAM	91.98 ± 0.007	89.24 ± 0.572	84.39 ± 0.445	64.82 ± 6.880
Agnostic-ASAM	$\mathbf{92.06 \pm 0.367}$	$\mathbf{90.01 \pm 0.282}$	$\mathbf{86.09 \pm 0.657}$	$\mathbf{73.25 \pm 0.353}$

All experiments are conducted using the ResNet32 architecture, with models trained from scratch for 200 epochs. The batch size is set to 512 for the training set and 128 for the validation set. Following Li et al. (2024) and Foret et al. (2021), we set $\rho = 0.1$ and $\rho_1 = 2\rho_2 = 0.2$ for SAM, FSAM, and Agnostic-SAM, $\rho_1 = 2\rho_2 = 2\rho = 2.0$ for ASAM and Agnostic-ASAM when training with all noise levels, except for the 80%, where we reduce the perturbation radius by half to ensure more stable convergence. In line with Li et al. (2024), we apply additional cutout techniques along with the basic augmentations outlined in Section 4.1. Each experiment is repeated three times with different random seeds, and we report the average and standard deviation of the results in Table 5. Note that training with SGD is prone to overfitting as the number of epochs increases. Therefore, we present the best results for SGD training at both 200 and 400 epochs.

4.4 Experiments on meta-learning

The concept of Agnostic-SAM is inspired by the agnostic approach in the MAML setting, where the meta-model is optimized on the meta-training set but aims to minimize loss on the validation set. The key difference is that Agnostic-SAM uses the gradient from the validation set as an indicator to close the generalization gap between the training and testing sets. Despite this difference, both approaches share the same underlying principle, making it reasonable to expect that applying Agnostic-SAM in the MAML setting will result in improved generalization performance.

Table 6: Meta-learning results on Mini-Imagenet dataset. All baseline results are taken from Abbas et al. (2022)

Method	Accuracy	
	5 ways 1 shot	5 ways 5 shots
MAML	47.13	62.20
SHARP-MAML	49.72	63.18
Agnostic-SAM	**50.08**	**64.29**

We compare our approach to standard MAML and Sharp-MAML (Abbas et al., 2022), which also addresses the loss-landscape flatness in bilevel models. The experiments follow the setup from

Table 7: Meta-learning results on Omniglot dataset. All baseline results are taken from Abbas et al. (2022)

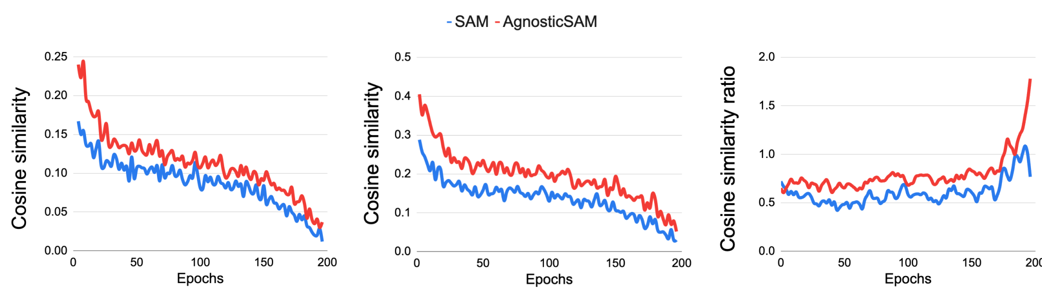
Method	Accuracy	
	20 ways 1 shot	20 ways 5 shots
MAML	91.77	96.16
SHARP-MAML	**92.89**	96.59
Agnostic-SAM	92.66	**97.28**

Abbas et al. (2022), specifically using the Sharp-MAML$_{low}$ variation, which focuses on minimizing the sharpness of meta-models trained on the meta-training set. Note that during the testing phase of MAML, only the meta-training set is used for a few update steps of the meta-model; and our Agnostic-SAM approach incorporates both the training and validation sets in the meta-model training process. Ideally, both the meta-training and meta-validation sets should be utilized to minimize the lower-level loss during training. However, this could introduce inconsistencies between the training and testing phases, potentially degrading performance during testing. To avoid this issue, we duplicate the meta-training set and use it as a validation set to minimize the lower-level loss of the meta-model, applying this procedure consistently during both the training and testing phases.

As with other experiments, we set $\rho_1 = 2\rho_2 = 2\rho$, with ρ as the perturbation radius for Sharp-MAML$_{low}$, and report the results in Tables 6 and 7. Our method consistently outperforms most baselines with significant improvements, demonstrating the effectiveness of Agnostic-SAM and its flexibility across various settings.

5 Ablation study

5.1 Cosine similarity of gradients



(a) Before updating model (b) After updating model (c) Improvement after updating

Figure 1: Cosine similarity of two gradients $\nabla_\theta \mathcal{L}_{B^t}(\theta_l)$ and $\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ (**a**) before updating model $cosine_b$, (**b**) after updating model $cosine_a$ and (**c**) the improvement of this score $change$

In Theorem 2, we prove that minimizing the loss function $\mathcal{L}_{B^t}$ could encourage two gradients $\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right)$ and $\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ to be more congruent since our update aims to maximize its lower bound, which is $\nabla_\theta \mathcal{L}_{B^t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$. In this subsection, we measure the cosine similarity between two gradients $\nabla_\theta \mathcal{L}_{B^t}(\theta_l)$ and $\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ before (denoted as $cosine_b$) and after (denoted as $cosine_a$) updating the model and measure the change of these two score (denoted as $change$).

$$cosine_b = \frac{\nabla_\theta \mathcal{L}_{B^t}(\theta_l) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)}{\|\nabla_\theta \mathcal{L}_{B^t}(\theta_l)\|_2 \|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\|_2}$$

$$
cosine_a = \frac{\nabla_\theta \mathcal{L}_{B^t}(\theta_{l+1}) \cdot \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}^v_{l+1}\right)}{\|\nabla_\theta \mathcal{L}_{B^t}(\theta_{l+1})\|_2 \|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}^v_{l+1}\right)\|_2}
$$

$$
change = \frac{cosine_a - cosine_b}{cosine_a}
$$

As shown in Figure 1c, both SAM and Agnostic-SAM improve the similarity after updating the model, this improvement also increases across training epochs. However, the similarity score of our Agnostic-SAM is always higher than SAM across the training process both before and after updating the model. This is evident that our Agnostic-SAM encourages gradient in training and validation set to be more similar during the training process.

5.2 Effectiveness of hyper-parameters

Momentum factor β. As mentioned in section 3.3, we use momentum with a factor β to estimate the gradients of the validation set. This approach helps stabilize the training process and ensures the model minimizes the loss across the entire validation set, rather than just a mini-batch. In this subsection, we examine the effect of the momentum factor on the model's performance. When setting $\beta = 0$, the perturbed model in each iteration maximizes the loss on a mini-batch of the training set while minimizing the loss on a mini-batch of the validation set. When $\beta > 0$, the perturbed model aims to minimize the loss over the entire validation set, while maximizing the loss on a mini-batch of the training set.

The experiments are set up with the same hyper-parameters as those of experiments on CIFAR100 under noisy labels settings in Section 4.3 with basic data augmentation but without the cutout technique. We set $\rho = 0.1$ for SAM and $\rho_1 = 2\rho_2 = 0.2$ for Agnostic-SAM. Results in Table 8 show that the value of β does not significantly affect model performance overall. As such, we simply set $\beta = 0.9$ in all experiments. With $\beta = 0$, our method still outperforms baselines consistently, strengthening our idea of using validation gradient to indicate the model into wider local minima while reducing the generalization gap of training and testing datasets.

Table 8: Effectiveness of momentum factor β on performance

Method	SAM	Agnostic-SAM				
		0.0	0.3	0.5	0.7	0.9
Accuracy	70.31 ± 0.2	71.14 ± 0.3	71.12 ± 0.1	70.865 ± 0.2	70.76 ± 0.3	70.91 ± 0.3

Validation batch size $|B^v|$ and complexity; sensitivity of perturbation radius ρ_1 and ρ_2. Detail of these experiment is presented in Appendix A.2

6 Conclusion and Limitation

In this paper, we explore the relationship between Sharpness-Aware Minimization (SAM) and the underlying principles of the Model-Agnostic Meta-Learning (MAML) algorithm, specifically in terms of their effects on model generalization. Building on this connection, we integrate sharpness-aware minimization with the agnostic perspective from MAML to develop a novel optimization framework, introducing the Agnostic-SAM approach. This method optimizes the model toward wider local minima using training data while ensuring low loss values on validation data. As a result, Agnostic-SAM demonstrates enhanced robustness against data shift issues. Through extensive experiments, we empirically show that Agnostic-SAM consistently outperforms baseline methods, delivering significant improvements in model performance across various datasets and challenging tasks. One limitation to note is that using an additional validation set when finding the perturbed model could potentially increase training time (depending on the size of the validation set). We consider this a trade-off between performance and training complexity. However, this issue could potentially be mitigated by reusing gradients from the training set in previous steps and we leave this as a direction for future work to reduce training complexity and still maintain performance.

Reproducibility Statement

We provide details of hyper-parameters for each experiment in Section 4 and Appendix A.2. Additionally, we open-source our code and provide instructions, scripts, and log files to reproduce experiments at `https://anonymous.4open.science/r/AgnosticSAM-F17F/README.md`

References

Momin Abbas, Quan Xiao, Lisha Chen, Pin-Yu Chen, and Tianyi Chen. Sharp-maml: Sharpness-aware model-agnostic meta learning. *arXiv preprint arXiv:2206.03996*, 2022.

Maruan Al-Shedivat, Trapit Bansal, Yura Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. Continuous adaptation via meta-learning in nonstationary and competitive environments. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=Sk2u1g-0-`.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(236), 2016a. URL `http://jmlr.org/papers/v17/15-290.html`.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016b.

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

Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7360–7371, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.508. URL `https://aclanthology.org/2022.acl-long.508`.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, mar 2003. ISSN 1532-4435.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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.

Jiaxing Chen, Weilin Yuan, Shaofei Chen, Zhenzhen Hu, and Peng Li. Evo-maml: Meta-learning with evolving gradient. *Electronics*, 12(18), 2023. ISSN 2079-9292. doi: 10.3390/electronics12183865. URL `https://www.mdpi.com/2079-9292/12/18/3865`.

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021.

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

Jiawei Du, Daquan Zhou, Jiashi Feng, Vincent YF Tan, and Joey Tianyi Zhou. Sharpness-aware training for free. *arXiv preprint arXiv:2205.14083*, 2022.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *UAI*. AUAI Press, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135. PMLR, 06–11 Aug 2017.

Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9537–9548, 2018. URL `https://proceedings.neurips.cc/paper/2018/hash/8e2c381d4dd04f1c55093f22c59c3a08-Abstract.html`.

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

Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *NIPS*, pp. 529–536. MIT Press, 1994.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos J. Storkey. Three factors influencing minima in sgd. *ArXiv*, abs/1711.04623, 2017.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *ICLR*. OpenReview.net, 2020.

Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. A fair comparison of two popular flat minima optimizers: Stochastic weight averaging vs. sharpness-aware minimization. *arXiv preprint arXiv:2202.00661*, 1, 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 *ICLR*. OpenReview.net, 2017.

Minyoung Kim, Da Li, Shell X Hu, and Timothy Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11148–11161. PMLR, 17–23 Jul 2022.

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*, pp. 5905–5914. PMLR, 2021.

Tao Li, Pan Zhou, Zhengbao He, Xinwen Cheng, and Xiaolin Huang. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5631–5640, 2024.

Thomas Möllenhoff and Mohammad Emtiyaz Khan. Sam as an optimal relaxation of bayes. *arXiv preprint arXiv:2210.01620*, 2022.

Sayan Mukherjee, Partha Niyogi, Tomaso A. Poggio, and Ryan M. Rifkin. Statistical learning: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. 2002. URL `https://api.semanticscholar.org/CorpusID:7478036`.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Van-Anh Nguyen, Tung-Long Vuong, Hoang Phan, Thanh-Toan Do, Dinh Phung, and Trung Le. Flat seeking bayesian neural network. In *Advances in Neural Information Processing Systems*, 2023.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *NeurIPS*, pp. 18420–18432, 2021.

Hoang Phan, Ngoc Tran, Trung Le, Toan Tran, Nhat Ho, and Dinh Phung. Stochastic multiple target sampling gradient descent. *Advances in neural information processing systems*, 2022.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

Zhe Qu, Xingyu Li, Rui Duan, Yao Liu, Bo Tang, and Zhuo Lu. Generalized federated learning via sharpness aware minimization. *arXiv preprint arXiv:2206.02618*, 2022.

Aravind Rajeswaran, Chelsea Finn, Sham M. Kakade, and Sergey Levine. Meta-learning with implicit gradients. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 113–124, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html`.

Tuan Truong, Hoang-Phi Nguyen, Tung Pham, Minh-Tuan Tran, Mehrtash Harandi, Dinh Phung, and Trung Le. Rsam: Learning on manifolds with riemannian sharpness-aware minimization, 2023.

Vladimir Naumovich Vapnik. Statistical learning theory. In *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. Wiley, 1998.

Bokun Wang, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Memory-based optimization methods for model-agnostic meta-learning and personalized federated learning. *Journal of Machine Learning Research*, 24(145):1–46, 2023. URL `http://jmlr.org/papers/v24/21-1301.html`.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pp. 10181–10192. PMLR, 2020.

A Appendix / supplemental material

In this appendix, we present the proofs in our paper and additional experiments. We open-source our code and provide instruction, scripts, and log files to reproduce experiments at `https://anonymous.4open.science/r/AgnosticSAM-F17F/README.md`

A.1 All Proofs

Proof of Theorem 1

Proof. We use the PAC-Bayes theory in this proof. In PAC-Bayes theory, θ could follow a distribution, says P, thus we define the expected loss over θ distributed by P as follows:

$$\mathcal{L}_{\mathcal{D}}(\theta, P) = \mathbb{E}_{\theta\sim P}\big[\mathcal{L}_{\mathcal{D}}(\theta)\big]$$
$$\mathcal{L}_{\mathcal{S}}(\theta, P) = \mathbb{E}_{\theta\sim P}\big[\mathcal{L}_{\mathcal{S}}(\theta)\big].$$

For any distribution $P = \mathcal{N}(\mathbf{0}, \sigma_P^2\mathbb{I}_k)$ and $Q = \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)$ over $\theta \in \mathbb{R}^k$, where P is the prior distribution and Q is the posterior distribution, use the PAC-Bayes theorem in Alquier et al. (2016a), for all $\beta > 0$, with a probability at least $1-\delta$, we have

$$\mathcal{L}_{\mathcal{D}}(\theta, Q) \le \mathcal{L}_{\mathcal{S}}(\theta, Q) + \frac{1}{\beta}\Big[\mathsf{KL}(Q\|P) + \log\frac{1}{\delta} + \Psi(\beta, N)\Big], \tag{9}$$

where Ψ is defined as

$$\Psi(\beta, N) = \log \mathbb{E}_P\mathbb{E}_{\mathcal{D}^N}\Big[\exp\big\{\beta\big[\mathcal{L}_{\mathcal{D}}(f_\theta) - \mathcal{L}_{\mathcal{S}}(f_\theta)\big]\big\}\Big].$$

When the loss function is bounded by L, then

$$\Psi(\beta, N) \le \frac{\beta^2L^2}{8N}.$$

The task is to minimize the second term of RHS of (9), we thus choose $\beta = \sqrt{8N}\frac{\mathsf{KL}(Q\|P)+\log\frac{1}{\delta}}{L}$. Then the second term of RHS of (9) is equal to

$$\sqrt{\frac{\mathsf{KL}(Q\|P) + \log\frac{1}{\delta}}{2N}} \times L.$$

The KL divergence between Q and P, when they are Gaussian, is given by formula

$$\mathsf{KL}(Q\|P) = \frac{1}{2}\left[\frac{k\sigma^2 + \|\theta\|^2}{\sigma_P^2} - k + k\log\frac{\sigma_P^2}{\sigma^2}\right].$$

For given posterior distribution Q with fixed σ^2, to minimize the KL term, the σ_P^2 should be equal to $\sigma^2 + \|\theta\|^2/k$. In this case, the KL term is no less than

$$k\log\Big(1 + \frac{\|\theta_0\|^2}{k\sigma^2}\Big).$$

Thus, the second term of RHS is

$$\sqrt{\frac{\mathsf{KL}(Q\|P) + \log\frac{1}{\delta}}{2N}} \times L \ge \sqrt{\frac{k\log\big(1+\frac{\|\theta\|^2}{k\sigma^2}\big)}{4N}} \times L \ge L$$

when $\|\theta\|^2 > \sigma^2\big\{\exp(4N/k) - 1\big\}$. Hence, for any $\|\theta\|_2 > \sigma^2\big\{\exp(4N/k) - 1\big\}$, we have the RHS is greater than the LHS, and the inequality is trivial. In this work, we only consider the case:

$$\|\theta\|^2 < \sigma^2\big(\exp\{4N/k\} - 1\big). \tag{10}$$

Distribution P is Gaussian centered around $\mathbf{0}$ with variance $\sigma_P^2 = \sigma^2 + \|\theta\|^2/k$, which is unknown at the time we set up the inequality, since θ is unknown. Meanwhile, we have to specify P in advance, since P is the prior distribution. To deal with this problem, we could choose a family of P such that its means cover the space of θ satisfying inequality (10). We set

$$c = \sigma^2\big(1 + \exp\{4N/k\}\big)$$
$$P_j = \mathcal{N}\big(0, c\exp\frac{1-j}{k}\mathbb{I}_k\big)$$
$$\mathfrak{P} := \big\{P_j : j = 1, 2, \dots\big\}$$

Then the following inequality holds for a particular distribution P_j with probability $1-\delta_j$ with $\delta_j = \frac{6\delta}{\pi^2 j^2}$

$$\mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\sigma^2)}\mathcal{L}_{\mathcal{D}}\big(f_{\theta'}\big) \le \mathbb{E}_{\theta'\sim\mathcal{N}(\theta,\sigma^2)}\mathcal{L}_{\mathcal{S}}\big(f_{\theta'}\big) + \frac{1}{\beta}\left[\mathsf{KL}(Q\|P_j) + \log\frac{1}{\delta_j} + \Psi(\beta, N)\right].$$

Use the well-known equation: $\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2}{6}$, then with probability $1-\delta$, the above inequality holds with every j. We pick

$$j^* := \left\lfloor 1 - k\log\frac{\sigma^2 + \|\theta\|^2/k}{c}\right\rfloor = \left\lfloor 1 - k\log\frac{\sigma^2 + \|\theta\|^2/k}{\sigma^2(1+\exp\{4N/k\})}\right\rfloor.$$

Therefore,

$$\begin{aligned}
&1 - j^* = \left\lceil k\log\frac{\sigma^2 + \|\theta\|^2/k}{c}\right\rceil \\
\Rightarrow\quad &\log\frac{\sigma^2 + \|\theta\|^2/k}{c} \le \frac{1-j^*}{k} \le \log\frac{\sigma^2 + \|\theta_0\|^2/k}{c} + \frac{1}{k} \\
\Rightarrow\quad &\sigma^2 + \|\theta\|^2/k \le c\exp\left\{\frac{1-j^*}{k}\right\} \le \exp(1/k)\big[\sigma^2 + \|\theta\|^2/k\big] \\
\Rightarrow\quad &\sigma^2 + \|\theta\|^2/k \le \sigma^2_{P_{j^*}} \le \exp(1/k)\big[\sigma^2 + \|\theta\|^2/k\big].
\end{aligned}$$

Thus the KL term could be bounded as follow

$$\begin{aligned}
\mathsf{KL}(Q\|P_{j^*}) &= \frac{1}{2}\left[\frac{k\sigma^2 + \|\theta\|^2}{\sigma^2_{P_{j^*}}} - k + k\log\frac{\sigma^2_{P_{j^*}}}{\sigma^2}\right] \\
&\le \frac{1}{2}\left[\frac{k(\sigma^2 + \|\theta\|^2/k)}{\sigma^2 + \|\theta\|^2/k} - k + k\log\frac{\exp(1/k)\big(\sigma^2 + \|\theta\|^2/k\big)}{\sigma^2}\right] \\
&= \frac{1}{2}\Big[k\log\frac{\exp(1/k)\big(\sigma^2 + \|\theta\|^2/k\big)}{\sigma^2}\Big] \\
&= \frac{1}{2}\Big[1 + k\log\Big(1 + \frac{\|\theta_0\|^2}{k\sigma^2}\Big)\Big]
\end{aligned}$$

For the term $\log\frac{1}{\delta_{j^*}}$, with recall that $c = \sigma^2\big(1 + \exp(4N/k)\big)$ and

$j^* = \left\lfloor 1 - k\log\frac{\sigma^2+\|\theta\|^2/k}{\sigma^2(1+\exp\{4N/k\})}\right\rfloor$, we have

$$\begin{aligned}
\log\frac{1}{\delta_{j^*}} = \log\frac{(j^*)^2\pi^2}{6\delta} &= \log\frac{1}{\delta} + \log\Big(\frac{\pi^2}{6}\Big) + 2\log(j^*) \\
&\le \log\frac{1}{\delta} + \log\frac{\pi^2}{6} + 2\log\Big(1 + k\log\frac{\sigma^2\big(1+\exp(4N/k)\big)}{\sigma^2 + \|\theta\|^2/k}\Big) \\
&\le \log\frac{1}{\delta} + \log\frac{\pi^2}{6} + 2\log\Big(1 + k\log\big(1 + \exp(4N/k)\big)\Big) \\
&\le \log\frac{1}{\delta} + \log\frac{\pi^2}{6} + 2\log\Big(1 + k\big(1 + \frac{4N}{k}\big)\Big) \\
&\le \log\frac{1}{\delta} + \log\frac{\pi^2}{6} + \log(1 + k + 4N).
\end{aligned}$$

Hence, the inequality

$$
\begin{aligned}
\mathcal{L}_{\mathcal{D}}\Big(\theta', \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)\Big) &\leq \mathcal{L}_{\mathcal{S}}\Big(\theta', \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)\Big) + \sqrt{\frac{\mathsf{KL}(Q\|P_{j^*}) + \log\frac{1}{\delta_{j^*}}}{2N}} \times L \\
&\leq \mathcal{L}_{\mathcal{S}}\Big(\theta', \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)\Big) \\
&+ \frac{L}{2\sqrt{N}}\sqrt{1 + k\log\Big(1 + \frac{\|\theta\|^2}{k\sigma^2}\Big) + 2\log\frac{\pi^2}{6\delta} + 4\log(N+k)} \\
&\leq \mathcal{L}_{\mathcal{S}}\Big(\theta', \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)\Big) \\
&+ \frac{L}{2\sqrt{N}}\sqrt{k\log\big(1 + \frac{\|\theta\|^2}{k\sigma^2}\big) + O(1) + 2\log\frac{1}{\delta} + 4\log(N+k)}.
\end{aligned}
$$

Since $\|\theta' - \theta\|^2$ is k chi-square distribution, for any positive t, we have

$$
\mathbb{P}\big(\|\theta' - \theta\|^2 - k\sigma^2 \geq 2\sigma^2\sqrt{kt} + 2t\sigma^2)\big) \leq \exp(-t).
$$

By choosing $t = \frac{1}{2}\log(N)$, with probability $1 - N^{-1/2}$, we have

$$
\|\theta' - \theta\|^2 \leq \sigma^2\log(N) + k\sigma^2 + \sigma^2\sqrt{2k\log(N)} \leq k\sigma^2\Big(1 + \sqrt{\frac{\log(N)}{k}}\Big)^2.
$$

By setting $\sigma = \rho \times \big(\sqrt{k} + \sqrt{\log(N)}\big)^{-1}$, we have $\|\theta' - \theta\|^2 \leq \rho^2$. Hence, we get

$$
\begin{aligned}
\mathcal{L}_{\mathcal{S}}\Big(\theta', \mathcal{N}(\theta, \sigma^2\mathbb{I}_k)\Big) &= \mathbb{E}_{\theta\sim\mathcal{N}(\theta, \sigma^2\mathbb{I}_k)}\mathbb{E}_{\mathcal{S}}\big[f_{\theta'}\big] = \int_{\|\theta'-\theta\|\leq\rho}\mathbb{E}_{\mathcal{S}}\big[f_{\theta'}\big]d\mathcal{N}(\theta, \sigma^2\mathbb{I}) \\
&+ \int_{\|\theta'-\theta\|>\rho}\mathbb{E}_{\mathcal{S}}\big[f_{\theta'}\big]d\mathcal{N}(\theta, \sigma^2\mathbb{I}) \\
&\leq \Big(1 - \frac{1}{\sqrt{N}}\Big)\max_{\|\theta'-\theta\|\leq\rho}\mathcal{L}_{\mathcal{S}}(\theta') + \frac{1}{\sqrt{N}}L \\
&\leq \max_{\|\theta'-\theta\|_2\leq\rho}\mathcal{L}_{\mathcal{S}}(\theta') + \frac{2L}{\sqrt{N}}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\theta) \leq \max_{\|\theta'-\theta\|\leq\rho}\mathcal{L}_{\mathcal{S}}(\theta') &+ \frac{4L}{\sqrt{N}}\Bigg[\sqrt{k\log\Big(1 + \frac{\|\theta\|^2}{\rho^2}\big(1 + \sqrt{\log(N)/k}\big)^2\Big)} \\
&+ 2\sqrt{\log\big(\frac{N+k}{\delta}\big) + O(1)}\Bigg] \\
&= \mathcal{L}_{\mathcal{D}}(\theta \mid \mathcal{S}) + \frac{4L}{\sqrt{N}}\Bigg[\sqrt{k\log\Big(1 + \frac{\|\theta\|^2}{\rho^2}\big(1 + \sqrt{\log(N)/k}\big)^2\Big)} \\
&+ 2\sqrt{\log\big(\frac{N+k}{\delta}\big) + O(1)}\Bigg].
\end{aligned}
$$

By choosing $\theta = \theta^*$ and $\mathcal{S} = S^v$ hence $N = N^v$, we reach the conclusion. □

Proof of Theorem 2

Proof. We have

$$
\mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) = \mathcal{L}_{B_t}\left(\theta_l\right) + \eta_1\|\nabla_\theta\mathcal{L}_{B^t}\left(\theta_l\right)\|_2^2 - \eta_2\nabla_\theta\mathcal{L}_{B^t}\left(\theta_l\right)\cdot\nabla_\theta\mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right).
$$

This follows that

$$\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) = \nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right) + 2\eta_1 H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) - \eta_2\left[H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) + H_{B^v}\left(\tilde{\theta}_l^v\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right)\right],$$

where $H_{B^t}\left(\theta_l\right) = \nabla_\theta^2 \mathcal{L}_{B_t}\left(\theta_l\right)$ and $H_{B^v}\left(\tilde{\theta}_l^v\right) = \nabla_\theta^2 \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)$ are the Hessian matrices.

$$\begin{aligned}\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\cdot\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) = {} & \nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) \\ & + 2\eta_1 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \\ & - \eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) \\ & - \eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^v}\left(\tilde{\theta}_l^v\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right).\end{aligned}$$

We now choose $\eta_1 \leq \frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{12\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}(\theta_l)\nabla_\theta \mathcal{L}_{B^t}(\theta_l)\right|}$, we then have

$$\eta_1\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right)\right| \leq \frac{1}{12}\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|.$$

This further implies

$$\eta_1 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \geq -\frac{1}{12}\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|.$$

Next we choose $\eta_2 \leq \min\left\{\frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}(\theta_l)\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}, \frac{\left|\nabla_\theta \mathcal{L}_{B_t}(\theta_l)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^v}\left(\tilde{\theta}_l^v\right)\nabla_\theta \mathcal{L}_{B^t}(\theta_l)\right|}\right\}$, we then have

$$\eta_2\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right| \leq \frac{\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6}.$$

$$-\eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^t}\left(\theta_l\right)\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) \geq -\frac{\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6}.$$

$$\eta_2\left|\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^v}\left(\tilde{\theta}_l^v\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right)\right| \leq \frac{\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6}.$$

$$-\eta_2 \nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)^T H_{B^v}\left(\tilde{\theta}_l^v\right)\nabla_\theta \mathcal{L}_{B^t}\left(\theta_l\right) \geq -\frac{\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|}{6}.$$

Finally, we yield

$$\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\cdot\nabla_\theta \mathcal{L}_{B^t}\left(\tilde{\theta}_l^t\right) \geq \nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right) - \frac{1}{2}\left|\nabla_\theta \mathcal{L}_{B_t}\left(\theta_l\right)\cdot\nabla_\theta \mathcal{L}_{B^v}\left(\tilde{\theta}_l^v\right)\right|.$$

□

A.2 Additional Experiments

Validation batch size $|B^v|$ and complexity Our method is to use a gradient on the validation set as a helper indicator to lead the model to wider local minima while maintaining low loss on the validation set, and the model should be updated mainly using training samples. Increasing validation mini-batch size could potentially increase performance and training time. In Table 9, we present the results of Agnostic-SAM with various validation batch sizes $|B^v|$ of CIFAR-100 with Resnet32 while maintaining a fixed training batch size $|B^t| = 512$, the other hyper-parameters are the same as above experiments with momentum factor β. We consider performance and training complexity to be the trade-off of Agnostic-SAM and find that setting $|B^t| = 4|B^v|$ works well for all experiments.

Table 9: Experiments on different sizes of validation mini-batch with a fixed size of training mini-batch is 512 samples

Method	Validation batch-size	Accuracy	Training time (s/epochs)
SAM	0	70.31 ± 0.233	11s
Agnostic-SAM	16	70.58 ± 0.219	11s
	32	71.07 ± 0.212	12s
	64	70.67 ± 0.049	13s
	128	71.21 ± 0.056	14s
	256	71.04 ± 0.219	15s

Sensitivity of perturbation radius ρ_1 and ρ_2 Throughout this paper, we used a consistent setting of $\rho_1 = 2\rho_2 = 2\rho$, where ρ represents the perturbation radius in the SAM method for all experiments. While these hyperparameters could be optimized for each experiment individually, we find that this configuration delivers good performance across most experiments. By setting $\rho_1 > \rho_2$, we ensure that the perturbed model prioritizes maximizing the loss on the training set rather than minimizing it on the validation set. This approach encourages the model to focus primarily on minimizing sharpness during the actual update step in Formula 6.

To verify the impact of these hyperparameters on model performance, we conduct experiments with varying perturbation radius and present the results in Figure 2. Notably, the configuration where $\rho_1 > \rho_2$ consistently yields higher accuracy compared to the setting where $\rho_1 < \rho_2$. When increasing ρ_2, the model places more emphasis on minimizing the validation set loss, rather than sharpness on the training set during the actual update step in Formula 6. This shift in focus can lead to overfitting, ultimately reducing performance.

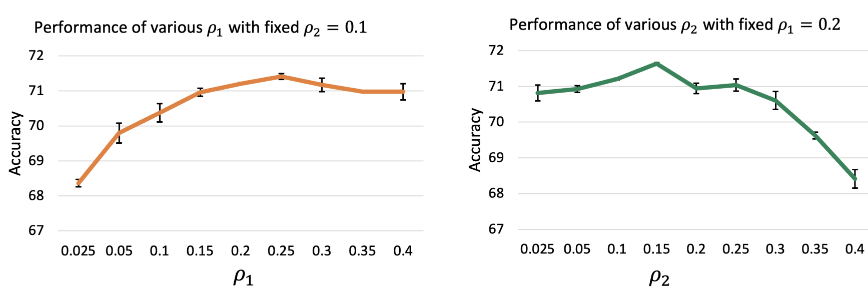


Figure 2: Experiments of various perturbation radius ρ_1 and ρ_2

Analysis of loss landscape and eigenvalues of the Hessian matrix We demonstrate the effectiveness of Agnostic-SAM in guiding models toward flatter regions of the loss landscape, as compared to both SAM and SGD, in Figures 3 and 4. The loss landscapes are visualized with the same setting, the blue areas represent lower loss values, while the red areas indicate higher loss values. Although SAM is shown to lead the model to a flatter region than SGD, Agnostic-SAM achieves an even smoother and significantly flatter loss landscape, especially in experiments with EfficientNet-B2 in Figure 3.

To further validate that Agnostic-SAM successfully locates minima with low curvature, we compute the Hessian of the loss landscape and report the five largest eigenvalues, sorted from λ_1 to λ_5, in Table 10. These eigenvalues provide insight into the curvature of the model at the optimized parameters. Larger eigenvalues indicate steeper curvature, meaning the model is more sensitive to small changes in its parameters. Conversely, smaller eigenvalues suggest flatter minima, which are typically associated with improved robustness, better generalization, and reduced sensitivity to overfitting. Negative eigenvalues indicate non-convex curvature in certain directions.

As shown in Table 10, Agnostic-SAM consistently achieves positive and lower eigenvalues compared to the baseline methods, suggesting that it effectively leads the model toward flatter regions of the loss

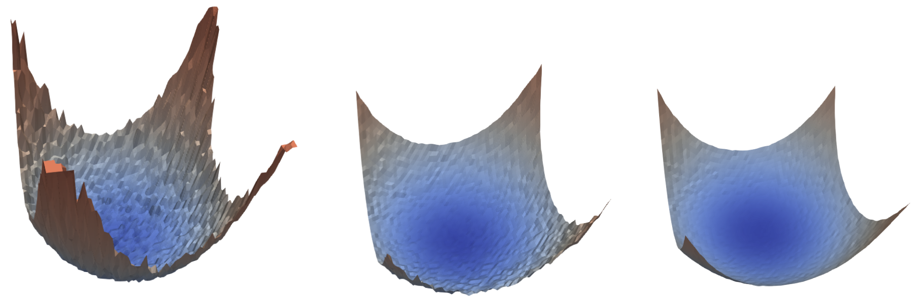


Figure 3: Loss landscape of **EffecientNet-B2** trained on Flower102 dataset with **(left)** SGD, **(middle)** SAM, and **(right)** Agnostic-SAM.

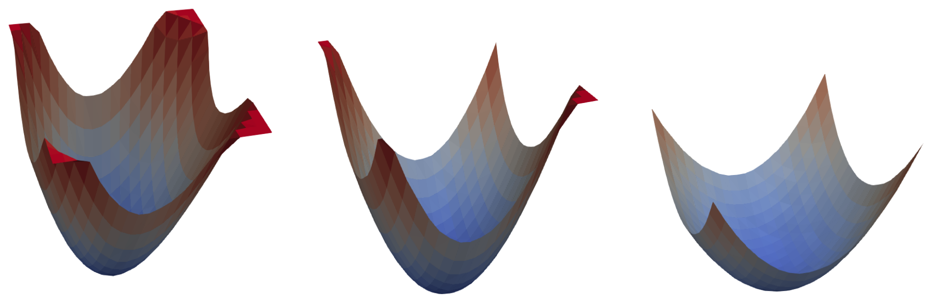


Figure 4: Loss landscape of **ResNet32** trained **(left)** SGD, **(middle)** SAM, and **(right)** Agnostic-SAM on Cifar100 dataset.

Methods	Top-5 eigenvalues of Hessian matrix				
	λ_1	λ_2	λ_3	λ_4	λ_5
EfficientNet-B2 on Flower102					
SGD	2.05×10^5	0.45×10^5	0.26×10^5	-0.47×10^5	-0.49×10^5
SAM	1.61×10^3	1.34×10^3	1.23×10^3	1.04×10^3	-0.97×10^3
Agnostic-SAM	0.61×10^3	0.41×10^3	0.37×10^3	0.32×10^3	0.31×10^3
Resnet32 on Cifar100					
SGD	3.07×10^6	2.40×10^6	2.10×10^6	1.64×10^6	1.46×10^6
SAM	1.50×10^6	1.14×10^6	0.96×10^6	0.87×10^6	0.81×10^6
Agnostic-SAM	1.04×10^6	0.79×10^6	0.66×10^6	0.58×10^6	0.57×10^6

Table 10: Eigenvalues of Hessian matrix

landscape. These results further support the efficacy of Agnostic-SAM in optimizing for smoother and more stable solutions across a variety of architectures and tasks.