ASAM: Adaptive Sharpness-Aware Minimization
for Scale-Invariant Learning of Deep Neural Networks

Jungmin Kwon [1] **Jeongseop Kim** [1] **Hyunseo Park** [1] **In Kwon Choi** [1]

Abstract

Recently, learning algorithms motivated from sharpness of loss surface as an effective measure of generalization gap have shown state-of-the-art performances. Nevertheless, sharpness defined in a rigid region with a fixed radius, has a drawback in sensitivity to parameter re-scaling which leaves the loss unaffected, leading to weakening of the connection between sharpness and generalization gap. In this paper, we introduce the concept of adaptive sharpness which is scale-invariant and propose the corresponding generalization bound. We suggest a novel learning method, adaptive sharpness-aware minimization (ASAM), utilizing the proposed generalization bound. Experimental results in various benchmark datasets show that ASAM contributes to significant improvement of model generalization performance.

1. Introduction

Generalization of deep neural networks has recently been studied with great importance to address the shortfalls of pure optimization, yielding models with no guarantee on generalization ability. To understand the generalization phenomenon of neural networks, many studies have attempted to clarify the relationship between the geometry of the loss surface and the generalization performance (Hochreiter et al., 1995; McAllester, 1999; Keskar et al., 2017; Neyshabur et al., 2017; Jiang et al., 2019). Among many proposed measures used to derive generalization bounds, loss surface sharpness and minimization of the derived generalization bound have proven to be effective in attaining state-of-the-art performances in various tasks (Hochreiter & Schmidhuber, 1997; Mobahi, 2016; Chaudhari et al., 2019; Sun et al., 2020; Yue et al., 2020).

[1]Samsung Research, Seoul, Republic of Korea. Correspondence to: Jeongseop Kim <jisean.kim@samsung.com>.

Especially, Sharpness-Aware Minimization (SAM) (Foret et al., 2021) as a learning algorithm based on PAC-Bayesian generalization bound, achieves a state-of-the-art generalization performance for various image classification tasks benefiting from minimizing sharpness of loss landscape, which is correlated with generalization gap. Also, they suggest a new sharpness calculation strategy, which is computationally efficient, since it requires only a single gradient ascent step in contrast to other complex generalization measures such as sample-based or Hessian-based approach.

However, even sharpness-based learning methods including SAM and some of sharpness measures suffer from sensitivity to model parameter re-scaling. Dinh et al. (2017) point out that parameter re-scaling which does not change loss functions can cause a difference in sharpness values so this property may weaken correlation between sharpness and generalization gap. We call this phenomenon scale-dependency problem.

To remedy the scale-dependency problem of sharpness, many studies have been conducted recently (Liang et al., 2019; Yi et al., 2019; Karakida et al., 2019; Tsuzuku et al., 2020). However, those previous works are limited to proposing only generalization measures which do not suffer from the scale-dependency problem and do not provide sufficient investigation on combining learning algorithm with the measures.

To this end, we introduce the concept of normalization operator which is not affected by any scaling operator that does not change the loss function. The operator varies depending on the way of normalizing, e.g., element-wise and filter-wise. We then define *adaptive sharpness* of the loss function, sharpness whose maximization region is determined by the normalization operator. We prove that adaptive sharpness remains the same under parameter re-scaling, i.e., *scale-invariant*. Due to the scale-invariant property, adaptive sharpness shows stronger correlation with generalization gap than sharpness does.

Motivated by the connection between generalization metrics and loss minimization, we propose a novel learning method, adaptive sharpness-aware minimization (ASAM),

which adaptively adjusts maximization regions thus acting uniformly under parameter re-scaling. ASAM minimizes the corresponding generalization bound using adaptive sharpness to generalize on unseen data, avoiding the scale-dependency issue SAM suffers from.

The main contributions of this paper are summarized as follows:

- We introduce adaptive sharpness of loss surface which is invariant to parameter re-scaling. In terms of rank statistics, adaptive sharpness shows stronger correlation with generalization than sharpness does, which means that adaptive sharpness is more effective measure of generalization gap.

- We propose a new learning algorithm using adaptive sharpness which helps alleviate the side-effect in training procedure caused by scale-dependency by adjusting their maximization region with respect to weight scale.

- We empirically show its consistent improvement of generalization performance on image classification and machine translation tasks using various neural network architectures.

The rest of this paper is organized as follows. Section 2 briefly describes previous sharpness-based learning algorithm. In Section 3, we introduce adaptive sharpness which is a scale-invariant measure of generalization gap after scale-dependent property of sharpness is explained. In Section 4, ASAM algorithm is introduced in detail using the definition of adaptive sharpness. In Section 5, we evaluate the generalization performance of ASAM for various models and datasets. We provide the conclusion and future work in Section 6.

2. Preliminary

Let us consider a model $f : X \rightarrow Y$ parametrized by a weight vector $\mathbf{w}$ and a loss function $l : Y \times Y \rightarrow \mathbb{R}_+$. Given a sample $S = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ drawn from data distribution D with i.i.d condition, the training loss can be defined as $L_S(\mathbf{w}) = \sum_{i=1}^n l(\mathbf{y}_i, f(\mathbf{x}_i; \mathbf{w}))/n$. Then, the generalization gap between the expected loss $L_D(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim D}[l(\mathbf{y}, f(\mathbf{x}; \mathbf{w}))]$ and the training loss $L_S(\mathbf{w})$ represents the ability of the model to generalize on unseen data.

Sharpness-Aware Minimization (SAM) (Foret et al., 2021) aims to minimize the following PAC-Bayesian generalization error upper bound

$$L_D(\mathbf{w}) \leq \max_{\|\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + h\left(\frac{\|\mathbf{w}\|_2^2}{\rho^2}\right) \quad (1)$$

for some strictly increasing function h. The domain of max operator, called maximization region, is an ℓ^p ball with radius ρ for $p \geq 1$. Here, sharpness of the loss function L is defined as

$$\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) - L_S(\mathbf{w}). \quad (2)$$

Because of the monotonicity of h in Equation 1, it can be substituted by ℓ^2 weight decaying regularizer, so the sharpness-aware minimization problem can be defined as the following minimax optimization

$$\min_{\mathbf{w}} \max_{\|\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

where λ is a weight decay coefficient.

SAM solves the minimax problem by iteratively applying the following two-step procedure for $t = 0, 1, 2, \ldots$ as

$$\begin{cases} \boldsymbol{\epsilon}_t = \rho \dfrac{\nabla L_S(\mathbf{w}_t)}{\|\nabla L_S(\mathbf{w}_t)\|_2} \\ \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t \left(\nabla L_S(\mathbf{w}_t + \boldsymbol{\epsilon}_t) + \lambda \mathbf{w}_t\right) \end{cases} \quad (3)$$

where α_t is an appropriately scheduled learning rate. This procedure can be obtained by a first order approximation of L_S and dual norm formulation as

$$\begin{aligned} \boldsymbol{\epsilon}_t &= \operatorname*{argmax}_{\|\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w}_t + \boldsymbol{\epsilon}) \\ &\approx \operatorname*{argmax}_{\|\boldsymbol{\epsilon}\|_p \leq \rho} \boldsymbol{\epsilon}^\top \nabla L_S(\mathbf{w}_t) \\ &= \rho \operatorname{sign}(\nabla L_S(\mathbf{w}_t)) \frac{|\nabla L_S(\mathbf{w}_t)|^{q-1}}{\|\nabla L_S(\mathbf{w}_t)\|_q^{q-1}} \end{aligned}$$

and

$$\begin{aligned} \mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}} L_S(\mathbf{w} + \boldsymbol{\epsilon}_t) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ &\approx \operatorname*{argmin}_{\mathbf{w}} (\mathbf{w} - \mathbf{w}_t)^\top \nabla L_S(\mathbf{w}_t + \boldsymbol{\epsilon}_t) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \\ &\approx \mathbf{w}_t - \alpha_t(\nabla L_S(\mathbf{w}_t + \boldsymbol{\epsilon}_t) + \lambda \mathbf{w}_t) \end{aligned}$$

where $1/p + 1/q = 1$ and $|\cdot|$ denotes element-wise absolute value function, and $\operatorname{sign}(\cdot)$ also denotes element-wise signum function. It is experimentally confirmed that the above two-step procedure produces the best performance when $p = 2$, which results in Equation 3.

As can be seen from Equation 3, SAM estimates the point $\mathbf{w}_t + \boldsymbol{\epsilon}_t$ at which the loss is approximately maximized around $\mathbf{w}_t$ in a rigid region with a fixed radius by performing gradient ascent, and performs gradient descent at $\mathbf{w}_t$ using the gradient at the maximum point $\mathbf{w}_t + \boldsymbol{\epsilon}_t$.

wt   wt+et

3. Adaptive Sharpness: Scale-Invariant Measure of Generalization Gap

In Foret et al. (2021), it is experimentally confirmed that the sharpness defined in Equation 2 is strongly correlated with the generalization gap. Also they show that SAM helps to find minima which show lower sharpness than other learning strategies and contributes to effectively lowering generalization error.

However, Dinh et al. (2017) show that sharpness defined in the rigid spherical region with a fixed radius can have a weak correlation with the generalization gap due to non-identifiability of rectifier neural networks, whose parameters can be freely re-scaled without affecting its output.

If we assume that A is a scaling operator on the weight space that does not change the loss function, as shown in Figure 1(a), the interval of the loss contours around $A\mathbf{w}$ becomes narrower than that around $\mathbf{w}$ but the size of the region remains the same, i.e.,

$$\max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) \neq \max_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(A\mathbf{w} + \boldsymbol{\epsilon}).$$

Thus, neural networks with $\mathbf{w}$ and $A\mathbf{w}$ can have arbitrarily different values of sharpness defined in Equation 2, although they have the same generalization gaps. This property of sharpness is a main cause of weak correlation between generalization gap and sharpness and we call this scale-dependency in this paper.

To solve the scale-dependency of sharpness, we introduce the concept of adaptive sharpness. Prior to explaining adaptive sharpness, we first define normalization operator. The normalization operator that cancels out the effect of A can be defined as follows.

Definition 1 (Normalization operator). *Let $\{T_{\mathbf{w}}, \mathbf{w} \in \mathbb{R}^k\}$ be a family of invertible linear operators on $\mathbb{R}^k$. Given a weight $\mathbf{w}$, if $T_{A\mathbf{w}}^{-1} A = T_{\mathbf{w}}^{-1}$ for any invertible scaling operator A on $\mathbb{R}^k$ which does not change the loss function, we say $T_{\mathbf{w}}^{-1}$ is a normalization operator of $\mathbf{w}$.*
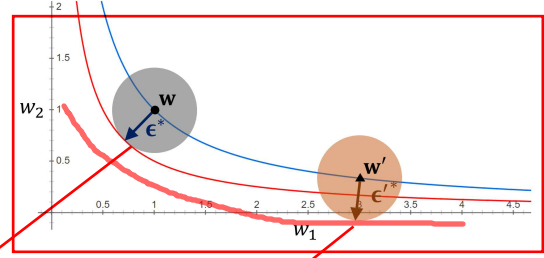
Using the normalization operator, we define adaptive sharpness as follows.

Definition 2 (Adaptive sharpness). *If $T_{\mathbf{w}}^{-1}$ is the normalization operator of $\mathbf{w}$ in Definition 1, adaptive sharpness of $\mathbf{w}$ is defined by*
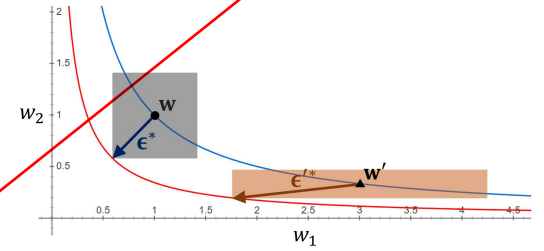
$$\max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) - L_S(\mathbf{w}) \qquad (4)$$

where $1 \leq p \leq \infty$.

Adaptive sharpness in Equation 4 has the following properties.

(a) $\|\boldsymbol{\epsilon}\|_2 \leq \rho$ and $\|\boldsymbol{\epsilon}'\|_2 \leq \rho$ (Foret et al., 2021)

(b) $\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_\infty \leq \rho$ and $\|T_{\mathbf{w}'}^{-1}\boldsymbol{\epsilon}'\|_\infty \leq \rho$ (Keskar et al., 2017)
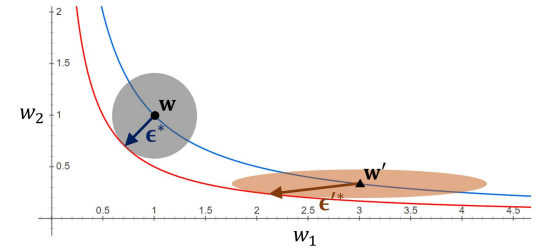
(c) $\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho$ and $\|T_{\mathbf{w}'}^{-1}\boldsymbol{\epsilon}'\|_2 \leq \rho$ (In this paper)

Figure 1. Loss contours and three types of maximization regions: (a) sphere, (b) cuboid and (c) ellipsoid. $\mathbf{w} = (1,1)$ and $\mathbf{w}' = (3, 1/3)$ are parameter points before and after multiplying a scaling operator $A = \mathrm{diag}(3, 1/3)$ and are expressed as dots and triangles, respectively. The blue contour line has the same loss at $\mathbf{w}$, and the red contour line has a loss equal to the maximum value of the loss in each type of region centered on $\mathbf{w}$. $\boldsymbol{\epsilon}^*$ and $\boldsymbol{\epsilon}'^*$ are the $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}'$ which maximize the loss perturbed from $\mathbf{w}$ and $\mathbf{w}'$, respectively.

Theorem 1. *For any invertible scaling operator A which does not change the loss function, values of adaptive sharpness at $\mathbf{w}$ and $A\mathbf{w}$ are the same as*

$$\max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) - L_S(\mathbf{w})$$

$$= \max_{\|T_{A\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_p \leq \rho} L_S(A\mathbf{w} + \boldsymbol{\epsilon}) - L_S(A\mathbf{w})$$

where $T_{\mathbf{w}}^{-1}$ and $T_{A\mathbf{w}}^{-1}$ are the normalization operators of $\mathbf{w}$ and $A\mathbf{w}$ in Definition 1, respectively.

Proof. From the assumption, it suffices to show that the

first terms of both sides are equal. By the definition of the normalization operator, we have $T_{A\mathbf{w}}^{-1}A = T_{\mathbf{w}}^{-1}$. Therefore,

$$\max_{\|T_{A\mathbf{w}}^{-1}\epsilon\|_p \leq \rho} L_S(A\mathbf{w} + \epsilon) = \max_{\|T_{A\mathbf{w}}^{-1}\epsilon\|_p \leq \rho} L_S(\mathbf{w} + A^{-1}\epsilon)$$

$$= \max_{\|T_{A\mathbf{w}}^{-1}A\epsilon'\|_p \leq \rho} L_S(\mathbf{w} + \epsilon')$$

$$= \max_{\|T_{\mathbf{w}}^{-1}\epsilon'\|_p \leq \rho} L_S(\mathbf{w} + \epsilon')$$

where $\epsilon' = A^{-1}\epsilon$. □

By Theorem 1, adaptive sharpness defined in Equation 4 is *scale-invariant* as with training loss and generalization loss. This property makes the correlation of adaptive sharpness with the generalization gap stronger than that of sharpness in Equation 2.

Figure 1(b) and 1(c) show how a re-scaled weight vector can have the same adaptive sharpness value as that of the original weight vector. It can be observed that the boundary line of each region centered on $\mathbf{w}'$ is in contact with the red line. This implies that the maximum loss within each region centered on $\mathbf{w}'$ is maintained when $\|T_{\mathbf{w}'}^{-1}\epsilon'\|_p \leq \rho$ is used for the maximization region. Thus, in this example, it can be seen that adaptive sharpness in 1(b) and 1(c) has scale-invariant property in contrast to sharpness of the spherical region shown in Figure 1(a).

The question that can be asked here is what kind of operators $T_{\mathbf{w}}$ can be considered as normalization operators which satisfy $T_{A\mathbf{w}}^{-1}A = T_{\mathbf{w}}^{-1}$ for any A which does not change the loss function. One of the conditions for the scaling operator A that does not change the loss function is that it should be node-wise scaling, which corresponds to row-wise or column-wise scaling in fully-connected layers and channel-wise scaling in convolutional layers. The effect of such node-wise scaling can be canceled using the inverses of the following operators:

- element-wise

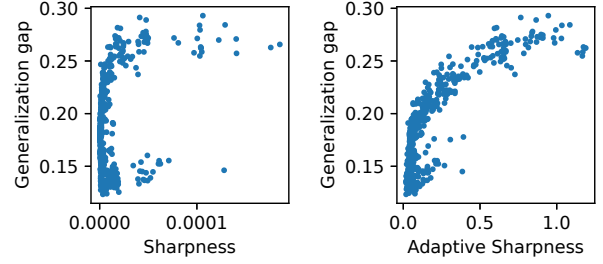$$T_{\mathbf{w}} = \mathrm{diag}(|w_1|, \ldots, |w_k|)$$

where

$$\mathbf{w} = [w_1, w_2, \ldots, w_k],$$

- filter-wise

$$T_{\mathbf{w}} = \mathrm{diag}(\mathrm{concat}(\|\mathbf{f}_1\|_2 \mathbf{1}_{n(\mathbf{f}_1)}, \ldots, \|\mathbf{f}_m\|_2 \mathbf{1}_{n(\mathbf{f}_m)},$$
$$|w_1|, \ldots, |w_q|)) \tag{5}$$

where

$$\mathbf{w} = \mathrm{concat}(\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m, w_1, w_2, \ldots, w_q).$$

(a) Sharpness ($p = 2$), $\tau = 0.174$.

(b) Adaptive sharpness ($p = 2$), $\tau = 0.636$.
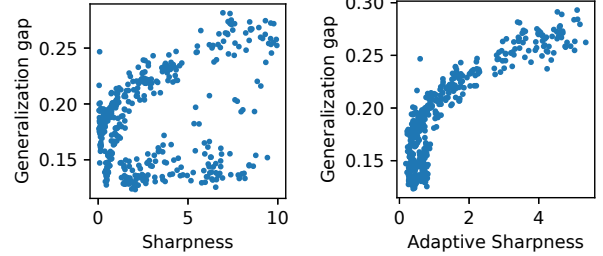
(c) Sharpness ($p = \infty$), $\tau = 0.257$.

(d) Adaptive sharpness ($p = \infty$), $\tau = 0.616$.

Figure 2. Scatter plots which show correlation of sharpness and adaptive sharpness with respect to generalization gap and their rank correlation coefficients τ.

Here, $\mathbf{f}_i$ is the i-th flattened weight vector of a convolution filter and w_j is the j-th weight parameter which is not included in any filters. And m is the number of filters and q is the number of other weight parameters in the model. If there is no convolutional layer in a model (i.e., $m = 0$), then $q = k$ and both normalization operators are identical to each other. Note that we use $T_{\mathbf{w}} + \eta I_k$ rather than $T_{\mathbf{w}}$ for sufficiently small $\eta > 0$ for stability. η is a hyper-parameter controlling trade-off between adaptivity and stability.

To confirm that adaptive sharpness actually has a stronger correlation with generalization gap than sharpness, we compare *rank statistics* which demonstrate the change of adaptive sharpness and sharpness with respect to generalization gap. For correlation analysis, we use 4 hyper-parameters: mini-batch size, initial learning rate, weight decay coefficient and dropout rate. As can be seen in Table 1, Kendall rank correlation coefficient (Kendall, 1938) of adaptive sharpness is greater than that of sharpness regardless of the value of p. Furthermore, we compare granulated coefficients (Jiang et al., 2019) with respect to different hyper-parameters to measure the effect of each hyper-parameter separately. In Table 1, the coefficients of adaptive sharpness are higher in most hyper-parameters and the average as well. Scatter plots illustrated in Figure 2 also show stronger correlation of adaptive sharpness. The dif-

Table 1. Rank statistics for sharpness and adaptive sharpness.

	$p = 2$		$p = \infty$	
	sharpness	adaptive sharpness	sharpness	adaptive sharpness
τ (rank corr.)	0.174	**0.636**	0.257	**0.616**
mini-batch size	0.667	**0.696**	0.777	**0.817**
learning rate	0.563	**0.577**	0.797	**0.806**
weight decay	−0.297	**0.534**	−0.469	**0.656**
dropout rate	**0.102**	−0.092	0.161	**0.225**
Ψ (avg.)	0.259	**0.429**	0.316	**0.626**

ference in correlation behaviors of adaptive sharpness and sharpness provides an evidence that scale-invariant property helps strengthen the correlation with generalization gap. The experimental details are described in Appendix B.

Although there are various normalization methods other than the normalization operators introduced above, this paper covers only element-wise and filter-wise normalization operators. Node-wise normalization can also be viewed as a normalization operator. Tsuzuku et al. (2020) suggest node-wise normalization method for obtaining normalized flatness. However, the method requires that the parameter should be at a critical point. Also, in the case of node-wise normalization using unit-invariant SVD (Uhlmann, 2018), there is a concern that the speed of the optimizer can be degraded due to the significant additional cost for scale-direction decomposition of weight tensors. Therefore the node-wise normalization is not covered in this paper. In the case of layer-wise normalization using spectral norm or Frobenius norm of weight matrices (Neyshabur et al., 2017), the condition $T_{A\mathbf{w}}^{-1} A = T_{\mathbf{w}}^{-1}$ is not satisfied. Therefore, it cannot be used for adaptive sharpness so we do not cover it in this paper.

Meanwhile, even though all weight parameters including biases can have scale-dependency, there remains more to consider when applying normalization to the biases. In terms of bias parameters of rectifier neural networks, there also exists translation-dependency in sharpness, which weakens the correlation with the generalization gap as well. Using the similar arguments as in the proof of Theorem 1, it can be derived that diagonal elements of $T_{\mathbf{w}}$ corresponding to biases must be replaced by constants to guarantee translation-invariance, which induces adaptive sharpness that corresponds to the case of not applying bias normalization. We compare the generalization performance based on adaptive sharpness with and without bias normalization, in Section 5.

There are several previous works which are closely related to adaptive sharpness. Li et al. (2018), which suggest a methodology for visualizing loss landscape, is related to

adaptive sharpness. In that study, filter-wise normalization which is equivalent to the definition in Equation 5 is used to remove scale-dependency from loss landscape and make comparisons between loss functions meaningful. In spite of their empirical success, Li et al. (2018) do not provide a theoretical evidence for explaining how filter-wise scaling contributes the scale-invariance and correlation with generalization. In this paper, we clarify how the filter-wise normalization relates to generalization by proving the scale-invariant property of adaptive sharpness in Theorem 1.

Also, sharpness suggested in Keskar et al. (2017) can be regarded as a special case of adaptive sharpness which uses $p = \infty$ and the element-wise normalization operator. Jiang et al. (2019) confirm experimentally that the adaptive sharpness suggested in Keskar et al. (2017) shows a higher correlation with the generalization gap than sharpness which does not use element-wise normalization operator. This experimental result implies that Theorem 1 is also practically validated.

Therefore, it seems that sharpness with $p = \infty$ suggested by Keskar et al. (2017) also can be used directly for learning as it is, but a problem arises in terms of generalization performance in learning. Foret et al. (2021) confirm experimentally that the generalization performance with sharpness defined in square region $\|\boldsymbol{\epsilon}\|_\infty \leq \rho$ result is worse than when SAM is performed with sharpness defined in spherical region $\|\boldsymbol{\epsilon}\|_2 \leq \rho$.

We conduct performance comparison tests for $p = 2$ and $p = \infty$, and experimentally reveal that $p = 2$ is more suitable for learning as in Foret et al. (2021). The experimental results are shown in Section 5.

4. Adaptive Sharpness-Aware Minimization

In the previous section, we introduce a scale-invariant measure called adaptive sharpness to overcome the limitation of sharpness. As in sharpness, we can obtain a generalization bound using adaptive sharpness, which is presented in the following theorem.

Theorem 2. *Let $T_{\mathbf{w}}^{-1}$ be the normalization operator on $\mathbb{R}^k$. If $L_D(\mathbf{w}) \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_D(\mathbf{w}+\boldsymbol{\epsilon})]$ for some $\sigma > 0$, then with probability $1 - \delta$,*

$$L_D(\mathbf{w}) \leq \max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + h\left(\frac{\|\mathbf{w}\|_2^2}{\eta^2 \rho^2}\right) \quad (6)$$

where $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a strictly increasing function, $n = |S|$ and $\rho = \sqrt{k}\sigma(1 + \sqrt{\log n/k})/\eta$.

Note that Theorem 2 still holds for $p > 2$ due to the monotonicity of p-norm, i.e., if $0 < r < p$, $\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_r$ for any $\mathbf{x} \in \mathbb{R}^n$. If $T_{\mathbf{w}}$ is an identity operator, Equation 6 is reduced equivalently to Equation 1. The proof of Equation 6 is described in detail in Appendix A.1.

Algorithm 1 ASAM algorithm ($p = 2$)

Input: Loss function l, training dataset $S :=$ $\cup_{i=1}^{n}\{(\mathbf{x}_i, \mathbf{y}_i)\}$, mini-batch size b, radius of maximization region ρ, weight decay coefficient λ, scheduled learning rate α, initial weight $\mathbf{w}_0$.
Output: Trained weight $\mathbf{w}$
Initialize weight $\mathbf{w} := \mathbf{w}_0$
while not converged **do**
    Sample a mini-batch B of size b from S
    $\boldsymbol{\epsilon} := \rho \dfrac{T_{\mathbf{w}}^2 \nabla L_B(\mathbf{w})}{\|T_{\mathbf{w}} \nabla L_B(\mathbf{w})\|_2}$
    $\mathbf{w} := \mathbf{w} - \alpha\left(\nabla L_B(\mathbf{w} + \boldsymbol{\epsilon}) + \lambda\mathbf{w}\right)^{\dagger}$
end while
return $\mathbf{w}$

Figure 3. Trajectories of SAM and ASAM.

The right hand side of Equation 6, i.e., generalization bound, can be expressed using adaptive sharpness as

$$\left(\max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) - L_S(\mathbf{w})\right) + L_S(\mathbf{w}) + h\left(\frac{\|\mathbf{w}\|_2^2}{\eta^2\rho^2}\right).$$

Since $h\left(\|\mathbf{w}\|_2^2/\eta^2\rho^2\right)$ is a strictly increasing function with respect to $\|\mathbf{w}\|_2^2$, it can be substituted with ℓ^2 weight decaying regularizer. Therefore, we can define adaptive sharpness-aware minimization problem as

$$\min_{\mathbf{w}} \max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_p \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2. \qquad (7)$$

To solve the minimax problem in Equation 7, it is necessary to find optimal $\boldsymbol{\epsilon}$ first. Analogous to SAM, we can approximate the optimal $\boldsymbol{\epsilon}$ to maximize $L_S(\mathbf{w} + \boldsymbol{\epsilon})$ using a first-order approximation as

$$\tilde{\boldsymbol{\epsilon}}_t = \underset{\|\tilde{\boldsymbol{\epsilon}}\|_p \leq \rho}{\arg\max}\, L_S(\mathbf{w}_t + T_{\mathbf{w}_t}\tilde{\boldsymbol{\epsilon}})$$

$$\approx \underset{\|\tilde{\boldsymbol{\epsilon}}\|_p \leq \rho}{\arg\max}\, \tilde{\boldsymbol{\epsilon}}^{\top} T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)$$

$$= \rho\, \text{sign}(\nabla L_S(\mathbf{w}_t)) \frac{|T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)|^{q-1}}{\|T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)\|_q^{q-1}}$$

where $\tilde{\boldsymbol{\epsilon}} = T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}$. Then, the two-step procedure for adaptive sharpness-aware minimization (ASAM) is expressed as

$$\begin{cases} \boldsymbol{\epsilon}_t = \rho T_{\mathbf{w}_t} \text{sign}(\nabla L_S(\mathbf{w}_t)) \dfrac{|T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)|^{q-1}}{\|T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)\|_q^{q-1}} \\ \mathbf{w}_{t+1} = \mathbf{w}_t - \alpha_t\left(\nabla L_S(\mathbf{w}_t + \boldsymbol{\epsilon}_t) + \lambda\mathbf{w}_t\right) \end{cases}$$

for $t = 0, 1, 2, \cdots$. Especially, if $p = 2$,

$$\boldsymbol{\epsilon}_t = \rho \frac{T_{\mathbf{w}_t}^2 \nabla L_S(\mathbf{w}_t)}{\|T_{\mathbf{w}_t} \nabla L_S(\mathbf{w}_t)\|_2}$$
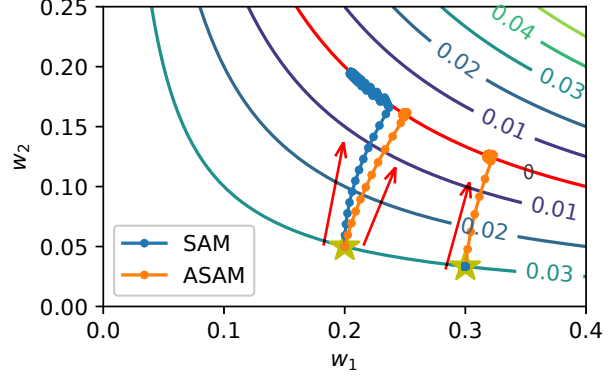
and if $p = \infty$,

$$\boldsymbol{\epsilon}_t = \rho T_{\mathbf{w}_t} \text{sign}(\nabla L_S(\mathbf{w}_t)).$$

In this study, experiments are conducted on ASAM in cases of $p = \infty$ and $p = 2$. The ASAM algorithm with $p = 2$ is described in detail on Algorithm 1. Note that the SGD (Nesterov, 1983) update marked with † in Algorithm 1 can be combined with momentum or be replaced by update of another optimization scheme such as Adam (Kingma & Ba, 2015).

5. Experimental Results

In this section, we evaluate the performance of ASAM. We first show how SAM and ASAM operate differently in a toy example. We then compare the generalization performance of ASAM with other learning algorithms for various model architectures and various datasets: CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), ImageNet (Deng et al., 2009) and IWSLT'14 DE-EN (Cettolo et al., 2014). Finally, we show how robust to label noise ASAM is.

5.1. Toy Example

As mentioned in Section 4, sharpness varies by parameter re-scaling even if its loss function remains the same, while adaptive sharpness does not. To elaborate this, we consider a simple loss function $L(\mathbf{w}) = |w_1 \text{ReLU}(w_2) - 0.04|$ where $\mathbf{w} = (w_1, w_2) \in \mathbb{R}^2$. Figure 3 presents the trajectories of SAM and ASAM with two different initial weights $\mathbf{w}_0 = (0.2, 0.05)$ and $\mathbf{w}_0 = (0.3, 0.033)$. The red line represents the set of minimizers of the loss function L, i.e., $\{(w_1, w_2); w_1 w_2 = 0.04, w_2 > 0\}$. As seen in Figure 1(a), sharpness is maximized when $w_1 = w_2$ within the same loss contour line, and therefore SAM tries to converge to $(0.2, 0.2)$. Here, we use $\rho = 0.05$ as in Foret et al. (2021). On the other hand, adaptive sharpness remains the same
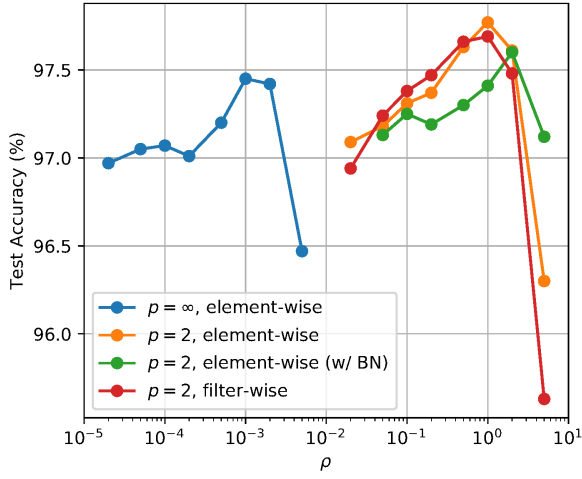
Figure 4. Test accuracy curves obtained from ASAM algorithm using a range of ρ with different factors: element-wise normalization with $p = \infty$, element-wise normalization with $p = 2$ with and without bias normalization (BN) and filter-wise normalization with $p = 2$.

along the same contour line which implies that ASAM converges to the point in the red line near the initial point as can be seen in Figure 3.

Since SAM uses a fixed radius in a spherical region for minimizing sharpness, it may cause undesirable results depending on the loss surface and the current weight. If $\mathbf{w}_0 = (0.3, 0.033)$, while SAM even fails to converge to the valley with $\rho = 0.05$, ASAM converges no matter which $\mathbf{w}_0$ is used if $\rho < \sqrt{2}$. In other words, appropriate ρ for SAM is dependent on the scales of $\mathbf{w}$ on the training trajectory, whereas ρ of ASAM is not.

5.2. Image Classification: CIFAR-10/100 and ImageNet

To confirm the effectiveness of ASAM, we conduct comparison experiments with SAM using CIFAR-10 and CIFAR-100 datasets. We use the same data split as the original paper (Krizhevsky et al., 2009). The hyper-parameters used in this test is described in Table 3. Before the comparison tests with SAM, there are three factors to be chosen in ASAM algorithm:

- normalization schemes: element-wise vs. filter-wise
- p-norm: $p = \infty$ vs. $p = 2$
- bias normalization: with vs. without

First, we perform the comparison test for filter-wise and element-wise normalization using WideResNet-16-8

Table 2. Maximum test accuracies for SGD, SAM and ASAM on CIFAR-10 dataset.

Model	SGD	SAM	ASAM
DenseNet-121	$91.00_{\pm 0.13}$	$92.00_{\pm 0.17}$	$\mathbf{93.33}_{\pm 0.04}$
ResNet-20	$93.18_{\pm 0.21}$	$93.56_{\pm 0.15}$	$\mathbf{93.82}_{\pm 0.17}$
ResNet-56	$94.58_{\pm 0.20}$	$95.18_{\pm 0.15}$	$\mathbf{95.42}_{\pm 0.16}$
VGG19-BN*	$93.87_{\pm 0.09}$	94.60	$\mathbf{95.07}_{\pm 0.05}$
ResNeXt29-32x4d	$95.84_{\pm 0.24}$	$96.34_{\pm 0.30}$	$\mathbf{96.80}_{\pm 0.06}$
WRN-28-2	$95.13_{\pm 0.16}$	$95.74_{\pm 0.08}$	$\mathbf{95.94}_{\pm 0.05}$
WRN-28-10	$96.34_{\pm 0.12}$	$96.98_{\pm 0.04}$	$\mathbf{97.28}_{\pm 0.07}$
PyramidNet-272[†]	$98.44_{\pm 0.08}$	$98.55_{\pm 0.05}$	$\mathbf{98.68}_{\pm 0.08}$

[*] Some runs completely failed, thus giving 10% of accuracy (success rate: SGD: 3/5, SAM 1/5, ASAM 3/5)
[†] PyramidNet-272 architecture is tested 3 times for each learning algorithm.

Table 3. Maximum test accuracies for SGD, SAM and ASAM on CIFAR-100 dataset.

Model	SGD	SAM	ASAM
DenseNet-121	$68.70_{\pm 0.31}$	$69.84_{\pm 0.12}$	$\mathbf{70.60}_{\pm 0.20}$
ResNet-20	$69.76_{\pm 0.44}$	$71.06_{\pm 0.31}$	$\mathbf{71.40}_{\pm 0.30}$
ResNet-56	$73.12_{\pm 0.19}$	$75.16_{\pm 0.05}$	$\mathbf{75.86}_{\pm 0.22}$
VGG19-BN*	$71.80_{\pm 1.35}$	$73.52_{\pm 1.74}$	$\mathbf{75.80}_{\pm 0.27}$
ResNeXt29-32x4d	$79.76_{\pm 0.23}$	$81.48_{\pm 0.17}$	$\mathbf{82.30}_{\pm 0.11}$
WRN-28-2	$75.28_{\pm 0.17}$	$77.25_{\pm 0.35}$	$\mathbf{77.54}_{\pm 0.14}$
WRN-28-10	$81.56_{\pm 0.13}$	$83.42_{\pm 0.04}$	$\mathbf{83.68}_{\pm 0.12}$
PyramidNet-272[†]	$88.91_{\pm 0.12}$	$89.36_{\pm 0.20}$	$\mathbf{89.90}_{\pm 0.13}$

[*] Some runs completely failed, thus giving 10% of accuracy (success rate: SGD: 5/5, SAM 4/5, ASAM 4/5)
[†] PyramidNet-272 architecture is tested 3 times for each learning algorithm.

model (Zagoruyko & Komodakis, 2016) and illustrate the results in Figure 4. As can be seen, both test accuracies are comparable across ρ, and element-wise normalization provides a slightly better accuracy at $\rho = 1.0$.

Similarly, Figure 4 shows how much test accuracy varies with the maximization region for adaptive sharpness. It can be seen that $p = 2$ shows better test accuracies than $p = \infty$, which is consistent with Foret et al. (2021). We could also observe that bias normalization does not contribute to the improvement of test accuracy in Figure 4. Therefore, we decide to use element-wise normalization operator and $p = 2$, and not to employ bias normalization in the remaining tests.

As ASAM has a hyper-parameter ρ to be tuned, we first conduct a grid search over {0.00005, 0.0001, 0.0002, ..., 0.5, 1.0, 2.0} for finding appropriate values of ρ. We use $\rho = 0.5$ for CIFAR-10 and $\rho = 1.0$ for CIFAR-100, because it gives moderately good performance across various models. We set ρ for SAM as 0.05 for CIFAR-10 and 0.1 for CIFAR-100 as in Foret et al. (2021). η for ASAM is set

to 0.01. We set mini-batch size to 128, and m-sharpness suggested by Foret et al. (2021) is not employed. The number of epochs is set to 200 for SAM and ASAM and 400 for SGD. Momentum and weight decay coefficient are set to 0.9 and 0.0005, respectively. Cosine learning rate decay (Loshchilov & Hutter, 2016) is adopted with an initial learning rate 0.1. Also, random resize, padding by four pixels, normalization and random horizontal flip are applied for data augmentation and label smoothing (Müller et al., 2019) is adopted with its factor of 0.1.

Using the hyper-parameters, we compare the best test accuracies obtained by SGD, SAM and ASAM for various rectifier neural network models: VGG (Simonyan & Zisserman, 2015), ResNet (He et al., 2016), DenseNet (Huang et al., 2017), WideResNet (Zagoruyko & Komodakis, 2016), and ResNeXt (Xie et al., 2017).

For PyramidNet-272 (Han et al., 2017), we additionally apply some latest techniques: AutoAugment (Cubuk et al., 2019), CutMix (Yun et al., 2019) and ShakeDrop (Yamada et al., 2019). We employ the m-sharpness strategy with $m = 32$. Initial learning rate and mini-batch size are set to 0.05 and 256, respectively. The number of epochs is set to 900 for SAM and ASAM and 1800 for SGD. We choose ρ for SAM as 0.05, as in Foret et al. (2021), and ρ for ASAM as 1.0 for both CIFAR-10 and CIFAR-100. Every entry in the tables represents mean and standard deviation of 5 independent runs. In both CIFAR-10 and CIFAR-100 cases, ASAM generally surpasses SGD and SAM, as can be seen in Table 2 and Table 3.

For the sake of evaluations at larger scale, we compare the performance of SGD, SAM and ASAM on ImageNet. We apply each method with ResNet-50 and use $\rho = 0.05$ for SAM and $\rho = 1.0$ for ASAM. The number of training epochs is 200 for SGD and 100 for SAM and ASAM. We use mini-batch size 512, initial learning rate 0.2, and SGD optimizer with weight decay coefficient 0.0001. Other hyper-parameters are the same as those of CIFAR-10/100 tests. We also employ m-sharpness with $m = 128$ for both SAM and ASAM.

Table 4 shows mean and standard deviation of maximum test accuracies over 3 independent runs for each method. As can be seen in the table, ASAM achieves higher accuracies than SGD and SAM. These results imply that ASAM can enhance generalization performance of rectifier neural network architectures in image classification task beyond CIFAR.

5.3. Machine Translation: IWSLT'14 DE-EN

To validate effectiveness of ASAM in tasks other than image classification, we apply SAM and ASAM to IWSLT'14

Table 4. Top1 and Top5 maximum test accuracies for SGD, SAM and ASAM on ImageNet dataset using ResNet-50.

	SGD	SAM	ASAM
Top1	$75.79_{\pm 0.22}$	$76.39_{\pm 0.03}$	$\mathbf{76.63}_{\pm 0.18}$
Top5	$92.62_{\pm 0.04}$	$92.97_{\pm 0.07}$	$\mathbf{93.16}_{\pm 0.18}$

DE-EN, a dataset on machine translation task.

In this test, we adopt Transformer architecture (Vaswani et al., 2017) and Adam optimizer as a base optimizer of SAM and ASAM instead of SGD. Learning rate, β_1 and β_2 for Adam are set to 0.0005, 0.9 and 0.98, respectively. Dropout rate and weight decay coefficient are set to 0.3 and 0.0001, respectively. Label smoothing is adopted with its factor 0.1. We choose $\rho = 0.1$ for SAM and $\rho = 0.2$ for ASAM as a result of a grid search over {0.005, 0.01, 0.02, ..., 0.5, 1.0, 2.0} using validation dataset. The results of the experiments are obtained from 3 independent runs.

As can be seen in Table 5, we could observe improvement even on IWSLT'14 in BLEU score when using Adam+ASAM instead of Adam or Adam+SAM.

Table 5. BLEU scores for Adam, Adam+SAM and Adam+ASAM on IWSLT'14 DE-EN dataset using Transformer.

BLEU score	Adam	Adam+SAM	Adam+ASAM
Validation	$35.34_{\pm <0.01}$	$35.52_{\pm 0.01}$	$\mathbf{35.66}_{\pm <0.01}$
Test	$34.86_{\pm <0.01}$	$34.78_{\pm 0.01}$	$\mathbf{35.02}_{\pm <0.01}$

5.4. Robustness to Label Noise

As shown in Foret et al. (2021), SAM is as robust to label noise in the training data as MentorMix (Jiang et al., 2020), which is a state-of-the-art method. We expect that ASAM would share the robustness to label noise. To confirm this, we compare the test accuracies of SGD, SAM and ASAM for ResNet-32 model and CIFAR-10 dataset whose labels in the training data are corrupted by symmetric label noise (Rooyen et al., 2015) with noise levels of 20%, 40%, 60% and 80%, and the test data is not touched. Hyperparameter settings are the same as that of previous CIFAR experiments. Table 6 shows test accuracies for SGD, SAM and ASAM obtained from 3 independent runs with respect to label noise levels. Compared to SGD and SAM, ASAM generally enhances the test accuracy across various noise level by retaining the robustness to label noise.

Table 6. Maximum test accuracies of ResNet-32 models trained on CIFAR-10 with label noise.

Noise rate	SGD	SAM	ASAM
0%	$94.50_{\pm 0.11}$	$94.80_{\pm 0.12}$	$\mathbf{94.88}_{\pm 0.12}$
20%	$91.32_{\pm 0.23}$	$92.94_{\pm 0.12}$	$\mathbf{93.21}_{\pm 0.10}$
40%	$87.68_{\pm 0.05}$	$90.62_{\pm 0.18}$	$\mathbf{90.89}_{\pm 0.13}$
60%	$82.50_{\pm 0.30}$	$86.58_{\pm 0.30}$	$\mathbf{87.41}_{\pm 0.16}$
80%	$68.35_{\pm 0.85}$	$\mathbf{69.92}_{\pm 0.98}$	$67.69_{\pm 1.34}$

6. Conclusions

In this paper, we have introduced adaptive sharpness with scale-invariant property that improves training path in weight space by adjusting their maximization region with respect to weight scale. Also, we have confirmed that this property, which ASAM shares, contributes in improvement of generalization performance. The superior performance of ASAM is notable from the comparison tests conducted against SAM, which is currently state-of-the-art learning algorithm in many image classification benchmarks. In addition to the contribution as a learning algorithm, adaptive sharpness can serve as a generalization measure with stronger correlation with generalization gap benefiting from their scale-invariant property. Therefore adaptive sharpness has a potential to be a metric for assessment of neural networks. We have also suggested the condition of normalization operator for adaptive sharpness but we did not cover all the normalization schemes which satisfy the condition. So this area could be further investigated for better generalization performance in future works.

7. Acknowledgements

References

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*, volume 57, 2014.

Chatterji, N., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. In *International Conference on Learning Representations*, 2019.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hochreiter, S. and Schmidhuber, J. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Hochreiter, S., Schmidhuber, J., et al. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, pp. 529–536, 1995.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jiang, L., Huang, D., Liu, M., and Yang, W. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International Conference on Machine Learning*, pp. 4804–4815. PMLR, 2020.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

Karakida, R., Akaho, S., and Amari, S.-i. The normalization method for alleviating pathological sharpness in wide neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., and Smelyanskiy, M. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 and CIFAR-100 datasets. 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 6391–6401. Curran Associates Inc., 2018.

Liang, T., Poggio, T., Rakhlin, A., and Stokes, J. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 888–896. PMLR, 2019.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

McAllester, D. A. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

Mobahi, H. Training recurrent neural networks by diffusion. *arXiv preprint arXiv:1601.04114*, 2016.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring generalization in deep learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Rooyen, B. v., Menon, A. K., and Williamson, R. C. Learning with symmetric label noise: the importance of being unhinged. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 10–18, 2015.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Sun, X., Zhang, Z., Ren, X., Luo, R., and Li, L. Exploring the vulnerability of deep neural networks: A study of parameter corruption. *CoRR*, abs/2006.05620, 2020.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020.

Uhlmann, J. A generalized matrix inverse that is consistent with respect to diagonal transformations. *SIAM Journal on Matrix Analysis and Applications*, 39(2):781–800, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS*, 2017.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yamada, Y., Iwamura, M., Akiba, T., and Kise, K. Shakedrop regularization for deep residual learning. *IEEE Access*, 7:186126–186136, 2019.

Yi, M., Meng, Q., Chen, W., Ma, Z.-m., and Liu, T.-Y. Positively scale-invariant flatness of relu neural networks. *arXiv preprint arXiv:1903.02237*, 2019.

Yue, X., Nouiehed, M., and Kontar, R. A. Salr: Sharpness-aware learning rates for improved generalization. *arXiv preprint arXiv:2011.05348*, 2020.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pp. 6023–6032, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

A. Proofs

A.1. Proof of Theorem 2

We first introduce the following concentration inequality.

Lemma 1. *Let* $\{\epsilon_i, i = 1, \dots, k\}$ *be independent normal variables with mean* 0 *and variance* σ_i^2. *Then,*

$$P\left(\sum_{i=1}^{k} \epsilon_i^2 \geq k\sigma_{\max}^2 \left(1 + \sqrt{\frac{\log n}{k}}\right)^2\right) \leq \frac{1}{\sqrt{n}}$$

where $\sigma_{\max} = \max\{\sigma_i\}$.

Proof. From Lemma 1 in Laurent & Massart (2000), for any $x > 0$,

$$P\left(\sum_{i=1}^{k} \epsilon_i^2 \geq \sum_{i=1}^{k} \sigma_i^2 + 2\sqrt{\sum_{i=1}^{k} \sigma_i^4 x} + 2\sigma_{\max}^2 x\right) \leq \exp(-x).$$

Since

$$\sum_{i=1}^{k} \sigma_i^2 + 2\sqrt{\sum_{i=1}^{k} \sigma_i^4 x} + 2\sigma_{\max}^2 x \leq \sigma_{\max}^2(k + 2\sqrt{kx} + 2x)$$

$$\leq \sigma_{\max}^2(\sqrt{k} + \sqrt{2x})^2,$$

plugging $x = \frac{1}{2}\log n$ proves the lemma.

$\square$

Theorem 3. *Let* $T_{\mathbf{w}}^{-1}$ *be a normalization operator of* $\mathbf{w}$ *on* $\mathbb{R}^k$. *If* $L_D(\mathbf{w}) \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_D(\mathbf{w} + \boldsymbol{\epsilon})]$ *for some* $\sigma > 0$, *then with probability* $1 - \delta$,

$$L_D(\mathbf{w}) \leq \max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \sqrt{\frac{1}{n-1}\left(k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{\eta^2\rho^2}\left(1 + \sqrt{\frac{\log n}{k}}\right)^2\right) + 4\log\frac{n}{\delta} + O(1)\right)}$$

where $n = |S|$ *and* $\rho = \sqrt{k}\sigma(1 + \sqrt{\log n/k})/\eta$.

Proof. The idea of the proof is given in Foret et al. (2021). From the assumption, adding Gaussian perturbation on the weight space does not improve the test error. Moreover, from Theorem 3.2 in Chatterji et al. (2019), the following generalization bound holds under the perturbation:

$$E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_D(\mathbf{w} + \boldsymbol{\epsilon})] \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_S(\mathbf{w} + \boldsymbol{\epsilon})] + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C(n,\sigma,k)\right)}.$$

Therefore, the left hand side of the statement can be bounded as

$$L_D(\mathbf{w}) \leq E_{\epsilon_i \sim \mathcal{N}(0,\sigma^2)}[L_S(\mathbf{w} + \boldsymbol{\epsilon})] + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C\right)}$$

$$\leq \left(1 - \frac{1}{\sqrt{n}}\right) \max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \frac{1}{\sqrt{n}} + \sqrt{\frac{1}{n-1}\left(\frac{1}{4}k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + \log\frac{n}{\delta} + C\right)}$$

$$\leq \max_{\|T_{\mathbf{w}}^{-1}\boldsymbol{\epsilon}\|_2 \leq \rho} L_S(\mathbf{w} + \boldsymbol{\epsilon}) + \sqrt{\frac{1}{n-1}\left(k\log\left(1 + \frac{\|\mathbf{w}\|_2^2}{k\sigma^2}\right) + 4\log\frac{n}{\delta} + 4C\right)}$$

where the second inequality follows from Lemma 1 and $\|T_{\mathbf{w}}^{-1}\|_2 \leq \frac{1}{\eta}$.

$\square$

B. Correlation Analysis

To capture the correlation between generalization measures, i.e., sharpness and adaptive sharpness, and actual generalization gap, we utilize Kendall rank correlation coefficient (Kendall, 1938). Formally, given the set of pairs of a measure and generalization gap observed $S = \{(m_1, g_1), \ldots, (m_n, g_n)\}$, Kendall rank correlation coefficient τ is given by

$$\tau(S) = \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(m_i - m_j)\text{sign}(g_i - g_j).$$

Since τ represents the difference between the proportion of concordant pairs, i.e., either both $m_i < m_j$ and $g_i < g_j$ or both $m_i > m_j$ and $g_i > g_j$ among the whole $\binom{n}{2}$ point pairs, and the proportion of discordant pairs, i.e., not concordant, the value of τ is in the range of $[-1, 1]$.

While the rank correlation coefficient aggregates the effects of all the hyper-parameters, granulated coefficient (Jiang et al., 2019) can consider the correlation with respect to the each hyper-parameter separately. If $\Theta = \prod_{i=1}^{N} \Theta_i$ is the Cartesian product of each hyper-parameter space Θ_i, granulated coefficient with respect to Θ_i is given by

$$\psi_i = \frac{1}{|\Theta_{-i}|} \sum_{\theta_1 \in \Theta_1} \cdots \sum_{\theta_{i-1} \in \Theta_{i-1}} \sum_{\theta_{i+1} \in \Theta_{i+1}} \cdots \sum_{\theta_N \in \Theta_N} \tau \left(\bigcup_{\theta_i \in \Theta_i} \{(m(\boldsymbol{\theta}), g(\boldsymbol{\theta}))\} \right)$$

where $\Theta_{-i} = \Theta_1 \times \cdots \Theta_{i-1} \times \Theta_{i+1} \times \Theta_N$. Then the average $\Psi = \sum_{i=1}^{N} \psi_i / N$ of ψ_i indicates whether the correlation exists across all hyper-parameters.

We vary 4 hyper-parameters, mini-batch size, initial learning rate, weight decay coefficient and dropout rate, to produce different models. It is worth mentioning that changing one or two hyper-parameters for correlation analysis may cause spurious correlation (Jiang et al., 2019). For each hyper-parameter, we use 5 different values in Table 7 which implies that $5^4 = 625$ configurations in total.

Table 7. Hyper-parameter configurations.

mini-batch size	32, 64, 128, 256, 512
learning rate	0.0033, 0.01, 0.033, 0.1, 0.33
weight decay	5e−7, 5e−6, 5e−5, 5e−4, 5e−3
dropout rate	0, 0.125, 0.25, 0.375, 0.5

By using the above hyper-parameter configurations, we train WideResNet-28-2 model on CIFAR-10 dataset. We use SGD as an optimizer and set momentum to 0.9. We set the number of epochs to 200 and cosine learning rate decay (Loshchilov & Hutter, 2016) is adopted. Also, random resize, padding by four pixels, normalization and random horizontal flip are applied for data augmentation and label smoothing (Müller et al., 2019) is adopted with its factor of 0.1. Using model parameters with training accuracy higher than 99.0% among the generated models, we calculate sharpness and adaptive sharpness with respect to generalization gap.

To calculate adaptive sharpness, we fix normalization scheme to element-wise normalization. We calculate adaptive sharpness and sharpness with both $p = 2$ and $p = \infty$. We conduct a grid search over $\{5\text{e}{-}6, 1\text{e}{-}5, 5\text{e}{-}5, \ldots, 5\text{e}{-}1, 1.0\}$ to obtain each ρ for sharpness and adaptive sharpness which maximizes correlation with generalization gap. As results of the grid search, we select 1e−5 and 5e−4 as ρs for sharpness of $p = 2$ and $p = \infty$, respectively, and select 5e−1 and 5e−3 as ρs for adaptive sharpness of $p = 2$ and $p = \infty$, respectively. To calculate maximizers of each loss function for calculation of sharpness and adaptive sharpness, we follow m-sharpness strategy suggested by Foret et al. (2021) and m is set to 8.