석 사 학 위 논 문
Master's Thesis

예리도 인지 최소화의 이해와 응용: 탈출 효율과 비대칭 경사면의 관점에서

On the Understanding of Sharpness-Aware Minimization and its Application: A Perspective on Escape Efficiency and Asymmetric Valleys

2022

박 시 환 (朴 時 煥 Park, Sihwan)

한 국 과 학 기 술 원

Korea Advanced Institute of Science and Technology

석 사 학 위 논 문

예리도 인지 최소화의 이해와 응용: 탈출 효율과 비대칭 경사면의 관점에서

2022

박 시 환

한 국 과 학 기 술 원

김재철AI대학원

예리도 인지 최소화의 이해와 응용: 탈출 효율과 비대칭 경사면의 관점에서

박 시 환

위 논문은 한국과학기술원 석사학위논문으로
학위논문 심사위원회의 심사를 통과하였음

2022년 06월 14일

심사위원장 양 은 호 (인)

심 사 위 원 신 진 우 (인)

심 사 위 원 황 성 주 (인)

On the Understanding of Sharpness-Aware Minimization and its Application: A Perspective on Escape Efficiency and Asymmetric Valleys

Sihwan Park

Advisor: Eunho Yang

A dissertation submitted to the faculty of
Korea Advanced Institute of Science and Technology in
partial fulfillment of the requirements for the degree of
Master of Science in AI

Daejeon, Korea
June 14, 2022

Approved by

Eunho Yang
Professor of Graduate School of AI

The study was conducted in accordance with Code of Research Ethics[1].

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of Korea Advanced Institute of Science and Technology, hereby declare that I have not committed any act that may damage the credibility of my research. This includes, but is not limited to, falsification, thesis written by someone else, distortion of research findings, and plagiarism. I confirm that my thesis contains honest conclusions based on my own careful research under the guidance of my advisor.

MAI

박시환. 예리도 인지 최소화의 이해와 응용: 탈출 효율과 비대칭 경사면의 관점에서. 김재철AI대학원 . 2022년. 32+iv 쪽. 지도교수: 양은호. (영문 논문)

Sihwan Park. On the Understanding of Sharpness-Aware Minimization and its Application: A Perspective on Escape Efficiency and Asymmetric Valleys. Kim Jaechul Graduate School of AI . 2022. 32+iv pages. Advisor: Eunho Yang. (Text in English)

초 록

예리도 인지 최소화는 편평한 최소 점을 찾아 좋은 일반화 성능을 갖도록 하는 학습 방법으로, 최근 다양한 분야에서 괄목할 만한 성과를 거두었다. 그럼에도, 예리도 인지 최소화에 대한 이론적 분석은 성공적인 성능 향상과 비교하면 많이 뒤떨어져 있는 상황이다. 본 연구에서는 예리도 인지 최소화에 대한 이론적 이해를 확장하기 위해 탈출 효율과 비대칭 경사면의 두 가지 새로운 관점에서 예리도 인지 최소화의 우수한 일반화 성능에 대해 분석한다. 본 연구에서는 다음의 두 가지를 증명한다. 우선, 예리도 인지 최소화는 기존의 확률적 경사 하강법보다 지역 최소 점을 더 빨리 탈출할 수 있다. 두 번째로, 예리도 인지 최소화는 비대칭 경사면에서 더 편평한 지역으로 수렴한다. 또한, 이러한 효과가 예리도 인지 최소화의 국소 최대화 반경이 커짐에 따라 증폭됨을 증명한다. 나아가, 제안된 이론에 기반하여 예리도 인지 최소화를 더욱 효율적으로 활용할 수 있는 새로운 학습 체계인 인색한 (Parsimonious) 예리도 인지 최소화를 제안한다. 본 연구에서는 다양한 데이터셋과 네트워크 구조에 대해 제안된 이론이 성립함을 실험적으로 검증하고, 제안된 학습 방법이 실제로 효과가 있음을 확인한다.

핵 심 낱 말 딥러닝, 일반화, 예리도 인지 최소화, 탈출 효율, 비대칭 경사면

Abstract

Sharpness-Aware Minimization (SAM) has emerged as a promising training scheme that leads to good generalization through finding flat minima. Despite its accomplishments in various fields, the existing theoretical understanding of SAM is far behind its successes. To extend the understanding of SAM, we theoretically analyze the SAM from two novel perspectives: escape efficiency and asymmetric valleys. First, we prove that SAM can escape a minimum faster than SGD. Hence the SAM can explore more minima than SGD and can converge to flatter minima by escaping minima where SGD would be stranded. Second, we show that SAM converges to a flatter region on asymmetric valleys than SGD and it leads to better generalization. Moreover, we prove that these effects are amplified by increasing the radius ρ of inner maximization. Based on the proposed theory, we further study an efficient way to utilize SAM, Parsimonious SAM (PSAM), which uses SAM periodically in the early phase of training. Finally, on various architectures and datasets, we empirically verify that the proposed theory holds well in practice, and PSAM presents comparable performance to SAM while it requires only 65% of the computational cost of SAM.

Keywords Deep Learning, Generalization, Sharpness-Aware Minimization, Escape Efficiency, Asymmetric Valleys

Contents

List of Tables

List of Figures

Chapter 1. Introduction

Recently, neural networks have achieved remarkable successes in various application areas including computer vision, natural language processing, and graph neural networks. While neural networks have evolved, their architecture became much more complicated and the number of parameters far exceeded the number of data points. Despite its overparameterized nature, it has been verified to have good generalization performance contrary to the conventional learning theory regarding the model complexity and overfitting. Such phenomenon is often referred to as the double-descent phenomenon [1]. Numerous efforts have attempted to explain this enigma of the generalizability of neural networks, as a consequence, there is a widely acknowledged theory known as *Flat Minima* hypothesis.

The flat minima hypothesis was first introduced in [2] and recently revisited by [3]. The hypothesis says that a neural network with a flatter minimum on its loss function tends to have better generalization than one with a sharper minimum. It stems from the fact that flat minima are robust to the distributional shift and thus tend to have a better fit on population loss. However, [4, 5] pointed out that the sharpness can be arbitrarily amplified by the network's parameter scale and a number of works [6, 7, 8, 9] studied to resolve this issue. With these progressive refinements and extensive empirical validation [10], the flat minima hypothesis became a promising narrative for explaining the neural network's generalizability.

Based on the hypothesis, several researchers studied optimization algorithms to find the flat minima [11, 12, 13] and among them, the Sharpness-Aware Minimization (SAM) [13] has emerged as a propitious training scheme. SAM finds a minimum where the entire neighborhood of minimum has uniformly low loss value and thus can find flat minima. However, some works [8, 14, 15, 16] pointed out the problems of SAM. [8] discussed that the sensitivity of SAM on parameter re-scaling [4], [14] tried to improve the efficiency of SAM, and [15] figured out that SAM can reluctantly converge to sharp minima and solved the issue by minimizing the surrogate loss gap. [16] proposed the look-ahead, layer-wise adaptive version of SAM for the scalability. In this line of advancements, SAM accomplished considerable success in various research areas.

However, theoretical understanding of SAM lags far behind empirical success. As pointed out by [17], the existing theoretical justification of SAM provided by [13] has a limited explanation. Though [17] provided some new aspects of theoretical analysis on SAM in terms of implicit biases [18, 19, 20], their results were limited to the diagonal neural networks. To extend our understanding of SAM, we aim to further investigate the behavior of SAM from a perspective of escape efficiency and asymmetric valleys [21].

The escape efficiency of stochastic gradient descent (SGD) [22, 23, 24, 25, 26] explains how can SGD find flat minima even the loss landscapes are highly non-convex and thus have infinitely many sharp minima. The escape efficiency is defined as an inverse of the mean exit time from a local minimum to outside of the valley of minimum. [25] showed that the escape efficiency of SGD is exponentially proportional to the sharpness of minima and inverse of depth of the valley. And [26] extended the results of [25] into the non-stationary regime. In other words, SGD can escape a minimum exponentially faster for a sharper and shallower minimum. We build our theory upon these results by showing that the loss function of SAM has sharper and shallower minima than SGD.

[21] introduced a new concept of minima beyond sharp and flat minima called asymmetric valleys. In asymmetric valleys, the loss function grows rapidly in one direction and relatively slowly in the

opposite direction. They empirically demonstrated that asymmetric valleys are prevalent in the neural network loss landscape. Furthermore, they theoretically proved that a bias toward the flatter side of the asymmetric valley leads to better generalization even if it slightly increases the training loss.

In this paper, to extend the theoretical understanding of how SAM can outperform SGD, we propose two main theorems: (a) **SAM can escape sharp minima faster than SGD**, and (b) **SAM generates biases into flatter region on asymmetric valleys and it leads to better generalization**. We prove the theorems and empirically verify the validity of our theorems on ResNet-18 [27], Preactivation ResNet-164 [28], WideResNet-28x10 [29], and PyramidNet-110 [30] on CIFAR-10 and 100 [31]. Furthermore, based on our theory, we propose an efficient way to utilize the SAM called *Parsimonious SAM (PSAM)*. PSAM is a periodical SAM in the exploration stage (i.e., early stage of training) and the same as the original SAM in the exploitation stage (i.e., latter stage of training). We demonstrate that PSAM presents comparable performance with SAM (83.54% vs 83.53% with WideResNet-28x10 on CIFAR-100) while it requires only 65% of computation cost compared to SAM. We summarize our contributions as follows:

- We propose novel theoretical results regarding how SAM outperforms SGD on two new aspects: escape efficiency and asymmetric valleys. First, we prove that SAM can escape a minimum faster than SGD and it implies that SAM can explore more minima, and can converge to the flatter minima by escaping minima that SGD would be stranded. Second, we prove that SAM converges to flatter region on asymmetric valleys than SGD and it leads to better generalization. We experimentally verify that the proposed theory holds well in practice. To the best of our knowledge, we are the first to theoretically and empirically investigate the behavior of SAM in terms of escape efficiency and asymmetric valleys.
- Based on the proposed theory, we also propose a novel way to utilize SAM in a more efficient manner, *Parsimonious SAM (PSAM)*. We demonstrate that PSAM presents comparable performance to SAM on various architectures and datasets while it requires only 65% of computational cost of SAM.

Chapter 2. Preliminaries

In this section, we provide some background knowledge for our main theory. Section 2.1 explains the details about the Sharpness-Aware Minimization [13], Section 2.2 gives existing results on the escape efficiency of SGD [26], and Section 2.3 offers definitions, assumptions and theorems regarding the asymmetric valleys [21].

Primitives Let $\mathcal{D}$ be an unknown data distribution on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, where the data point $\mathbf{x} \in \mathbb{R}^{d_x}$ and the label $\mathbf{y} \in \mathbb{R}^{d_y}$. From the data distribution, we have n i.i.d. samples $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n) \overset{\text{iid}}{\sim} \mathcal{D}$. In supervised learning, we want to find a minimum $w^* \triangleq \operatorname{argmin}_{w \in \mathbb{R}^d} \mathcal{L}(w)$ where $\mathcal{L}(w) \triangleq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell(\mathbf{x}, \mathbf{y}; w)] \in \mathbb{R}^d \to \mathbb{R}$ is the population loss, $w \in \mathbb{R}^d$ denotes the model parameter, and $\ell \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^d \to \mathbb{R}$ is the loss function. Since the data distribution $\mathcal{D}$ is unknown in most cases, instead of directly optimizing $\mathcal{L}$, we often find the empirical risk minimizer $\hat{w}^* \in \mathbb{R}^d$ where $\hat{w}^* \triangleq \operatorname{argmin}_{w \in \mathbb{R}^d} \hat{\mathcal{L}}(w)$, and $\hat{\mathcal{L}}(w) \triangleq \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \mathbf{y}_i; w)$. An open ball centered at $w \in \mathbb{R}^d$ with radius $r \in \mathbb{R}_{>0} (= \{x \in \mathbb{R} : x > 0\})$ is defined as $B_r(w) \triangleq \{v \in \mathbb{R}^d : \|v - w\|_2 < r\}$.

2.1 Sharpness-Aware Minimization

SAM [13] finds parameter values whose entire neighborhoods have uniformly low training loss value instead of finding simply have low training loss value. SAM problem can be represented as follows:

$$\min_{w \in \mathbb{R}^d} \hat{\mathcal{L}}^{SAM}(w) + \lambda \|w\|_2^2 \quad \text{where} \quad \hat{\mathcal{L}}^{SAM}(w) \triangleq \max_{\|\epsilon\|_p \le \rho} \hat{\mathcal{L}}(w + \epsilon), \tag{2.1}$$

where $\lambda \ge 0$ is an ℓ_2 regularization magnitude, $\rho \ge 0$ is a hyperparameter that decides the diameter of neighborhood around w, and $p \in [1, \infty)$ indicates p-norm for maximization over ϵ.

Since directly solving the inner maximization in Equation 2.1 involves another optimization problem, SAM detours the maximization problem by considering the first-order Taylor expansion of $\hat{\mathcal{L}}(w+\epsilon)$ with respect to ϵ around 0:

$$\begin{aligned} \epsilon^*(w) &\triangleq \operatorname{argmax}_{\|\epsilon\|_p \le \rho} \hat{\mathcal{L}}(w + \epsilon) \approx \operatorname{argmax}_{\|\epsilon\|_p \le \rho} \hat{\mathcal{L}}(w) + \epsilon^\top \nabla_w \hat{\mathcal{L}}(w) \\ &= \operatorname{argmax}_{\|\epsilon\|_p \le \rho} \epsilon^\top \nabla_w \hat{\mathcal{L}}(w), \end{aligned}$$

since $\hat{\mathcal{L}}(w)$ in the second equality is independent to ϵ. The solution $\hat{\epsilon}(w)$ of the approximated maximization problem can be found by the classical dual norm problem:

$$\hat{\epsilon}(w) = \rho \times \operatorname{sign}(\nabla_w \hat{\mathcal{L}}(w)) \frac{|\nabla_w \hat{\mathcal{L}}(w)|^{q-1}}{\left(\|\nabla_w \hat{\mathcal{L}}(w)\|_q^q\right)^{\frac{1}{p}}}, \tag{2.2}$$

where $\frac{1}{p} + \frac{1}{q} = 1$. By substituting Equation 2.2 into the first term of Equation 2.1, we obtain an approximated SAM problem

$$\min_{w \in \mathbb{R}^d} \hat{\mathcal{L}}(w + \hat{\epsilon}(w)).$$

However, computing its gradient involves Hessian computation which requires a very high computational cost. Therefore, SAM simply drops the second-order derivative terms in $\nabla_w \hat{\mathcal{L}}(w+\hat{\epsilon})$ and finally obtains the approximated gradient of SAM:

$$\nabla_w \hat{\mathcal{L}}^{SAM}(w) \approx \nabla_w \hat{\mathcal{L}}(w)|_{w+\hat{\epsilon}}.$$

Note that even though SAM goes through several approximation steps, they empirically demonstrated that SAM performs well in practice. In addition, they showed that including second-order derivative terms does not always lead to better performance.

2.2 Escape Efficiency of SGD

Here we provide some theoretical results on the escape efficiency of SGD in [26]. Let's start from the following refined definition of the SGD.

SGD Formulation With the initial parameter $w_0 \in \mathbb{R}^d$, learning rate $\eta > 0$, and batch size B, SGD generates a sequence of parameters $\{w_t\}_{t\in\mathbb{N}}$ by the following update rule:

$$w_{t+1} = w_t - \eta \nabla_w \mathcal{L}_B(w_t), \forall t \in \mathbb{N}, \tag{2.3}$$

where $\mathcal{L}_B(w) = \frac{1}{B}\sum_{i=1}^{B} \ell(\mathbf{x}_i, \mathbf{y}_i; w)$.

Since the mini-batches are randomly sampled from data distribution $\mathcal{D}$, Equation 2.3 has an inherent randomness. To decouple the deterministic part and stochastic part of Equation 2.3, we decompose the batch gradient $\nabla \mathcal{L}_B(w_t)$ into (a) true gradient term $\nabla_w \mathcal{L}(w)$ and (b) the noise term $\nabla \mathcal{L}(w_t) - \nabla_w \mathcal{L}_B(w_t)$. We model the noise term (b) as a Gaussian noise. Then we can rewrite the Equation 2.3 as follows:

$$w_{t+1} = w_t - \eta \nabla_w \mathcal{L}(w_t) + \sqrt{\frac{\eta}{B}}\omega_t, \tag{2.4}$$

where $\omega_t \sim \mathcal{N}(0, \eta C(w_t))$ is a parameter-dependent Gaussian noise with its covariance.

Mean Exit Time Let $w^* \in \mathbb{R}^d$ be a local minimum of the loss function $\mathcal{L}(w)$ and $w^* \in D \subseteq \mathbb{R}^d$ be a open neighborhood around w^*. Then the mean exit time of SGD from w^* to outside of D is defined as follows.

Definition 2.2.1. (Mean Exit Time from D, Definition 2 in [26]) Consider an refined SGD (Equation 2.4) starting from $w_0 \in D$. Then, the mean exit time of SGD from D is defined as

$$\mathbb{E}[\nu] \triangleq \mathbb{E}\left[\min\{t\eta : w_t \notin D\}\right].$$

This means that the mean exit time is the minimum step t times learning rate η such that w_t is outside of D. Hence, SGD with a small mean exit time rapidly escapes the minimum. In this manner, we define the escape efficiency of SGD as an inverse of the mean exit time:

$$\text{Escape Efficiency} \triangleq \mathbb{E}[\nu]^{-1}.$$

Next, we present some widely used assumptions on this literature [22, 24, 25].

Assumption 2.2.2. (Locally Quadratic, Assumption 1 of [26]) There exists a matrix $H^* \in \mathbb{R}^{d\times d}$ such that for any $w \in D$, the following equality holds:

$$\forall w \in D, \mathcal{L}(w) = L(w^*) + \nabla_w \mathcal{L}(w^*)(w - w^*) + \frac{1}{2}(w - w^*)^\top H^*(w - w^*).$$

The assumption 2.2.2 means that the loss function around local minimum is quadratic. Although this is a quite strong condition, it is accepted as the inevitable minimal assumption necessary to theoretically analyze the complicated loss function of a modern neural networks.

Assumption 2.2.3. (Hesse Covariance Matrix, Assumption 2 of [26]) For any $w \in D$, $C(w)$ is approximately equal to H^*.

The assumption 2.2.3 means that the noise of SGD in Equation 2.4 follows a Gaussian distribution where the covariance matrix is the Hessian of loss function at the local minimum. This was first introduced by [24] to reflect the anisotropic nature of SGD noise and since then, most studies have used this assumption.

For the next assumption, we need an additional object : Let $\phi = \{\phi_t\}_{t\in[0,T]} \subset \mathbb{R}^d$ be a trajectory in the parameter space over a time interval $[0, T]$ with a terminal time T, where $\phi_t \in \mathbb{R}^d$ is a parameter which continuously changes in t. ϕ is regarded as a continuous map $[0, T] \to \mathbb{R}^d$, i.e., is an element of $\mathcal{C}_T\left(\mathbb{R}^d\right)$ (a set of continuous trajectories in $\mathbb{R}^d$). With this object, we have the following assumption:

Assumption 2.2.4. (Assumption 3 of [26]) There exists $K > 0$ such that for any $\phi \in \mathcal{C}_T\left(\mathbb{R}^d\right)$ and $t \in [0, T]$, $\dot{\phi}_t \leq K$ holds.

The assumption 2.2.4 means that for any escaping trajectory from w_0 (inside of D) to w_T (outside of D), the trajectory does not change drastically. This is required to eliminate some pathological escaping trajectories. With these assumptions, we introduce the main result of [26].

Theorem 2.2.5. *(Thoerem 3 of [26]) Consider the discrete Gaussian SGD (Equation 2.4) whose initial point is the local minimum $w_0 = w^*$. Suppose that assumption 2.2.2, 2.2.3, 2.2.4 hold. Then, the mean exit time from the neighborhood D has the following limit:*

$$2\lambda_{\max}^{-\frac{1}{2}}\Delta\mathcal{L} - A\lambda_{\min}^{\frac{1}{2}}\left(\kappa^{\frac{1}{2}} - 1\right) \leq \lim_{\eta\to 0}\frac{\eta}{B}\ln\mathbb{E}[\nu] \leq 2\lambda_{\max}^{-\frac{1}{2}}\Delta\mathcal{L} + A\lambda_{\min}^{\frac{1}{2}}\left(\kappa^{\frac{1}{2}} - 1\right), \tag{2.5}$$

where $\Delta\mathcal{L} \triangleq \min_{w\in\partial D}\mathcal{L}(w) - \mathcal{L}(w^*)$ is the depth of minimum, some constant A, $\kappa \triangleq \frac{\lambda_{\max}}{\lambda_{\min}}$ is the condition number of $C(w^*)$, and $\lambda_{\max}, \lambda_{\min}$ are the maximum, minimum eigenvalues of H^* respectively.

Theorem 2.2.5 implies that the escape efficiency of SGD $\sim \exp\left[-\frac{B}{\eta}\Delta\mathcal{L}\lambda_{\max}^{-\frac{1}{2}}\right]$ since the second term of both sides (i.e., $A\lambda_{\min}^{\frac{1}{2}}(\kappa^{\frac{1}{2}} - 1)$) is dominated by the first term. Thus we can say that SGD can escape the minimum faster as (a) the minimum is sharper (larger $\lambda_{\max}$) and (b) the minimum is shallower (smaller depth $\Delta\mathcal{L}$).

2.3 Asymmetric Valleys

Before we discuss the asymmetric valleys, we formally define the asymmetric valley.

Definition 2.3.1. (Asymmetric Valleys, definition 1 and 2 of [21]) Given constants $p > 0, r > \zeta > 0, c \geq 1$, a direction u (i.e., $\|u\|_2 = 1$) is (r, p, c, ζ)-**asymmetric direction** with respect to a point $w \in \mathbb{R}^d$ and loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ if $0 < \nabla_\ell \mathcal{L}(w + \ell u) < p$ and $\nabla_\ell \mathcal{L}(w - \ell u) < -cp$ for any $\ell \in (\zeta, r)$. A local minimum $w^* \in \mathbb{R}^d$ of loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ is in (r, p, c, ζ)-**asymmetric valley** if there exists at least one direction $u \in \mathbb{R}^d$ such that (r, p, c, ζ)-asymmetric direction with respect to w^* and $\mathcal{L}$.

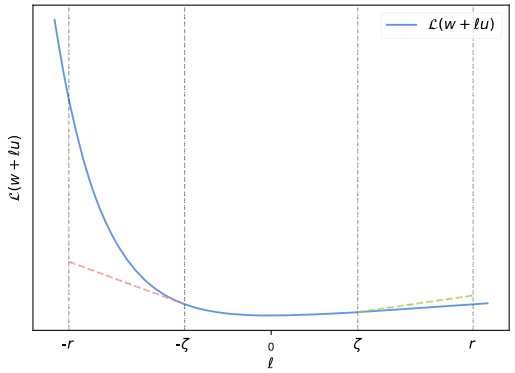


Figure 2.1: Visualization of (r, p, c, ζ)-asymmetric direction. $\nabla_\ell \mathcal{L}(w + \ell u) < -cp$ (above the red dashed line) for $\ell \in (-r, -\zeta)$ and $0 < \nabla_\ell \mathcal{L}(w + \ell u) < p$ (below the green dashed line) for $\ell \in (\zeta, r)$.

In words, as shown in the Figure 2.1, (r, p, c, ζ)-asymmetric direction u means that the loss function $\mathcal{L}$ grows slowly along u and grows rapidly towards $-u$. Thus the loss function has asymmetric gradients along u and its opposite direction. In the definition, ζ handles a very small region where the loss function has near-zero gradients around the local minimum since the loss function is often assumed as smooth.

Now, we introduce some theoretical results in [21] regarding the generalization and bias into the flat side with several definitions and assumptions.

Definition 2.3.2. ((δ, R)-shift gap, Definition 3 of [21]) For $\xi \geq 0, \delta \in \mathbb{R}^d$, and fixed functions $\mathcal{L}$ and $\hat{\mathcal{L}}$, we define the (δ, R)-shift gap between $\mathcal{L}$ and $\hat{\mathcal{L}}$ with respect to a point w as $\xi_\delta(w) := \max_{v \in B(R)} |\mathcal{L}'(w + v + \delta) - \hat{\mathcal{L}}(w + v)|$ where $\mathcal{L}'(w) := \mathcal{L}(w) - \min_w \mathcal{L}(w) + \min_w \hat{\mathcal{L}}(w)$.

Here, $\mathcal{L}'$ has introduced to eliminate the vertical gap between the population loss $\mathcal{L}$ and empirical risk $\hat{\mathcal{L}}$. Thus $\xi_\delta(w)$ can measure the true shift gap (i.e., horizontal differences) without considering vertical differences. As shown in the Figure 2.2, low $\xi_\delta(w)$ value means that after δ-shifting the population loss $\mathcal{L}$, population loss and empirical risk match well. For instance, $\xi_\delta(w) = 0$ implies $\mathcal{L}$ is locally identical to $\hat{\mathcal{L}}$ after the shift δ.

Assumption 2.3.3. (Random shift assumption, Assumption 1 of [21]) For a given population loss $\mathcal{L}$ and a random empirical loss $\hat{\mathcal{L}}$, constants $R > 0, r \geq \zeta > 0, \xi \geq 0$, a vector $\bar{\delta} \in \mathbb{R}^d$ with $r \geq \bar{\delta}_i \geq \zeta$ for all $i \in [d]$, a minimizer $\hat{w}^*$, we assume that there exists a random variable $\delta \in \mathbb{R}^d$ correlated with $\hat{\mathcal{L}}$ such that $P(\delta_i = \bar{\delta}_i) = P(\delta_i = -\bar{\delta}_i) = 1/2$ for all $i \in [d]$, and the (δ, R)-shift gap between $\mathcal{L}$ and $\hat{\mathcal{L}}$ with respect to $\hat{w}^*$ is bounded by ξ.

Assumption 2.3.3 means that there exists a random shift $\delta \in \mathbb{R}^d$ such that (δ, R)-shift is bounded by some constant ξ. More precisely, the population loss $\mathcal{L}$ and empirical loss $\hat{\mathcal{L}}$ are almost identical

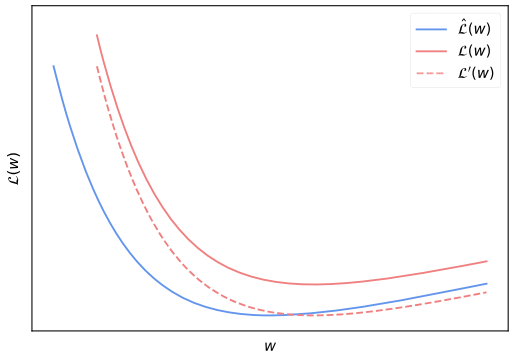


Figure 2.2: Visualization of $\hat{\mathcal{L}}, \mathcal{L}$, and $\mathcal{L}'$ on $\mathbb{R}$. $\mathcal{L}'$ is the vertically shifted population loss $\mathcal{L}$ to eliminate the vertical differences between $\mathcal{L}$ and $\hat{\mathcal{L}}$. By considering $\hat{\mathcal{L}}$ and $\mathcal{L}'$, we can measure the effect of shift δ more accurately. Note that in this figure, $\hat{\mathcal{L}}$ is identical to $\mathcal{L}'$ for some shift δ.

(bounded by ξ) under the random shift δ. This assumption is required to derive the specific lower bound of the population risk difference in Theorem 2.3.5. Note that since $\hat{\mathcal{L}}$ is defined on a set of random samples from $\mathcal{D}$, the shift δ should be also a random variable. And since the samples defining $\hat{\mathcal{L}}$ were independently and identically sampled, the expectation of δ is essentially zero. Assumption 2.3.3 reflects this observation.

Assumption 2.3.4. (Locally asymmetric, Assumption 2 of [21]) For a given empirical loss $\hat{\mathcal{L}}$, and a minimizer $\hat{w}^*$, there exist orthogonal directions $u_1, u_2, \ldots, u_k \in \mathbb{R}^d$ such that u_i is (r, p_i, c_i, ζ)-asymmetric with respect to $\hat{w}^* + v - \langle v, u_i \rangle u_i$ for all $v \in B(R')$ and $i \in [k]$.

Assumption 2.3.4 means that there are orthogonal asymmetric directions $u_1, \ldots, u_k$ which are asymmetric with respect to all points in the R'-neighborhood of the minimum w^*. This assumption is required to preserve the asymmetricity of a biased solution. More specifically, suppose that (orthogonal) asymmetric directions $u_1, \ldots, u_k$ are asymmetric with respect to the local minimum w^* but not for the near points around w^*. Then, some biased solutions around w^* may have different asymmetric directions and hence we cannot directly compare the two solutions. Note that both assumptions were verified to fit in well with practice by [21].

Under these assumptions, [21] introduced the main theoretical result that says a bias into flatter side on asymmetric valley leads to the better generalization:

Theorem 2.3.5. *(Theorem 1 of [21]) For any $l \in \mathbb{R}^k$, if assumption 2.3.3 holds for $R = \|l\|_2$, assumption 2.3.4 holds for $R' = \|\bar{\delta}\|_2 + \|l\|_2$, and $\frac{4\xi}{(c_i - 1)p_i} < l_i < \min\{r - \bar{\delta}_i, \bar{\delta}_i - \zeta\}$ for all $i \in [k]$, then we have*

$$\mathbb{E}_\delta \mathcal{L}(\hat{w}^*) - \mathbb{E}_\delta \mathcal{L}(\hat{w}^* + \sum_{i=1}^{k} l_i u_i) \geq \sum_{i=1}^{k} (c_i - 1) l_i p_i / 2 - 2k\xi > 0.$$

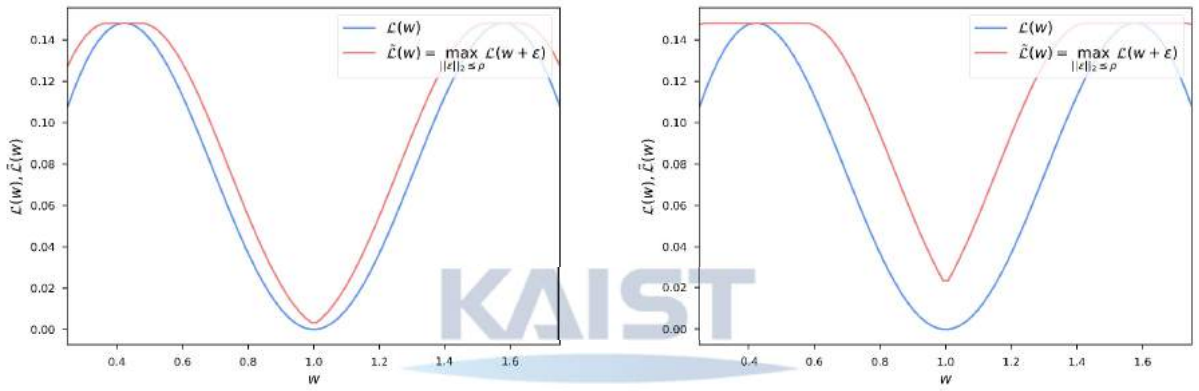
Theorem 2.3.5 says that the biased solution $\hat{w}^* + \sum_{i=1}^{k} l_i u_i$ into flatter side on asymmetric valleys has better generalization than the original one $\hat{w}^*$. Also implies that there are proper ranges on biases (l_i's) and in that range, a larger bias provides greater generalization improvement.

Chapter 3. Proposed Theory

The common belief of the SAM's empirical success is simply that SAM can find flat minima than SGD. Here, based on the straightforward observations (Section 3.1), we provide new perspectives of the SAM's successes which are one that extends the common belief into a more sophisticated manner (Section 3.2) and another one that reveals the biasing effect of SAM on asymmetric valleys and its consequences (Section 3.3).

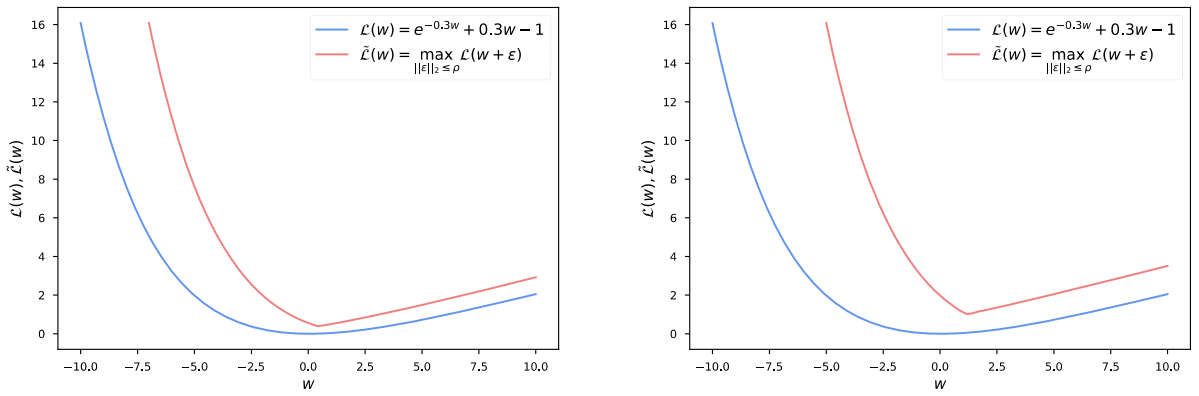
3.1 Motivation

Our theory is inspired by the following simple observations:



(a) Symmetric case, small ρ

(b) Symmetric case, large ρ



(c) Asymmetric case, small ρ

(d) Asymmetric case, large ρ

Figure 3.1: Observations on $\tilde{\mathcal{L}}$ for symmetric loss and asymmetric loss on $\mathbb{R}$.

The Figure 3.1 presents the observations on $\mathcal{L}(w)$ in $\mathbb{R}$ for symmetric $\mathcal{L}(w)$ and asymmetric $\mathcal{L}(w)$. For symmetric loss, we use a simple sextic function $\mathcal{L}(w) = w^2(w-1)^2(w-2)^2$ and for asymmetric loss, we use the LINEX loss function $\mathcal{L}(w) = e^{\alpha w} - \alpha w - 1$ with $\alpha = -0.3$. In all cases, we can observe that $\tilde{\mathcal{L}}(w)$ is floating above the $\mathcal{L}(w)$, i.e., the shape of $\tilde{\mathcal{L}}(w)$ is more narrow and shallow than $\mathcal{L}(w)$. Also, for asymmetric cases (Figure 3.1c and 3.1d), we can observe that the local minimum of $\tilde{\mathcal{L}}(w)$ is biased

to flat side on the asymmetric valley. Furthermore, we can see that the aforementioned effects become greater as ρ increases. In the following subsections, we formalize these observations into general cases.

3.2 Escape Efficiency of SAM

According to the Theorem 2.2.5, SGD escapes a minimum quickly as it is sharper and shallower. Given this result, we show that SAM loss (Equation 2.1) converts a minimum into a more sharp and shallow one and thus SAM can escape the minimum faster than SGD. We state our theorem as follows:

Theorem 3.2.1. *(SAM converts a minimum into sharper and shallower minimum) Let $\mathcal{L}(w) : \mathbb{R}^d \to \mathbb{R}$ be the loss function and $w^* \in \mathbb{R}^d$ be a local minimum. Assume that there exists an open neighborhood $D \subseteq \mathbb{R}^d$ containing w^* such that $\mathcal{L}$ is locally quadratic on D. And let $B_r(w^*) \subseteq \mathbb{R}^d$ be the largest open ball such that $B_r(w^*) \subseteq D$. Then, when we let $\tilde{\mathcal{L}}(w) \triangleq \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}(w+\epsilon)$ with $\rho \in (0, r)$, we have $\tilde{w}^* = w^*$ and*

(i) $\Delta\mathcal{L} = \Delta\tilde{\mathcal{L}} + \frac{\lambda_{\max}}{2}\rho^2 > \Delta\tilde{\mathcal{L}}$

(ii) $\tilde{\lambda}_{\max} \ge \left(1 + \frac{\rho}{r-\rho}\right)\lambda_{\max}$

where $\tilde{w}^* \in \mathbb{R}^d$ is the local minimum of $\tilde{\mathcal{L}}$, $\Delta\mathcal{L} = \min_{w \in \partial D} \mathcal{L}(w) - \mathcal{L}(w^*)$ is the depth of minimum and $\lambda_{\max}$ and $\tilde{\lambda}_{\max}$ are the maximum eigenvalues of $\nabla^2_w \mathcal{L}(w^*)$ and $\nabla^2_w \tilde{\mathcal{L}}(\tilde{w}^*)$, respectively.

Theorem 3.2.1 says that SAM loss ($\tilde{\mathcal{L}}(w)$, Equation 2.2) converts a valley of minimum (i.e., basin of attraction) into a sharper and shallower valley than the original one. And the effect increases as ρ increases. Since we know that SGD can escape a minimum faster as the valley of minimum is sharper and shallower (Theorem 2.2.5), as a consequence of Theorem 3.2.1, we can conclude that SAM can escape the minimum faster than SGD. This result implies that **(a) SAM can escape more flat minima where SGD would be stranded**, and **(b) SAM can explore more minima than SGD within the same time steps.** Hence, we can say that SAM is more likely to find a flat minimum than SGD. For the proof of the theorem, please refer to Section 7.1.1.

Meanwhile, there is another reason why SAM can find flat minima. Since SAM considers the maximum over ρ-neighborhood around a point, if a valley of minimum is narrower than ρ, it would be eliminated by taking the maximum. Hence, SAM can exclude such extremely sharp (narrow) minima.

3.3 Biasing Effect of SAM on Asymmetric Valleys

We start with the definition of symmetric direction as a counterpart of asymmetric direction:

Definition 3.3.1. (Symmetric Direction) Given constants $b > a \ge 0$, $r > \zeta \ge 0$, a direction $v \in \mathbb{R}^d$ is (r, a, b, ζ)-**symmetric direction** with respect to a point $w \in \mathbb{R}^d$ and loss function $\mathcal{L}$ if $a < \nabla_\ell \mathcal{L}(w+\ell v) < b$ and $-b < \nabla_\ell \mathcal{L}(w-\ell v) < -a$ for any $\ell \in (\zeta, r)$.

A symmetric direction v means that the scale of gradients along v are similar on both sides of v. From the Equation 2.1, we can regard SAM as SGD on the perturbed loss $\tilde{\mathcal{L}}(w) \triangleq \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}(w+\epsilon)$. Based on this interpretation, we prove that SAM generates a bias into the flatter side of the asymmetric valley by showing the perturbed loss $\tilde{\mathcal{L}}(w)$ has shifted local minima to the flatter side than the original loss $\mathcal{L}(w)$. With the definition 2.3.1 and 3.3.1, we state the following theorem:

Theorem 3.3.2. *(SAM generates a bias into flatter side on asymmetric valleys) Let $w^* \in \mathbb{R}^d$ be a local minimum of $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$. Suppose that there is an $(r, p, c, 0)$-asymmetric direction u with respect to w^* and $\mathcal{L}$. That is, there are constants $p > 0, r > 0, c \geq 1$ such that $0 < \nabla_\ell \mathcal{L}(w^* + \ell u) < p$ and $\nabla_\ell \mathcal{L}(w^* - \ell u) < -cp$ for any $\ell \in (0, r)$. Then, when we let $\tilde{\mathcal{L}}(w) = \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(w+\epsilon)$ for some $\rho \in (0, r)$ and $\tilde{w}^*$ be the local minimum of $\tilde{\mathcal{L}}$, we have*

$$\langle \tilde{w}^*, u \rangle \in \left[\langle w^*, u \rangle + \frac{c-1}{c+1}\rho, \langle w^*, u \rangle + \rho \right).$$

Furthermore, $\langle \tilde{w}^*, v \rangle \in \left(\langle w^*, v \rangle - \frac{b-a}{a+b}\rho, \langle w^*, v \rangle + \frac{b-a}{a+b}\rho \right)$ for any $(r, a, b, 0)$-symmetric direction v.

Theorem 3.3.2 says that for an asymmetric direction u, the perturbed loss $\tilde{\mathcal{L}}$ has new local minimum $\tilde{w}^*$ near the original one w^* which is shifted to the flatter side along u. And also implies that for any other symmetric direction v, SAM does not generate any bias on both sides. Note that in the theorem, we only provide the interval where $\tilde{w}^*$ lies. This is because the only information we have is upper bounds of gradients on both sides along u (i.e., $\nabla_\ell \mathcal{L}(w^* - \ell u) < -cp$ and $\nabla_\ell \mathcal{L}(w^* + \ell u) < p$). The exact location of $\tilde{w}^*$ depends on the actual value of the loss function. Nevertheless, it is sufficient to conclude that SAM generates bias toward the flat side of asymmetric valleys.

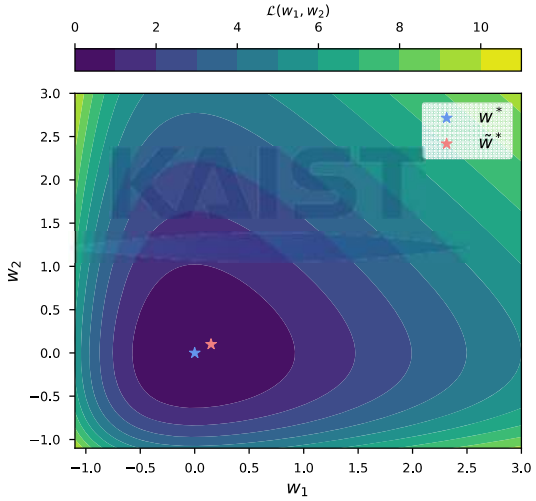


Figure 3.2: An example of applying Theorem 3.3.2 on $\mathbb{R}^2$. SAM solution $\tilde{w}^*$ is marked as the red star and SGD solution w^* is marked as the blue star.

Note that although the Theorem 3.3.2 considers one asymmetric direction, it can be also applied for several (orthonormal) asymmetric directions. Suppose that there are $k \leq d$ orthonormal asymmetric directions $u_1, \ldots, u_k \in \mathbb{R}^d$ (such orthonormal set of asymmetric directions can be attained by the Gram-Schmidt process). Then for each u_i, SAM generates a bias into the flatter side along the u_i. By summing up the biases, we can say that the SAM generates a bias into a flatter region that is aware of all the asymmetric directions. Figure 3.2 gives an example of applying Theorem 3.3.2 for two asymmetric directions on $\mathbb{R}^2$. In Figure 3.2, the loss function has two asymmetric directions: along with the first coordinate (along with w_1) and along with the second coordinate (along with w_2). As we can see, SAM generates bias into the flatter side for each asymmetric direction and thus $\tilde{w}^*$ is shifted to the diagonal direction (sum of two asymmetric directions). Please refer to Section 7.1.2 for the proof of the theorem.

Next, we theoretically demonstrate that the bias generated by SAM (Theorem 3.3.2) leads to better generalization. We provide the following theorem as a corollary of Theorem 2.3.5:

Theorem 3.3.3. *(SAM leads to better generalization) Suppose that there exists a $(r, p, c, 0)$-asymmetric direction $u \in \mathbb{R}^d$ with respect to a local minimum $\hat{w}^* \in \mathbb{R}^d$ of empirical risk $\hat{\mathcal{L}} : \mathbb{R}^d \to \mathbb{R}$. Let $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ be the population loss. If assumption 2.3.3 holds for $R = \frac{c-1}{c+1}\rho$ and assumption 2.3.4 holds for $R' = \|\bar{\delta}\|_2 + \frac{c-1}{c+1}\rho$, then we have*

$$\mathbb{E}_\delta \mathcal{L}(\hat{w}^*) - \mathbb{E}_\delta \mathcal{L}(\tilde{w}^*) \geq \frac{(c-1)^2 p}{2(c+1)}\rho - 2\xi > 0 \tag{3.1}$$

for the SAM solution $\tilde{w}^* \in \mathbb{R}^d$ with $\rho \in \left(\frac{4(c+1)\xi}{(c-1)^2 p}, \frac{c+1}{c-1}\min\{r - \langle\bar{\delta}, u\rangle, \langle\bar{\delta}, u\rangle\}\right)$.

Theorem 3.3.3 says that the bias generated by SAM leads to better generalization with the certain condition on ρ. From the theorem, we obtain some insights on the value of ρ. First, the existence of a lower bound of ρ implies that using SAM with small ρ does not guarantee that SAM generalizes better. Second, the Equation 3.1 suggests using as large as possible ρ would provide maximal generalization benefit. For the proof of the theorem, please refer to Section 7.1.3.

However, since the upper bound of ρ is closely related to the specific asymmetric direction u (i.e., c and $\langle\bar{\delta}, u\rangle$), in practice, it is hard to obtain the maximal generalization benefit. Since there are many different asymmetric directions and thus many different conditions (ranges) on ρ, we have to choose the minimal one among the possible maximal ρ's. Nonetheless, if we can use different ρ for each asymmetric direction, we may reach a maximal generalization benefit. Thus this observation gives rise to the possibility that further improving direction of SAM through properly designed ρ-scheduling in some adaptive manner.

Remarks on $\zeta = 0$ In the Theorems, we assume that $\zeta = 0$ for simplicity. However, letting $\zeta = 0$ necessarily means that there is a 'jump' in the gradient at $\ell = 0$ and this condition cannot hold for smooth functions. If we consider the ReLU networks, there's no problem since its loss landscape is non-smooth. Nevertheless, even in the case of neural networks with smooth activation functions (such as tanh and sigmoid), letting $\zeta = 0$ does not severely harm the validity of our theory since ζ was introduced to handle the very small neighbor around the local minima that have near-zero gradients. Please note that the original work [21] also used this simplifying assumption.

Chapter 4. Proposed Framework

In this section, we propose a novel framework utilizing the SAM in a more efficient manner, *Parsimonious SAM (PSAM)*. Section 4.1 introduces our motivation for the framework, and section 4.2 provides the detailed method and algorithm of the framework.

4.1 Motivation

SAM [13] has suffered from the high computational cost involved in solving the inner maximization problem (Equation 2.1). According to the Equation 2.2, SAM essentially requires a doubled computational cost compared to the vanilla SGD training since the SAM computes two backpropagation operations for each update step. This expensive computational cost has become a major impediment to the scalability of SAM. To mitigate the issue, [14] proposed a more efficient SAM (referred to as ESAM) by considering stochastic weight perturbation and sharpness-sensitive data selection. However, it still requires an extra 40% computational cost than SGD (while SAM requires 100% extra cost). To further make the SAM efficient, we propose a novel framework with the following motivation.

Based on our previous result (Theorem 3.2.1), we raise the following question: ***If the role of SAM at the early phase of training is all that supports escaping a minimum faster, is it necessary to use SAM for every update?*** As an answer to this question, we speculate that at the early phase of training, periodically using SAM update may be sufficient to quickly escape a minimum since all we need for escaping minima is some noisy gradient. However, in the latter phase, SAM update should be used densely (i.e., for every update step) to guarantee that (a) consider as many as possible asymmetric directions and (b) converge to the biased solution.

In addition to our theory, some experimental results also support our vision. Table 5.2 shows that SGD $\rightarrow$ SAM outperforms SAM $\rightarrow$ SGD, which means that SAM is more effective on the exploitation phase. With these intuitions, we develop a new scheme exploiting SAM in the next subsection.

4.2 Method

We propose a novel framework utilizing the SAM in a more efficient way called ***Parsimonious SAM (PSAM)***. The *PSAM* consists of two phases. In the early phase of training ($\sim E$ steps), it uses SAM update once for $n-1$ vanilla SGD updates (i.e., period n). In the later phase of training (E steps $\sim$), it uses SAM for every update. Here, the threshold E determines whether the current time step is in the early phase or latter phase and the period n are hyperparameters.

Computational Cost of PSAM Since PSAM periodically requires two backpropagation, its computational cost can be drastically reduced. The computation cost of PSAM depends on the threshold E and period n. Let the extra computational cost of SAM be 1. Then, the computational cost of PSAM with E, n is

$$\left(1-\frac{E}{T}\right)+\frac{E}{T}\times\frac{1}{n}=1-\frac{(n-1)E}{nT},$$

where T is the number of total training steps. For instance, if we let $E = 0.75T$ and $n = 20$, the extra computational cost of PSAM is 0.2875 (i.e., 28.75% of original SAM). Note that the extra computation cost of ESAM [14] is about 0.4. In the experiment section, we verify that $E = 0.75T$ and $n = 20$ is actually a feasible option (does not significantly hurt the performance).

We provide the algorithm of PSAM as follows:

Algorithm 1 PSAM Algorithm

Require: Loss function $\ell(\mathbf{x}, \mathbf{y}; w)$, Model parameter $w \in \mathbb{R}^d$, Batch size B, Learning rate $\eta > 0$, Neighborhood size $\rho > 0$, Threshold E, Period n.

Initialize parameters w_0.

for $t = 0$ to T **do**

  Sample Batch $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_B, \mathbf{y}_B)\}$

  Empirical Risk $\hat{\mathcal{L}}(w_t) \triangleq \frac{1}{B} \sum_{i=1}^{B} \ell(\mathbf{x}_i, \mathbf{y}_i; w_t)$

  if $t < E$ **then**

    if $t \bmod n = 0$ **then**

      $\hat{\epsilon}(w_t) = \rho \frac{\nabla_w \hat{\mathcal{L}}(w_t)}{\|\nabla_w \hat{\mathcal{L}}(w_t)\|_2}$ ▷ Equation 2.2 with $p = 2$

      $w_t \leftarrow w_t - \eta \nabla_w \hat{\mathcal{L}}(w)|_{w_t + \hat{\epsilon}(w_t)}$ ▷ SAM update

    else

      $w_t \leftarrow w_t - \eta \nabla_w \hat{\mathcal{L}}(w_t)$ ▷ SGD update

    end if

  else

    $\hat{\epsilon}(w_t) = \rho \frac{\nabla_w \hat{\mathcal{L}}(w_t)}{\|\nabla_w \hat{\mathcal{L}}(w_t)\|_2}$ ▷ Equation 2.2 with $p = 2$

    $w_t \leftarrow w_t - \eta \nabla_w \hat{\mathcal{L}}(w)|_{w_t + \hat{\epsilon}(w_t)}$ ▷ SAM update

  end if

end for

return w_T

Remarks on PSAM Note that PSAM is orthogonal to the existing efficient version of SAM, Efficient SAM (ESAM) [14]. While PSAM manages the frequency of SAM updates, ESAM considers efficient ways (SWP and SDS, please refer to [14] for the details) for each SAM update. Thus, PSAM becomes much further efficient if the ESAM method is applied to every SAM update simultaneously.

Chapter 5. Empirical Analysis

In this section, we provide an empirical analysis of our theory and the proposed method. Section 5.1 presents the experimental setup including datasets, architectures, and hyperparameters. Section 5.2 provides empirical evidence that SAM can escape a minimum faster than SGD, and Section 5.3 supports that SAM generates biases into the flatter region. Section 5.4 provides experimental results on the proposed method *PSAM*.

5.1 Experimental Setup

We evaluate the validity of our proposed theory and method on various architectures and datasets. Here we present an experimental setup including dataset, model architectures, and hyperparameters. All the experiments are conducted on a single machine equipped with 8 RTX 3090 GPUs and JAX [32] is used for the experimental framework. We use CIFAR-10,100 [31] which are the widely used computer vision benchmark datasets. We investigate our theory and method with four common CNN architectures: ResNet-18 [27], Preactivation ResNet-164 [28], WideResNet-28x10 [29], and PyramidNet-110 [30].

We use batch size 256 for all experiments to follow the m-sharpness strategy of SAM with $m = 32$. We train the networks using SGD with Nesterov momentum [33] with momentum 0.9 and cosine learning rate scheduling [34] (No warmup). We use learning rate 0.1 and weight decay coefficient 5×10^{-4} for all architectures. The training epochs are set to 200 epochs for all architectures except PyramidNet-110 (300 epochs). We set $\rho = 0.15$ for ResNet-18, $\rho = 0.05$ for Preactivation ResNet-164 and WideResNet-28x10 on CIFAR-10 and set $\rho = 0.3$ for ResNet-18 and $\rho = 0.1$ for Preactivation ResNet-164 and WideResNet-28x10 on CIFAR-100. For PyramidNet-100, we use $\rho = 0.2$ for both on CIFAR-10 and 100. All the details are applied identically to both SGD and SAM.

5.2 Escape Efficiency of SAM

To verify that SAM escapes a minimum faster than SGD, we conduct the following experiment: start from the local minimum, and measure how many update steps are required to escape the minimum. More precisely, to obtain the local minima, we train the model with SGD for 200 epochs. And start from the pre-trained model, run SGD and SAM with the same learning rate (constant learning rate) and same batch size, we count how many steps are required until the training accuracy falls below 90%.

For the pre-training, we use the same configurations as mentioned in Section 5.1. For escape test, we use batch size 256 for all architectures and learning rate 0.03 for ResNet-18 and WideResNet-28x10, 0.022 for PreActivation ResNet-164, and 0.025 for PyramidNet-110. The learning rates are selected by grid search with $\{0.02, 0.021, \ldots, 0.03\}$ to avoid that escape too quickly or failing to escape.

Table 5.1 shows that SAM requires much smaller steps to escape the minimum than SGD. And as ρ increases, the escaping speed becomes much faster. For the typical value of ρ (0.15 for ResNet-18, 0.05 for PreActResNet-164 and WideResNet-28x10, 0.2 for PyramidNet-110), we can see that SAM escapes the minimum approximately 8 to 50 times faster than SGD. Hence, we can conclude that the results support our theory (Theorem 3.2.1) very well.

Table 5.1: Number of steps required to escape the local minimum on CIFAR-10. We repeat the experiment 100 times with different random seeds and report mean ± std.

Method	Architecture			
	ResNet18	**PreActResNet164**	**WideResNet28x10**	**PyramidNet110**
SGD	139.32 ±54.83	100.50 ±41.25	366.05 ±145.38	104.43 ±37.77
SAM ($\rho = 0.05$)	136.50 ±161.04	50.98 ±62.91	67.81 ±45.65	21.82 ±9.17
SAM ($\rho = 0.1$)	18.91 ±7.26	14.06 ±2.64	15.30 ±4.18	7.16 ±4.08
SAM ($\rho = 0.2$)	4.18 ±2.95	7.88 ±1.62	6.40 ±2.46	2.19 ±1.00
SAM ($\rho = 0.3$)	2.23 ±0.55	3.30 ±2.03	2.61 ±1.14	1.16 ±0.39
SAM ($\rho = 0.5$)	1.21 ±0.41	1.07 ±0.26	1.08 ±0.27	1.00 ±0.00
SAM ($\rho = 0.7$)	1.01 ±0.10	1.01 ±0.10	1.00 ±0.00	1.00 ±0.00

5.3 SAM on Asymmetric Valleys

To evaluate the biasing effect of SAM on asymmetric valleys, we design two experiments: (a) (SGD → SAM) Training the model with vanilla SGD for the early phase and then switching to SAM for the latter phase, and (b) (SAM → SGD) Training the model with SAM for early phase and then switch to vanilla SGD for latter phase. In both experiments, we switch the training scheme after 75% of total training epochs. Note that this criterion is widely used [35, 36] to divide the exploration phase and exploitation phase.

Table 5.2: Test accuracy (%) of SGD, SAM, SGD → SAM, and SAM → SGD on CIFAR-10 and 100. We repeat the experiment five times with different random seeds and report mean ± std.

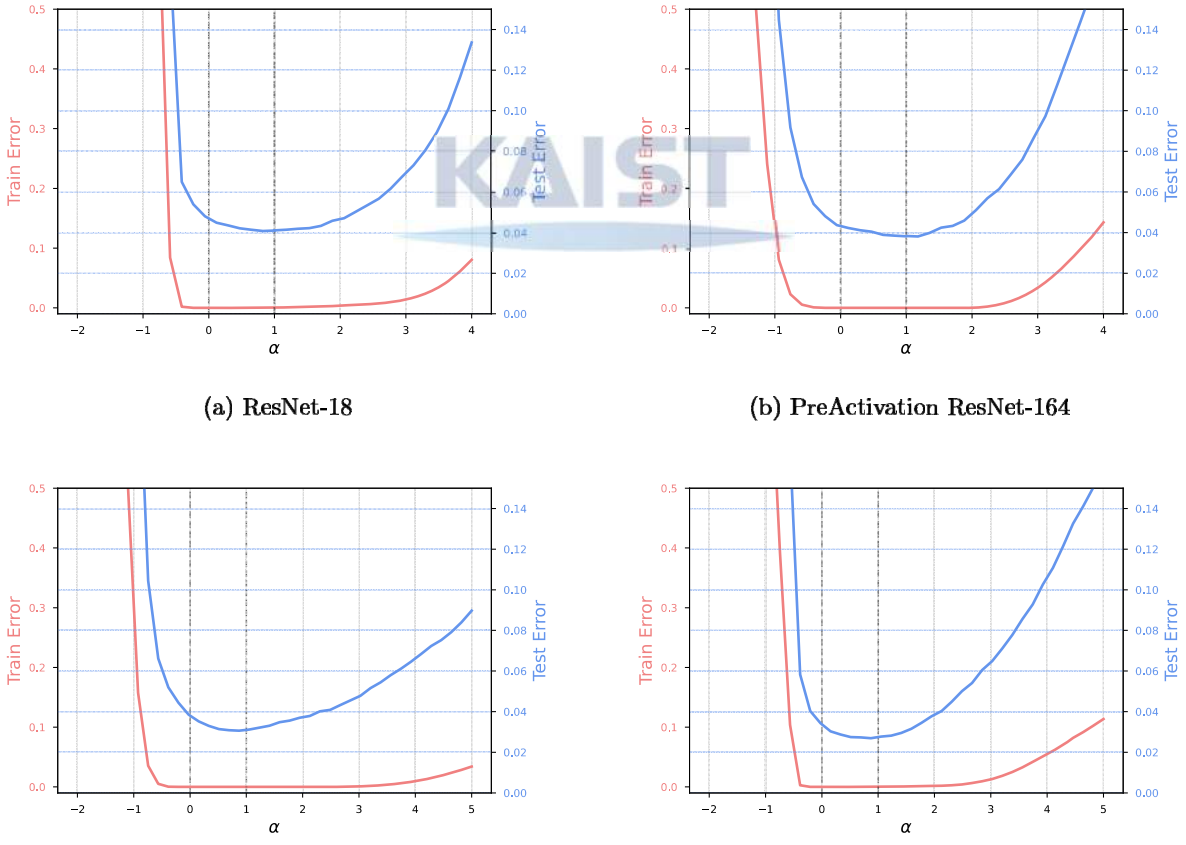
	Method	Architecture			
		ResNet18	**PreActResNet164**	**WideResNet28x10**	**PyramidNet110**
CIFAR-10	SGD	95.40 ±0.14	95.58 ±0.12	96.31 ±0.15	96.57 ±0.13
	SAM	96.13 ±0.08	96.45 ±0.11	97.21 ±0.06	97.39 ±0.07
	SGD → SAM	96.03±0.10	96.23 ±0.05	96.92 ±0.05	97.27 ±0.08
	SAM → SGD	96.00 ±0.13	96.02 ±0.12	96.94 ±0.08	96.88 ±0.06
CIFAR-100	SGD	78.78 ±0.21	78.48 ±0.25	81.23 ±0.18	82.47 ±0.23
	SAM	79.75 ±0.14	81.01 ±0.24	83.62 ±0.16	85.35 ±0.14
	SGD → SAM	80.15 ±0.10	80.26 ±0.17	83.01 ±0.18	84.62 ±0.20
	SAM → SGD	79.54 ±0.12	79.16 ±0.28	82.97 ±0.09	83.22 ±0.15

As seen in Table 5.2, SGD → SAM and SAM → SGD provide better generalization performances than vanilla SGD. It can be interpreted as SAM is effective not only at the early stage of training but also at the later stage of training. This is a noteworthy observation. The general belief in the performance improvement of SAM is that it finds a wide valley of minima, and where it converges in the found valley is not considered. However, this observation gives a novel insight that the performance improvement

of SAM does not only come from looking for a wide valley but also converges to a flatter region in the found valley.

To demonstrate that the effect of SAM on asymmetric valleys matches well with our theory (Theorem 3.3.2 and 3.3.3), we visualize the loss landscape between SGD solution and SGD → SAM (Figure 5.1) and between SAM solution and SAM → SGD solution (Figure 5.2). In the figures, we plot the train error rate curve (red curve) and test error rate curve (blue curve) along the line through two solutions $(1-\alpha)w_1 + \alpha w_2$.

In Figure 5.1, when we switch to SAM after SGD training, we can see that SGD → SAM solution is located in the more flat region in the valley compared to the corresponding SGD solution (does not switch to SAM). And in the Figure 5.2, when we switch to SGD after SAM training, we can see that SAM → SGD solution is on the sharper side on asymmetric valleys while the corresponding SAM solution (which does not switch to SGD) is in a flatter region. These results support that SAM generates biases into the flatter side of the asymmetric valley and it leads to better generalization. We provide additional loss landscape visualizations on CIFAR-100 in Section 7.2.



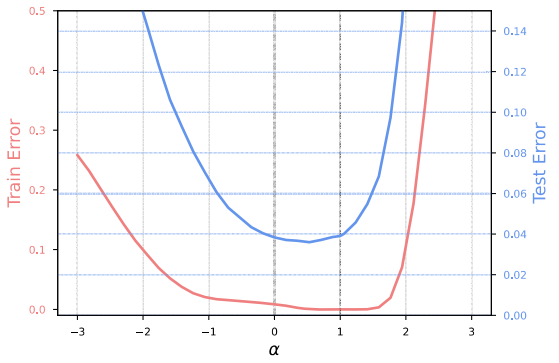
(a) ResNet-18

(b) PreActivation ResNet-164

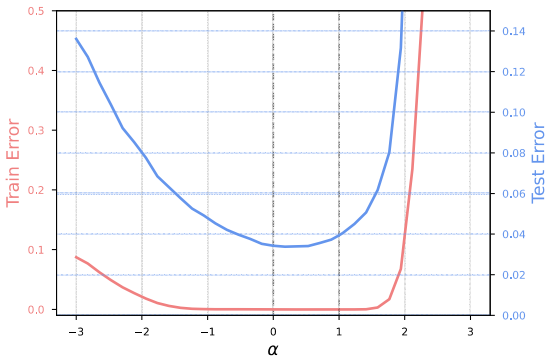
(c) WideResNet-28x10

(d) PyramidNet-110

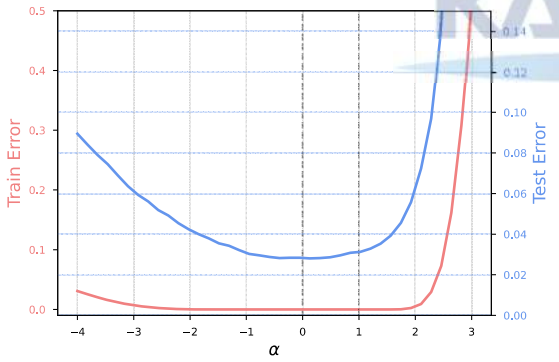
Figure 5.1: Loss landscape visualization between SGD solution and SGD → SAM solution on CIFAR-10. We visualize the train error rate curve (red curve) and test error rate curve (blue curve) between SGD solution and SGD → SAM solution. $\alpha = 0$ is the SGD solution and $\alpha = 1$ is the SGD → SAM solution.



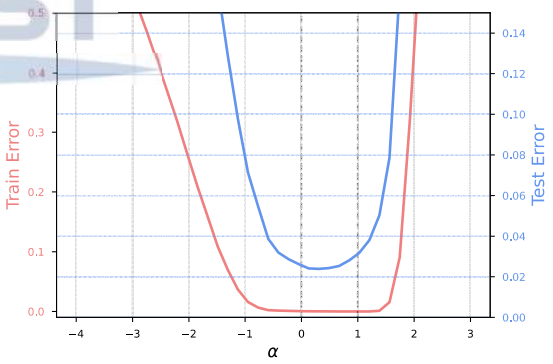
(a) ResNet-18



(b) PreActivation ResNet-164

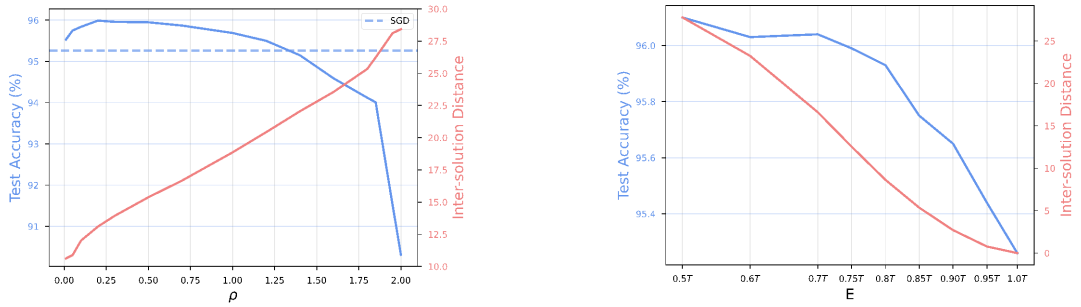


(c) WideResNet-28x10



(d) PyramidNet-110

Figure 5.2: Loss landscape visualization between SAM solution and SAM $\rightarrow$ SGD solution on CIFAR-10. We visualize the train error rate curve (red curve) and test error rate curve (blue curve) between SAM solution and SAM $\rightarrow$ SGD solution. $\alpha = 0$ is the SAM solution and $\alpha = 1$ is the SAM $\rightarrow$ SGD solution.



(a) Different ρ (b) Different switch epochs E

Figure 5.3: Test accuracy (%) of SGD $\rightarrow$ SAM and inter-solution ℓ_2 distances between SGD solution and SGD $\rightarrow$ SAM solution for different ρ and switch epochs with ResNet-18 on CIFAR-10. Blue dashed line in the left figure indicates the averaged test accuracy of SGD.

Ablative Study on ρ and switch epoch According to the Theorem 3.3.2, the size of bias increases as ρ increases. Figure 5.3a shows that as ρ increases, the size of bias (inter-solution distance) increases as well and it leads to more better generalization. However, it is confirmed that the bias rather damages the performance when using too large ρ. Such ρ's seem to have gone beyond the upper bound of ρ is provided in the Theorem 3.3.3 which ensures the better generalization. Figure 5.3b shows that as more training with SAM (as switch epoch become smaller), the inter solution distance increases and has better generalization. This supports that the more SAM is used, the more asymmetric directions are considered and thus has greater bias (larger inter-solution distance) and has better generalization.

Experimental Motivation of PSAM On the other hand, we can observe an interesting tendency in Table 5.2 : SGD $\rightarrow$ SAM leads to better performance compared to SAM $\rightarrow$ SGD in 7 out of 8 cases. This means that the effect of SAM in the exploitation phase is greater than in the exploration phase. In other words, it can be seen that the main effect of SAM is to find a wider region from the found valley rather than to find wide valleys. This observation motivates our method, *PSAM*.

5.4 Experimental Results on PSAM

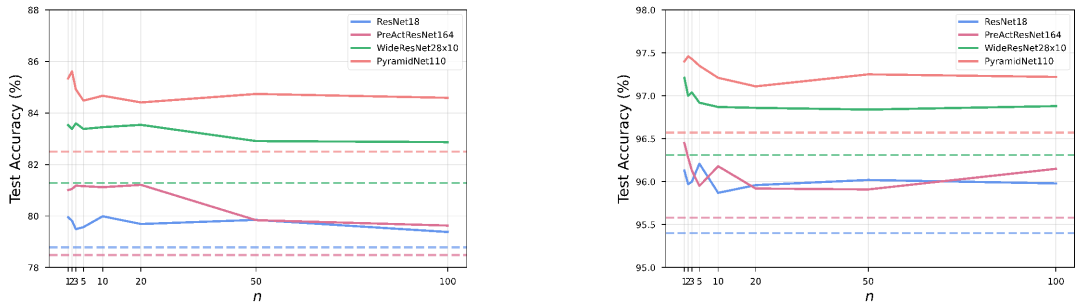
We evaluate the effectiveness of PSAM on CIFAR-10 and 100 with various architectures. All training environments are the same as mentioned in Section 5.1.

Table 5.3: Test accuracy (%) of PSAM on CIFAR-10 and CIFAR-100. We repeat the experiment five times with different random seeds and report mean ± std.

	Method	Architecture			
		ResNet18	**PreActResNet164**	**WideResNet28x10**	**PyramidNet110**
CIFAR-10	SGD	95.40 ±0.14	95.58 ±0.12	96.31 ±0.15	96.57 ±0.13
	SAM	96.13 ±0.08	96.45 ±0.11	97.21 ±0.06	97.39 ±0.07
	PSAM ($n = 20$)	95.96 ±0.10	95.91 ±0.05	96.85 ±0.05	97.25 ±0.08
CIFAR-100	SGD	78.78 ±0.21	78.48 ±0.25	81.23 ±0.18	82.47 ±0.23
	SAM	79.75 ±0.14	81.01 ±0.24	83.62 ±0.16	85.35 ±0.14
	PSAM ($n = 20$)	79.69 ±0.10	81.22 ±0.23	83.53 ±0.17	84.41 ±0.20

Table 5.3 shows that PSAM provides a comparable performance with the original SAM while it requires only $\approx 65\%$ computational cost of original SAM (relative computational cost : 1 for SGD, 2 for SAM, 1.4 for ESAM [14], and 1.2875 for PSAM). However, in a few cases, it is confirmed that the PSAM's performance is significantly lower than SAM, but it is much better than SGD.

Ablative Study on n Through the ablative study on the period n, we find that it is possible to obtain the desired performance by tuning the period n. Figure 5.4 shows that smaller n tend to have better performance while it hurts the efficiency of PSAM. And also we can see that the correlation between the period and performance is highly dependent on the dataset and architecture. Hence, there is a trade-off between the efficiency and performance in PSAM and the sweet spot should be found by the hyperparameter tuning.



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

Figure 5.4: Test accuracy (%) of PSAM with different n. Dashed lines indicate the test accuracy of SGD.

Chapter 6. Concluding Remark

In this paper, beyond the common belief of SAM, we demonstrate that the performance improvement of SAM originated from two fundamental causes. First, SAM can escape minima faster than SGD. As consequence, SAM may explore more minima than SGD under the same time constraint and can escape from more flat minima where SGD would be trapped. Accordingly, SAM is more likely to find wider valleys of minimum than SGD. Second, SAM biases a solution toward the flatter side of asymmetric valleys and it leads to better generalization. On the basis of our theory, we also suggest an efficient way of using SAM (Parsimonious SAM) and corroborate its applicability on a variety of architectures and datasets.

To the best of our knowledge, we are the first to conceptually scrutinize the SAM in terms of escape efficiency and asymmetric valleys. Our work fills the gap between qualitative knowledge and quantitative theoretical analysis regarding the generalizability of SAM. We believe the proposed theory not only clarifies how SAM finds flat minima but also offers researchers a novel perspective for future research on the behavior of SAM and the design of better sharpness-aware minimization algorithms.

Chapter 7. Supplementary Materials

7.1 Proofs of Theorems

7.1.1 Proof of Theorem 3.2.1

Proof. We begin with the following lemmas:

Lemma 7.1.1. *For a quadratic function $f(x) : \mathbb{R} \to \mathbb{R}$, $\frac{d^2}{dx^2} f(x^*) = \frac{2}{\delta^2} \left(f(x^* + \delta) - f(x^*) \right)$ where x^* is the minimum of f.*

Proof. Consider the second-order Taylor expansion of f at x^*:

$$\begin{aligned} f(x^* + \delta) &= f(x^*) + \frac{d}{dx} f(x^*)\delta + \frac{1}{2} \left(\frac{d^2}{dx^2} f(x^*) \right) \delta^2 \\ &= f(x^*) + \frac{1}{2} \left(\frac{d^2}{dx^2} f(x^*) \right) \delta^2. \end{aligned}$$

since x^* is the minimum of f and thus $\frac{d}{dx} f(x^*) = 0$. Note that since f is quadratic, the second-order Talyor expansion is exactly same as f. Hence we have

$$f(x^* + \delta) - f(x^*) = \frac{1}{2} \left(\frac{d^2}{dx^2} f(x^*) \right) \delta^2,$$

and thus

$$\frac{d^2}{dx^2} f(x^*) = \frac{2}{\delta^2} \left(f(x^* + \delta) - f(x^*) \right).$$

□

Lemma 7.1.2. *Let $\mathcal{L}(w) : \mathbb{R} \to \mathbb{R}$ and w^* be a local minimum of $\mathcal{L}(w)$. Assume that there exists an open neighborhood $D \subseteq \mathbb{R}$ containing w^* such that $\mathcal{L}$ is locally quadratic on D. And let $B_r(w^*) \subseteq \mathbb{R}$ be the largest open ball such that $B_r(w^*) \subseteq D$. Then for $\tilde{\mathcal{L}}(w) = \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(w + \epsilon) : \mathbb{R} \to \mathbb{R}$ with $\rho \in (0, r)$, we have $\Delta\mathcal{L} = \Delta\tilde{\mathcal{L}} + \frac{\rho^2}{2} \frac{d^2}{dw^2} \mathcal{L}(w^*) > \Delta\tilde{\mathcal{L}}$ where $\tilde{w}^* \in \mathbb{R}$ is the minimum of $\tilde{\mathcal{L}}$ and $\Delta\mathcal{L} = \min_{w \in \partial D} \mathcal{L}(w) - \mathcal{L}(w^*)$.*

Proof. Without loss of the generality, we simply let $w^* = 0$ and $\mathcal{L}(w) = \alpha w^2$ with $\alpha > 0$ for $w \in (-r, r)$ since any quadratic function can be obtained by parallel translation of αw^2. And also, since $\rho < r$, we observe that

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w + \rho) = \alpha(w + \rho)^2, & w \in [0, r - \rho) \\ \mathcal{L}(w - \rho) = \alpha(w - \rho)^2, & w \in (-r + \rho, 0) \end{cases}$$

And hence, we know that $\tilde{w}^* = 0$. Since taking maximum over some neighborhood does not changes maximum values, we know that $\min_{w \in \partial D} \mathcal{L}(w) = \min_{w \in \partial D} \tilde{\mathcal{L}}(w)$. However, since $\mathcal{L}(w^*) = 0$ and $\tilde{\mathcal{L}}(\tilde{w}^*) = \tilde{\mathcal{L}}(0) = \mathcal{L}(\rho) = \alpha\rho^2$, we obtain

$$\begin{aligned} \Delta\mathcal{L} &= \min_{w \in \partial D} \mathcal{L}(w) - \mathcal{L}(w^*) = \min_{w \in \partial D} \mathcal{L}(w) - \tilde{\mathcal{L}}(\tilde{w}^*) + \tilde{\mathcal{L}}(\tilde{w}^*) - \mathcal{L}(w^*) \\ &= \min_{w \in \partial D} \tilde{\mathcal{L}}(w) - \tilde{\mathcal{L}}(\tilde{w}^*) + \tilde{\mathcal{L}}(\tilde{w}^*) - \mathcal{L}(w^*) \\ &= \Delta\tilde{\mathcal{L}} + \alpha\rho^2 = \Delta\tilde{\mathcal{L}} + \frac{\rho^2}{2} \frac{d^2}{dw^2} \mathcal{L}(w^*) > \Delta\tilde{\mathcal{L}}. \end{aligned}$$

since $\alpha = \frac{1}{2}\frac{d^2}{dw^2}\mathcal{L}(w^*)$. □

Lemma 7.1.3. *Let $\mathcal{L}(w) : \mathbb{R} \to \mathbb{R}$ and w^* be a local minimum of $\mathcal{L}(w)$. Assume that there exists an open neighborhood $D \subseteq \mathbb{R}$ containing w^* such that $\mathcal{L}$ is locally quadratic on D. And let $B_r(w^*) \subseteq \mathbb{R}$ be the largest open ball such that $B_r(w^*) \subseteq D$. Then for $\tilde{\mathcal{L}}(w) = \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}(w+\epsilon) : \mathbb{R} \to \mathbb{R}$ with $\rho \in (0, r)$, we have $\left(1 + \frac{\rho}{r-\rho}\right)\frac{d^2}{dw^2}\mathcal{L}(w^*) \le \frac{d^2}{dw^2}\tilde{\mathcal{L}}(\tilde{w}^*)$ where $\tilde{w}^* \in \mathbb{R}$ is the minimum of $\tilde{\mathcal{L}}$.*

Proof. Similar to the proof of Lemma 7.1.2, since $\rho < r$, we observe that

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w+\rho) = \alpha(w+\rho)^2, & w \in [0, r-\rho) \\ \mathcal{L}(w-\rho) = \alpha(w-\rho)^2, & w \in (-r+\rho, 0) \end{cases}$$

While $\tilde{\mathcal{L}}$ is not differentiable at the minimum $\tilde{w}^*$, we can approximate the second-order derivative $\frac{d^2}{dw^2}\tilde{\mathcal{L}}(\tilde{w}^*)$ through the function value sensitivity $\tilde{\mathcal{L}}(\tilde{w}^* + \delta) - \tilde{\mathcal{L}}(\tilde{w}^*)$ by the Lemma 7.1.1. Note that the Lemma 7.1.1 only holds for twice differentiable functions. However, even though the $\tilde{\mathcal{L}}(w)$ is not differentiable at the minimum $\tilde{w}^* = 0$, it is a continuous convex function (in fact, quadratic almost everywhere), thus approximating the $\tilde{\mathcal{L}}(w)$ as a quadratic function would not involve any severe approximation error.

Since $\tilde{w}^* = 0$, we have

$$\begin{aligned} \tilde{\mathcal{L}}(\tilde{w}^* + \delta) - \tilde{\mathcal{L}}(\tilde{w}^*) &= \begin{cases} \mathcal{L}(\rho+\delta) - \mathcal{L}(\rho) = \alpha(\rho+\delta)^2 - \alpha\rho^2, & \delta \in [0, r-\rho) \\ \mathcal{L}(-\rho+\delta) - \mathcal{L}(\rho) = \alpha(-\rho+\delta)^2 - \alpha\rho^2, & \delta \in (-r+\rho, 0) \end{cases} \\ &= \begin{cases} \alpha\delta^2 + 2\alpha\delta\rho, & \delta \in [0, r-\rho) \\ \alpha\delta^2 - 2\alpha\delta\rho, & \delta \in (-r+\rho, 0) \end{cases} \\ &= \alpha\delta^2 + 2\alpha|\delta|\rho, \quad |\delta| < r - \rho \end{aligned}$$

Hence, we have

$$\frac{d^2}{dw^2}\tilde{\mathcal{L}}(\tilde{w}^*) = \frac{2}{\delta^2}(\tilde{\mathcal{L}}(\tilde{w}^* + \delta) - \tilde{\mathcal{L}}(\tilde{w}^*)) = \frac{2}{\delta^2}(\alpha\delta^2 + 2\alpha|\delta|\rho) = 2\alpha\left(1 + \frac{\rho}{|\delta|}\right) \ge 2\alpha\left(1 + \frac{\rho}{r-\rho}\right).$$

Then, since $\frac{d^2}{dw^2}\mathcal{L}(w^*) = 2\alpha$, we can conclude that $\frac{d^2}{dw^2}\tilde{\mathcal{L}}(\tilde{w}^*) \ge \left(1 + \frac{\rho}{r-\rho}\right)\frac{d^2}{dw^2}\mathcal{L}(w^*)$.

□

Now, let $v_1, \ldots, v_d \in \mathbb{R}^d$ be the orthonormal eigenvectors of H^* ($\nabla_w^2\mathcal{L}(w^*)$, Hessian of $\mathcal{L}(w)$ at the minimum w^*) that spans $\mathbb{R}^d$ (i.e., orthonormal basis of $\mathbb{R}^d$) which are corresponds to eigenvalues $\lambda_1, \ldots, \lambda_d$ where $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_d$. This is possible by the spectral theorem since H^* is a Hermitian matrix. Since SAM considers the direction which $\mathcal{L}$ grows the fastest, we can consider only v_1 that corresponds to the largest eigenvalue λ_1. Then by Lemma 7.1.2, we have that $\Delta\mathcal{L} = \Delta\tilde{\mathcal{L}} + \frac{\lambda_1}{2}\rho^2 = \Delta\tilde{\mathcal{L}} + \frac{\lambda_{\max}}{2}\rho^2$.

Next, let's prove that $\lambda_{\max} < \tilde{\lambda}_{\max}$. From our previous definitions, we know that v_1 is the eigenvector of $\lambda_1 = \lambda_{\max}$. Let $H^* = \nabla_w\mathcal{L}(w^*)$ and $\tilde{H}^* = \nabla_w\tilde{\mathcal{L}}(\tilde{w}^*)$. If we consider the second-order derivative along direction v_1, by Lemma 7.1.3, we can say that

$$v_1^\top \tilde{H}^* v_1 \ge \left(1 + \frac{\rho}{r-\rho}\right) v_1^\top H^* v_1 = \left(1 + \frac{\rho}{r-\rho}\right)\lambda_1 \tag{7.1}$$

since the second order derivative along v is $v^\top Hv$. Then, since the $\tilde{\lambda}_{\max} = \max_{\|v\|_2=1} v^\top \tilde{H}^* v$ and (7.1) means that $v_1^\top \tilde{H}^* v_1 \geq \left(1 + \frac{\rho}{r-\rho}\right) \lambda_1$ with $\|v_1\|_2 = 1$, we can conclude that

$$\tilde{\lambda}_{\max} \geq v_1^\top \tilde{H}^* v_1 \geq \left(1 + \frac{\rho}{r-\rho}\right) \lambda_1 = \left(1 + \frac{\rho}{r-\rho}\right) \lambda_{\max}.$$

□

7.1.2 Proof of Theorem 3.3.2

Proof. We begin with the following lemmas:

Lemma 7.1.4. *Let $w^* = 0$ be a local minima of function $\mathcal{L} : \mathbb{R} \to \mathbb{R}$. Suppose that w^* is $(r, p, c, 0)$-asymmetric valley. That is, there are constants $p > 0, r > 0, c \geq 1$ such that $0 < \frac{d}{dw}\mathcal{L}(w) < p$ for $w \in (0, r)$ and $\frac{d}{dw}\mathcal{L}(w) < -cp$ for $w \in (-r, 0)$. Then, when we let $\tilde{\mathcal{L}}(w) = \max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(w + \epsilon)$ for $\rho \in (0, r)$, $\tilde{\mathcal{L}}$ has a local minima $\tilde{w}^* \in [\frac{c-1}{c+1}\rho, \rho)$.*

Proof. First, note that the asymmetric direction u is defined on $w \in (-r, r)$ and we consider $\tilde{\mathcal{L}}(w)$ only for $w \in (-r + \rho, r - \rho)$ to ensure that $w + \epsilon$ is inside the valley around the local minimum $w = 0$. From the definition of $\tilde{\mathcal{L}}(w)$, we observe that

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w + \rho) & \text{if } \mathcal{L}(w + \rho) \geq \mathcal{L}(w - \rho) \\ \mathcal{L}(w - \rho) & \text{if } \mathcal{L}(w + \rho) < \mathcal{L}(w - \rho) \end{cases}$$

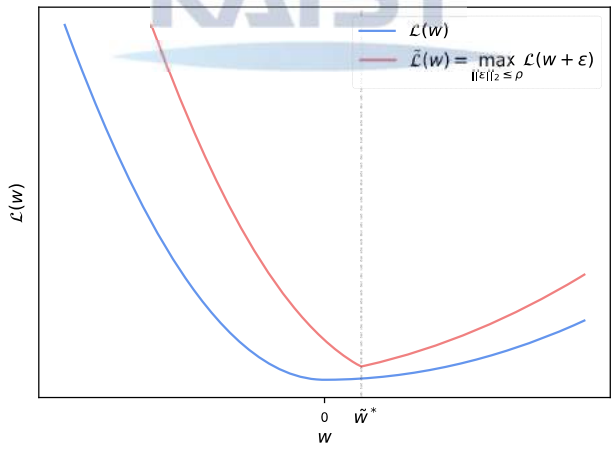


Figure 7.1: Visualization of $\mathcal{L}$ and $\tilde{\mathcal{L}}$ on $\mathbb{R}$.

Then we know that when $w \geq \rho$, we have $\mathcal{L}(w + \rho) \geq \mathcal{L}(w - \rho)$ since $\mathcal{L}$ is a increasing function on $(0, r)$ by assumption. Similarly, when $w \leq -\rho$, $\mathcal{L}(w + \rho) < \mathcal{L}(w - \rho)$ since $\mathcal{L}$ is a decreasing function on $(-r, 0)$.

For $-\rho < w < \rho$, we have $w + \rho > 0$, $w - \rho < 0$ and thus $\mathcal{L}(w + \rho) < p(w + \rho)$, $\mathcal{L}(w - \rho) > -cp(w - \rho)$ by the definition of asymmetric direction. Then if $w \in \mathbb{R}$ satisfies $p(w + \rho) < -cp(w - \rho)$(which is equivalent to $w < \frac{c-1}{c+1}\rho$), we have $\mathcal{L}(w + \rho) < \mathcal{L}(w - \rho)$.

Hence we obtain that

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w + \rho) & \text{if } w \geq \rho \\ \mathcal{L}(w - \rho) & \text{if } w < \frac{c-1}{c+1}\rho \end{cases}$$

For $\frac{c-1}{c+1}\rho \le w < \rho$, we cannot guarantee that $\mathcal{L}(w+\rho) \ge \mathcal{L}(w-\rho)$ or not since we only have upper bound of $\mathcal{L}(w+\rho)$ and lower bound of $\mathcal{L}(w-\rho)$. However, since $\tilde{\mathcal{L}}(w) = \mathcal{L}(w+\rho)$ is a increasing function on $w \ge \rho$ and $\tilde{\mathcal{L}}(w) = \mathcal{L}(w-\rho)$ is a decreasing function on $w < \frac{c-1}{c+1}\rho$, we can conclude that $\tilde{\mathcal{L}}(w)$ has a minimum $\tilde{w} \in [\frac{c-1}{c+1}\rho, \rho)$. Note that the exact position of $\tilde{w}$ within $[\frac{c-1}{c+1}\rho, \rho)$ depends on the actual function value. □

Lemma 7.1.5. *Let $w^* = 0$ be a local minima of function $\mathcal{L} : \mathbb{R} \to \mathbb{R}$. Suppose that w^* is $(r, a, b, 0)$-symmetric valley. That is, there are constants $b > a > 0, r > 0$ such that $a < \frac{d}{dw}\mathcal{L}(w) < b$ for $w \in (0, r)$ and $-b < \frac{d}{dw}\mathcal{L}(w) < -a$ for $w \in (-r, 0)$. Then, when we let $\tilde{\mathcal{L}}(w) = \max_{\|\epsilon\|_2 \le \rho} \mathcal{L}(w+\epsilon)$ for $\rho \in (0, r)$, $\tilde{\mathcal{L}}$ has a local minima $\tilde{w}^* \in (-\frac{b-a}{a+b}\rho, \frac{b-a}{a+b}\rho)$. Furthermore, if w^* is perfectly symmetric valley, i.e., $\mathcal{L}(-w) = \mathcal{L}(w)$ for any $w \in (-r, r)$, then $\tilde{w}^* = w^*$.*

Proof. Analogous to the proof of Lemma 7.1.4, we consider $\tilde{\mathcal{L}}(w)$ only for $w \in (-r+\rho, r-\rho)$ and assume that $\rho < r$. We observe that

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w+\rho) & \text{if } \mathcal{L}(w+\rho) \ge \mathcal{L}(w-\rho) \\ \mathcal{L}(w-\rho) & \text{if } \mathcal{L}(w+\rho) < \mathcal{L}(w-\rho) \end{cases}$$

And we know that $\mathcal{L}(w+\rho) > \mathcal{L}(w-\rho)$ for $w \ge \rho$ and $\mathcal{L}(w+\rho) < L(w-\rho)$ for $w \le -\rho$ since $\mathcal{L}(w)$ is a increasing function on $w > 0$ and decreasing function on $w < 0$. For $-\rho < w < \rho$, we have $w+\rho > 0$, $w-\rho < 0$ and thus we obtain

$$a(w+\rho) < \mathcal{L}(w+\rho) < b(w+\rho), \quad -a(w-\rho) < \mathcal{L}(w-\rho) < -b(w-\rho)$$

from the definition of $(r, a, b, 0)$-symmetric direction.

Then if w satisfies that $b(w+\rho) \le -a(w-\rho)$, then we have $\mathcal{L}(w+\rho) < b(w+\rho) \le -a(w-\rho) < \mathcal{L}(w-\rho)$ and thus $\mathcal{L}(w+\rho) < \mathcal{L}(w-\rho)$. Also, if w satisfies that $-b(w-\rho) \le a(w+\rho)$, we have $\mathcal{L}(w-\rho) < b(w-\rho) \le a(w+\rho) < \mathcal{L}(w+\rho)$ and thus $\mathcal{L}(w-\rho) < \mathcal{L}(w+\rho)$.

Hence, we obtain

$$\tilde{\mathcal{L}}(w) = \begin{cases} \mathcal{L}(w+\rho) & \text{if } w \ge \frac{b-a}{a+b}\rho > 0 \\ \mathcal{L}(w-\rho) & \text{if } w \le \frac{a-b}{a+b}\rho < 0 \end{cases}$$

Since $\tilde{\mathcal{L}}(w)$ is increasing on $w \ge \frac{b-a}{a+b}\rho > 0$ and $\tilde{\mathcal{L}}(w)$ is decreasing on $w \le \frac{a-b}{a+b}\rho < 0$, we can conclude that $\tilde{\mathcal{L}}(w)$ has a minimum at $\tilde{w} \in \left(\frac{a-b}{a+b}\rho, \frac{b-a}{a+b}\rho\right)$. Here, note that the exact position of $\tilde{w}$ depends on the actual value of function.

Furthermore, if $\mathcal{L}(w) = \mathcal{L}(-w)$ for any $w \in [0, r-\rho)$, we have $\mathcal{L}(w+\rho) \ge \mathcal{L}(w-\rho)$ for $w \ge 0$ and $\mathcal{L}(w+\rho) < \mathcal{L}(w-\rho)$ for $w < 0$. Thus $\tilde{\mathcal{L}}(w)$ has minimum at $\tilde{w}^* = w^* = 0$. □

Now we prove the main theorem. Let $\mathcal{L}_{w,u}(\ell) \triangleq \mathcal{L}(w+\ell u) : \mathbb{R} \to \mathbb{R}$ be the function $\mathcal{L}(w)$ along the asymmetric direction u. Then if we consider $\mathcal{L}_{w^*,u}(\ell)$, we know that $\mathcal{L}_{w^*,u}(\ell)$ has a minimum at $\ell^* = 0$ since w^* is a minimum of $\mathcal{L}(w)$ on $\mathbb{R}^d$. Since u is a $(r, p, c, 0)$-asymmetric direction with respect to the local minimum $w^* \in \mathbb{R}^d$ and function $\mathcal{L}(w)$, u is also a $(r, p, c, 0)$-asymmetric direction with respect to the local minimum $\ell^* = 0$ and function $\mathcal{L}_{w^*,u}(\ell)$ by the definition of $\mathcal{L}_{w^*,u}(\ell)$.

Then by the Lemma 7.1.4, we obtain that the perturbed loss $\tilde{\mathcal{L}}_{w^*,u}(\ell)$ has a local minimum $\tilde{\ell}^* \in \left[\frac{c-1}{c+1}\rho, \rho\right)$. Since $\tilde{\mathcal{L}}_{w^*,u}(\ell) = \tilde{\mathcal{L}}(w^* + \ell u)$ and $\tilde{\mathcal{L}}_{w^*,u}(\ell)$ has local minima $\tilde{\ell}^* > 0$, we can say that $\tilde{\mathcal{L}}(w)$ has a local minima $\tilde{w}^* = w^* + \tilde{\ell}^* u$. This is equivalent to

$$\langle \tilde{w}^*, u \rangle = \langle w^*, u \rangle + \tilde{\ell}^* \in \left[\langle w^*, u \rangle + \frac{c-1}{c+1}\rho, \langle w^*, u \rangle + \rho \right),$$

where $\tilde{w}^* \in \mathbb{R}^d$ is a local minimum of $\tilde{\mathcal{L}}(w)$.

Now, let's consider a $(r, a, b, 0)$-symmetric direction $v \in \mathbb{R}^d$. Then by the Lemma 7.1.5, we obtain that the perturbed loss $\tilde{\mathcal{L}}_{w^*,v}(\ell)$ has a local minimum $\tilde{\ell}^* \in \left(\frac{a-b}{a+b}\rho, \frac{b-a}{a+b}\right)$. Similar to the previous argument, we can conclude that $\tilde{\mathcal{L}}(w)$ has a local minimum $\tilde{w}^* = w^* + \tilde{\ell}^* v$ which is equivalent to

$$\begin{aligned}\langle \tilde{w}^*, v \rangle = \langle w^*, v \rangle + \tilde{\ell}^* &\in \left[\langle w^*, v \rangle + \frac{a-b}{a+b}\rho, \langle w^*, v \rangle + \frac{b-a}{a+b}\rho \right) \\ &= \left[\langle w^*, v \rangle - \frac{b-a}{a+b}\rho, \langle w^*, v \rangle + \frac{b-a}{a+b}\rho \right).\end{aligned}$$

□

7.1.3 Proof of Theorem 3.3.3

Proof. By the Theorem 3.3.2, we know that $\tilde{w}^* = \hat{w}^* + lu$ for $l \in \left[\frac{c-1}{c+1}\rho, \rho\right)$. In the Theorem 2.3.5, since larger l_i gives greater gap, simply letting $\tilde{w}^* = \hat{w}^* + \left(\frac{c-1}{c+1}\rho\right) u$ is enough to obtain the lower bound of loss gap between $\hat{w}^*$ and $\tilde{w}^*$. To apply the Theorem 2.3.5 for $\tilde{w}^*$, we let $l = \frac{c-1}{c+1}\rho \in \mathbb{R}$, Assumption 2.3.3 holds for $R = \frac{c-1}{c+1}\rho$ and Assumption 2.3.4 holds for $R' = \|\bar{\delta}\|_2 + \frac{c-1}{c+1}\rho$.

From the equality $l = \frac{c-1}{c+1}\rho$, the condition on l_i in Theorem 2.3.5 is translated to

$$\frac{4\xi}{(c-1)p} < \frac{c-1}{c+1}\rho < \min\{r - \langle \bar{\delta}, u \rangle, \langle \bar{\delta}, u \rangle\}$$

which is equivalent to

$$\frac{4(c+1)\xi}{(c-1)^2 p} < \rho < \frac{c+1}{c-1}\min\{r - \langle \bar{\delta}, u \rangle, \langle \bar{\delta}, u \rangle\}$$

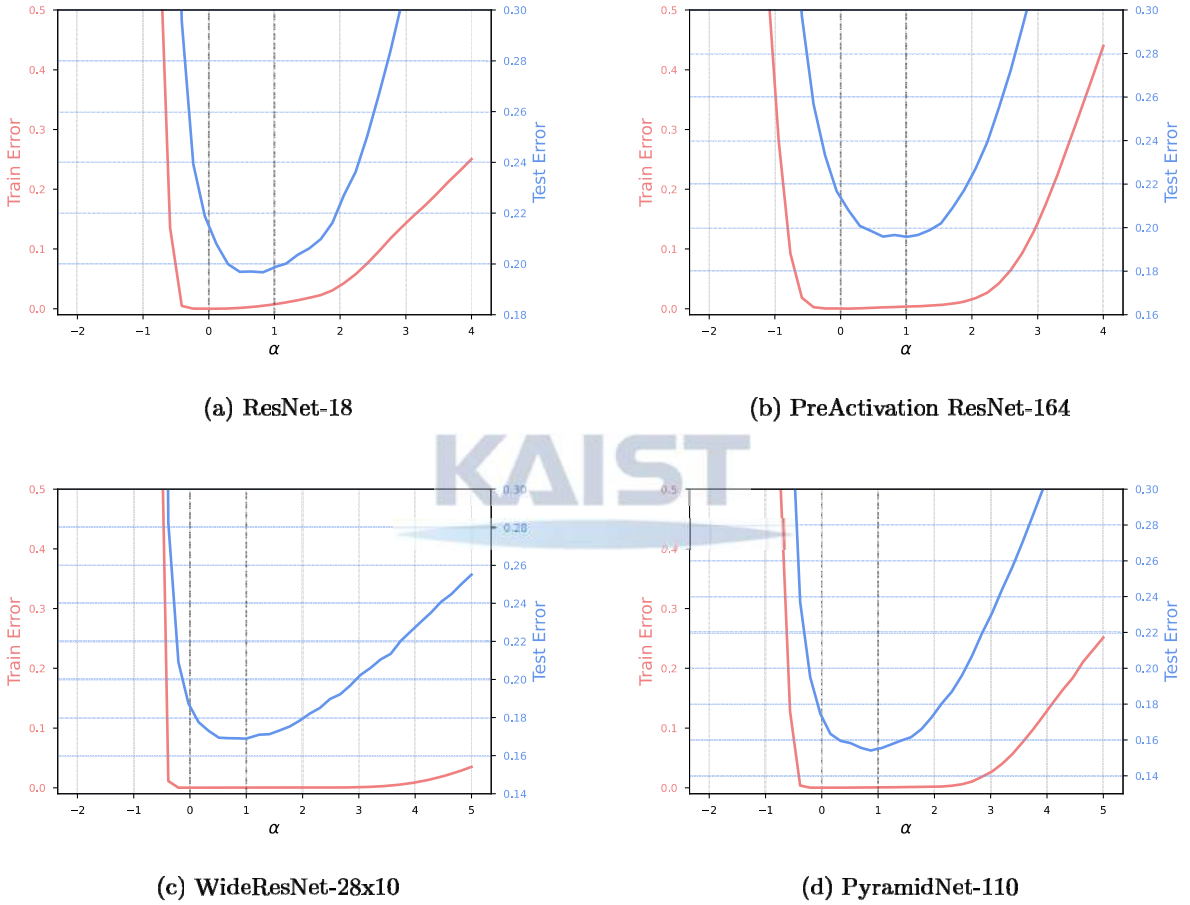
by substituting $\rho = \frac{c+1}{c-1}l$ and $\delta_1 = \langle \delta, u \rangle$ since δ_1 is shift amount along the direction u in Assumption 2.3.3. Then, by the Theorem 2.3.5, we obtain

$$\mathbb{E}_\delta \mathcal{L}(\hat{w}^*) - \mathbb{E}_\delta \mathcal{L}(\tilde{w}^*) \geq \frac{(c-1)^2 p}{2(c+1)}\rho - 2\xi > 0.$$

□

7.2 Additional Experimental Results

In addition to the results on Section 5.3, here we provide loss visualization results on CIFAR-100.



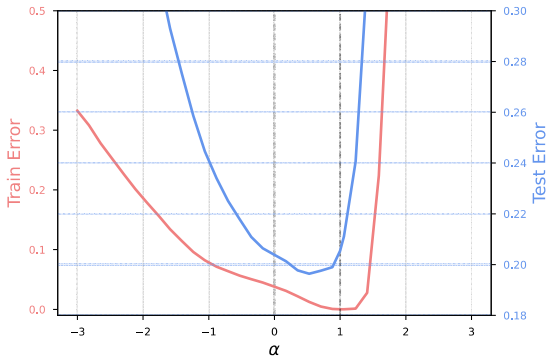
(a) ResNet-18

(b) PreActivation ResNet-164

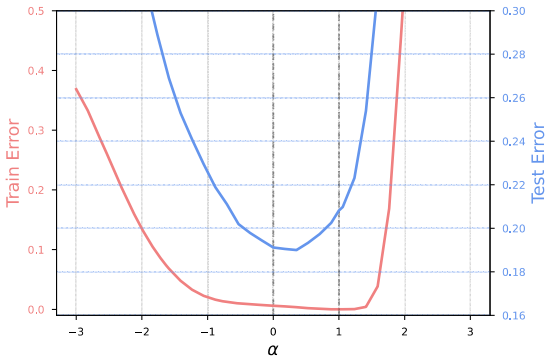
(c) WideResNet-28x10

(d) PyramidNet-110

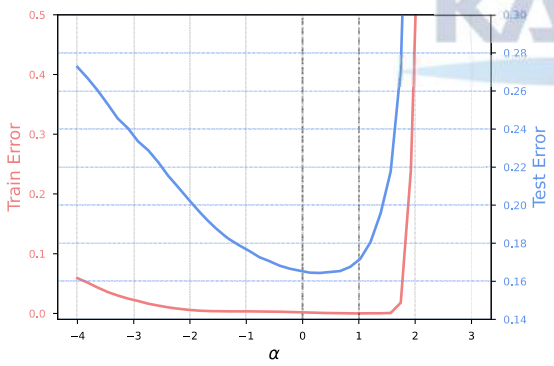
Figure 7.2: Loss landscape visualization between SGD solution and SGD → SAM solution on CIFAR-100. We visualize the train error rate curve (red curve) and test error rate curve (blue curve) between SGD solution and SGD → SAM solution. $\alpha = 0$ is the SGD solution and $\alpha = 1$ is the SGD → SAM solution.



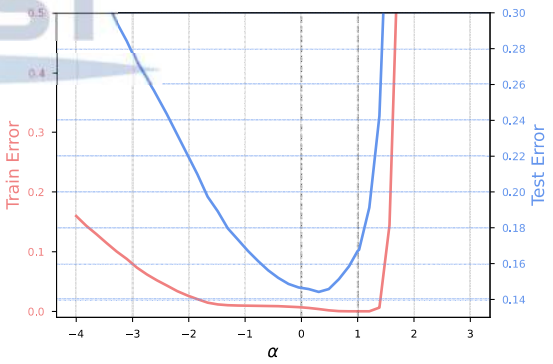
(a) ResNet-18



(b) PreActivation ResNet-164



(c) WideResNet-28x10



(d) PyramidNet-110

Figure 7.3: Loss landscape visualization between SAM solution and SAM $\rightarrow$ SGD solution on CIFAR-100. We visualize the train error rate curve (red curve) and test error rate curve (blue curve) between SAM solution and SAM $\rightarrow$ SGD solution. $\alpha = 0$ is the SAM solution and $\alpha = 1$ is the SAM $\rightarrow$ SGD solution.

Bibliography

[1] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021:124003, 12 2021.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[3] Nitish Shirish Keskar, Jorge Nocedal, Ping Tak Peter Tang, Dheevatsa Mudigere, and Mikhail Smelyanskiy. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1019–1028. PMLR, 06–11 Aug 2017.

[5] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[6] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 888–896. PMLR, 2019.

[7] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9636–9647. PMLR, 13–18 Jul 2020.

[8] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5905–5914. PMLR, 18–24 Jul 2021.

[9] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

[11] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[12] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[14] Jiawei Du, Hanshu Yan, Jiashi Feng, Joey Tianyi Zhou, Liangli Zhen, Rick Siow Mong Goh, and Vincent Tan. Efficient sharpness-aware minimization for improved training of neural networks. In *International Conference on Learning Representations*, 2022.

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

[16] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12360–12370, June 2022.

[17] Maksym Andriushchenko and Nicolas Flammarion. Towards understanding sharpness-aware minimization. *arXiv preprint arXiv:2206.06232*, 2022.

[18] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

[19] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

[21] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in Neural Information Processing Systems*, 32:2553–2564, 2019.

[22] Stanisław Jastrzebski, Zac Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Amos Storkey, and Yoshua Bengio. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[23] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

[24] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7654–7663. PMLR, 09–15 Jun 2019.

[25] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

[26] Hikaru Ibayashi and Masaaki Imaizumi. Quasi-potential theory for escape problem: Quantitative sharpness effect on sgd's escape from local minima. *arXiv preprint arXiv:2111.04004*, 2021.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[29] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

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

[31] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

[32] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs. *github*, 2018.

[33] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[35] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 876–885. AUAI Press, 2018.

[36] 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, 2021.