硕士学位论文

基于离散时间随机动力系统的常步长随机梯度下降的优化与泛化分析

UNDERSTANDING OPTIMIZATION AND GENERALIZATION OF CONSTANT STEP-SIZE SGD VIA DISCRETE-TIME RANDOM DYNAMICAL SYSTEMS

研　　究　　生：张海宇

指　导　教　师：朱一飞助理教授

南方科技大学

二〇二六年三月

国内图书分类号：O29　　学校代码：14325
国际图书分类号：519.2　　密级：公开

理学硕士学位论文

基于离散时间随机动力系统的常步长随机梯度下降的优化与泛化分析

学 位 申 请 人：张海宇
指　导　教　师：朱一飞助理教授
学　科　名　称：数学
答　辩　日　期：2026 年 5 月
培　养　单　位：数学系
学位授予单位：南方科技大学

UNDERSTANDING OPTIMIZATION AND GENERALIZATION OF CONSTANT STEP-SIZE SGD VIA DISCRETE-TIME RANDOM DYNAMICAL SYSTEMS

A dissertation submitted to
Southern University of Science and Technology
in partial fulfillment of the requirement
for the degree of
Master of Science
in
Mathematics

by
Zhang Haiyu

Supervisor: Assistant Prof. Zhu Yifei

May, 2026

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

刘 XX	教授	南方科技大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	南方科技大学
委员	刘 XX	教授	南方科技大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	南方科技大学

DECLARATION OF ORIGINALITY AND AUTHORIZATION OF THESIS, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Declaration of Originality of Thesis

I hereby declare that this thesis is my own original work under the guidance of my supervisor. It does not contain any research results that others have published or written. All sources I quoted in the thesis are indicated in references or have been indicated or acknowledged. I shall bear the legal liabilities of the above statement.

Signature: Date:

Declaration of Authorization of Thesis

I fully understand the regulations regarding the collection, retention and use of thesis of Southern University of Science and Technology.

1. Submit the electronic version of thesis as required by the University.

2. The University has the right to retain and send the electronic version to other institutions that allow the thesis to be read by the public.

3. The University may save all or part of the thesis in certain databases for retrieval, and may save it with digital, cloud storage or other methods for the purpose of teaching and scientific research. I agree that the full text of the thesis can be viewed online or downloaded within the campus network.

（1）I agree that once submitted, the thesis can be retrieved online and the first 16 pages can be viewed within the campus network.

（2）I agree that ☐ upon submission/ ☐ 12 months after submission, the full text of the thesis can be viewed and downloaded by the public.

4. This authorization applies to decrypted confidential thesis.

Signature of Author: Date:

Signature of Supervisor: Date:

摘 要

本文从离散时间随机动力系统的角度研究常步长随机梯度下降算法（SGD）的优化与泛化行为。通过将 SGD 建模为一个迭代函数系统，我们从吸收集、平稳分布以及随机吸引子的角度分析其渐近行为与泛化性。

在强凸且光滑的损失函数情形下，我们证明 SGD 对应的 Markov 链指数收敛到唯一的平稳分布，同时给出关于步长、批量大小，以及梯度噪声的显式优化误差界。在非凸情形下，参考 Shirokoff 和 Zaleski 的工作，我们考虑可分离损失函数这一简化情形。该工作证明系统存在多个互不相交的吸收集，每个吸收集上具有唯一遍历平稳测度，并给出描述整体收敛行为的遍历分解定理。我们在文中进一步探讨这一分布层面的结果与已有扩散近似框架下得到的结果的差异，并指出后者在刻画 SGD 长期行为上的局限性。同时，在此基础上，我们证明在每个吸收集内部存在随机单点吸引子，且在任意固定的迭代映射序列下，吸收集内部任意初始点出发的轨道最终都会以指数速率同步到同一条轨道。我们进一步证明沿同步轨道的 Lyapunov 指数为负，并据此推导出与步长、梯度噪声以及损失景观几何结构相关的局部优化误差界。

在泛化方面，我们以 Camuto 等人基于平稳分布分形维数所建立的泛化误差界为基础，将其应用于监督学习任务，从而得到关于算法超参数的显式泛化误差界。与仅依赖假设空间容量的经典泛化界相比，该理论结果能够刻画 SGD 动力学对泛化性能的影响，并在实际情况下给出更紧的界估计。我们通过实验验证了平稳分布分形维数作为泛化指标在简单深度学习任务中的有效性。此外，我们还发现权重轨迹的分形维数可以追踪随机优化算法迭代过程中不同的训练阶段，从而帮助我们更好地理解顿悟（Grokking）等现象。

这一框架将 SGD 动力学与其优化误差及泛化误差之间建立起联系，为刻画常步长 SGD 的渐近行为与泛化性提供了一种新的分析方法，同时为隐式正则化等其他问题的研究提供新的思路。

关键词：常步长随机梯度下降，随机动力系统，优化误差，泛化误差

ABSTRACT

We study the optimization and generalization behavior of constant step-size stochastic gradient descent (SGD) through the lens of discrete-time random dynamical systems. By modeling SGD as an iterated function system (IFS), we analyze its asymptotic behavior and generalization in terms of absorbing sets, stationary distributions, and random attractors.

In the setting of strongly convex and smooth loss function, we show that the associated Markov chain is independent of initializations and converges exponentially to a unique stationary distribution, and derive explicit optimization bounds in terms of step size, batch size, and gradient noise. In the non-convex setting, we consider a simplified case with separable loss function, building on the work of Shirokoff and Zaleski. They establish the existence of multiple disjoint absorbing sets each supporting a unique ergodic stationary measure and give an ergodic decomposition theorem to characterize the global convergence behavior. We discuss the discrepancies between this result and those obtained under existing diffusion approximation framework, and point out the limitations of the latter in characterizing the long-term behavior of SGD. Furthermore, we show that within each absorbing set there exists a random singleton attractor, and trajectories from arbitrary initializations eventually synchronize under any fixed sequence of iterative maps. We prove that the Lyapunov exponent along the synchronized trajectory is negative and based on it, derive corresponding local optimization error bounds related to the step size, gradient noise, and geometry of loss landscape.

For generalization, we follow the work of Camuto et al., which proposed a generalization bound based on the fractal dimension of the stationary distribution. We apply the theory to supervised learning tasks to obtain explicit bounds in terms of algorithmic hyperparameters. This bound goes beyond classical generalization bounds which rely solely on the capacity of the hypothesis class and are often overly loose. We conduct corresponding experiments to validate the effectiveness of the fractal dimension of the stationary distribution as a generalization measure for simple deep learning tasks. As a further extension, we find that the fractal dimension of the weight trajectory can serve as a progress indicator for tracking distinct training phases of stochastic iterative algorithms, offering new insights into phenomena such as Grokking.

This framework bridges the dynamics of SGD with its optimization and generalization performance, providing a promising approach to characterizing the asymptotic behavior and generalization properties of constant step-size SGD, and may offer new ideas for potential future research directions such as implicit regularization.

Keywords: Constant step-size SGD; Random dynamical system; Optimization error; Generalization error

TABLE OF CONTENTS

摘 要.....I
ABSTRACT.....II
CHAPTER 1 INTRODUCTION.....1
CHAPTER 2 PRELIMINARIES.....3
2.1 Random dynamical system and iterated function system.....3
2.2 Dimension theory of iterated function system.....7
CHAPTER 3 PROBLEM SETTING.....10
3.1 Optimization and generalization error in deep learning.....10
3.2 Constant step-size SGD as an iterated function system.....11
CHAPTER 4 OPTIMIZATION DYNAMICS AND CONVERGENCE OF THE ASSOCIATED MARKOV CHAIN.....13
4.1 SGD with strongly convex loss.....14
4.2 SGD with non-convex and separable loss.....18
4.2.1 Construction of absorbing sets.....19
4.2.2 Stationary distributions and exponential convergence rates of the associated Markov chain.....21
4.2.3 Random attractor and stability analysis on absorbing sets.....25
4.2.4 Asymptotic optimization error bounds for local and global minima.....32
CHAPTER 5 FRACTAL DIMENSION OF THE STATIONARY DISTRIBUTION AND GENERALIZATION.....37
5.1 Generalization error bound via fractal dimension of the stationary distribution37
5.2 Applications to supervised learning tasks.....40
5.3 Experiments.....41
5.3.1 Correlation between fractal dimension of stationary distribution and generalization gap.....42
5.3.2 Evolution of the fractal dimension of local weight trajectories during training process.....46
CONCLUSION AND FUTURE WORK.....49
REFERENCES.....52

APPENDIX A ADDITIONAL TECHNICAL BACKGROUND 56
APPENDIX B EXPERIMENT DETAILS 57
ACKNOWLEDGEMENTS 58
RESUME 59

CHAPTER 1 INTRODUCTION

As deep learning models and tasks grow increasingly complex and diverse, stochastic gradient descent (SGD) remains the dominant optimization method in modern deep learning [11]. Nevertheless, the optimization dynamics and generalization behavior of SGD are still not fully understood. This gap is particularly pronounced for constant step-size SGD. While constant step-size SGD is frequently employed in practical training, the majority of theoretical work has focused on SGD with decaying step sizes [24, 40], leaving the theoretical understanding of constant step-size SGD comparatively underdeveloped.

To bridge this gap, existing theoretical analyses of constant step-size SGD have largely been conducted within the stochastic differential equation (SDE) approximation framework [36, 39, 32]. For instance, [32] shows that a larger ratio of step size to batch size drives SGD toward wider, flatter minima, which correlates with improved generalization. These works provide valuable insights that align well with empirical observations. However, this framework suffers from significant limitations: rigorous theoretical justification for the SDE approximation is available only for SGD with vanishingly small learning rates and only over finite time horizons [36, 37].

As a complementary approach, some work has studied the convergence of constant step-size SGD through the lens of discrete-time Markov chain theory. It has been shown that constant step-size SGD does not converge to a single point, but rather to a stationary distribution, which has also been corroborated empirically [59]. This framework was systematically developed under quadratic loss functions by [20], proving that the iterates converge at an exponential rate to a unique stationary distribution and providing an explicit expansion of the mean squared error of the averaged SGD iterates as a power series in the step size η, which characterizes the influence of initialization, gradient noise, and step size on the limiting behavior. It has also been extended to nonconvex and non-smooth loss functions satisfying a dissipativity condition by [57]. Under this setting, they establish the uniqueness of the stationary distribution and prove asymptotic normality of the Polyak-Ruppert averaged iterates around the mean of the stationary distribution. They further characterize the bias between this mean and the critical points of the loss function, providing explicit bounds in terms of the step size.

Motivated by the theoretical gaps in understanding constant step-size SGD, and in-

spired by the aforementioned work, this paper studies the optimization and generalization behavior of constant step-size SGD from the perspective of discrete-time random dynamical systems (RDS). This perspective is more natural than the SDE approximation and goes beyond the Markov chain perspective by characterizing both the distributional and the pathwise behavior of SGD. Specifically, building on the work of [53, 13], we model SGD as an iterated function system (IFS), study the distributional convergence to a stationary distribution and the pathwise synchronization phenomenon, and give the corresponding optimization and generalization bounds. We conduct a theoretical analysis of SGD under certain conditions, complemented by empirical experiments, aiming to provide insights into the following questions:

（1）What are the optimization dynamics and asymptotic behavior of the constant step-size SGD in different regimes?

（2）How can we characterize the optimization error and generalization error using the algorithm hyperparameters?

（3）How do initialization and randomness arising from the selection of the iterated map (i.e., mini-batch selection) affect the optimization and generalization behavior of constant step-size SGD?

This paper is organized as follows. In Chapter 2, we introduce the background on RDS, IFS, and fractal dimensions, providing the key concepts and theorems which will be used throughout the subsequent analysis. In Chapter 3, we present the relevant background of deep learning, including optimization error and generalization error, and provide a formal IFS-based mathematical formulation of constant step-size mini-batch SGD to establish the basic setup for theoretical analysis.

Chapters 4 and 5 are the core parts of the paper, devoted to the analysis of optimization and generalization behavior of constant step-size SGD, respectively. In Chapter 4, we study the optimization dynamics and convergence behavior of constant step-size SGD at both the distributional and pathwise levels, under two settings: strongly convex loss functions and non-convex separable loss functions, and establish corresponding upper bounds on the optimization error. In Chapter 5, we introduce the generalization error bound based on the fractal dimension of the stationary distribution proposed in [13], and apply it to specific supervised learning task to establish an explicit relationship between the generalization error and the algorithm hyperparameters. We further conduct extended experiments to empirically characterize the generalization behavior.

CHAPTER 2 PRELIMINARIES

In this chapter, we introduce random dynamical systems [4, 22, 50] and dimension theory in fractal geometry [23, 45]. These concepts and results will be used in the subsequent analysis in Chapter 4 and 5.

2.1 Random dynamical system and iterated function system

We begin with some basic concepts of random dynamical system.

Definition 2.1 (Random dynamical system (RDS)): Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let $(\Omega, \mathcal{F}, \mathbb{P}, (\vartheta^t)_{t\in\mathbb{T}})$ be a metric dynamical system. A random dynamical system on $\mathcal{X}$ over ϑ is a mapping

$$\Phi : \mathbb{T} \times \Omega \times \mathcal{X} \to \mathcal{X}, \qquad (t, \omega, x) \mapsto \Phi(t, \omega, x),$$

such that:

(i) (Measurability) Φ is $(\mathcal{B}(\mathbb{T}) \otimes \mathcal{F} \otimes \mathcal{B},\ \mathcal{B})$-measurable.

(ii) (Cocycle property) For all $s, t \in \mathbb{T}$ and $\omega \in \Omega$,

$$\Phi(t + s, \omega, \cdot) = \Phi\big(t, \vartheta^s\omega, \cdot\big) \circ \Phi(s, \omega, \cdot),$$

and if $0 \in \mathbb{T}$ then $\Phi(0, \omega, \cdot) = \mathrm{id}_{\mathcal{X}}$ for all $\omega \in \Omega$.

To facilitate the study of random dynamics, it is often convenient to represent a random dynamical system by an induced deterministic dynamical system on the extended space $\Omega \times X$, namely the skew product.

Remark 2.1 (Skew product): Let Φ be an RDS on $(\mathcal{X}, \mathcal{B})$ over the base dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, (\vartheta^t)_{t\in\mathbb{T}})$. Then the mapping

$$\Theta^t(\omega, x) := \big(\vartheta^t\omega,\ \Phi(t, \omega, x)\big), \qquad t \in \mathbb{T}.$$

is a dynamical system on $(\Omega \times \mathcal{X}, \mathcal{F} \otimes \mathcal{B})$ which is called the skew product of the metric dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, (\vartheta^t)_{t\in\mathbb{T}})$ and the cocycle $\Phi(t, \omega, \cdot)$ on $\mathcal{X}$.

We introduce relevant concepts and key theorems of RDS below. Many of these notions and results can be viewed as generalizations of their counterparts in deterministic dynamical systems, while randomness in the iterated maps gives rise to a richer structure in RDS. The skew product allows one to apply classical results from deterministic ergodic

theory (e.g., the theory of invariant measures and the Oseledets multiplicative ergodic theorem) to extend the study of invariant measures and Lyapunov exponents to the random setting under appropriate measurability and integrability assumptions. These concepts and theorems will be employed in the analysis of Chapter 4.

Definition 2.2 (Invariant measure for an RDS): Let ν^* be a probability measure on $(\Omega \times \mathcal{X}, \mathcal{F} \otimes \mathcal{B})$. We call ν^* a invariant measure for Φ if

$$(\Theta^t)_{\#}\nu^* = \nu^* \text{ for all } t \in \mathbb{T}, \qquad \text{and} \qquad (\pi_\Omega)_{\#}\nu^* = \mathbb{P},$$

where $\pi_\Omega(\omega, x) = \omega$ and $(\cdot)_{\#}$ denotes the pushforward of measures.

A set $R \in \mathcal{F} \otimes \mathcal{B}$ is called a random set. The ω-section of a random set R is defined by

$$R(\omega) = \{x : (\omega, x) \in R\}, \quad \omega \in \Omega.$$

R is compact if $R(\omega)$ is a compact subset of $\mathcal{X}$ for $\mathbb{P}$-a.e. ω.

Definition 2.3 (Absorbing set for bounded sets): Let $\mathcal{B}$ denote a class of bounded deterministic subsets of $\mathcal{X}$. A random compact set $K(\omega)$ is called an absorbing set (w.r.t. $\mathcal{B}$) if for every bounded deterministic set $B \in \mathcal{B}$, there exists $t_B(\omega) > 0$ such that for $\mathbb{P}$-a.e. ω,

$$\Phi(t, (\vartheta^{-1})^t\omega, B) \subset K(\omega), \quad \forall t \geq t_B(\omega).$$

Definition 2.4 (Random attractor for bounded sets): Let $\mathcal{B}$ denote a class of bounded deterministic subsets of $\mathcal{X}$. A compact random set A is called strictly Φ-invariant if

$$\Phi(t, \omega, A(\omega)) = A(\vartheta^t\omega) \quad \text{for all } t \geq 0, \text{ for } \mathbb{P}\text{-a.e. } \omega.$$

It is called a random pullback attractor (w.r.t. $\mathcal{B}$) if it is strictly invariant and for every $B \in \mathcal{B}$,

$$\lim_{t\to\infty} \operatorname{dist}\big(\Phi(t, (\vartheta^{-1})^t\omega, B),\ A(\omega)\big) = 0 \quad \text{for } \mathbb{P}\text{-a.e. } \omega,$$

where $\operatorname{dist}(E, E') := \sup_{x\in E} \inf_{y\in E'} d(x, y)$ is the Hausdorff semi-distance.

It is called a random forward attractor (w.r.t. $\mathcal{B}$) if it is strictly invariant and for every $B \in \mathcal{B}$,

$$\lim_{t\to\infty} \operatorname{dist}\big(\Phi(t, \omega, B),\ A(\theta^t\omega)\big) = 0 \quad \text{for } \mathbb{P}\text{-a.e. } \omega.$$

If $A(\omega) = \{a(\omega)\}$ is a singleton for $\mathbb{P}$-a.e. ω, then A is called a random (pullback/-forward) singleton attractor. We also say that $a(\omega)$ is a stable random fixed point.

Theorem 2.1 (Multiplicative Ergodic Theorem for a discrete-time RDS): Let $(\mathcal{X}, \mathcal{B})$ be a d-dimensional C^1 Riemannian manifold. Assume there exists an invariant measure ν^* on $(\Omega \times \mathcal{X}, \mathcal{F} \otimes \mathcal{B})$ and that the integrability condition

$$\int_{\Omega\times\mathcal{X}} \log^+ \left\| D_x\Phi(1,\omega,x) \right\| d\nu^*(\omega,x) < \infty$$

holds. Then for ν^*-a.e. (ω, x), there exist an integer $k = k(\omega, x) \in \{1, \dots, d\}$, Lyapunov exponents

$$\lambda_1(\omega,x) > \lambda_2(\omega,x) > \cdots > \lambda_k(\omega,x),$$

and a flag of subspaces of the tangent space

$$T_x\mathcal{X} = V^1_{\omega,x} \supset V^2_{\omega,x} \supset \cdots \supset V^k_{\omega,x} \supset \{0\},$$

such that for all $i = 1, \dots, k$:

$$k(\Theta^1(\omega,x)) = k(\omega,x), \qquad \lambda_i(\Theta^1(\omega,x)) = \lambda_i(\omega,x),$$

and

$$D_x\Phi(1,\omega,x)\, V^i_{\omega,x} = V^i_{\Theta^1(\omega,x)}.$$

For every $v \in V^i_{\omega,x} \setminus V^{i+1}_{\omega,x}$ (with $V^{k+1}_{\omega,x} = \{0\}$), the limit exists and

$$\lim_{n\to\infty} \frac{1}{n} \log \left\| D_x\Phi(n,\omega,x)\, v \right\| = \lambda_i(\omega,x).$$

In particular, the maximal Lyapunov exponent is

$$\lambda_{\max}(\omega,x) := \lim_{n\to\infty} \frac{1}{n} \log \left\| D_x\Phi(n,\omega,x) \right\| = \lambda_1(\omega,x).$$

If ν^* is ergodic for Θ, then $k(\omega, x)$ and each $\lambda_i(\omega, x)$ are ν^*-a.e. constant, and so are the dimensions $\dim V^i_{\omega,x}$.

In this paper, our main mathematical object is the iterated function system (IFS). An IFS is a discrete-time random dynamical system, which is generated by products of random mappings.

Definition 2.5 (Iterated function system (IFS)[①]): Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Fix a finite index set $S = \{1, \dots, r\}$, maps $T_i : \mathcal{X} \to \mathcal{X}$, and probabilities $(p_i)_{i\in S}$. Let $\Omega = S^{\mathbb{N}}$, equip it with the product σ-algebra $\mathcal{F}$. The components of $\omega \in \Omega$ are i.i.d. and the product measure $\mathbb{P}$ on Ω is defined by $\mathbb{P} = p^{\otimes\mathbb{N}}$ with $p(\{i\}) = p_i$. Let $\sigma : \Omega \to \Omega$ be the left shift. Then $(\Omega, \mathcal{F}, \mathbb{P}, (\sigma^t)_{t\in\mathbb{N}})$ is a metric dynamical system. Define

$$\Phi(0,\omega,x) = x, \qquad \Phi(n,\omega,x) = T_{\omega_{n-1}} \circ \cdots \circ T_{\omega_0}(x), \quad n \geq 1.$$

① Here we only consider place-independent iterated function system, that is, p_i is indenpendent of x.

Then Φ is an RDS called iterated function system on $\mathcal{X}$, denoted by **IFS**$(\{T_i, p_i\}_{i\in S})$, and the associated skew product is $\Theta_n(\omega, x) = (\sigma^n\omega, \Phi(n, \omega, x))$.

An IFS is a Markov RDS as shown in Remark 2.2. Therefore, the analysis of the distributional properties of an IFS reduces naturally to the framework of Markov chain theory.

Remark 2.2 (IFS as a Markov chain): Since the components of ω are i.i.d. at each step, if we marginalize ω, the resulting process $\{X_k\}_{k\geq 0}$ is naturally a Markov chain with the following transition

$$X_{k+1} = T_{i_k}(X_k),$$

where i_k are i.i.d. distributed over S with probability p.

The associated Markov operator $\mathcal{P}$ on probability distributions is defined by

$$\mu_{k+1} = \mathcal{P}\mu_k, \tag{2-1}$$

where

$$\mathcal{P}\mu(A) = \sum_{i=1}^{r} \int_{T_i^{-1}(A)} p_i \mu(dx) = \sum_{i=1}^{r} p_i \mu(T_i^{-1}(A)) \quad \text{for } A \in \mathcal{B}. \tag{2-2}$$

Let μ^* be a stationary distribution with respect to the operator $\mathcal{P}$, i.e.,

$$\mathcal{P}\mu^* = \mu^*.$$

Then there exists a Markov invariant measure ν^* (See definition 1.5.5 [22]) for Φ on $\Omega\times\mathcal{X}$ s.t.

$$\mu^*(A) = \nu^*(\Omega \times A), \qquad \forall A \in \mathcal{B}.$$

The one-to-one correspondence theorem for μ^* and ν^* is stated in Theorem 1.5.6 [22].

Here we introduce two types of IFS, which will be used in the following chapters.

An IFS is average-contractive [18] on a metric space $(\mathcal{X}, d)$ if

$$\sum_{i=1}^{r} p_i \log \frac{d(T_i(x), T_i(x'))}{d(x, x')} < 0, \tag{2-3}$$

for all $x, x' \in \mathcal{X}$.

An IFS is monotone [16] on a Banach space $\mathcal{X}$ if it satisfies the following property:

$$\forall x, y \in \mathcal{X}, \text{ if } x \leq y \text{ then } T_i(x) \leq T_i(y), \tag{2-4}$$

where the partial order on $\mathcal{X}$ is defined by

$$\forall x, y \in \mathcal{X}, \quad x \leq y \quad \Longleftrightarrow \quad y - x \in \mathcal{X}_+.$$

Here $\mathcal{X}^+ \subset \mathcal{X}$ is a closed convex cone which satisfies $\mathcal{X}^+ \cap -\mathcal{X}^+ = \{0\}$. In other words, each iteration map T_i preserves the partial order defined by the cone $\mathcal{X}_+$.

2.2 Dimension theory of iterated function system

IFSs are commonly used to construct fractals. In this section, we introduce the following concepts in fractal geometry and geometric measure theory, which will be used in Chapter 5.

Definition 2.6 (Hausdorff dimension): Let X be a subset of a metric space $(\mathcal{X}, d)$ and μ be a finite Borel probability measure on $\mathcal{X}$. For any $s \geq 0$ and $\delta > 0$, the s-dimensional Hausdorff outer measure is

$$\mathcal{H}^s_\delta(X) := \inf\left\{\sum_{i=1}^{\infty}(\operatorname{diam} U_i)^s : X \subset \bigcup_{i=1}^{\infty} U_i,\ \operatorname{diam}(U_i) < \delta\right\}.$$

Then, the s-dimensional Hausdorff measure is

$$\mathcal{H}^s(X) := \lim_{\delta \to 0} \mathcal{H}^s_\delta(X).$$

The Hausdorff dimension of X is

$$\dim_{\mathrm{H}}(X) := \inf\{s \geq 0 : \mathcal{H}^s(X) = 0\} = \sup\{s \geq 0 : \mathcal{H}^s(X) = \infty\}.$$

The Hausdorff dimension of μ is

$$\dim_{\mathrm{H}} \mu := \inf\{\dim_{\mathrm{H}} Z \mid Z \subset \mathcal{X}, \mu(Z) = 1\}.$$

Definition 2.7 (Box dimension): Let X be a bounded subset of a metric space $(\mathcal{X}, d)$ and μ be a finite Borel probability measure on $\mathcal{X}$. For each $\delta > 0$, let $N^d_\delta(X)$ be the covering number of X, i.e., the cardinality of a minimal set of points N such that $X \subseteq \bigcup_{x \in N} \bar{B}_\delta(x)$. The upper box dimension of X is

$$\overline{\dim}_{\mathrm{Box}}(X) := \limsup_{\delta \to 0}\left(\frac{\log N^d_\delta(X)}{\log(\frac{1}{\delta})}\right)$$

and the upper box dimension of μ is

$$\overline{\dim}_{\mathrm{Box}}(\mu) := \lim_{\epsilon \to 0} \inf\{\dim_{\mathrm{Box}}(Z) : Z \subset \mathcal{X}, \mu(Z) \geq 1 - \epsilon\}.$$

Similarly, the lower box dimension $\underline{\dim}_{\mathrm{Box}}$ is defined in the same form, with sup replaced by inf.

We introduce two other fractal dimension concepts related to the upper box dimension: the minimal spanning tree dimension and the persistent homology dimension. The

latter provides a numerical method for computing fractal dimension using tools from topological data analysis and we will use it in Section 5.3. For the basic concepts of persistent modules and persistent homology in topological data analysis (TDA), please refer to [14, 41].

Definition 2.8 (Minimum spanning tree dimension): Let X be a bounded subset of a metric space $(\mathcal{X}, d)$ and μ be a finite Borel probability measure on $\mathcal{X}$. Let $\mathbf{x} \subset X$ be a finite set and $\mathcal{T}(\mathbf{x})$ be the corresponding minimum spanning tree. The α-weighted lifetime sum of $\mathbf{x}$ is

$$E_\alpha(\mathbf{x}) := \sum_{e \in \mathcal{T}(\mathbf{x})} |e|^\alpha,$$

with $\alpha \geq 0$. The minimal spanning tree dimension is

$$\dim_{\mathrm{MST}}(X) := \inf\{\alpha : \exists C \text{ so that } E_\alpha(\mathbf{x}) < C \ \forall \text{ finite } \mathbf{x} \subset X\}.$$

Definition 2.9 (Persistent homology dimension): Let X be a bounded subset of a metric space $(\mathcal{X}, d)$ and μ be a finite Borel probability measure on $\mathcal{X}$. Let $\mathbf{x} \subset X$ be a finite set and $\mathrm{PH}_i(\mathcal{VR}(\mathbf{x}))$ be the i-dimensional persistence module of the Vietoris–Rips complex on $\mathbf{x}$ and $|I(\gamma)|$ is the persistence of some persistent generator γ. The weighted i^{th} homology lifetime sum is

$$E^i_\alpha(\mathbf{x}) = \sum_{\gamma \in \mathrm{PH}_i(\mathcal{VR}(\mathbf{x}))} |I(\gamma)|^\alpha.$$

The PH_i-dimension of X is the smallest exponent α for which E^i_α is uniformly bounded for all finite subsets of X, i.e.

$$\dim^i_{\mathrm{PH}}(X) := \inf\left\{\alpha \ : \ E^i_\alpha(\mathbf{x}) < C \text{ for some constant } C > 0, \text{ for all finite } \mathbf{x} \subset X\right\}.$$

Let $x_1, \ldots, x_n$ be n random samples drawn from $\mathcal{X}$ with distribution μ. The PH_i-dimension of μ is

$$\dim^i_{\mathrm{PH}}(\mu) = \frac{1}{1-\beta},$$

where

$$\beta = \limsup_{n \to \infty} \frac{\log(\mathbb{E}(E^i_1(x_1, \ldots, x_n)))}{\log(n)}.$$

Remark 2.3: Note that there is a bijection between the edges of the minimal spanning tree $\mathcal{T}(\mathbf{x})$ and the intervals in the canonical decomposition of $\mathrm{PH}_0(\mathcal{VR}(\mathbf{x}))$. Hence we have

$$\dim^0_{\mathrm{PH}}(X) = \dim_{\mathrm{MST}}(X)$$

for all X bounded subset of $\mathcal{X}$.

The above concepts of fractal dimension are equivalent under certain constraints on the sets and measures, as shown in Thm 2.2 and 2.3.

Theorem 2.2 ([33]): Let X be a bounded subset of a metric space $(\mathcal{X}, d)$. Then

$$\dim_{\mathrm{PH}}^{0}(X) = \dim_{\mathrm{MST}}(X) = \overline{\dim}_{\mathrm{Box}}(X)$$

for all X bounded subset of $\mathcal{X}$.

Theorem 2.3 ([23, 52]): Let μ be s-Ahlfors regular measure on a metric space $(\mathcal{X}, d)$ for some $s > 1$, i.e., there exist constants $c_1, c_2 > 0$ and $r_0 > 0$ such that

$$c_1 r^s \le \mu(B(x,r)) \le c_2 r^s, \quad \forall x \in \mathrm{supp}(\mu), \quad \forall r \in (0, r_0).$$

Then

$$\dim_{\mathrm{H}}(\mu) = \overline{\dim}_{\mathrm{Box}}(\mu) = \underline{\dim}_{\mathrm{Box}}(\mu) = \dim_{\mathrm{PH}}^{0}(\mu) = s.$$

We now consider the fractal dimension of the stationary distribution for IFS. Currently, the results on the exact representation of fractal dimension are mostly limited to conformal fractals. Here we state an upper bound for the fractal dimension of the stationary distribution of an IFS with certain contractivity.

Theorem 2.4 ([18, 43, 46]): Assume that the IFS defined in Definition 2.5 satisfies

$$\lambda_1 := \lim_{n\to\infty} \frac{1}{n} \log\Big(\sup_{x,x' \in \mathcal{X}} \Big\{ \frac{d(\mathcal{T}_n(x), \mathcal{T}_n(x'))}{d(x,x')} \Big\} \Big) < 0. \tag{2-5}$$

where $\mathcal{T}_n = T_{\omega_0} \cdots T_{\omega_{n-1}}$. Then there exists an unique stationary distribution μ^* and

$$\dim_{\mathrm{H}}(\mu^*) \le \frac{\sum_{i=1}^{r} p_i \log p_i}{\lambda_1}.$$

Moreover, if the IFS is average-contractive (2-3), which is stronger than (2-5), following from Kingman's subadditive ergodic theorem, then we have

$$\dim_{\mathrm{H}}(\mu^*) \le \frac{\sum_{i=1}^{r} p_i \log p_i}{\lambda_2},$$

where $\lambda_2 = \sum_{i=1}^{r} p_i \int \log \|DT_i(x)\| \, d\mu^*(x)$.

CHAPTER 3 PROBLEM SETTING

In this chapter, we introduce key concepts in deep learning, including optimization error and generalization error, and provide a mathematical formulation of the constant step-size SGD algorithm. The notation introduced in this chapter will continue to be used in Chapter 4 and 5.

3.1 Optimization and generalization error in deep learning

Let the data probability space be $(\mathcal{Z}, \mathcal{F}, \mu_z)$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $x \in \mathcal{X}$ are features and $y \in \mathcal{Y}$ are labels. Our goal is to solve a population risk minimization problem

$$\min_{w \in \mathbb{R}^d} \left\{ \mathcal{R}(w) := \mathop{\mathbb{E}}_{z \sim \mu_z} [\ell(w, z)] := \mathop{\mathbb{E}}_{(x,y) \sim \mu_z} [\mathcal{L}(h_w(x), y)] \right\},$$

where ℓ is the composition of the loss function $\mathcal{L} : Y \times Y \to \mathbb{R}$ and $\ell(w, z) = \ell(w, (x, y)) = \mathcal{L}(h_w(x), y)$. Let the training dataset be $S_{n_{\text{data}}} := (z_i)_{i=1}^{n_{\text{data}}} \sim \mu_z^{\otimes n_{\text{data}}}$, which consists of n_{data} i.i.d. data points.

In practice, since μ_z is unknown, we minimize the empirical risk

$$\widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w) := \frac{1}{n_{\text{data}}} \sum_{i=1}^{n_{\text{data}}} \ell(w, z_i)$$

through a stochastic optimization algorithm $\mathcal{A}$.

Let w^* be a global minimizer of $\widehat{\mathcal{R}}_{S_{n_{\text{data}}}}$, i.e.,

$$w^* \in \arg\min_{w \in \mathbb{R}^d} \left\{ \widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w) \right\}.$$

We then have the following population risk decomposition bound with respect to the weight w

$$\mathcal{R}(w) \leq \widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w^*) + \underbrace{(\widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w) - \widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w^*))}_{\text{optimization error}} + \underbrace{|\mathcal{R}(w) - \widehat{\mathcal{R}}_{S_{n_{\text{data}}}}(w)|}_{\text{generalization error}}.$$

Here, the optimization error arises from the limitations of the optimization algorithm, whereas the generalization error stems from the finite samples of data. Since we do not analyze approximation error induced by the model architecture in this paper, we separate the empirical optimization suboptimality and the generalization error, which is different from the classical excess risk decomposition. We will analyze the optimization error and

generalization error in Chapter 4 and Chapter 5, respectively.

3.2 Constant step-size SGD as an iterated function system

In this paper, we focus on the constant step-size mini-batch SGD algorithm [48] without replacement within a batch. Let η be the step size and b be the batch size. Then there are $\binom{n_{\text{data}}}{b}$ possible ways to select a mini-batch from n_{data} training samples. For simplicity, we rewrite the loss function in the following form:

$$F(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{b}\sum_{j\in B_i}\ell(w, z_j)\right),$$

where $n = \binom{n_{\text{data}}}{b}$, $B_i \subseteq \{1, \ldots n_{\text{data}}\}$ satisfying $|B| = b$ and $f_i(w) = \frac{1}{b}\sum_{j\in B_i}\ell(w, z_j)$. Define the maps $g_i : \mathbb{R}^d \to \mathbb{R}^d$ as

$$g_i(w) := w - \eta\nabla f_i(w) \quad (1 \le i \le n). \tag{3-1}$$

We model the SGD as an iterated function system **IFS**$(\{g_i, p_i\}_{i\in\{1,2,\ldots n\}})$ with

$$\Phi(k, \omega, w) = g_{\omega_{k-1}} \circ \cdots \circ g_{\omega_0}(w) \tag{3-2}$$

and uniform selection probability $p_i = \frac{1}{n}$ as defined in Definition 2.5.

We also formulate it as a Markov chain given in Remark 2.2, which will be employed in the asymptotic distributional analysis in Chapter 4. Let W_k be the random variable of weight after k iteration. $\{W_k\}_{k=0}^{\infty}$ forms a time-homogeneous Markov chain with the initialization distribution W_0 and the following transition:

$$W_k = g_{i_k}(W_{k-1}), \tag{3-3}$$

where $i_k \in \{1, 2, 3, \ldots, n\}$ is drawn i.i.d. from the uniform distribution $I \sim \text{Uniform}\{1, \ldots, n\}$. The associated Markov operator is denoted by $\mathcal{P}$.

We decompose the gradient into a deterministic term and gradient noise $\xi_I(w) = \nabla F(w) - \nabla f_I(w)$ induced by mini-batch sampling and obtain

$$g_I(w) := w - \eta\nabla F(w) + \eta\xi_I(w).$$

It's easy to see that the expectation of gradient noise $\mathbb{E}\|\xi_I(w)\| = 0$. Here we clarify that if the variance of gradient noise of a single sample $\mathbb{E}\|\nabla\ell(w, z_i) - \nabla F(w)\|^2 \le \sigma^2$, then the variance of gradient noise of a batch

$$\mathbb{E}\|\xi_I(w)\|^2 \le \frac{\sigma^2}{b}\cdot\frac{n_{\text{data}} - b}{n_{\text{data}} - 1}$$

as proved in [47]. If $b \ll n_{\text{data}}$, $\frac{n_{\text{data}}-b}{n_{\text{data}}-1} \to 1$ and we have $\mathbb{E}\|\xi_I(w)\|^2 \leq \frac{\sigma^2}{b}$. This implies that increasing the batch size reduces the variance of the gradient noise. Hence, under our formulation, although the batch size b does not appear explicitly in the optimization error bounds derived in Chapter 4, its effect is implicitly reflected therein. We will also mention this point in the analysis of optimization error in Chapter 4.

CHAPTER 4 OPTIMIZATION DYNAMICS AND CONVERGENCE OF THE ASSOCIATED MARKOV CHAIN

In this chapter, we investigate the optimization dynamics of constant step-size SGD from the perspective of IFS and give an asymptotic analysis of the optimization error.

In Section 4.1, we focus on the convergence of the Markov chain induced by SGD under strongly convex loss functions. Specifically, we prove that the IFS is average-contractive and the corresponding Markov chain converges exponentially to a unique stationary distribution, and give the asymptotic upper bound of the expectation of the optimization error.

For non-convex loss functions, the situation becomes significantly more complex, and it is difficult to obtain results of broad generality. Hence, in Section 4.2, inspired by [53], we consider a simplified setting in which the loss function is separable, show the iterated maps are monotone and analyze the dynamics of the monotone IFS induced under this assumption.

It has been established in [53] that globally, the state space admits a decomposition into a transient set and a union of mutually disjoint absorbing sets, from which trajectories cannot escape once entered. At the distributional level, each attracting set admits a unique ergodic stationary distribution, and the global convergence of the Markov chain depends on the initial distribution, converging to a stationary distribution given by a convex combination of the ergodic staitonary distribution. We will review these results in Sections 4.2.1 and 4.2.2, accompanied by Example 4.1 for better understanding. Meanwhile, by contrasting with the SDE framework, we point out the failure of the latter in predicting the long-term distributional behavior of SGD.

In Section 4.2.3, we extend their analysis to the pathwise level and prove the existence of a random singleton attractor and synchronization in each absorbing set, which implies locally, trajectories starting from different initial points will eventually synchronize to the same trajectory at an exponential rate when the sequence of iterated mapsis fixed. This indicates that the absorbing set into which a trajectory falls is jointly determined by the random initialization and the iteration maps during the early phase of training, whereas once the trajectory enters an absorbing set, all remaining randomness originates solely from the realization of the iteration maps. As a result, within each absorbing set,

two trajectories with different initializations but the same random seed asymptotically attain the same training performance.

We further establish that the Lyapunov exponent along the synchronized trajectory is negative, and based on this, derive an asymptotic bound on the local optimization error in terms of the Lyapunov exponent in Section 4.2.4. This dynamical system framework offers a new perspective for deriving optimization bounds with negative Lyapunov exponents, which is fundamentally different from the classical framework for non-convex optimization analysis. We give a comparative discussion of the optimization error bounds obtained under the two frameworks in Section 4.2.4.

Throughout this chapter, we only consider α-smooth loss functions. We make the following assumptions:

Assumption 4.1 (α-smooth): Each f_i is α_i-smooth for some $\alpha_i > 0$, i.e.,

$$\|\nabla f_i(w) - \nabla f_i(w)\| \leq \alpha_i \|w - w'\|$$

for $w, w' \in \mathbb{R}^d$ and $1 \leq i \leq n$.

4.1 SGD with strongly convex loss

In this section, we focus on the SGD with strongly convex loss which admits a unique stationary distribution. We make the following additional assumption:

Assumption 4.2 (β-strongly convex): Each f_i is β_i-strongly convex for some $\beta_i > 0$, i.e.,

$$f_i(\lambda w + (1-\lambda)w') \leq \lambda f(w) + (1-\lambda) f(w') - \frac{\beta_i}{2}\lambda(1-\lambda)\|w - w'\|^2.$$

for $\lambda \in [0,1], w, w' \in \mathbb{R}^d$, and $1 \leq i \leq n$.

This assumption is usually satisfied in regression problems with regularization, such as ridge regression and logistic regression with an L_2 regularization. The convexity of f_i typically implies the corresponding IFS has a certain form of contractivity, as stated in lemma 4.1. Note that the "contractivity" established here ($\gamma = \frac{1}{n}\sum_{i=1}^n \gamma_i < 1$) is stronger than the standard definition of average-contractivity (2-3) since $\mathbb{E}(\log a) \leq \log(\mathbb{E}a)$ for any $a > 0$. The latter does not require convexity of f_i and allows some of the iterated mapsto be expansive. A more general statement for the existence and uniqueness of stationary distribution follows from the theory on average-contractive IFS. See Theorem 1 [18], Theorem 2.1 [6] and Theorem 2.1 [1] for more details.

Lemma 4.1 (Average-contractive): Let $\alpha = \frac{\sum_{i=1}^n \alpha_i \beta_i}{\sum_{i=1}^n \beta_i}$. Assume that Assumption 4.1

and 4.2 hold. Then for any $\eta \in (0, \frac{2}{\alpha})$, the corresponding IFS (3-1)-(3-3) is average-contractive as defined in (2-3).

Proof: By Assumption 4.1 and 4.2, for each i, we have

$$(\nabla f_i(w) - \nabla f_i(w'))^T(w - w') \geq \frac{1}{\alpha_i}\|\nabla f_i(w) - \nabla f_i(w')\|^2, \tag{4-1}$$

$$(\nabla f_i(w) - \nabla f_i(w'))^T(w - w') \geq \beta_i\|w - w'\|^2. \tag{4-2}$$

Hence, for any $w, w' \in \mathbb{R}^d$,

$$\begin{aligned}\|g_i(w) - g_i(w')\|^2 &= \|w - w'\|^2 - 2\eta(\nabla f_i(w) - \nabla f_i(w'))^T(w - w') + \eta^2\|\nabla f_i(w) - \nabla f_i(w')\|^2 \\ &\leq \|w - w'\|^2 - 2\eta\left(1 - \frac{\alpha_i\eta}{2}\right)\beta_i\|w - w'\|^2 - 2\eta\frac{\alpha_i\eta}{2}\frac{1}{\alpha_i}\|\nabla f_i(w) - \nabla f_i(w')\|^2 \\ &\quad + \eta^2\|\nabla f_i(w) - \nabla f_i(w')\|^2 = (1 - 2\beta_i\eta + \alpha_i\beta_i\eta^2)\|w - w'\|^2.\end{aligned}$$

Let $\gamma_i = 1 - 2\beta_i\eta + \alpha_i\beta_i\eta^2 > 0$ since $\beta_i < \alpha_i$ and $\gamma = \frac{1}{n}\sum_{i=1}^n \gamma_i$. Then

$$\frac{1}{n}\sum_{i=1}^n \log\frac{\|g_i(w) - g_i(w')\|}{\|w - w'\|} \leq \frac{1}{2n}\sum_{i=1}^n \log\gamma_i \leq \frac{1}{2}\log\gamma < 0.$$

∎

By Assumption 4.1 and 4.2, there exists a unique global minimizer w^* with $\nabla F(w^*) = 0$. The following theorem shows (3-3) converges exponentially to a unique stationary distribution and provide a upper bound for the optimization error of SGD algorithm by the covariance of stochastic gradient noise. The convergence in Theorem 4.1 under the 1-Wasserstein distance relies on the strong average-contraction induced by our convexity assumption as shown in Lemma 4.1. A similar argument can be found in [28] for the contractive case and in [20] for the quadratic loss functions.

Theorem 4.1: Assume Assumption 4.1 and 4.2 hold and $\mathbb{E}[\|\nabla f_I(\omega^*)\|^2] < \infty$. Let $\gamma' = \frac{1}{n}\sum_{i=1}^n \sqrt{\gamma_i}$ and $\gamma^* = \max_i \sqrt{\gamma_i}$ where γ_i is defined in Lemma 4.1.

(i) There exists a unique stationary distribution μ^* for the induced Markov chain (3-3). Moreover, the Markov chain is exponentially contractive in the $1-$Wasserstein distance with stationary ditribution μ^*, i.e.,

$$W_1(\mathcal{P}^k\mu, \mu^*) \leq (\gamma')^k W_1(\mu, \mu^*), \qquad \forall k \geq 0, \qquad \forall \mu \in \mathcal{P}_1(\mathbb{R}^d).$$

Here the $1-$Wasserstein distance W_1 are defined in Definition A.1.

(ii) Let Σ be the covariance matrix of $\xi_I(w^*)$ and $\sigma = \sqrt{\mathrm{tr}(\Sigma)}$. Then

$$W_1(\mu^*, \delta_{w^*}) \leq \frac{\eta\sigma}{1 - \gamma'}.$$

(iii) Let $\alpha' = \sup_{i\in\{1,\dots n\}} \alpha_i$. Then the expectation of optimization error

$$\mathbb{E}_{\mu^*}\big[F(W) - F(w^*)\big] \le \frac{\alpha'\eta^2\sigma^2}{2(1-\gamma')(1-\gamma^*)}.$$

Proof: (i) Fix $\mu, \mu' \in \mathcal{P}_1(\mathbb{R}^d)$ and $\pi \in C(\mu, \mu')$. Let $W \sim \mu$ and $W \sim \mu'$ be two random variables with the joint distribution $(W, W') \sim \pi$. Define $W_+ \sim \mathcal{P}\mu, W'_+ \sim \mathcal{P}\mu'$ and with the joint distribution $\mathcal{P}\pi$ s.t. $\mathcal{P}\pi((W_+, W'_+) \in A) = \frac{1}{n}\sum_{i=1}^n \pi((g_i(W), g_i(W')) \in A)$ for all measurable set A.

By lemma 4.1,

$$W_1(\mathcal{P}\mu, \mathcal{P}\mu') \le \mathbb{E}_{\mathcal{P}\pi}\big[\|W_+ - W'_+\|\big] = \mathbb{E}_\pi\left[\frac{1}{n}\sum_{i=1}^n \|g_i(W) - g_i(W')\|\right]$$
$$\le \gamma'\mathbb{E}_\pi[\|W - W'\|].$$

Taking the infimum over all π, we have

$$W_1(\mathcal{P}\mu, \mathcal{P}\mu') \le \gamma' W_1(\mu, \mu'). \tag{4-3}$$

It's easy to see $\gamma' \le \sqrt{\gamma} < 1$. Hence, the Markov operator $\mathcal{P}$ is a γ'-contraction on $(\mathcal{P}_1(\mathbb{R}^d), W_1)$. Note that $(\mathcal{P}_1(\mathbb{R}^d), W_1)$ is a complete metric space as shown in proposition 7.1.5 [2]. By Banach fixed point theorem, there exists a unique stationary measure μ^* and moreover for any $\mu \in \mathcal{P}_1(\mathbb{R}^d)$,

$$W_1(\mathcal{P}^k\mu, \mu^*) \le (\gamma')^k W_1(\mu, \mu^*).$$

(ii) We use the notations in the proof of (i). For each $i \in \{1, \dots n\}$, by the triangle inequality,

$$\|g_i(W) - w^*\| \le \|g_i(W) - g_i(w^*)\| + \|g_i(w^*) - w^*\| \le \sqrt{\gamma_i}\|W - w^*\| + \|g_i(w^*) - w^*\|.$$

Then we have

$$\mathbb{E}_{\mathcal{P}\mu}\|W_+ - w^*\| \le \gamma'\mathbb{E}_\mu\|W - w^*\| + \eta\mathbb{E}[\|\nabla f_I(\omega^*)\|].$$

Since δ_{w^*} is a Dirac measure, the joint distribution of μ and δ_{w^*} is exactly $\mu \cdot \delta_{w^*}$. Then

$$W_1(\mathcal{P}\mu, \delta_{w^*}) = \mathbb{E}_{\mathcal{P}\mu}\|W_+ - w^*\| \le \gamma' W_1(\mu, \delta_{w^*}) + \eta\mathbb{E}[\|\nabla f_I(\omega^*)\|].$$

By iterations, we have

$$W_1(\mathcal{P}^k\mu, \delta_{w^*}) \le (\gamma')^k W_1(\mu, \delta_{w^*}) + \frac{1-(\gamma')^k}{1-\gamma'}\eta\mathbb{E}[\|\nabla f_I(\omega^*)\|].$$

Note that

$$\mathbb{E}[f_I(w^*)] = \nabla F(w^*) = 0.$$

By Jensen's inequality,

$$\mathbb{E}[\|\nabla f_I(w^*)\|] \le \sqrt{\mathbb{E}\|\nabla f_I(w^*)\|^2} = \sigma.$$

Letting $k \to \infty$ and using the conclusion in (i), we obtain the inequality in (ii).

(iii) Note that

$$\|g_i(W) - w^*\|^2 \le (1 + \frac{1-\sqrt{\gamma_i}}{\sqrt{\gamma_i}})\|g_i(W) - g_i(w^*)\|^2 + (1 + \frac{\sqrt{\gamma_i}}{1-\sqrt{\gamma_i}})\|g_i(w^*) - w^*\|^2$$
$$\le \sqrt{\gamma_i}\|W - w^*\|^2 + \frac{1}{1-\sqrt{\gamma_i}}\eta\|\nabla f_i(w^*)\|^2.$$

Hence,

$$\mathbb{E}_{\mathcal{P}\mu}\|W_+ - w^*\|^2 \le \gamma'\mathbb{E}_\mu\|W - w^*\|^2 + \eta\frac{1}{1-\gamma^*}\mathbb{E}[\|\nabla f_I(w^*)\|^2].$$

By iterations, we have

$$\mathbb{E}_{\mathcal{P}^k\mu}\|W - w^*\|^2 \le (\gamma')^k\mathbb{E}_\mu\|W - w^*\|^2 + \eta\frac{1}{1-\gamma^*}\frac{1-(\gamma')^k}{1-\gamma'}\mathbb{E}[\|\nabla f_I(w^*)\|]^2.$$

Letting $k \to \infty$, and combining the $\alpha-$ smoothness of F, we have

$$\mathbb{E}_{\mu^*}\big[F(W) - F(w^*)\big] \le \mathbb{E}_{\mu^*}\big[\langle\nabla F(w^*), W - w^*\rangle + \frac{\alpha'}{2}\|W - w^*\|^2\big] \le \frac{\alpha'\eta^2\sigma^2}{2(1-\gamma')(1-\gamma^*)}.$$

∎

Remark 4.1 (W is an unbiased estimator of w^* for quadratic loss under μ^*): An observation is that for quadratic loss (not necessarily strongly convex), since we have

$$\mathbb{E}_{\mu^*}[\nabla F(W)] = 0,$$

then

$$\nabla F(\mathbb{E}_{\mu^*}[W]) = \mathbb{E}_{\mu^*}[\nabla F(W)] = 0,$$

which implies

$$\mathbb{E}_{\mu^*}[W] = w^*.$$

Remark 4.2 (Upper bound of pathwise time-average optimization error): Let $\{w_k\}_{k=0}^\infty$ be a weight trajectories with initialization w_0 and a sequence of iteration maps ω, where $w_k = \Phi(k, \omega, w_0)$. Define $\nu^* := \mathbb{P} \otimes \mu^*$. Using the Birkhoff ergodic theorem and Theorem 4.1 (i) for the metric dynamical system $(\Omega \times \mathbb{R}^d, \nu^*, \Theta)$, we have for $\mathbb{P}$-a.e. ω and μ^*-a.e. w_0 and any $H \in L^1(\nu^*)$,

$$\frac{1}{N}\sum_{k=0}^{N-1} H\big(\Theta^k(\omega, w_0)\big) \xrightarrow[N\to\infty]{\nu^*\text{-a.s.}} \int_{\Omega\times\mathbb{R}^d} H\, d\nu^*.$$

Let $H(\omega, w) = F(w) - F(w^*)$. By Fubini's theorem, we have

$$\frac{1}{N}\sum_{k=0}^{N-1}\big(F(w_k))-F(w^*)\big) \xrightarrow[N\to\infty]{\nu^*\text{-a.s.}} \int_{\mathbb{R}^d}\big(F(w)-F(w^*)\big)\mu^*(dw) = \mathbb{E}_{W\sim\mu^*}\big[F(W)-F(w^*)\big].$$

Hence, combining with Theorem 4.1 (iii) yields a pathwise bound on the long-run time-average suboptimality along a typical SGD trajectory :

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\big(F(w_k)-F(w^*)\big) \le \frac{\alpha'\eta^2\sigma^2}{2(1-\gamma')(1-\gamma^*)}.$$

Theorem 4.1 illustrates that in the strongly convex setting, iterates do not converge to the global minimizer but continue to fluctuate around it and the limiting behavior of SGD is independent of initializations. It also shows that the expectation of optimization error $\le O(\eta\sigma^2)$, which indicates that larger variance of gradient-noise at w^* leads to a larger upper bound of the optimization error, while stronger contraction (from strong convexity or small η) makes the error smaller. Note that a larger batch size corresponds to smaller gradient noise as discussed at the end of Section 3.2. Therefore, in the strongly convex setting, a larger batch size combined with a smaller learning rate leads to a smaller optimization error. We also note that the convergence rate improves (i.e., the factor γ' becomes smaller) as η increases from 0. This gives rise to a tradeoff: a smaller η gives a tighter upper bound on the optimization error, but may lead to slower convergence.

4.2 SGD with non-convex and separable loss

In this section, we study constant step-size SGD under separable non-convex losses, inspired by a recent insightful work [53]. We adopt their setup and continue the analysis under the following assumptions.

Assumption 4.3 (Separable): Each f_i is separable, i.e.,

$$f_i(w) = \sum_{j=1}^{d} f_i^{(j)}(w_j) \quad \text{where} \quad w = (w_1, w_2, \dots, w_d) \in \mathbb{R}^d.$$

This assumption arises naturally in a range of practically relevant settings where features are independent and parameters are decoupled across dimensions, such as Naive Bayes classification and linear regression on data with orthogonalized features.

Assumption 4.4 (Coerciveness): Each $f_i^{(j)}$ has a finite For each $i \in \{1, ..n\}$ and $j \in \{1, ...d\}$,

$$f_i^{(j)}(w) \to \infty \text{ as } \|w\| \to \infty.$$

Assumption 4.5: Let

$$A_i^{(j)} = \left\{ w_j \in \mathbb{R} : (f_i^{(j)})'(w_j) = 0 \right\}.$$

Then $|A_i^{(j)}| < \infty$ and $\bigcap_{i=1}^n A_i^{(j)} = \varnothing$.

Assume $0 < \eta < \frac{1}{\sup_i \alpha_i}$, then $g_i^{(j)}$ is injective and monotone increasing since for all $w_j < w_j'$, we have

$$f_i^{(j)}(w_j') - f_i^{(j)}(w_j) < \eta^{-1}(w_j' - w_j),$$

which implies

$$g_i^{(j)}(w_j) < g_i^{(j)}(w_j')$$

by Assumption 4.1. We focus our analysis on the monotone IFS (defined in (2-4)) considered here.

4.2.1 Construction of absorbing sets

We first construct the absorbing set and give a decomposition of the state space in Proposition 4.1. The basic idea is to begin with the one-dimensional case, determine the possible direction of each iterate by examining the critical points of the loss function and its sign and identify a one-dimensional absorbing set. Assumption 4.3 then allows us to extend this construction to higher dimensions, producing a well-structured absorbing set in the form of a Cartesian product of intervals.

Fix $j \in \{1, \dots, d\}$. Let

$$E = \prod_{j=1}^{d} [c_{\min}^{(j)}, c_{\max}^{(j)}], \tag{4-4}$$

where $c_{\min}^{(j)}$ and $c_{\max}^{(j)}$ are the minimum and maximum of the set of critical point $\mathscr{C}^{(j)} = \bigcup_{i=1}^n \left\{ w_j \in \mathbb{R} : (f_i^{(j)})'(w_j) = 0, f_i^{(j)} \neq 0 \right\}$. The absorbing sets of E are constructed as follows. For each $j \in \{1, \dots d\}$, define

$$L^{(j)} := \bigcup_{i=1}^{n} \left\{ w_j \in \mathbb{R} : (f_i^{(j)})'(w_j) > 0 \right\},$$

$$R^{(j)} := \bigcup_{i=1}^{n} \left\{ w_j \in \mathbb{R} : (f_i^{(j)})'(w_j) < 0 \right\}.$$

By Assumption 4.4, we have

$$(-\infty, c_{\min}^{(j)}) \subset R^{(j)} \setminus L^{(j)}, \quad (-\infty, c_{\max}^{(j)}) \subset R^{(j)} \setminus L^{(j)}$$

and by Assumption 4.5,

$$L^{(j)} \cup R^{(j)} = \mathbb{R}, \quad \partial L^{(j)} \cap \partial R^{(j)} = \varnothing.$$

Let

$$U^{(j)} := \bigcup_{m=1}^{M_j} U_m^{(j)} := \bigcup_{m=1}^{M_j} [l_m^{(j)}, r_m^{(j)}],$$

where each

$$[l_m^{(j)}, r_m^{(j)}] \in L^{(j)} \cap R^{(j)}$$

with

$$l_m^{(j)} \in \partial L^{(j)}, r_m^{(j)} \in \partial R^{(j)} \text{ and } l_m^{(j)} < r_m^{(j)}.$$

The construction of the one-dimensional interval $[l, r]$ is illustrated in Fig. 4-1. Intuitively, at each iteration, any point w_j in $R^{(j)} \setminus L^{(j)}$ can only move to the right or remain unchanged since $(f_i^{(j)})'(w_j) \leq 0$. Analogously, any point in $L^{(j)} \setminus R^{(j)}$ can only move to the left or remain unchanged. Points in the overlap $L^{(j)} \cap R^{(j)}$ have positive probability to move either to the right or to the left. Hence, as indicated by the arrows in Fig. 4-1, the interval $[l, r]$ attracts nearby points into it and can served as a one-dimensional absorbing set.

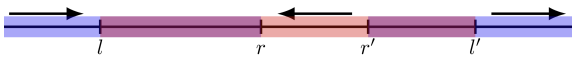

Figure 4-1 Schematic diagram of the construction of a one-dimensional absorbing set. The blue segment indicates the interval containing $R^{(j)}$, and the red segment indicates the interval containing $L^{(j)}$. The purple segment represents their intersection, and we only take the interval $[l, r]$ as the absorbing set.

Let $M = \prod_{j=1}^{d} \{1, .., M_j\}$. For each $\mathbf{m} = (m_1, ...m_d) \in M$, define the $d-$dimensional compact set

$$U_{\mathbf{m}} = \prod_{j=1}^{d} U_{m_j}^{(j)}. \tag{4-5}$$

The rigorous proof that $U_{\mathbf{m}}$ is an absorbing set is given in Proposition 4.1.

Proposition 4.1 (Absorbing sets): Assume Assumption 4.1 and Assumption 4.3-4.5 hold. E and $U_{\mathbf{m}}$ are defined in (4-4) and (4-5). Let $U = \bigcup_{m \in M} U_{\mathbf{m}}$ and $T = E \setminus U$. Then each $U_{\mathbf{m}}$ is an absorbing set. Specifically, there exists an integer $\ell_0 \geq 1$ such that for every probability distribution μ on E,

$$(\mathcal{P}^{\ell_0}\mu)(T) \leq \left(1 - \frac{1}{n^{\ell_0}}\right)\mu(T). \tag{4-6}$$

For $\mathbb{P}$-a.e. ω and all $w_0 \in E$,

$$\tau_U(\omega, w_0) := \inf\{k \geq 0 : \Phi(k, \omega, w_0) \in U\} < \infty, \tag{4-7}$$

and

$$g_i(U_{\mathbf{m}}) \subset U_{\mathbf{m}} \tag{4-8}$$

for each $m \in M$ and $i \in \{1, ..n\}$.

Proof: See Theorem 2.2 (a) [53] for (4-6) and (4-8).

For (4-7), taking $\mu = \delta_{w_0}$,

$$\mathbb{P}\{\omega : \Phi(k\ell_0, \omega, w_0) \in T\} = (\mathcal{P}^k \delta_{w_0})(T) \leq \left(1 - \frac{1}{n^{\ell_0}}\right)^k \delta_{w_0}(T) \xrightarrow[k\to\infty]{} 0.$$

It follows that

$$\mathbb{P}\{\omega : \tau_U(\omega, w_0) = \infty\} = \lim_{k\to\infty} \mathbb{P}\{\omega : \tau_U(\omega, w_0) > k\ell_0\} \leq \lim_{k\to\infty} \mathbb{P}\{\omega : \Phi(k\ell_0, \omega, w_0) \in T\} = 0.$$

Hence, $\tau_U(\omega, w_0) < \infty$ for $\mathbb{P}$-a.e. ω. ∎

Remark 4.3: Here the absorbing sets $U_{\mathbf{m}}$ are deterministic. Specifically, U is a global absorbing set for $\mathbb{P}$-a.e. ω w.r.t the power set of E and each $U_{\mathbf{m}}$ is an absorbing set for $\mathbb{P}$-a.e. ω w.r.t. the power set of the subset $V_{\mathbf{m}} = \{w \in E : e_{\mathbf{m}}(w) = 1\} \subset E$ as shown in Theorem 4.2.

Proposition 4.1 shows that the state space decomposes into a transient set T and a disjoint union of absorbing sets $U = \bigcup_{m\in M} U_{\mathbf{m}}$ and for $\mathbb{P}$-a.e. ω, the trajectory starting from any $w_0 \in E$ enters one absorbing components $U_{\mathbf{m}}$ in finite time and thereafter stays in it.

4.2.2 Stationary distributions and exponential convergence rates of the associated Markov chain

We now restate the main results of [53] in Theorem 4.2, which establishes the convergence of the associated Markov chain at the distributional level. It shows that locally, any distrbution whose support lies in an absorbing set $U_{\mathbf{m}}$ converges exponentially fast to the unique stationary distribution $\mu^*_{\mathbf{m}}$ on $U_{\mathbf{m}}$ as given in (i). Globally, given an initial distribution on E, it converges exponentially to the corresponding stationary distribution , which is the convex combination of the $\mu^*_{\mathbf{m}}$ as given in the ergodic decomposition in (ii), with coefficients given by the probabilities $p_{\mathbf{m}}(\mu)$ of entering each absorbing set $U_{\mathbf{m}}$.

Theorem 4.2 (Convergence of the Markov chain and ergodic decomposition): Under the same assumptions as in Proposition 4.1, we have

(i) for each $\mathbf{m} \in M$, there exists a unique stationary distribution $\mu_{\mathbf{m}}^*$ such that for any $\mu \in \mathcal{P}(U_{\mathbf{m}})$,

$$d_{\alpha_{\mathbf{m}}}\left(\mathcal{P}^k\mu, \mu_{\mathbf{m}}^*\right) \le \left(1 - \frac{1}{n^{\ell_{\mathbf{m}}}}\right)^{\lfloor k/\ell_{\mathbf{m}} \rfloor} \quad k > 0 \tag{4-9}$$

for some $\ell_{\mathbf{m}} \in \mathbb{N}$, $\ell_{\mathbf{m}} \ge d$ and $\alpha_{\mathbf{m}} \in \{+1, -1\}^d$.

Here the metrics $d_{\alpha_{\mathbf{m}}}$ are defined in Definition A.3 in Appendix A.

(ii) given a probability distribution $\mu \in \mathcal{P}(E)$, there exists a stationary distribution

$$\mu^* = \sum_{\mathbf{m} \in M} p_{\mathbf{m}}(\mu)\mu_m^*$$

such that

$$\tilde{d}(\mathcal{P}^k\mu, \mu^*) \le 3\left(1 - \frac{1}{n^{\ell}}\right)^{[k/\ell]} \quad k > 0 \tag{4-10}$$

with $\ell := 2(\ell_0 \vee \max_{\mathbf{m} \in M} \ell_{\mathbf{m}}) \ge d$ (ℓ_0 is defined in Proposition 4.1) and

$$\tilde{d}(\mu_1, \mu_2) := d_{\mathrm{TV}}(\mu_1|_T, \mu_2|_T) + \sum_{m \in \mathcal{M}} d_{\alpha_{\mathbf{m}}}(\mu_1|_{U_{\mathbf{m}}}, \mu_2|_{U_{\mathbf{m}}}), \quad \mu_1, \mu_2 \in \mathcal{P}(E),$$

defined in Definition A.2 and A.3 in Appendix A.

Here $p_{\mathbf{m}}(\mu) := \lim_{k\to\infty}(\mathcal{P}^k\mu)(U_{\mathbf{m}}) = \int_E e_{\mathbf{m}}(w)\, d\mu(w)$, where $\{e_{\mathbf{m}}\}_{\mathbf{m} \in M}$ are the continuous left eigenfunctions of $\mathcal{P}$, i.e.,

$$\int e_{\mathbf{m}}(w)\, d\mathcal{P}\mu' = \int e_{\mathbf{m}}(w)\, d\mu' \text{ for any } \mu' \in \mathcal{P}(E),$$

satisfying

$$e_{\mathbf{m}}(w) \ge 0, \sum_{\mathbf{m} \in M} e_{\mathbf{m}}(w) = 1$$

and

$$\int_E e_{\mathbf{m}}(w)\, d\mu_{\mathbf{m}'}^*(w) = \delta_{\mathbf{m}\mathbf{m}'}.$$

Proof: See Theorem 2.2 (b) [53] for (i) and Theorem 2.2 (c) [53] for (ii). ∎

Remark 4.4: The proof of Theorem 4.2 (i) is mainly derived from Thm 5.2 [8]. It is shown that the monotone IFS satisfies the splitting condition on each $U_{\mathbf{m}}$, from which the uniqueness of the staitonary distribution follows. Note that the stationary distribution μ_m^* is generally not a product measure and Theorem 4.2 (i) can not be deduced by simply reducing to the one-dimensional case since the same iterate map is applied across all coordinates and the coordinate dynamics are coupled.

Remark 4.5: Theorem 4.2 can be interpreted as a spectral analysis of the Markov operator $\mathcal{P}$. The invariant measures $\mu_{\mathbf{m}}^*$ are right eigenvectors of $\mathcal{P}$ corresponding to eigen-

value 1, while the functions $e_{\mathbf{m}}$ are left eigenvectors. The coefficients $c_{\mathbf{m}}(\mu_0) = \int e_{\mathbf{m}}\, d\mu_0$ represent the spectral projection of the initial measure μ_0 onto the eigenspace of eigenvalue 1, using the left eigenvectors $e_{\mathbf{m}}$. The spectral gap controls the exponential rate of convergence in (4-9) and (4-10).

Theorem 4.2 demonstrates that globally, the long-time limiting distribution depends on the initialization, as different initial conditions may lead to convergence in different absorbing sets with probabilities determined by the functions $e_{\mathbf{m}}$ satisfying $e_{\mathbf{m}}(w) = \frac{1}{n}\sum_{i=1}^{n} e_{\mathbf{m}}(g_i(w))$. Intuitively, $e_{\mathbf{m}}(w)$ is determined by the direction and magnitude of the gradients at every point along the trajectories from w to the boundaries of the absorbing sets. If along most of these trajectories, a majority of the maps push toward $U_{\mathbf{m}}$ with stronger force, then $e_{\mathbf{m}}(w)$ is large, meaning that the trajectories reaching $U_{\mathbf{m}}$ from w are both numerous and short.

We give a concrete example below to better illustrate Theorem 4.2. The corresponding one-dimensional case can be found in Section 3.2 of [53]. Notably, the results may differ from the distributional properties of SGD predicted under the SDE framework, which reflects the failure of the latter in capturing the long-term behavior of SGD. We show the failure of diffusion approximation in Remark 4.6.

Example 4.1: Given $d \geq 1$ and $\lambda \leq \frac{2\sqrt{3}}{9}$. Let $n = 2d$. Let

$$F(w) = \sum_{j=1}^{d} \frac{1}{4}(1 - w_j^2)^2.$$

The graph of $F(w)$ in the two-dimensional case is shown in the first panel of Fig. 4-2, from which we can see that the loss landscape exhibits four basins. Let

$$f_{2j-1}(w) = F(w) + \lambda_j w_j, \quad f_{2j}(w) = F(w) - \lambda_j w_j, \quad j = 1, \dots, d.$$

For each j, there exists three critical points $c_3^{(j)} < 0 < c_1^{(j)} \leq c_2^{(j)}$ for $f_{2j-1}^{(j)}$ and $-c_2^{(j)} \leq -c_1^{(j)} < 0 < -c_3^{(j)}$ are the critical points of $f_{2j}^{(j)}$ as shown in the second and third panels of Figure 4-2, where the equality holds if and only if $\lambda = \frac{2\sqrt{3}}{9}$. Let $U_1^{(j)} = [c_3^{(j)}, -c_2^{(j)}]$ and $U_2^{(j)} = [c_2^{(j)}, -c_3^{(j)}]$. By the construction of absorbing sets in Section 4.2.1, there exist 2^d absorbing sets $U_{\mathbf{m}} = \prod_{j=1}^{d} U_{m_j}^{(j)}$ for any $\mathbf{m} \in [1,2]^n$. Then by theorem 4.2, the associated Markov chain has 2^d ergodic invariant distributions $\mu_{\mathbf{m}}^*$ supported on $U_{\mathbf{m}}$ for each $\mathbf{m} \in [1,2]^n$ and any initial distribution μ converges to the stationary distribution $\mu^* = \sum_{\mathbf{m}\in[1,2]^n} p_{\mathbf{m}}(\mu)\mu_m^*$.

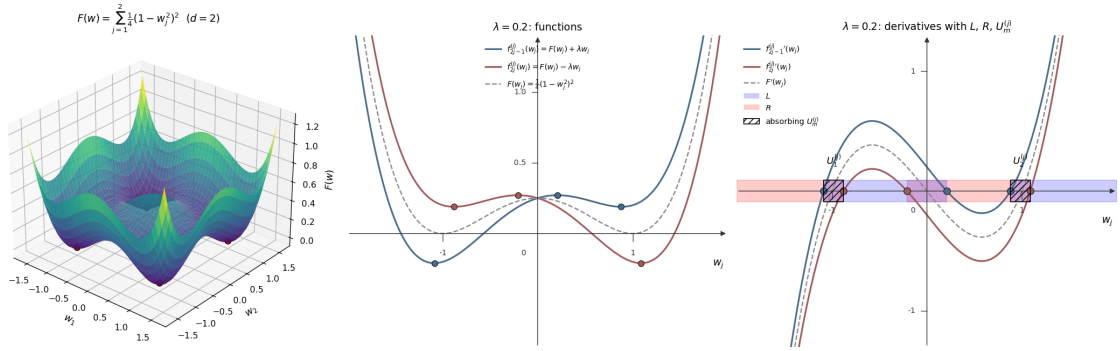

Figure 4-2 Visualization of the loss landscape in $d = 2$ and a schematic illustration of the absorbing sets $U_1^{(j)}$ and $U_2^{(j)}$ with $\lambda = 0.2$.

Remark 4.6 (Failure of the diffusion approximation): Recall that the diffusion approximation of SGD given in Theorem 9 [36] satisfies the following Itô SDE

$$dW_t = -\nabla\left(F(W_t) + \frac{1}{4}\eta|\nabla F(W_t)|^2\right)dt + \sqrt{\eta}\,\Sigma(W_t)^{1/2}\,dB_t, \quad W_0 = w_0,$$

where B_t is a standard d-dimensional Brownian motion, and

$$\Sigma(w) = \mathbb{E}\left[(\nabla f_I(w) - \nabla F(w))(\nabla f_I(w) - \nabla F(w))^{\mathsf{T}}\right].$$

Then the distribution $\mu_t(w)$ satisfies the corresponding Fokker-Planck equation

$$\frac{\partial \mu_t(w)}{\partial t} = \nabla\cdot\left(\nabla\left(F(W_t) + \frac{1}{4}\eta|\nabla F(W_t)|^2\right)\mu_t(w)\right) + \frac{\eta}{2}\sum_{j_1,j_2=1}^{d}\frac{\partial^2}{\partial w_{j_1}\partial w_{j_2}}\left[\Sigma_{j_1 j_2}(w)\,\mu_t(w)\right],$$

with initial condition $\mu_t(w) = \delta(w - w_0)$.

In Example 4.1, since

$$\Sigma(w) = \frac{1}{2d}\sum_{j=1}^{d} 2\lambda_j^2 e_j e_j^T = \frac{1}{d}\,\mathrm{diag}(\lambda_1^2, \lambda_2^2, \ldots, \lambda_d^2)$$

is positive definite, the SDE is elliptic and it admits a unique stationary distribution

$$\mu^*(w) \propto \exp\left(-\frac{2d}{\eta}\sum_j \frac{F^{(j)}(w_j) + \frac{\eta}{4}(F'^{(j)}(w_j))^2}{\lambda_j^2}\right)$$

with full support on $\mathbb{R}^d$ (See Chapter 4 [44]).

Remark 4.7: For the case $n < d$, the diffusion matrix $\Sigma(w)$ becomes degenerate, and the support of the invariant measure in the SDE framework becomes more complicated and the validity of the SDE framework requires further verification.

We can see the discrepancies of the distributional properties predicted by the two frameworks through Example 4.1 and Remark 4.6. The diffusion approximation predicts a unique, everywhere-positive stationary distribution and the ergodic theorem guarantees

that every region of positive stationary measure is visited infinitely often and no local minimum is permanently inaccessible. However, Theorem 4.2 shows that SGD trajectories are confined to certain trapping regions of the loss landscape once entering, from which escape is impossible and transitions between two local minima occur only if both lie within the same absorbing set.

The reason the diffusion approximation fails to capture the long-term distributional behavior of SGD in this case is that it replaces the discrete one-step transition of SGD with a Gaussian transition supported on all of $\mathbb{R}^d$. This is an unavoidable consequence of requiring the approximating Markov process to have continuous sample paths. Consequently, the absorbing-set structure from the Doeblin decomposition, which depends on the discrete support of the one-step transition, is entirely erased.

4.2.3 Random attractor and stability analysis on absorbing sets

We further explore the pathwise properties of the IFS on each absorbing set $U_{\mathbf{m}}$ in Proposition 4.2 and 4.3. Our main idea is to show that the corresponding IFS (3-2) satisfies the more general splitting condition introduced in [19], from which strong synchronization follows. We therefore prove, via the following lemma, that (3-2) indeed satisfies the strict splitting condition.

Lemma 4.2 (Strict splitting condition on $U_{\mathbf{m}}$): Let $\Phi(n,\omega,w) = g_{\omega_{n-1}}\circ\cdots\circ g_{\omega_0}(w)$. Under the same assumptions as in Proposition 4.1, for each $U_{\mathbf{m}}$, there exist two sequences $(p_1,\dots,p_\ell)$ and $(q_1,\dots,q_r)$ with $p_\ell = q_r$, such that $R_1 := g_{p_\ell}\circ\cdots\circ g_{p_1}(U_{\mathbf{m}})$ and $R_2 := g_{q_r}\circ\cdots\circ g_{q_1}(U_{\mathbf{m}})$ satisfy

$$\pi_j(\Phi(n,\omega,R_1)) \cap \pi_j(\Phi(n,\omega,R_2)) = \varnothing$$

for every $\omega \in \Omega$, every $n \geq 0$, and every projection $\pi_j : \mathbb{R}^d \to \mathbb{R}$, $j = 1,\dots,d$. That is, Φ splits on $U_{\mathbf{m}}$ as defined in Definition 2.1 [19].

Proof: We proceed with the discussion by building on the results of Theorem 2.2(b) [53]. With translation and scaling transformation, assume $U_{\mathbf{m}} = [0,1]^d$. Denote $g_{\vec{p}} = g_{p_\ell}\circ\cdots\circ g_{p_1}$ if $\vec{p} = (p_1,\dots,p_l)$. By the proof of Theorem 2.2(b) [53], there exist $\vec{p} = (p_1,\dots,p_m)$, $\vec{q} = (q_1,\dots,q_m)$ and $\tilde{w} \in [0,1]$ s.t.

$$g_{\vec{p}}([0,1]^d) \preceq_\alpha \tilde{w} \quad \text{and} \quad \tilde{w} \preceq_\alpha g_{\vec{q}}([0,1]^d).$$

Note that g_i is injective, then $0 < \tilde{w} < 1$. WLOG, for each dimension j, we have

$$g_{\vec{p}}^{(j)}([0,1]) \subseteq [0,\tilde{w}_j] \quad \text{and} \quad g_{\vec{q}}^{(j)}([0,1]) \subseteq [\tilde{w}_j,1].$$

Since $g_{\vec{p}}^{(j)}$ is strictly increasing and $\tilde{w}_j < 1$, we have $g_{\vec{p}}^{(j)}(\tilde{w}_j) < g_{\vec{p}}^{(j)}(1) \leq \tilde{w}_j$. Similarly, we have $g_{\vec{q}}^{(j)}(\tilde{w}_j) > g_{\vec{q}}^{(j)}(0) \geq \tilde{w}_j$. Hence,

$$g_{\vec{p}\circ\vec{p}}^{(j)}([0,1]) \subseteq g_{\vec{p}}([0,\tilde{w}_j]) \subset [0,\tilde{w}_j] \quad \text{and} \quad g_{\vec{q}\circ\vec{q}}^{(j)}([0,1]) \subseteq g_{\vec{q}}[\tilde{w}_j,1] \subset [\tilde{w}_j,1].$$

Therefore,

$$g_{\vec{p}\circ\vec{p}}^{(j)}([0,1]) \cap g_{\vec{q}\circ\vec{q}}^{(j)}([0,1]) = \varnothing.$$

It is clear that every g_i preserves this strict separation, which leads to the conclusion that Φ splits. ∎

The following proposition shows the existence of a random singleton attractor and provides a characterization of pathwise synchronization. To define pullback attractors, We give a natural extension of base dynamical system to an invertible measure preserving dynamical system $(\Omega = \{1,...n\}^{\mathbb{Z}}, \mathcal{F}', \mathbb{P} = p^{\otimes\mathbb{Z}}, (\sigma^k)_{k\in\mathbb{Z}})$ with the left shift $(\sigma\omega)_k = \omega_{k+1}$ and right shift $(\sigma^{-1}\omega)_k = \omega_{k-1}$ for all $k \in \mathbb{Z}$. It is worth noting that this extension does not alter the forward dynamics. Indeed, for each $k \geq 1$ the forward k-step cocycle

$$\Phi(k,\omega,w) = g_{\omega_{k-1}} \circ \cdots \circ g_{\omega_0}(w)$$

depends only on the coordinates $(\omega_0, \dots, \omega_{n-1})$ and the additional "past" coordinates $(\omega_{-1}, \omega_{-2}, \dots)$ have no influence on the distribution or realization of forward orbits. Such a construction is implicit in [19], though no explicit two-sided extension is carried out there. For the equivalence, see Remark 4.8.

Proposition 4.2 (Existence of random singleton attractor and synchronization): Under the same assumptions as in Proposition 4.1, in each $U_{\mathbf{m}}$, there exists a random attractor $A_{\mathbf{m}}$ such that for $\mathbb{P}$-a.e. ω, $A_{\mathbf{m}}(\omega) = \{a_{\mathbf{m}}(\omega)\}$ and

$$\operatorname{dist}\Big(\Phi\big(k,\omega,U_{\mathbf{m}}\big), A_{\mathbf{m}}(\sigma^k\omega)\Big) \leq c(\omega)q^k, \qquad \forall k \geq 1, \tag{4-11}$$

for some $0 < c(\omega) < \infty$ and $q \in (0,1)$.

Proof: For each $k \geq 0$, define

$$C_k(\omega) := \Phi(k, \sigma^{-k}\omega, U_{\mathbf{m}}).$$

Since $U_{\mathbf{m}}$ is compact and for each i, g_i is continuous and injective on $U_{\mathbf{m}}$, $C_k(\omega)$ is compact and nonempty for every $n \geq 0$ and $\omega \in \Omega$. By the forward invariance of $U_{\mathbf{m}}$, we have

$$C_{k+1}(\omega) = \Phi(k, \sigma^{-k}\omega, \Phi(1, \sigma^{-(k+1)}\omega, U_{\mathbf{m}})) \subseteq \Phi(k, \sigma^{-k}\omega, U_{\mathbf{m}}) = C_k(\omega).$$

Hence, $(C_k(\omega))_{k\geq 0}$ is nonincreasing, i.e., $C_{k+1}(\omega) \subseteq C_k(\omega)$. Define

$$A_{\mathbf{m}}(\omega) := \bigcap_{k\geq 0} C_k(\omega).$$

By Lemma 4.2 and Theorem 4.1 [19], $A_{\mathbf{m}}(\omega)$ is a singleton for $\mathbb{P}$-a.e. ω. Let

$$A_{\mathbf{m}}(\omega) = \{a_{\mathbf{m}}(\omega)\} \qquad \text{for } \mathbb{P}\text{-a.e. } \omega. \tag{4-12}$$

Note that for each $k \geq 0$, by the cocycle property we have

$$\Phi(1, \omega, C_k(\omega)) = \Phi\big(1, \omega, \Phi(k, \sigma^{-k}\omega, U_{\mathbf{m}})\big) = \Phi(k+1, \sigma^{-k}\omega, U_{\mathbf{m}}) = C_{k+1}(\sigma\omega).$$

Therefore,

$$\Phi(1, \omega, a_{\mathbf{m}}(\omega)) = \Phi(1, \omega, (\bigcap_{k\geq 0} C_k(\omega)) = \bigcap_{k\geq 0} \Phi(1, \omega, C_k(\omega)) = \bigcap_{k\geq 0} C_{k+1}(\sigma\omega) = a_{\mathbf{m}}(\sigma\omega),$$

where the second equality holds since each g_i is injective on $U_{\mathbf{m}}$. Hence $A_{\mathbf{m}}$ is strictly φ-invariant.

By Lemma 4.2 and Theorem 3 in [19], due to the equivalence of the 1-norm and 2-norm in Euclidean space, for $\mathbb{P}$-a.e. ω there exists $0 < c(\omega) < \infty$ and $q \in (0,1)$ s.t.

$$\operatorname{diam}\big(\Phi(k, \omega, U_{\mathbf{m}})\big) \leq c(\omega) q^k, \qquad \forall k \geq 1, \tag{4-13}$$

where $\operatorname{diam}(D) := \sup_{x,y\in D} \|x - y\|$ for some subset $D \subseteq U_{\mathbf{m}}$. For any $B \subseteq U_{\mathbf{m}}$, by (4-12), we have

$$\operatorname{dist}\Big(\Phi\big(k, \sigma^{-k}\omega, B\big), A(\omega)\Big) \leq \operatorname{diam}\big(C_k(\omega)\big) \xrightarrow[k\to\infty]{} 0,$$

and by (4-13), we have

$$\begin{aligned} \operatorname{dist}\Big(\Phi\big(k, \omega, B\big), A_{\mathbf{m}}(\sigma^k\omega)\Big) &= \operatorname{dist}\Big(\Phi\big(k, \omega, B\big), \Phi\big(k, \omega, A_{\mathbf{m}}(\omega)\big)\Big) \\ &\leq \operatorname{diam}\big(\Phi\big(k, \omega, U_{\mathbf{m}}\big)\big) \leq c(\omega) q^k \xrightarrow[k\to\infty]{} 0. \end{aligned} \tag{4-14}$$

for $\mathbb{P}$-a.e. ω. This proves strong pullback and forward attraction and completes the proof of the existence of random singleton attractor, and (4-11) follows directly from (4-14).

∎

Remark 4.8 (Random pullback attractor and the one-sided "inverse" coding map in [19]): The proof of (4-12) originates from the coding map defined in [19] which is based on the one-sided space $\Sigma_k = S^{\mathbb{N}}$ equipped with the inverse Markov measure $\mathbb{P}^-$. Since in our case, the iteration maps are i.i.d. chosen, we have $\mathbb{P}^- = \mathbb{P}$. Define the fibre

$$I_\xi = \bigcap_{k\geq 0} g_{\xi_0} \circ \cdots \circ g_{\xi_k}(U_{\mathbf{m}})$$

for $\xi \in \Sigma_k$ as shown in Section 4 [19]. On our two-sided extension, taking $\xi(\omega) =$

$(\omega_{-1}, \omega_{-2}, \dots)$ turns this one-sided fibre into the pullback intersection

$$\bigcap_{k\geq 0} \Phi(k, \sigma^{-k}\omega, U_{\mathbf{m}}).$$

Hence, the singleton fibre given by Theorem 4.1 in [19] is precisely the singleton random pullback attractor in our setting. It is worth noting that $A_{\mathbf{m}}(\omega) = \{a_{\mathbf{m}}(\omega)\}$ is measurable with respect to the past σ-algebra generated by $(\omega_{-1}, \omega_{-2}, \dots)$ and independent of $(\omega_0, \omega_1, \dots)$.

Remark 4.9: The existence of a random singleton attractor proved in Proposition 4.2 implies the uniqueness of the stationary distribution given in Theorem 4.1 (i). The coding map $a_{\mathbf{m}}$ establishes a correspondence between the symbolic space Ω and $U_{\mathbf{m}}$: The unique stationary distribution $\mu^*_{\mathbf{m}}$ for the Markov chain (3-3) on $U_{\mathbf{m}}$ is exactly the pushforward of the base measure under the random attractor $a(\omega)$, i.e.,

$$\mu^*_{\mathbf{m}} = (a_{\mathbf{m}})_{\#}\mathbb{P}.$$

Remark 4.10: If $a_{\mathbf{m}}$ is a trivial map, i.e., $a_{\mathbf{m}}(\omega) = w^*$ for any $\omega \in \Omega$, then w^* is a global minimizer satisfying $\nabla f_i(w^*) = 0$ for all i.

We further demonstrate the linearized stability of the random fixed point by showing that the Lyapunov exponent is negative in Proposition 4.3. To relate the differential of the iterated mapsto the pathwise exponential contraction of trajectories, we introduce the following additional assumption. Note that a sufficient condition for Assumption 4.6 is that each f_i is three-times continuously differentiable and its third derivative has bounded operator norm. We also give a more general statement without Assumption 4.6 in Remark 4.11.

Assumption 4.6 (Lipschitz Hessian on $U_{\mathbf{m}}$): Each $f_i \in C^2(E)$ and the Hessian of each f_i is Lipschitz on $U_{\mathbf{m}}$, i.e., there exists a constant $H_{\mathbf{m}} > 0$ such that for all $w, w' \in U_{\mathbf{m}}$ and all $i = 1, \dots, n$,

$$\|\nabla^2 f_i(w) - \nabla^2 f_i(w')\| \leq H_{\mathbf{m}} \|w - w'\|,$$

where $\|\cdot\|$ denotes the operator norm.

Proposition 4.3 (Negative Lyapunov exponents): Based on the assumptions in Proposition 4.1, further assume that Assumption 4.6 holds. Then the maximal Lyapunov exponent along this random singleton attractor obtained in Proposition 4.2 satisfies

$$\lambda_{\mathbf{m}\max} = \lim_{k\to\infty} \frac{1}{k} \log \left\| D\Phi(k, \omega, a_{\mathbf{m}}(\omega)) \right\| < 0 \qquad \text{for } \mathbb{P}\text{-a.e. } \omega.$$

Proof: WLOG, let $U_{\mathbf{m}} = [0,1]^d$. We define the maximal Lyapunov exponent along the

random singleton attractor by

$$\lambda_{\mathbf{m}\max}(\omega) := \lim_{k\to\infty}\frac{1}{k}\log\left\|D\Phi(k,\omega,a_{\mathbf{m}}(\omega))\right\|$$

Note that $\mathbb{P}(d\omega)\otimes\delta_{a_{\mathbf{m}}(\omega)}(dx)$ is ergodic on $\Omega\times U_{\mathbf{m}}$. Since $a_{\mathbf{m}}(\omega)\in U_{\mathbf{m}}$ which is compact, we have $\int_\Omega \log\left\|Dg_{\omega_0}(a_{\mathbf{m}}(\omega))\right\| d\mathbb{P}(\omega) < \infty$. By Theorem 2.1, $\lambda_{\mathbf{m}\max}(\omega)$ is a.s. constant and we denote it by $\lambda_{\mathbf{m}\max}$.

Note that by Assumption 4.3,

$$D\Phi(k,\omega,a_{\mathbf{m}}(\omega)) = \operatorname{diag}\Big(\Phi'^{(1)}(k,\omega,\pi_1(a_{\mathbf{m}}(\omega))),\dots,\Phi'^{(d)}(k,\omega,\pi_d(a_{\mathbf{m}}(\omega)))\Big),$$

and

$$\left\|D\Phi(k,\omega,a_{\mathbf{m}}(\omega))\right\| = \max_{1\le j\le d}\Phi'^{(j)}(k,\omega,\pi_j(a_{\mathbf{m}}(\omega))),$$

where

$$\Phi'^{(j)}(k,\omega,\pi_j(w)) = (g^{(j)}_{\omega_{k-1}}\circ\cdots\circ g^{(j)}_{\omega_0})'(\pi_j(w)).$$

Hence, we only need to prove that for each $0\le j\le d$,

$$\lim_{k\to\infty}\frac{1}{k}\log\left(\Phi'^{(j)}(k,\omega,\pi_j(a_{\mathbf{m}}(\omega)))\right) < 0 \qquad \text{for } \mathbb{P}\text{-a.e. } \omega.$$

Note that on $[0,1]$,

$$\operatorname{diam}\big(\Phi^{(j)}(k,\omega,[0,1])\big) \ge |\Phi'^{(j)}(k,\omega,w_j^k)|\times 1,$$

where

$$w_j^k = \inf_{w_j\in[0,1]}|\Phi'^{(j)}(k,\omega,[0,1])(w_j)|.$$

Let $\Omega'\subset\Omega$ and $\mathbb{P}(\Omega')=1$ satisfying (4-13). Then for all $\omega\in\Omega'$,

$$0 < \Phi'^{(j)}(k,\omega,w_j^k) \le c(\omega)q^k. \tag{4-15}$$

By Assumption 4.6, we have

$$\|Dg_i(w)-Dg_i(w')\| = \eta\|\nabla^2 f_i(w)-\nabla^2 f_i(w')\| \le \eta H_{\mathbf{m}}\|w-w'\|.$$

By Assumption 4.1 and $\eta < \frac{1}{\sup_i \alpha_i}$, $\|\nabla g_i(w)\| = \|I-\eta\nabla^2 f_i(w)\| > 0$ for all $w\in E$. We may assume that $g_i^{(j)'}(w_j) > c_j > 0$ for all i. Then for any $w,w'\in[0,1]$,

$$\left|\log g_i^{(j)'}(w)-\log g_i^{(j)'}(w')\right| \le \frac{1}{c_j}\left|g_i^{(j)'}(w)-g_i^{(j)'}(w')\right| \le \frac{H_{\mathbf{m}}}{c_j}|w-w'|.$$

Hence, for any $k \in \mathbb{N}$ and all $\omega \in \Omega'$,

$$\begin{aligned} &|\log \Phi'^{(j)}(k,\omega,\pi_j(a_{\mathbf{m}}(\omega)) - \log \Phi'^{(j)}(k,\omega,w_j^k)| \\ &\le \frac{H_{\mathbf{m}}}{c_j}\sum_{k=0}^{\infty} \operatorname{diam}\big(\Phi^{(j)}(k,\omega,[0,1])\big) \\ &\le \frac{H_{\mathbf{m}}}{c_j}\sum_{k=0}^{\infty} c(\omega)q^k = \frac{H_{\mathbf{m}}}{c_j}\cdot\frac{c(\omega)}{1-q} =: K(\omega). \end{aligned} \tag{4-16}$$

Combining (4-15) and (4-16), we have for all $\omega \in \Omega'$,

$$\begin{aligned} &\lim_{k\to\infty}\frac{1}{k}\log\big(\Phi'^{(j)}(k,\omega,\pi_j(a_{\mathbf{m}}(\omega)))\big) \\ &\le \lim_{k\to\infty}\frac{1}{k}\big(\log \Phi'^{(j)}(k,\omega,w_j^k)) + |\log \Phi'^{(j)}(k,\omega,\pi_j(a_{\mathbf{m}}(\omega)) - \log \Phi'^{(j)}(k,\omega,w_j^k)|\big) \\ &\le \lim_{k\to\infty}\frac{1}{k}(\log c(\omega) + k\log q + K(\omega)) = \log q < 0. \end{aligned}$$

∎

Remark 4.11: Without Assumption 4.6, we can still establish that the Lyapunov exponent at the random fixed point is non-positive, i.e., $\lambda_{\mathbf{m}\max} \le 0$. Suppose that on the contrary, $\lambda_{\mathbf{m}\max} > 0$. This will lead to a contradiction by the unstable manifold theorem (Theorem 6.1 [49]). To avoid ambiguity in notation, we use $\mathcal{M} := U_{\mathbf{m}}$ and $p(\omega) := a_{\mathbf{m}}(\omega)$. Choose $r(\omega) > 0$ such that the exponential map

$$\exp_{p(\omega)} : B_{T_{p(\omega)}\mathcal{M}}(0, r(\omega)) \to \mathcal{M}$$

is a diffeomorphism onto its image, and likewise for $\exp_{p(\sigma^n\omega)}$. Define the conjugated cocycle in exponential coordinates by

$$F_\omega^n(v) := \exp_{p(\sigma^n\omega)}^{-1}\Big(\Phi\big(n,\omega,\exp_{p(\omega)}(v)\big)\Big), \qquad v \in B_{T_{p(\omega)}\mathcal{M}}(0, r(\omega)).$$

Then $F_\omega^n(0) = 0$ and the derivative cocycle satisfies

$$D_0F_\omega^n = D_{p(\omega)}\Phi(n,\omega,\cdot)\big|_{p(\omega)} : T_{p(\omega)}\mathcal{M} \to T_{p(\sigma^n\omega)}\mathcal{M}.$$

In particular, the maximal Lyapunov exponent along the invariant graph $\{(\omega, p(\omega))\}$ is

$$\lambda_{\mathbf{m}\max} = \lim_{k\to\infty}\frac{1}{k}\log\Big\|D_0F_\omega^n\Big\| = \lim_{k\to\infty}\frac{1}{k}\log\Big\|D_{p(\omega)}\Phi(n,\omega,p(\omega))\Big\| > 0.$$

By Theorem 6.1 [49], for $\mathbb{P}$-a.e. ω there exists a nontrivial local unstable manifold $W_{\mathrm{loc}}^u(\omega) \ni 0$. We can choose a measurable point $v(\omega) \in W_{\mathrm{loc}}^u(\omega) \setminus \{0\}$ which admits a backward orbit converging exponentially to 0. Hence, translating back, the unstable set is nontrivial, which contradicts with the strong synchronization on $U_{\mathbf{m}}$ by Proposition 2 [51]. Therefore, $\lambda_{\mathbf{m}\max} \le 0$.

Corollary 4.1: For every $\varepsilon \in (0, -\lambda_{\mathbf{m}\max})$, there exists a measurable random variable $c_\varepsilon(\omega) \in (0, \infty)$ such that for $\mathbb{P}$-a.e. ω and all $k \geq 0$,

$$\left\| D\Phi(k, \theta^{-k}\omega, a_{\mathbf{m}}(\theta^{-k}\omega)) \right\| \leq c_\varepsilon(\omega)\, e^{k(\lambda_{\mathbf{m}\max}+\varepsilon)}. \tag{4-17}$$

Proposition 4.2 and 4.3 characterize the asymptotic pathwise behavior and stability of the system dynamics. Proposition 4.2 shows that, for a fixed sequence of iteration maps $\omega \in \Omega$, once the dynamics has been run long enough within each absorbing set $U_{\mathbf{m}}$ to reach stationary state, trajectories starting from different weight initializations synchronize to the same point and subsequently evolve along a single trajectory $\{a_{\mathbf{m}}(\sigma^k\omega)\}_{k\geq 0}$. In particular, the dependence of training trajectories on the initialization decays exponentially over time and the randomness in the asymptotic behavior lies solely in the selection of the iterated maps. This characterize the effect of initialization and randomness arising from the selection of the iterated maps to optimization behavior of constant step-size SGD: In the early stage of training, the random initialization and the iteration maps jointly determine which absorbing set the trajectory falls into. Once inside an absorbing set, the randomness stems entirely from the choice of iteration maps. Consequently, within each absorbing set, two trajectories with different initializations but the same random seed will eventually achieve the same training performance. Moreover, the Lyapunov exponent $\lambda_{\mathbf{m}\max} < 0$ given in Proposition 4.3 further indicates that the random fixed point is stable and infinitesimal perturbations are not exponentially amplified. This will be used to derive the optimization error bound in Section 4.2.4.

It is worth noting that the existence of a random singleton attractor and synchronization are stronger pathwise statements than the distributional statement in Theorem 4.2 (i), and the relationship between these two viewpoints is discussed in Remark 4.9. Therefore, Propositions 4.2 and 4.3 can be regarded as an extension and strengthening of Theorem 4.2 (i), providing a more refined description of the optimization dynamics at the pathwise level. Compared to distributional-level results, synchronization decouples the randomness from initialization and the randomness from the selection of iterated maps, and establishes asymptotic pathwise agreement of trajectories with different initializations within the same absorbing set.

4.2.4 Asymptotic optimization error bounds for local and global minima

Based on Proposition 4.3, we establish an upper bound of local optimization error on $U_{\mathbf{m}}$ in Theorem 4.3. Our main approach is to bound the distance between the random singleton attractor and the local minimum via recursion, subsequently deriving the optimization error through Assumption 4.1. Specifically, the existence of the random attractor yields a trajectory along which we can perform recursion to obtain a series and equation (4-17) ensures the convergence of it.

Theorem 4.3 (Upper bound of local optimization error): Assume the assumptions in Proposition 4.2 hold. Then each $U_{\mathbf{m}}$ contains at least one local minimizer. Take $w_{\mathbf{m}}^*$ to be the local minimizer with the lowest function value of F in $U_{\mathbf{m}}$. Assume that Assumption 4.6 holds and $c_{0,\varepsilon} := \mathbb{E}_{\mathbb{P}}[c_\varepsilon^2(\omega)] < \infty$ for some $\varepsilon \in (0, -\lambda_{\mathbf{m}\max})$ in (4-17). Denote $R_{\mathbf{m}} = \operatorname{diam}(U_{\mathbf{m}})$, $\sigma_{\mathbf{m}} := \sup_{i\in\{1,\dots n\}} \|\xi_i(w_{\mathbf{m}}^*)\|$ and $\alpha' := \sup_{i\in\{1,\dots n\}} \alpha_i$. Then the local optimization error on each $U_{\mathbf{m}}$

$$\mathbb{E}_{\mu_{\mathbf{m}}^*}\big[F(W) - F(w_{\mathbf{m}}^*)\big] = \mathbb{E}_{\mathbb{P}}\big[F(a_{\mathbf{m}}(\omega)) - F(w_{\mathbf{m}}^*)\big] \le \frac{\alpha'\eta^2 c_{0,\varepsilon}^2 (2\sigma_{\mathbf{m}} + H_{\mathbf{m}} R_{\mathbf{m}}^2)^2}{8(1 - e^{(\lambda_{\mathbf{m}\max}+\varepsilon)})^2}.$$

Proof: The existence of local minimizer in each $U_{\mathbf{m}}$ has been proved in Theorem 2.2(a) [53]. Now we define the error vector $b(\omega) = a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*$. Then

$$b(\theta\omega) = a_{\mathbf{m}}(\theta\omega) - w_{\mathbf{m}}^* = a_{\mathbf{m}}(\omega) - \eta\nabla f_{\omega_0}(a_{\mathbf{m}}(\omega)) - w_{\mathbf{m}}^* = b(\omega) - \eta\nabla f_{\omega_0}(a_{\mathbf{m}}(\omega)).$$

By fundamental theorem of calculus, we have

$$\nabla f_i(a_{\mathbf{m}}(\omega)) - \nabla f_i(w_{\mathbf{m}}^*) = \Big(\int_0^1 \nabla^2 f_i(w_{\mathbf{m}}^* + t(a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*))\, dt\Big)(a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*).$$

Then

$$\nabla f_{\omega_0}(a_{\mathbf{m}}(\omega)) = \nabla f_{\omega_0}(w_{\mathbf{m}}^*) + \nabla^2 f_{\omega_0}(a_{\mathbf{m}}(\omega))\, b(\omega) + r_{\omega_0}(w_{\mathbf{m}}^*, a_{\mathbf{m}}(\omega)),$$

where

$$r_i(w_{\mathbf{m}}^*, a_{\mathbf{m}}(\omega)) := \left(\int_0^1 (\nabla^2 f_i(a_{\mathbf{m}}(\omega) + t(a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*)) - \nabla^2 f_i(a_{\mathbf{m}}(\omega)))\, dt\right)(a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*). \tag{4-18}$$

Hence,

$$b(\theta\omega) = \big(I - \eta\nabla^2 f_{\omega_0}(a_{\mathbf{m}}(\omega))\big)b(\omega) - \eta\nabla f_{\omega_0}(w_{\mathbf{m}}^*) - \eta r_{\omega_0}(w_{\mathbf{m}}^*, a_{\mathbf{m}}(\omega)).$$

By (4-18) and Assumption 4.6,

$$\begin{aligned}\|r_i(w_{\mathbf{m}}^*, a_{\mathbf{m}}(\omega))\| &\le \|a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*\| \int_0^1 H_{\mathbf{m}} \|(1-t)(a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*)\|\, dt \\ &= \frac{H_{\mathbf{m}}}{2} \|a_{\mathbf{m}}(\omega) - w_{\mathbf{m}}^*\|^2 \le \frac{H_{\mathbf{m}} R^2}{2}.\end{aligned} \tag{4-19}$$

Let

$$D(\omega) = I - \eta \nabla^2 f_{\omega_0}(a_{\mathbf{m}}(\omega)), \zeta(\omega) := -\nabla f_{\omega_0}(w_{\mathbf{m}}^*) \text{ and } \rho(\omega) := -r_{\omega_0}(w_{\mathbf{m}}^*, a_{\mathbf{m}}(\omega)).$$

We have

$$b(\theta\omega) = D(\omega) b(\omega) + \eta\, \zeta(\omega) + \eta \rho(\omega).$$

For $N \ge 1$, denote

$$b_{1,N}(\omega) = \eta(\zeta(\theta^{-1}\omega) + \sum_{k=2}^{N} D(\theta^{-1}\omega) \cdots D(\theta^{-k+1}\omega)\, \zeta(\theta^{-k}\omega))$$

and

$$b_{2,N}(\omega) = \eta(\rho(\theta^{-1}\omega) + \sum_{k=2}^{N} D(\theta^{-1}\omega) \cdots D(\theta^{-k+1}\omega)\, \rho(\theta^{-k}\omega)).$$

By recursion,

$$b(\omega) = \prod_{k=1}^{N} D(\theta^{-k}\omega) b(\theta^{-N}\omega) + b_{1,N} + b_{2,N}.$$

Let $N \to \infty$, by (4-17) and (4-19),

$$\begin{aligned}\|b(\omega)\| &\le \|b_{1,\infty}(\omega)\| + \|b_{2,\infty}(\omega)\| \\ &\le \frac{\eta(2\sigma_{\mathbf{m}} + H_{\mathbf{m}} R^2)}{2} \left(\frac{c_{0,\epsilon}(\omega)}{1 - e^{(\lambda_{\mathbf{m}\max} + \varepsilon)}}\right).\end{aligned} \tag{4-20}$$

By Assumption 4.1,

$$F(a(\omega)) - F(w_{\mathbf{m}}^*) \le \langle \nabla F(w_{\mathbf{m}}^*), a(\omega) - w_{\mathbf{m}}^* \rangle + \frac{\alpha'}{2} \|a(\omega) - w_{\mathbf{m}}^*\|^2 \le \frac{\alpha' \eta^2 c_{0,\epsilon}^2(\omega)(2\sigma_{\mathbf{m}} + H_{\mathbf{m}} R^2)^2}{8(1 - e^{(\lambda_{\mathbf{m}\max} + \varepsilon)})^2}.$$

Taking expectation we obtain the conclusion. ∎

Remark 4.12 (Relation between the maximal Lyapunov exponent and the geometry of the loss landscape): Note that

$$\begin{aligned}\lambda_{\mathbf{m}\max} &\le \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k} \log \|Dg_{\omega_i}(a_{\mathbf{m}}(\sigma^i\omega)\| = \int_{\Omega} \log \|Dg_{\omega_0}(a_{\mathbf{m}}(\omega)\|\, d\mathbb{P}(\omega) \\ &= \frac{1}{n} \sum_{i=0}^{n} \int \log \|Dg_i(w)\|\, d\mu_{\mathbf{m}}^*(w) = \frac{1}{n} \sum_{i=0}^{n} \int \log \|I - \eta \nabla^2 f_i(w)\|\, d\mu_{\mathbf{m}}^*(w).\end{aligned}$$

The second equality is derived by the Birkhoff ergodic theorem on the skew product by

setting $H(\omega, w) = \log \|Dg_{\omega_0}(w)\|$ and the third equality is due to the independence between ω_0 and $a_{\mathbf{m}}(\omega)$. This implies that if we fix η, the Lyapunov exponent is negatively correlated with the norm of the Hessian matrix of the loss function on the staionary distribution $\mu^*_{\mathbf{m}}$ (Note that the case we discussed here is under the assumption that the Hessian matrix of loss is positive definite and the Lyapunov exponent is negative. That is, we still consider the optimization dynamics below the edge of stability).

Remark 4.13 (Upper bound of local optimization error with Łojasiewicz condition): Assume

$$\sigma'^2_{\mathbf{m}} := \int_{U_{\mathbf{m}}} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(w) - \nabla F(w)\|^2 d\mu^*_{\mathbf{m}} < \infty.$$

If the absorbing set $U_{\mathbf{m}}$ contains only one local minimizer $w^*_{\mathbf{m}}$ and the landscape of $U_{\mathbf{m}}$ geometrically satisfies the Łojasiewicz condition commonly used in non-convex analysis [10, 5, 17, 57], i.e.,

$$\|\nabla F(w)\| \geq c_{\mathbf{m}} |F(w) - F(w^*_{\mathbf{m}})|^{\delta}, \quad \delta \in [1/2, 1), \tag{4-21}$$

we can obtain the following upper bound on the optimization error

$$\mathbb{E}_{\mu^*_{\mathbf{m}}}\big[F(W) - F(w^*_{\mathbf{m}})\big] \leq \left(\sqrt{\frac{\alpha'\eta}{2-\alpha'\eta}}\frac{|\sigma'_{\mathbf{m}}|}{c_m}\right)^{\frac{1}{\delta}}.$$

Proof: For each $w \in U_{\mathbf{m}}$ and $i \in \{1, \dots, n\}$, by Assumption 4.1,

$$F(g_i(w)) \leq F(w) - \eta\langle \nabla F(w), \nabla f_i(w)\rangle + \frac{\alpha'\eta^2}{2}\|\nabla f_i(w)\|^2.$$

Then

$$(\mathscr{P}F)(w) = \frac{1}{n}\sum_{i=0}^{n} F(g_i(w)) \leq F(w) - \eta\|\nabla F(w)\|^2 + \frac{\alpha'\eta^2}{2}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w)\|^2.$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w)\|^2 = \|\nabla F(w)\|^2 + \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w) - \nabla F(w)\|^2.$$

We have

$$\int (\mathscr{P}F)(w)\, d\mu^*_m = \int F(w)\, d\mu^*_m \leq \int \left(F(w) - \eta(1 - \frac{\alpha'\eta}{2})\|\nabla F(w)\|^2\right) d\mu^*_m + \frac{\alpha'\eta^2}{2}\sigma'^2_{\mathbf{m}}.$$

Hence,

$$\int \|\nabla F(w)\|^2 d\mu^*_m \leq \frac{\alpha'\eta}{2-\alpha'\eta}\sigma'^2_{\mathbf{m}}.$$

Combining with (4-21), we obtain the upper bound

$$\mathbb{E}_{\mu^*_{\mathbf{m}}}\left[F(W)-F(w^*_{\mathbf{m}})\right] \leq \left(\frac{1}{c_{\mathbf{m}}}\right)^{\frac{1}{\delta}} \mathbb{E}_{\mu^*}\left[\|\nabla F(W)\|^{\frac{1}{\delta}}\right] \leq \left[\mathbb{E}_{\mu^*}\left[\|\nabla F(W)\|^2\right]\right]^{\frac{1}{2\delta}} \leq \left(\sqrt{\frac{\alpha'\eta}{2-\alpha'\eta}}\frac{|\sigma'_{\mathbf{m}}|}{c_{\mathbf{m}}}\right)^{\frac{1}{\delta}}.$$

∎

Theorem 4.3 gives an upper bound of optimization error from the perspective of stability of the random dynamical system, which shows that the local optimization error is positively correlated with the step size and the magnitude of the gradient noise, and negatively correlated with the Lyapunov exponent $|\lambda_{\mathbf{m}}|$. The Lyapunov exponent is closely related to the geometry of the loss landscape, as discussed in Remark 4.12. In this bound, the step size, gradient noise and geometric scale of the absorbing set captures how strongly the noise can push the system away from the minimizer $w^*_{\mathbf{m}}$, while the Lyapunov exponent captures how strongly the dynamics pull the system toward the random attractor, which is located in the vicinity of local minimizers of F.

We compare the optimization error bound derived under our random dynamical systems framework with the optimization bound obtained under the Łojasiewicz condition framework in Remark 4.13, which is commonly used in nonconvex analysis. Interestingly, although the two optimization bounds are derived using different techniques, they exhibit a certain consistency: In the case of a single local minimizer, a smaller δ in (4-21) usually corresponds to a steeper valley, which leads to a larger $|\lambda_{\mathbf{m}}|$, and they both lead to a tighter optimization bound. Moreover, these factors also influence the convergence rate of the optimization process, which should be reflected in the finite-time analysis of the optimization error. This shows that although the Lyapunov exponent is in general difficult to quantify, the bound is qualitatively consistent with the optimization error bound obtained in classical frameworks.

Note that an absorbing set $U_{\mathbf{m}}$ may contain more than one local minimizer, and the support of the ergodic stationary distribution $\mu^*_{\mathbf{m}}$ may simultaneously cover multiple local minimizers (see the example in Section 3.2 of [53]). This generality renders the Łojasiewicz condition inapplicable, as it is typically formulated for a neighborhood of a single minimizer. Our optimization bound remains valid in this multiple minimizer setting and the stationary distribution may not concentrate near the minimizer with the lowest function value within $U_{\mathbf{m}}$, but instead concentrate near other local minimizers with higher function values. This also explains why, unlike the Łojasiewicz-based bound in Remark 4.13, our bound does not reduce to zero as the gradient noise vanishes. However, we

should acknowledge that this generality comes at a cost: the bound may be loose when the stationary distribution concentrates around local minimizers with higher function values.

Now we consider the global optimization error in this nonconvex setting. Assuming w^* is the global minimizer, given the initial measure μ, we obtain the global optimization error by combining Theorem 4.3 and 4.2 (ii), which can be can be decomposed into two parts: the local optimization error after the trajectory enters a specific absorbing set, and the gap between the corresponding local minima and the global minima.

$$\mathbb{E}_{\mu^*}\big[F(W) - F(w^*)\big] = \sum_{\mathbf{m}\in M} p_{\mathbf{m}}(\mu)\Big(\int F\, d\mu^*_{\mathbf{m}} - F(w^*_{\mathbf{m}})\Big) + \sum_{\mathbf{m}\in M} p_{\mathbf{m}}(\mu)(F(w^*_{\mathbf{m}}) - F(w^*))$$
$$\leq \sum_{\mathbf{m}\in M} p_{\mathbf{m}}(\mu)\frac{\eta^2 c_{0,\epsilon}^2(2\sigma_{\mathbf{m}} + H_{\mathbf{m}}R_{\mathbf{m}}^2)^2}{8(1 - e^{(\lambda_{\mathbf{m}\max}+\varepsilon)})^2} + \sum_{\mathbf{m}\in M} p_{\mathbf{m}}(\mu)(F(w^*_{\mathbf{m}}) - F(w^*)),$$

where μ^* is the stationary distribution. Notably, the global minimizer may not belong to any absorbing set $U_{\mathbf{m}}$, as discussed in Section 3.3 of [53]. Hence, it is also an interesting question when one can guarantee that the global minimum lies in an absorbing set, or more specifically, falls within the support of the stationary distribution. This would require imposing certain assumptions on the gradient noise and we leave this as part of the future work.

CHAPTER 5 FRACTAL DIMENSION OF THE STATIONARY DISTRIBUTION AND GENERALIZATION

In this chapter, we investigate the generalization behavior of the SGD algorithm from the perspective of IFS.

Related work has established generalization bounds for stochastic optimization algorithms based on hypothesis space complexity or algorithm stability [7, 12]. However, traditional worst-case bounds derived from complexity measures such as VC dimension or Rademacher complexity typically fail to account for the influence of the specific algorithm, yielding pessimistic estimates in overparametrized neural networks [58]. Moreover, such bounds are difficult to connect with the implicit regularization phenomena observed in stochastic optimization, such as preference for flat minima and low-norm solutions [42, 26].

Motivated by this problem and building on the study of the asymptotic behavior of SGD in Chapter 4, we introduce algorithm-dependent generalization bounds for SGD using the fractal dimension of the stationary distribution in Section 5.1. This bound was proposed in [13]. Building on this, we further establish the connection between generalization error and algorithm hyperparameters in specific supervised learning tasks in Section 5.2. We also extend new experiments to characterize generalization behavior of SGD and other stochastic iterated algorithms in Section 5.3.

5.1 Generalization error bound via fractal dimension of the stationary distribution

Classical theory of iterated function systems (IFS) [31] establishes that a finite family of contractions $\{f_1, \dots, f_n\}$ on a complete metric space admits a unique compact attractor $K = \bigcup_i f_i(K)$, and a unique stationary distribution μ^* with $\mathrm{supp}(\mu^*) = K$. For self-affine IFS, the attractor is generically a fractal [23]. For more general settings, such as the average contractive and order-preserving IFS discussed in Chapter 4, the cases are more varied and intricate [25]. Modeling SGD as an IFS, this motivates us to think about the fractal structure of the stationary distribution of the weights and relate it to the generalization behavior of the model. The optimization trajectory of SGD lives in an extremely high-

dimensional space. However, the weight iterates that actually converge may concentrate on a low-dimensional subspace, as shown in [34]. Here, we introduce the generalization bound established in [13], which employs the fractal dimension of the stationary distribution as an intrinsic dimension of the weight space to quantify the effective complexity of the model and the algorithm.

[13] assumes that the IFS corresponding to SGD is average contractive to guarantee the existence of a unique ergodic stationary distribution. In fact, for a more general setting, it suffices to assume that the Markov chain corresponding to SGD converges to some stationary distribution, denoted by $\mu^*_{S_{n_{\text{data}}}}$. Different initializations may converge to different stationary distributions, and exhibit different generalization properties.

We present the following assumptions used in [13].

Assumption 5.1: for $\mu_z^{\otimes n_{\text{data}}}$-a.e. $S_{n_{\text{data}}}$, the local dimension

$$\lim_{r\to 0}\left|\frac{\log \mu^*_{S_{n_{\text{data}}}}(B_r(w))}{\log r}\right| \text{ exists for } \mu^*_{S_{n_{\text{data}}}}\text{-a.e. } w.$$

This is a common regularity condition to establish the connection between the Hausdorff dimension of the measure and the upper box-counting dimension of the set nearly full measure in proposition 1 [13].

Assumption 5.2: For all $n_{\text{data}} \in \mathbb{N}_+$, Let $\mathcal{F}$ and $\mathcal{G}$ be the sub-σ-algebras of $\mathcal{B}^{\otimes n_{\text{data}}}$ generated by

$$\left\{\hat{\mathcal{R}}_{S_{n_{\text{data}}}}(w) : w \in \mathbb{R}^d\right\}$$

and

$$\left\{\mathbf{1}\{w \in N_\beta(A_{S_{n_{\text{data}}},\delta})\},\ \mu_{W|S_{n_{\text{data}}}}(A_{S_{n_{\text{data}}},\delta}),\ \dim_{\text{H}}(\mu^*_{S_{n_{\text{data}}}}) : \delta,\beta \in \mathbb{Q},\ \delta,\beta>0, w \in N_\beta\right\}$$

respectively, where $A_{S_{n_{\text{data}}}}$ is defined in Proposition 1 [13] and

$$N_\beta := \left\{\left(\frac{(2j_1+1)\beta}{2\sqrt{d}},\dots,\frac{(2j_d+1)\beta}{2\sqrt{d}}\right) : j_i \in \mathbb{Z},\ i=1,\dots,d\right\},$$

and $N_\beta(A_{S_{n_{\text{data}}}}) := \{x \in N_\beta : B_\beta(x) \cap A_{S_{n_{\text{data}}}} \neq \emptyset\}$. There exists a constant $M \geq 1$ such that for any $F \in \mathcal{F}$ and $G \in \mathcal{G}$,

$$\mathbb{P}(F \cap G) \leq M\,\mathbb{P}(F)\,\mathbb{P}(G).$$

Assumption 5.3: ℓ is L-Lipschitz continuous in w for all z. Moreover, $\ell(w,z)$ is (a,b)-sub-exponential for all w, i.e.,

$$\log \mathbb{E}_{\mu_z}\left[\exp(\lambda\,(\ell(w,z)-\mathcal{R}(w)))\right] \leq \frac{\lambda^2 a^2}{b},$$

for all $|\lambda| < 1/b$.

Now we restate the theorem.

Theorem 5.1: Given the training data $S_{n_{\text{data}}}$, assume that the associated Markov chain (3-3) converges into a stationary distribution $\mu^*_{S_{n_{\text{data}}}} \subset \mathcal{P}(\mathbb{R}^d)$ and Assumption 5.1, 5.2 and 5.3 hold. Then for sufficiently large n,

$$\left|\hat{\mathcal{R}}_{S_{n_{\text{data}}}}(w) - \mathcal{R}(w)\right| \leq C\sqrt{\frac{\dim_{\text{H}}(\mu^*_{S_{n_{\text{data}}}})\log^2(n_{\text{data}}L^2)}{n_{\text{data}}} + \frac{\log(13M/\zeta)}{n_{\text{data}}}},$$

with probability at least $1 - 2\zeta$ over the joint distribution of $S_{n_{\text{data}}} \sim \mu_z^{\otimes n_{\text{data}}}, w \sim \mu^*_{S_{n_{\text{data}}}}$.

Proof: See Theorem 1 [13]. ∎

Remark 5.1: Since the stationary distribution of w depends on the data, the constant M introduced in Assumption 5.2 serves as a technical condition that allows one to decouple w from $S_{n_{\text{data}}}$ when applying Assumption 5.3. The constant M measures the correlation between the geometric properties of the stationary distribution and $\hat{\mathcal{R}}_{S_{n_{\text{data}}}}$, and can be viewed as a form of algorithmic stability [12], which also influences generalization. For example, as an extreme case, a deterministic algorithm may converge to a single point with the dimension term zero, while M may become unbounded. Some works have focused on providing a deeper understanding of the term M [21] and establishing generalization bounds free of this term via stability assumptions [56]. But in this chapter, we focus primarily on the fractal dimension term.

Remark 5.2: In certain cases, the Hausdorff dimension of sets and measures is equivalent to other notions of fractal dimension as indicated in Theorem 2.2 and 2.3, which provides a useful tool for the numerical estimation of fractal dimensions in Section 5.3.

This bound is inspired by generalization bounds based on covering numbers, where the fractal dimension of the stationary distribution appears as a measure of capacity and complexity. Intuitively, the fractal dimension of the stationary distribution quantifies the space-filling capacity of the support of the limiting distribution of the model parameters, serving as an effective measure of capacity. On the other hand, it captures the complexity and roughness of the support's boundary and interior, which mirrors the intricacy of the loss landscape geometry. A heuristic intuition is that flat minima tend to be locally smoother, corresponding to a lower fractal dimension, and a rigorous conclusion requires further investigation.

It is worth noting that compared with traditional generalization bounds based on

covering numbers, this quantity is closely related to the data and the SGD algorithm. This addresses the problem of overly loose generalization bounds arising from hypothesis class capacity measures based solely on model architecture and data. It allows us to establish a connection between generalization error and algorithmic hyperparameters (e.g., step size, batch size) through the lens of dynamical systems. We will discuss this in detail with concrete examples in Section 5.2.

5.2 Applications to supervised learning tasks

In this section, based on the generalization bound established in Theorem 5.1, we further relate the fractal dimension of the stationary distribution to the algorithm parameters and the training data in supervised learning tasks to obtain a more explicit generalization bound.

The basic idea is that we obtain the contractivity of the SGD algorithm by constraining η, which guarantees the existence and uniqueness of the stationary μ^*. Then by Theorem 2.4, we have

$$\dim_{\mathrm{H}}(\mu^*) \leq \frac{-\frac{1}{n}\log(n)}{-\sum_{i=1}^{n}\frac{1}{n}\int \log\left\|I-\frac{\eta}{b}\sum_{j\in B_i}\nabla^2\ell(w,z_j)\right\| d\mu^*(x)}, \tag{5-1}$$

where $n = \binom{n_{\text{data}}}{b}$ as formulated in Section 3.2. We then compute the Hessian matrix of the loss function to derive the upper bound of fractal dimension of the stationary distribution.

Here we take regularized multi-class logistic regression as an example in Example 5.1 to analyze the upper bound of the Hausdorff dimension of the stationary distribution. Other examples can be found in Section 4 of [13].

Example 5.1 (Multi-class Logistic regression with L_2 regularization): Consider a C-class classification problem with data points $z_i = (x_i, y_i)$, where $x_i \in \mathbb{R}^d$ are input features and $y_i \in \{1, \dots, C\}$ are class labels. The model parameter is $w := [w_1, \dots, w_C] \in \mathbb{R}^{d\times C}$. The regularized softmax loss is

$$\ell(w, z_i) = -\log\frac{e^{w_{y_i}^T x_i}}{\sum_{c=1}^{C} e^{w_c^T x_i}} + \frac{\lambda}{2}\|w\|_F^2, \quad \lambda > 0,$$

where $\|w\|_F$ is the Frobenius norm. Then for $\eta < \frac{4}{4\lambda+\max_i\|x_i\|^2}$,

$$\dim_{\mathrm{H}}(\mu^*) \leq \frac{\log\binom{n_{\text{data}}}{b}}{\log\left(\frac{1}{1-\eta\lambda}\right)}.$$

Proof: The Hessian matrix

$$\nabla^2 \ell(w, z_i) = (\mathrm{diag}(p_i) - p_i p_i^T) \otimes x_i x_i^T + \lambda I_{dC},$$

where $p_i = [p_{i,1}, \dots, p_{i,C}] \in \mathbb{R}^C$ with $p_{i,c} = \frac{e^{w_c^T x_i}}{\sum_{c'} e^{w_{c'}^T x_i}}$ and $\otimes$ denotes the Kronecker product. For any unit vector $v \in \mathbb{R}^C$,

$$v^T(\mathrm{diag}(p_i) - p_i p_i^T)v = \sum_{c=1}^{C} p_{i,c}(v_c)^2 - (\sum_{c=1}^{C} p_{i,c} v_c)^2 \geq 0.$$

Note that the last equation is exactly the variance of V with $P(V = v_c) = p_{i,c}$, then by Popoviciu's inequality on variances, we have

$$\|\mathrm{diag}(p_i) - p_i p_i^T\| \leq \frac{(\max_c v_c - \min_c v_c)^2}{4} \leq \frac{2(\max_c v_c)^2 + 2(\min_c v_c)^2}{4} \leq \frac{1}{2}.$$

Hence, $\lambda \leq \|\nabla^2 \ell(w, z_i)\| \leq \lambda + \frac{1}{2}\max_i \|x_i\|^2$. Then

$$\|I - \frac{\eta}{b}\sum_{i \in B} \nabla^2 \ell(w, z_i)\| = \max\{|1 - \eta\left(\lambda + \frac{\max_i \|x_i\|^2}{2}\right)|, |1 - \eta\lambda|\}.$$

Assume that $\eta < \frac{4}{4\lambda + \max_i \|x_i\|^2}$, we have $|1 - \eta\left(\lambda + \frac{\max_i \|x_i\|^2}{2}\right)| < |1 - \eta\lambda| < 1$, then the IFS is contractive. Hence, by (5-1),

$$\dim_{\mathrm{H}}(\mu^*) \leq \frac{\log\binom{n_{\mathrm{data}}}{b}}{-\sum_{i=1}^{n} \frac{1}{n} \int \log \|I - \frac{\eta}{b}\sum_{j \in B_i} \nabla^2 \ell(w, z_j)\| \, d\mu^*(x)} \leq \frac{\log\binom{n_{\mathrm{data}}}{b}}{\log\left(\frac{1}{1-\eta\lambda}\right)}.$$

∎

From Example 5.1, we obtain the following insights into the generalization behavior of SGD in regularized multi-class logistic regression. As the stepsize $\eta < 1/\lambda$ and the regularization parameter λ increases , the dimension bound decreases, indicating improved generalization. The decrease of mini-batch size $b \ll n_{\mathrm{data}}$ also leads to a tighter dimension bound and hence implies better generalization. These qualitative predictions are consistent with commonly observed empirical phenomena in machine learning, and hence provide a theoretical perspective for explaining these experimental findings.

5.3 Experiments

In this section, we provide experiments to validate the results in Sections 5.1. Our experiments consist of two main parts: In Section 5.3.1, we investigate the correlation between the fractal dimension of stationary distribution and the generalization performance under different hyperparameters (e.g., learning rate, batch size) and different initializa-

tions. Specifically, we approximate the stationary distribution using the converged weight trajectory, compute its persistent homology (PH) dimension using tools from topological data analysis (TDA), and visualize its relationship with generalization gap. The experiments show that the fractal dimension of the stationary distribution can indeed serve as a characterization of generalization on simple datasets, while on more complex datasets this metric needs to be further refined. Moreover, in Section 5.3.2, we observe that during training, the fractal dimension of the local weight trajectory can be viewed as a progress indicator for tracking the generalization behavior of neural networks when use iterative optimization algorithms (including but not limited to SGD). We conduct experiments across different training phases to illustrate this phenomenon. More experiment details are provided in Appendix B.

5.3.1 Correlation between fractal dimension of stationary distribution and generalization gap

In this section, we conduct three experiments to investigate the relationship between the fractal dimension of the stationary distribution and generalization and systematically study the effects of key parameters such as step size and batch size.

In the first two experiments, we train shallow neural networks using SGD on MNIST, FashionMNIST, and CIFAR-10 under varying hyperparameter configurations and random initializations. We approximate the stationary distribution of the SGD iterates via the weight trajectory, estimate its PH dimension, and perform Pearson correlation analysis between the estimated PH dimension and the generalization gap of accuracies (train acc-test acc). The results of these two experiments are shown in Figure 5-1 and 5-2, respectively.

Figure 5-1 illustrates the relationship between the fractal dimension of the SGD stationary distribution and the generalization gap of accuracies across varying learning rates and batch sizes. The results show that on both MNIST and FashionMNIST datasets, the two quantities exhibit a significantly positive correlation ($p \leq 0.05$), which validates the theoretical results presented in Section 5.1. Notably, the correlation is not significant on CIFAR-10, which may be attributed to the greater complexity of both the dataset and the resulting generalization behavior. Specifically, certain combinations of learning rate and batch size yield very low training and test accuracies in our CIFAR-10 experiments (e.g., learning rate = 1, batch size = 64, train acc = 0.437, test acc = 0.467), suggesting that the model fails to fit the data adequately. In such underfitting regimes, the relationship

between fractal dimension and generalization gap of accuracies becomes less pronounced.

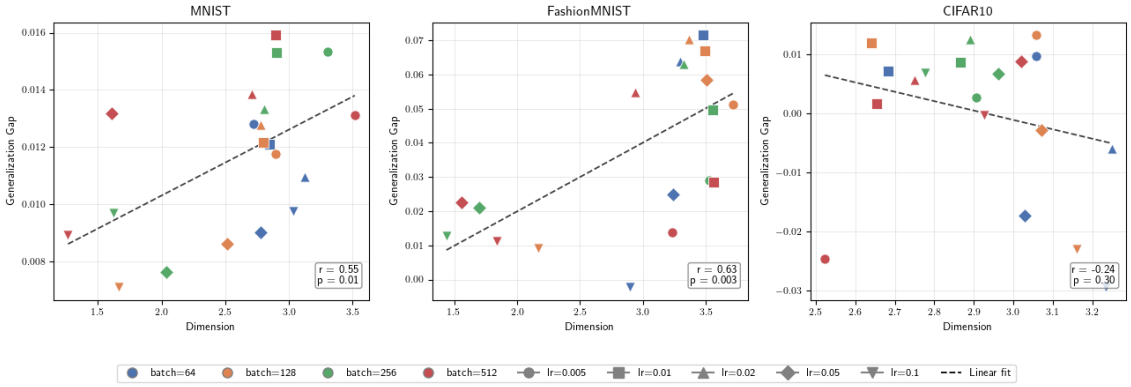

Figure 5-1 Correlation between the PH dimension of the stationary distribution and the generalization gap (train acc - test acc) for shallow neural networks trained with varying learning rates and batch sizes. Each subplot reports the Pearson correlation coefficient r and its associated p-value p in the bottom-right corner.

Figure 5-2 visualizes the relationship between the fractal dimension of the SGD stationary distribution and the generalization gap across different initializations. By varying the random seed, we observe that the trajectories converge to different regions, as shown in the upper figure of Figure 5-2. This is because in complex tasks, the iterated function system (IFS) induced by SGD can hardly satisfy a contraction property, and hence the absorbing sets are generally non-unique, as illustrated in the simplified setting in Section 4.2. The lower figure of Figure 5-2 illustrates the relationship between the fractal dimension of the SGD stationary distribution and the generalization gap. A clear positive correlation is observed on MNIST, while the relationship is not significant on the other two datasets. The stationary distribution seems to be less sensitive to the random seed on these two datasets, especially on CIFAR-10.

Across these two experiments, we observe that the PH dimension of the stationary distribution serves as a reliable indicator of generalization on simpler datasets such as MNIST, while no significant correlation is observed on CIFAR-10. As we discussed earlier, a reasonable explanation is that the shallow networks employed in these experiments lack sufficient capacity to learn meaningful representations on CIFAR-10, resulting in poor performance across all hyperparameter configurations. This in turn leads to minimal variance in the generalization gap, making it difficult to detect any meaningful correlation with the PH dimension. Motivated by this limitation, we conduct the third experiment using a deeper convolutional network on CIFAR-10, as described below.

In the third experiment, we train a deeper convolutional neural network using SGD on

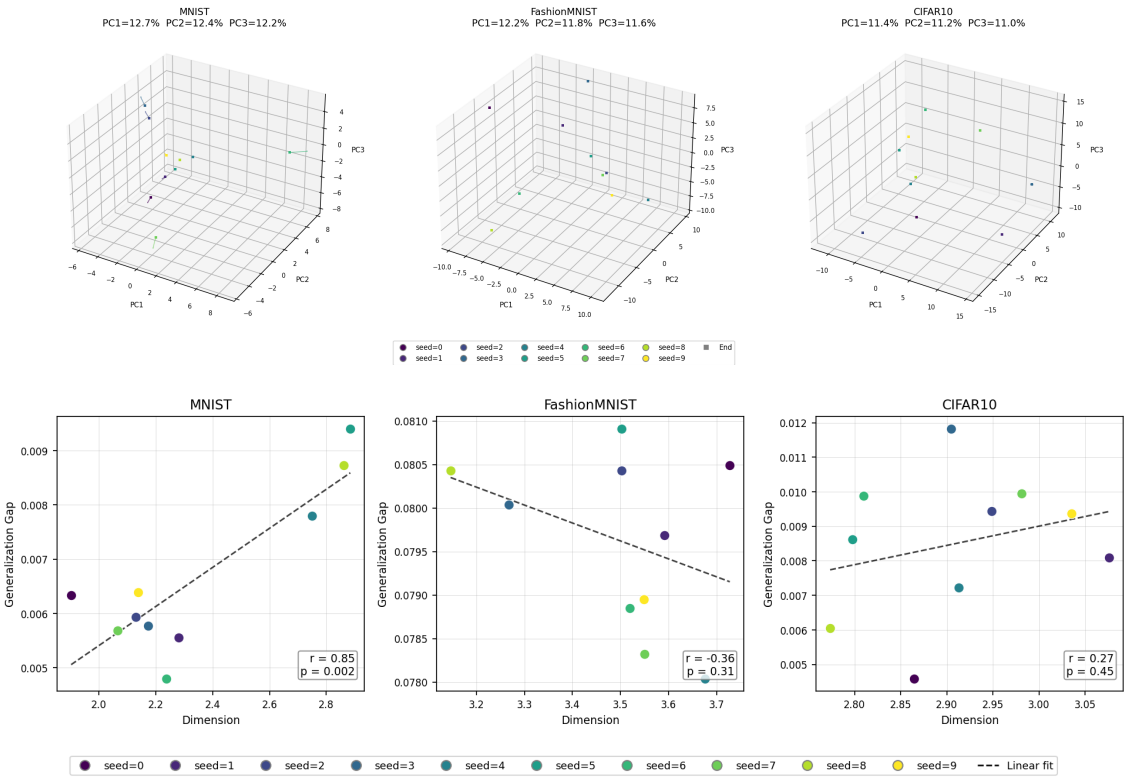

Figure 5-2 The upper figure visualizes high-dimensional SGD weight trajectories from different initializations, projected into three dimensions via PCA, where the small squares denote the regions of convergence at the end of training. The percentages indicate the explained variance ratio. The lower figure shows the correlation between the PH dimension of the stationary distribution and the generalization gap (train acc - test acc) for shallow neural networks across multiple random initializations. Each subplot reports the Pearson correlation coefficient r and the corresponding p-value p in the bottom-right corner.

CIFAR-10 under different learning rates and batch sizes. We conduct Pearson correlation analysis between the PH dimension of the stationary distribution and the generalization error of losses (|train loss- test loss|), and examine the individual relationships between each hyperparameter (learning rate and batch size) and both the PH dimension and generalization error, in order to disentangle their respective effects on the fractal dimension of the stationary distribution and the resulting generalization performance.

The results are shown in Figure 5-3. These results validate the positive correlation between the fractal dimension of the stationary distribution and the generalization error in deeper neural networks. Furthermore, the consistent relationships between learning rate, batch size, fractal dimension, and generalization error align with the well-established empirical observation that larger learning rates and smaller batch sizes tend to yield better generalization. Taken together, the fractal dimension can be viewed as a unified complexity measure that integrates the effects of both learning rate and batch size into a single

quantity predictive of generalization, therefore establishing an explicit link between algorithm hyperparameters and generalization performance that can inform principled model selection and hyperparameter tuning.

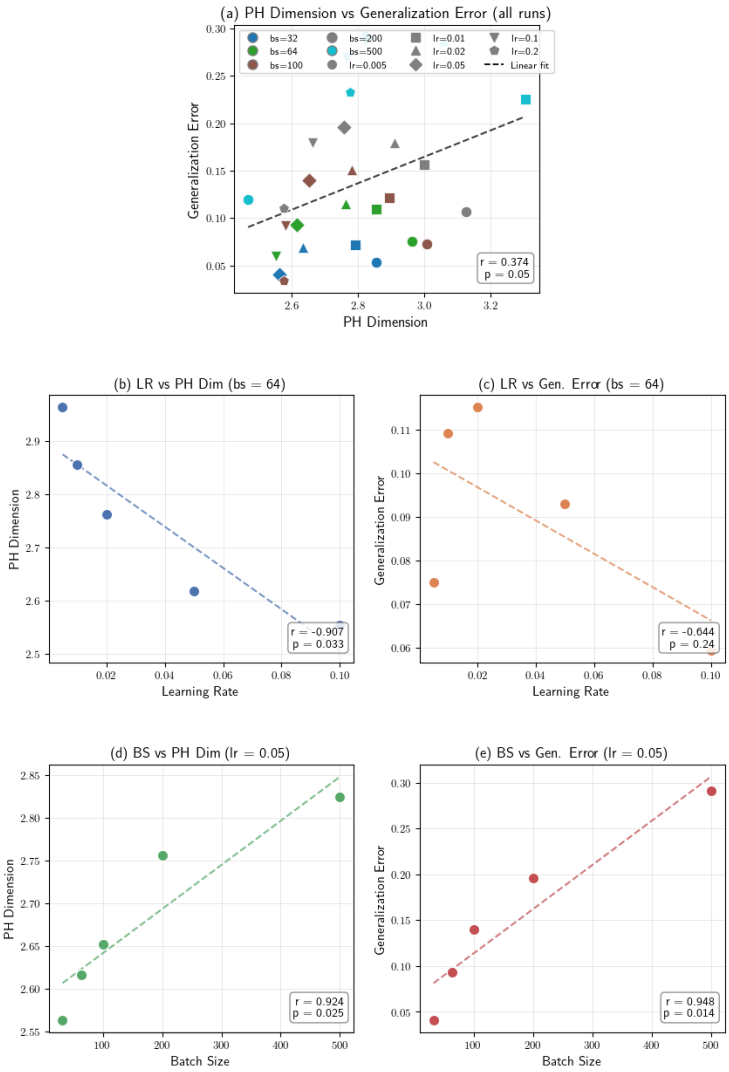

Figure 5-3 Correlation analysis between PH dimension of stationary distribution, generalization error (|train loss- test loss|), and hyperparameters for a deep neural network trained on CIFAR-10 with constant step-size SGD. (a) PH dimension of the stationary distribution versus generalization error across all learning rate and batch size configurations. (b) Learning rate versus PH dimension with batch size fixed at 64. (c) Learning rate versus generalization error with batch size fixed at 64. (d) Batch size versus PH dimension with learning rate fixed at 0.05. (e) Batch size versus generalization error with learning rate fixed at 0.05. Each subplot reports the Pearson correlation coefficient r and its associated p-value p in the bottom-right corner.

The three experiments above validate the effectiveness of the fractal dimension of the stationary distribution as an indicator of generalization. Meanwhile, we also point out potential limitations. One concern is that the strong negative correlation between learning rate and PH dimension may partly reflect a geometric artifact: smaller step sizes produce

denser weight trajectories, which naturally yield higher dimension estimates regardless of the loss landscape geometry. This confounding effect makes it difficult to isolate the generalization-relevant component of the fractal dimension from the mechanical influence of the step size. Future work could further disentangle these effects to more rigorously establish the validity of the fractal dimension of the stationary distribution as an intrinsic generalization indicator.

5.3.2 Evolution of the fractal dimension of local weight trajectories during training process

As an extension experiment, our motivation is that the fractal dimension of the weight trajectory can, to some extent, capture geometric information about the loss landscape, and may therefore serve as a reflection of the dynamic training process to help us better understand phenomena in deep learning. We conduct two experiments to investigate the evolution of the fractal dimension of local weight trajectories across different training phases and explore whether this perspective can provide a dynamic explanation for model training and generalization. Specifically, we employ a sliding window to collect local weight trajectories and plot the evolution curve of the fractal dimension of each windowed local weight trajectory throughout the entire training process.

In the first experiment, as one of the most standard experimental setups, we train a small fully connected neural network on the MNIST training dataset consisting of 60,000 samples using SGD optimizer and evaluate it on the test dataset of 10,000 samples. The evolution curve of the fractal dimension of local weight trajectories is shown in Figure 5-4. The results indicate that during training, as the model transitions from the underfitting phase to the fitting phase, the test accuracy increases while the fractal dimension of the local weight trajectory decreases.

The result is not surprising. Intuitively, as the model transitions from the underfitting phase to the fitting phase, it settles into the vicinity of a local minimum and the geometric structure of the trajectory simplifies accordingly. The generalization error keeps decreasing throughout the entire process.

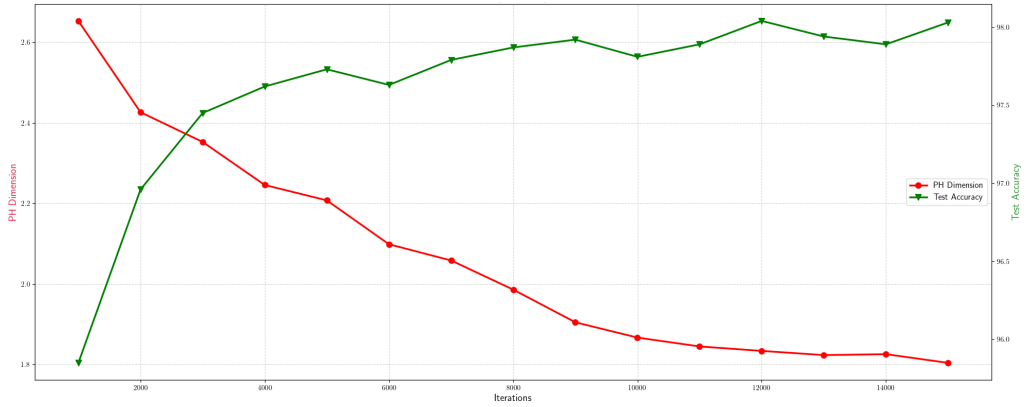

Figure 5-4 Evolution of test accuracy and PH dimension of the local weight trajectories during the training process in the standard underfitting-fitting experiment.

As an extension, in the second experiment, we train a fully connected neural network on a subset of the MNIST training dataset consisting of 1,000 samples using AdamW optimizer and test it on the test dataset with 10,000 samples. This setup is adopted from [38] to induce the Grokking phenomenon (also known as delayed generalization). Specifically, the training process is divided into a memorization phase and a generalization phase, as illustrated in the top panel of Figure 5-5.

In Figure 5-5, we visualize the fractal dimension of the local weight trajectories throughout the training process. As a complement, inspired by [3], we additionally visualize the fractal dimension of the data representations from the last hidden layer of the model. Interestingly, we observe that the fractal dimension of the local weight trajectories starts relatively low and gradually increases during the overfitting phase, with the rate of increase accelerating markedly as the generalization error begins to rise, before subsequently declining. Meanwhile, the fractal dimension of the data representations remains nearly flat with fluctuations in the early stages, and then begins to decrease simultaneously with the fractal dimension of the weight trajectories during the generalization phase.

These results suggest that the fractal dimensions of local weight trajectories and data representations can indeed correspond to phase transitions in model training and generalization. One possible explanation is that during the memorization phase, the model rapidly finds a memorizing solution close to the initialization with low trajectory complexity. During the overfitting phase (10^4-10^5 iterations), under the effect of weight decay regularization in AdamW, the model begins moving toward task-relevant directions and explores more complex regions, leading to increasing trajectory complexity. Finally, as the model slides into a valley containing a generalizing solution and settles near a flat minimum, the fractal dimension of the weight trajectories first rises briefly before gradually

decreasing.

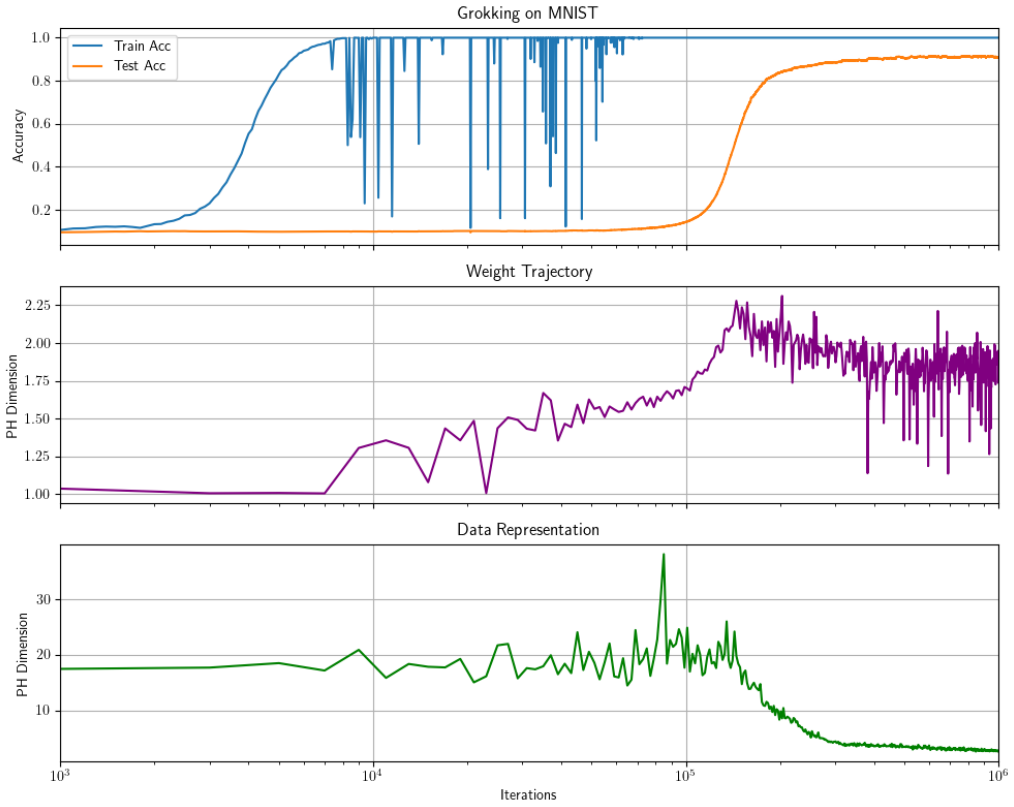

Figure 5-5 Evolution of test accuracy, PH dimension of the local weight trajectories and PH dimension of data representations during the training process in the Grokking experiment.

It is worth noting that, as discussed earlier, the Grokking phenomenon is closely tied to the underlying mechanism of AdamW optimizer (similar phenomenons are difficult to observe under the same setup with SGD). Since AdamW employs an adaptive learning rate, the fractal dimension of the weight trajectory is jointly influenced by both the loss landscape geometry and the optimizer's curvature adaptation, making the causal interpretation of Figure 5-5 non-trivial. Whether the fractal dimension of the local weight trajectory can serve as a rigorous generalization indicator under adaptive optimizers therefore remains an open question. Nevertheless, our results suggest that the fractal dimension of the stationary distribution captures meaningful geometric structure that correlates with generalization behavior, motivating a more systematic investigation into the relationship between optimization dynamics, trajectory geometry, and generalization.

CONCLUSION AND FUTURE WORK

Summary

This paper analyzes the optimization and generalization behavior of constant step-size SGD from the perspective of discrete-time random dynamical systems.

We analyze the optimization behavior of SGD under strongly convex and non-convex separable loss functions in Chapter 4. Strongly convex loss functions typically ensure that the IFS induced by SGD satisfies the average contractivity. This contractivity property guarantees that the corresponding Markov chain converges exponentially to a unique stationary distribution, which implies the limiting behavior of SGD is independent of initializations. The expectation of the optimization error on stationary distribution is bounded by the constant step size and the gradient noise at the minimizer.

For non-convex and separable losses, the results are totally different. Globally, there exist multiple absorbing sets and every trajectory eventually settles into one of them depending on the initialization, potentially converging to different stationary distributions as shown in [53]. Locally, we prove that on each absorbing set, different trajectories synchronize to the same trajectory under fixed sequence of iteration maps and the random behavior of trajectories do not depend on the initialization. We further prove that the Lyapunov exponent along the limiting orbit is negative, and based on this derive a local optimization error bound with the step size, gradient noise, and the local loss geometry within the absorbing set.

In Chapter 5, we demonstrate that from the perspective of random dynamical systems, the generalization error of the model is positively correlated with the fractal dimension of the stationary distribution, as shown in the generalization bound proposed in [13] and validated in the experiments in Section 5.3.1. Here, the fractal dimension of the stationary distribution serves as an algorithm-dependent measure of effective capacity and complexity, linking algorithmic hyperparameters to the generalization performance. For instance, as shown in the example of multi-class logistic regression with L_2 regularization presented in Section 5.2, SGD with a larger step size and smaller batch size tends to correspond to a smaller fractal dimension of the stationary distribution and yield better generalization, which is consistent with empirical findings. We also find that the fractal

dimension of the weight trajectory can serve as a progress indicator for tracking generalization performance during training in the empirical experiments. In the Grokking experiment, we observe that changes in the fractal dimension of the weight trajectory are synchronized with the phase transitions in the training process, which helps us better understand the optimization and generalization behavior of stochastic iterative algorithms during training. This motivates further investigation into the underlying reasons for counterintuitive phenomena in machine learning.

This framework offers a new perspective for analyzing the long-term qualitative behavior of constant step-size SGD and establishing optimization and generalization error bounds by leveraging tools from discrete-time random dynamical systems. It goes beyond the SDE and Markov chain framework commonly adopted in the existing literature, providing a more refined framework that enables us to gain interesting insights into the optimization and generalization of constant step-size SGD.

Future work

Some directions for future work building on this paper have already been discussed in Section 4.2.4 and 5.3.1. Here we highlight several potential directions for extension.

A natural direction for future work is whether the analysis and results under nonconvex separable loss functions can be extended to more general non-convex settings. Here the separability assumption simplifies the construction of absorbing sets and the verification of monotonicity and the splitting condition for the iterated maps, which serves as a key technical tool for proving convergence of the Markov chain and establishing synchronization in each absorbing set. For general non-convex loss functions, there are two main challenges to establish similar Doeblin decompsition. The first concerns whether absorbing sets can be effectively identified and constructed. Explicit construction of the boundaries of absorbing sets typically requires accounting for correlations among variables across different dimensions, and an absorbing set can be a closed set with much more complex geometry, necessitating case-by-case treatment. A possible direction to avoid explicit construction is to adopt a Lyapunov function approach. The most straightforward idea is to use sublevel sets of the loss function F as candidate absorbing sets, decomposed into connected components associated with distinct local minima. This would requires restrictions on the step size and the gradient noise to ensure that the sublevel sets are positively invariant and more general case would require the construction of other

Lyapunov functions, which remains a highly nontrivial challenge. The second concerns proving uniqueness of the stationary distribution within each absorbing set and establishing a spectral gap of the Markov operator. This may require richer mathematical tools, such as Lyapunov exponents analysis and other coupling methods. Furthermore, under general non-convex loss functions, exploring pathwise properties (e.g., Lyapunov exponents, synchronization), which are closely related to the step size and the geometry of the loss landscape, may yield interesting insights. For instance, in the setting of this paper, synchronization within each absorbing set can be established and the Lyapunov exponents are shown to be strictly negative. In [15], also within a discrete-time RDS framework, it is shown that the sign of the Lyapunov exponents determines whether SGD can converge to the corresponding global minimum in overparametrized neural networks. Under large step sizes and more complex loss landscape geometry, chaotic phenomena may emerge [30]. These insights can contribute to a deeper understanding and improvement of stochastic optimization algorithms.

If analogous results can be established for general non-convex loss landscapes, a promising direction for future work would be to leverage the connection between optimization dynamics and generalization performance within the random dynamical systems framework to further investigate the implicit bias of SGD. As shown in Theorem 4.2, SGD may converge to different absorbing sets, and the probability of converging to any particular one and the geometry of it, to some extent, reflect the inherent preference of SGD. This may be a worthwhile entry point for investigating the implicit bias in the future.

There remain several related topics that have not been covered in this paper, which could serve as directions for future work. The heavy-tailed phenomenon of SGD [29, 54, 35], where the second moment of gradient noise is unbounded, has not been addressed here and remains an open challenge. More broadly, the discrete-time random dynamical system framework can potentially be extended to other stochastic optimization algorithms, such as SGD with momentum. For these variants of SGD, the weight iterates are no longer Markovian and a preliminary idea is to augment the state space with auxiliary variables so that the framework of discrete-time random dynamical systems can be applied. We hope these directions will inspire further investigation to help us better understand the optimization and generalization of stochastic optimization algorithms.

REFERENCES

[1] ALSMEYER G, FUH C D. Limit theorems for iterated random functions by regenerative methods[J]. Stochastic processes and their applications, 2001, 96(1): 123-142.

[2] AMBROSIO L, GIGLI N, SAVARÉ G. Gradient flows: in metric spaces and in the space of probability measures[M]. Springer, 2005.

[3] ANSUINI A, LAIO A, MACKE J H, et al. Intrinsic dimension of data representations in deep neural networks[J]. Advances in Neural Information Processing Systems, 2019, 32.

[4] ARNOLD L. Random dynamical systems[G]//Dynamical Systems: Lectures Given at the 2nd Session of the Centro Internazionale Matematico Estivo (CIME) held in Montecatini Terme, Italy, June 13–22, 1994. Springer, 2006: 1-43.

[5] ATTOUCH H, BOLTE J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features[J]. Mathematical Programming, 2009, 116(1): 5-16.

[6] BARNSLEY M F, DEMKO S G, ELTON J H, et al. Invariant measures for Markov processes arising from iterated function systems with place-dependent probabilities[C]//Annales de l'IHP Probabilités et statistiques: vol. 24: 3. 1988: 367-394.

[7] BARTLETT P L, MENDELSON S. Rademacher and gaussian complexities: Risk bounds and structural results[J]. Journal of machine learning research, 2002, 3(Nov): 463-482.

[8] BHATTACHARYA R N, LEE O. Asymptotics of a class of Markov processes which are not in general irreducible[J]. The Annals of Probability, 1988: 1333-1347.

[9] BIRDAL T, LOU A, GUIBAS L J, et al. Intrinsic dimension, persistent homology and generalization in neural networks[J]. Advances in neural information processing systems, 2021, 34: 6776-6789.

[10] BOLTE J, DANIILIDIS A, LEWIS A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems[J]. SIAM Journal on Optimization, 2007, 17(4): 1205-1223.

[11] BOTTOU L, CURTIS F E, NOCEDAL J. Optimization methods for large-scale machine learning[J]. SIAM review, 2018, 60(2): 223-311.

[12] BOUSQUET O, ELISSEEFF A. Stability and generalization[J]. Journal of machine learning research, 2002, 2(Mar): 499-526.

[13] CAMUTO A, DELIGIANNIDIS G, ERDOGDU M A, et al. Fractal structure and generalization properties of stochastic optimization algorithms[J]. Advances in neural information processing systems, 2021, 34: 18774-18788.

[14] CARLSSON G, VEJDEMO-JOHANSSON M. Topological data analysis with applications[M]. Cambridge University Press, 2021.

[15] CHEMNITZ D, ENGEL M. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning[J]. Journal of Machine Learning Research, 2025, 26(134): 1-46.

[16] CHUESHOV I. Monotone random systems theory and applications[M]. Springer, 2004.

[17] DEREICH S, KASSING S. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes[J]. arXiv preprint arXiv:2102.09385, 2021.

[18] DIACONIS P, FREEDMAN D. Iterated random functions[J]. SIAM review, 1999, 41(1): 45-76.

[19] DÍAZ L J, MATIAS E. Stability of the Markov operator and synchronization of Markovian random products[J]. Nonlinearity, 2018, 31(5): 1782-1806.

[20] DIEULEVEUT A, DURMUS A, BACH F. Bridging the gap between constant step size stochastic gradient descent and Markov chains[J]. 2020.

[21] DUPUIS B, VIALLARD P, DELIGIANNIDIS G, et al. Uniform generalization bounds on data-dependent hypothesis sets via pac-bayesian theory on random sets[J]. Journal of Machine Learning Research, 2024, 25(409): 1-55.

[22] ENGEL M. Lecture Notes on Random Dynamical Systems[J].

[23] FALCONER K. Fractal geometry: mathematical foundations and applications[M]. John Wiley & Sons, 2013.

[24] FEHRMAN B, GESS B, JENTZEN A. Convergence rates for the stochastic gradient descent method for non-convex objective functions[J]. Journal of Machine Learning Research, 2020, 21(136): 1-48.

[25] FENG D J, HU H. Dimension theory of iterated function systems[J]. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 2009, 62(11): 1435-1500.

[26] FORET P, KLEINER A, MOBAHI H, et al. Sharpness-aware minimization for efficiently improving generalization[J]. International Conference on Learning Representations, 2020.

[27] GEWEKE J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments[R]. Federal Reserve Bank of Minneapolis, 1991.

[28] GUPTA A, HASKELL W B. Convergence of recursive stochastic algorithms using Wasserstein divergence[J]. SIAM Journal on Mathematics of Data Science, 2021, 3(4): 1141-1167.

[29] GURBUZBALABAN M, SIMSEKLI U, ZHU L. The heavy-tail phenomenon in SGD[C]// International Conference on Machine Learning. 2021: 3964-3975.

[30] HERRMANN L, GRANZ M, LANDGRAF T. Chaotic dynamics are intrinsic to neural network training with SGD[J]. Advances in Neural Information Processing Systems, 2022, 35: 5219-5229.

[31] HUTCHINSON J E. Fractals and self similarity[J]. Indiana University Mathematics Journal, 1981, 30(5): 713-747.

[32] JASTRZĘBSKI S, KENTON Z, ARPIT D, et al. Three factors influencing minima in sgd[J]. arXiv preprint arXiv:1711.04623, 2017.

[33] KOZMA G, LOTKER Z, STUPP G. The minimal spanning tree and the upper box dimension [J]. Proceedings of the American Mathematical Society, 2006, 134(4): 1183-1187.

[34] LI C, FARKHOOR H, LIU R, et al. Measuring the intrinsic dimension of objective landscapes [J]. arXiv preprint arXiv:1804.08838, 2018.

[35] LI J, LOU Z, RICHTER S, et al. The stochastic gradient descent from a nonlinear time series perspective[J]. preprint, 2024.

[36] LI Q, TAI C, et al. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations[J]. Journal of Machine Learning Research, 2019, 20(40): 1-47.

[37] LI Z, MALLADI S, ARORA S. On the validity of modeling sgd with stochastic differential equations (sdes)[J]. Advances in Neural Information Processing Systems, 2021, 34: 12712-12725.

[38] LIU Z, MICHAUD E J, TEGMARK M. Omnigrok: Grokking beyond algorithmic data[J]. International Conference on Learning Representations, 2022.

[39] MANDT S, HOFFMAN M D, BLEI D M. Stochastic gradient descent as approximate bayesian inference[J]. Journal of Machine Learning Research, 2017, 18(134): 1-35.

[40] MOULINES E, BACH F. Non-asymptotic analysis of stochastic approximation algorithms for machine learning[J]. Advances in neural information processing systems, 2011, 24.

[41] NANDA V. Computational algebraic topology lecture notes[J]. URL: https://people. maths. ox. ac. uk/nanda/cat/TDANotes. pdf, 2021.

[42] NEYSHABUR B, BHOJANAPALLI S, MCALLESTER D, et al. Exploring generalization in deep learning[J]. Advances in neural information processing systems, 2017, 30.

[43] NICOL M, SIDOROV N, BROOMHEAD D. On the fine structure of stationary measures in systems which contract-on-average[J]. Journal of Theoretical Probability, 2002, 15(3): 715-730.

[44] PAVLIOTIS G A. Stochastic processes and applications[J]. Texts in applied mathematics, 2014, 60: 41-43.

[45] PRZYTYCKI F, URBAŃSKI M. Conformal fractals: ergodic theory methods: vol. 371[M]. Cambridge University Press, 2010.

[46] RAMS M. Dimension estimates for invariant measures of contracting-on-average iterated function systems[J]. arXiv preprint math/0606420, 2006.

[47] ROBB R. w. G. Cochran, Sampling Techniques (John Wiley & Sons, 1963), ix+ 413 pp., 72s. [J]. Proceedings of the Edinburgh Mathematical Society, 1963, 13(4): 342-343.

[48] ROBBINS H, MONRO S. A stochastic approximation method[J]. The annals of mathematical statistics, 1951: 400-407.

[49] RUELLE D. Characteristic exponents and invariant manifolds in Hilbert space[J]. Annals of Mathematics, 1982: 243-290.

[50] SCHEUTZOW M. Comparison of various concepts of a random attractor: A case study[J]. Archiv der Mathematik, 2002, 78(3): 233-240.

[51] SCHEUTZOW M, VORKASTNER I. Synchronization, Lyapunov exponents and stable manifolds for random dynamical systems[C]//International Conference on Stochastic Partial Differential Equations and Related Fields. 2016: 359-366.

[52] SCHWEINHART B. Fractal dimension and the persistent homology of random geometric complexes[J]. Advances in Mathematics, 2020, 372: 107291.

[53] SHIROKOFF D, ZALESKI P. Convergence of Markov Chains for Constant Step-Size Stochastic Gradient Descent with Separable Functions[J]. SIAM Journal on Applied Dynamical Systems, 2025, 24(3): 2005-2043.

[54] SIMSEKLI U, SENER O, DELIGIANNIDIS G, et al. Hausdorff dimension, heavy tails, and generalization in neural networks[J]. Advances in Neural Information Processing Systems, 2020, 33: 5138-5151.

[55] SOKAL A. Monte Carlo methods in statistical mechanics: foundations and new algorithms[G]//Functional integration: Basics and applications. Springer, 1997: 131-192.

[56] TUCI M, BASTIAN L, DUPUIS B, et al. Mutual Information Free Topological Generalization Bounds via Stability[J]. arXiv preprint arXiv:2507.06775, 2025.

[57] YU L, BALASUBRAMANIAN K, VOLGUSHEV S, et al. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias[J]. Advances in Neural Information Processing Systems, 2021, 34: 4234-4248.

[58] ZHANG C, BENGIO S, HARDT M, et al. Understanding deep learning requires rethinking generalization[J]. International Conference on Learning Representations, 2016.

[59] ZHANG J, LI H, SRA S, et al. Neural network weights do not converge to stationary points: An invariant measure perspective[C]//International Conference on Machine Learning. 2022: 26330-26346.

APPENDIX A ADDITIONAL TECHNICAL BACKGROUND

This appendix provides additional background for Theorem 4.1 and Theorem 4.2 in Chapter 4.

Definition A.1: Let $(\mathcal{X}, d)$ be a complete and separable space. Define

$$\mathcal{P}_1(\mathcal{X}) \triangleq \left\{\mu \text{ probability on } \mathcal{X} \,:\, \int_{\mathcal{X}} d(x, x_0)\,\mu(dx) < \infty, \text{for some } x_0 \in \mathcal{X}\right\}.$$

For $\mu_1, \mu_2 \in \mathcal{P}_1(X)$, let $C(\mu_1, \mu_2)$ be the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ_1 and μ_2. Then the 1-Wasserstein distance between μ_1 and μ_2 is

$$W_1(\mu_1, \mu_2) := \inf_{\nu \in C(\mu_1, \mu_2)} \int_{\mathcal{X}\times\mathcal{X}} d(x, x')\, d\nu(x, x').$$

Definition A.2: Let $A \subset \mathbb{R}^d$ be a closed and bounded Borel set. For probability measure $\mu_1, \mu_2 \in \mathcal{P}(A)$, define the total variation distance between μ_1 and μ_2

$$d_{TV}(\mu_1, \mu_2) := \sup_{C \in \mathcal{B}(\mathbb{R}^d)} |\mu_1(C \cap A) - \mu_2(C \cap A)|.$$

The following metric is defined on the partially ordered structure and is weaker than d_{TV}.

Definition A.3: Let $\{e_1, e_2, ..., e_d\}$ be the basis vectors of $\mathbb{R}^d$. Define the cone

$$\mathbb{R}^d_\alpha := \left\{c_1(\alpha_1 e_1) + c_2(\alpha_2 e_2) \ldots + c_d(\alpha_d e_d) \,:\, c_j \geq 0 \text{ for } 1 \leq j \leq d\right\},$$

where $\alpha = (\alpha_1, \alpha_2, .., \alpha_d) \in \{+1, -1\}^d$. We call a map f is monotone with respect to $\mathbb{R}^d_\alpha$ if

$$x \preceq_\alpha x' \quad \Longrightarrow \quad f(x) \preceq_\alpha f(x'),$$

for all $x, x' \in \mathbb{R}^d$, where $x \preceq_\alpha y$ means $y - x \in \mathbb{R}^d_\alpha$.

Let $A \subset \mathbb{R}^d$ be a closed and bounded Borel set. For probability measure $\mu_1, \mu_2 \in \mathcal{P}(A)$, define

$$d_\alpha(\mu_1, \mu_2) := \sup_{C \in \mathcal{C}_\alpha} |\mu_1(C \cap A) - \mu_2(C \cap A)|,$$

where $\mathcal{C}_\alpha = \left\{\{x \in \mathbb{R}^d \,:\, f(x) \preceq_\alpha c\}, f \text{ is monotone with respect to } \mathbb{R}^d_\alpha, c \in \mathbb{R}^d\right\}$.

APPENDIX B EXPERIMENT DETAILS

Model architecture and algorithm hyperparameters

In the first two experimemts of Section 5.3.1, we use two convolutional layers with 8 and 16 filters respectively, each followed by max pooling, and a fully connected head with one hidden layer of 64 units for MNIST and FashionMNIST. For CIFAR-10, we use three convolutional layers with 16, 32, and 32 filters respectively, followed by a fully connected head with one hidden layer of 128 units. We use the SGD optimizer and Cross entropy loss, and a learning rate of 0.02 and a batch size of 128 in the second experiment. We train 5 convolutional layers with 64 channels each and 3 fully connected layers on CIFAR-10 for the last experiment.

In the first experiment of Section 5.3.2, we use three fully connected layers of 100 units with ReLU activations, and a final linear head mapping to 10 output classes. We use the SGD optimizer and Cross entropy loss with a learning rate of 0.1, a batch size of 50. For the second experiment, we use three fully connected layers of 200 units with ReLU activations, and a final linear head mapping to 10 output classes. We use the AdamW optimizer and MSE loss with a learning rate of 0.001 and a batch size of 200.

Stationary distribution approximation and PH dimension estimation

We use the weight trajectory to approximate the stationary distribution under the assumption of ergodicity. Specifically, We collect weight samples from the constant step-size SGD trajectory in two phases: A warmup phase runs until the chain reaches stationarity, detected via a sliding-window Geweke diagnostic [27], followed by a collection phase which records a weight snapshot every 5 gradient steps. To obtain approximately i.i.d. draws from the stationary distribution, we thin the chain by retaining every τ-th snapshot, where τ is the integrated autocorrelation time estimated via Sokal's automatic windowing algorithm [55]. We estimate the PH dimension of these samples using tools from TDA [9] and report the train and test losses and accuracies as averages over all collected samples.

ACKNOWLEDGEMENTS

我非常感谢朱一飞老师在本科与硕士期间对我的指导、鼓励和教诲。他鼓励我去探索更多有趣的问题，为我提供了很多机会，去拓展自己的边界。与朱老师的许多谈话，令我受益良多。

我还要感谢吴开亮老师，苏耀峰老师对我学业规划与科研上的指导，课题组同学对我的热心帮助。最后感谢我的家人与朋友们一路以来的陪伴与支持。

RESUME

Zhang Haiyu was born in 2002, in Jiaxing, Zhejiang, China.

In September 2020, she was admitted to Southern University of Science and Technology (SUSTech). In June 2024, she obtained her bachelor's degree in science from the Department of Mathematics, SUSTech.

From September 2024, she started to pursue a master degree of science in the Department of Mathematics, SUSTech.