©Copyright 2018
Xiaofeng Ye

Stochastic Dynamics: Markov Chains, Random Transformations and Applications

Xiaofeng Ye

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Hong Qian, Chair

Panos Stinis

Eric Shea-Brown

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Stochastic Dynamics: Markov Chains, Random Transformations and Applications

Xiaofeng Ye

Chair of the Supervisory Committee:
Professor Hong Qian
Department of Applied Mathematics

Stochastic dynamical systems, as a rapidly growing area in applied mathematics, has been a successful modeling framework for biology, chemistry and data science. Depending upon the origin of uncertainties in an application problem, the theory of stochastic dynamics has two different mathematical representations: stochastic processes and random dynamical systems (RDS). RDS is a more refined mathematical description of the reality; it provides not only the stochastic trajectory following one initial condition, but also describes how the entire phase space, with all initial conditions, simultaneously changes with time. Stochastic processes represent the stochastic movement of individual system. RDS, however, describes the motions of many systems that experience a common deterministic law that is randomly changing with time due to extrinsic noises, which represent fluctuating environment or complex external signal. The dynamics of an RDS may exhibit a quite counterintuitive phenomenon called noise-induced synchronization: the stochastic motions of noninteracting systems under a common noise synchronize; their trajectories become close to each other, while the individual one remains stochastic. In Chapter 2, I establish some elementary contradistinctions between Markov chain (MC) and RDS descriptions of stochastic dynamical systems with discrete time and discrete state space setting. In particular I study the linear representation of the RDS and show the expectation of the matrix-valued random variable is in fact the transition probability matrix of the corresponding MC induced by i.i.d. RDS. In Chapter 3,

I study the metric entropy of MC and its corresponding RDS, and establish several inequalities about entropies and entropy productions. Next in Chapter 4 and Chapter 5, the theory of noise-induced synchronization is introduced together with a more intuitive version of the multiplicative ergodic theory, and then is applied to hidden Markov models for developing an efficient algorithm of parameter inference. In Chapter 6, The multi-dimensional Ornstein-Uhlenbeck process is used to study the dynamics of a free-draining polymer, in particular, the mean looping time. This work points to a future direction for stochastic model reduction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge many people who make this dissertation possible.

First, I would like to express my gratitude to my advisor, Professor Hong Qian, for his support and mentorship over my graduate study at University of Washington. His thought-provoking instruction motivates me to dive deep into the problem and to pursue the convincing answer. His foresight and vision helps me foresee the future of the field. I would like to thank my thesis committee. I appreciate Dr. Panos Stinis for his encouragement and insightful discussions. Many ideas were inspired during our deep conversations. I am also grateful to Professor Eric Shea-Brown, Professor Ioana Dumitriu and Professor Eli Shlizerman for providing valuable suggestions.

I am very fortunate to work with my group mates, as well as my collaborators, Dr. Yue Wang and Dr. Yian Ma. I would like to thank Dr. Wang to discuss rigorous mathematics many times with me on whiteboard and Dr. Ma to introduce a fascinating application to my theory. I would like to express my thanks to Dr. Nathan Baker for his mentorship during my internship in PNNL. I appreciate Professor Kevin Lin and Professor Anthony Quas for teaching me MET and writing letters for me. I also thank our department chair Professor Bernard Deconinck for his help and encouragement.

Most importantly, I would like to attribute my accomplishment to my family. My parents are always by my side to support me mentally and financially. Thank forever to my wonderful wife, Candy Yu, who has offered me great support and understanding. Our mutual trust overcomes all types of difficulties. Without them, I would not be able to accomplish this dissertation.

DEDICATION

This work is dedicated to Candy Yu, my wonderful wife,
to my parents for supporting me all the way and to my dreams.

Chapter 1

INTRODUCTION

Dynamic descriptions of a natural phenomenon is the foundation of modern science. Mathematical theories of dynamics are usually divided into deterministic and stochastic models. While the theory of deterministic dynamics, first articulated by Isaac Newton, has great impact on our understanding of natural and engineering world, there are growing interests in the latter descriptions of complex natural phenomena such as many-body physics, chemical kinetics, and biology. [80]

The modern theory of deterministic, nonlinear dynamical system provides a qualitative view of global understanding of dynamics by looking at the evolution of the whole phase space. The goal of the theory of dynamical system is to understand the ensemble of solutions, as a function of either initial conditions, or as a function of parameters arising in the system. Dynamical systems with discrete or continuous times are usually referred to as *iterative maps* or *flows*, respectively. Phase portrait and local linear stability analysis of maps and differential equations are now routines of analyzing dynamics. Important concepts emerging from this type of studies are the notions of attractors, invariant manifolds, local and global vector field bifurcations, topological conjugacies and canonical forms, linear representations of nonlinear dynamics, Lyapunov exponents [48, 120], among others. In particular, the "attractor" of a dynamical system should ideally contain the asymptotic dynamics of simultaneous trajectories with different initial conditions. It has been extensively studied in autonomous and non-autonomous dynamical system [22, 46]. In the latter case, a distinction between push forward attractor and pullback attractor has arisen. At the same time, the field of nonlinear deterministic dynamics has also witnessed a surge of activities in terms of Perron-Frobenius-Ruelle operator (or transfer operator) [9, 67] and Koopman operator [3] as

the linear representations of nonlinear dynamics.

The stochastic processes perspective, particularly in terms of Markov processes, also have successful applications in engineering, chemistry, biology and physics [35, 97, 115]. For example, continuous time Markov process, like the multi-dimensional Ornstein-Uhlenbeck (OU) process, is one of the most frequently used mathematical models to describe the dynamics of free-draining polymers. It is not an overstatement that the theory of polymer dynamics is founded on the applied mathematics of stochastic processes. While the general theory of an OU process is well developed (see [93] and the references cited within), explicitly analytical results on the kinetics of the formation of a end-to-end loop are still highly sought after in theoretical chemistry and biochemistry [111, 126].

On the other hand, a large class of stochastic models is based on discrete state discrete time *Markov Chain* (MC), with either finite or countable state space. Not only this setting reduces the technicality of the mathematical ideas for their introduction to a much broader audiences, but also many measurements in science and engineering are intrinsically discrete. One can find extensive literature on studies of the statistical properties of the sample trajectory, such as, invariant distribution, metric entropy, entropy production rate, mean first passage time and mixing time, etc [1, 51, 67, 118].

The stochastic counterpart of dynamical systems theory is known as *random transformation* [60] or *random dynamical system*(RDS) [4, 61]. The RDS theory has a mathematical setup that is rather different from the theory of Markov processes and it is currently accessible only to a small group of professional mathematicians. The existing theory is mostly based on continuous phase space. One of the goals of my work is to initiate an applied mathematical study of RDS with discrete state space. In particular, we like to establish the relation, as intuitively as possible, between the theory of MC and the theory of random transformation. In the mean time, many results in continuous state are compared.

The main idea of RDS is to pick deterministic transformation, one by one, from a set of possible transformations with certain probability and to apply the transformation on the current states. The simplest case is the iterations of random transformations are chosen

independently with identical distribution (i.i.d.). In this case, the one-point motion of RDS is statistically equivalent with the MC. However, the difference arises when simultaneous trajectories with different initial conditions are considered. These simultaneous motion of two or more points in RDS are not independent, and actually once they collide at some instance they will be together forever. In the continuous state space, trajectories of RDS subjected to the same randomness, but starting from different initial conditions may converge to a single random solution almost surely. This phenomenon is called *noise-induced synchronization*. Synchronization has been widely discovered as a relevant property in modeling of external noises [76]. In neurosciences, one observes this synchronization by common noise as a reliable response of one single neural oscillator on a repeatedly applied external pre-recorded input, which may be seen as a dynamical system driven by the same noise path but different initial conditions [66, 73]. We note not every RDS possesses the synchronization property. Crudely speaking, in order to see synchronization, one needs two ingredients: local contraction so that nearby points approach each other; along with a global irreducibility condition. The local contraction condition is equivalent with the negative maximum Lyapunov exponents in RDS. The Lyapunov exponents in RDS that characterize the rate of separation of infinitesimally close trajectories, are defined by the Oseledets multiplicative ergodic theorem (MET) [86]. This theorem is the most fundamental theorem in theory of RDS. Not only in neuroscience, but also other fields, including the data science, have discovered behavior and properties relevant to MET [8]. For example, in Chapter 5, it is utilized to develop an efficient statistical inference algorithm [124].

Through studying synchronization property in an RDS, one appreciates the fact that RDS formulation is a more refined model of stochastic dynamics than an MC. In other words, RDS itself contains more information (or randomness) than its corresponding MC. From statistical physics, entropy is exactly the quantity to measure the information (or randomness) of the system. It could reflect the information in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ if the random variable is properly chosen, such as $-\log(\mathbb{P}(\omega))$. But in the dynamical system, the set of possible outcome Ω in the probability space is all possible sequences. It is distinctly different from

physicists' notion of dynamics as "step-by-step" motion. This difference brings two different choices of random variables: the current state X_t for physicists and the sequence up to time t, $\{X_s\}_{0\leq s\leq t}$ for mathematicians. In the mean time, entropy itself also has two choices: the Shannon entropy $S(\mathbb{P}(\omega)) = \mathbb{E}^{\mathbb{P}}(-\log \mathbb{P}(\omega))$ and the relative entropy $H(\mathbb{P}, \mathbb{P}') = \mathbb{E}^{\mathbb{P}}(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{P}'}(\omega))$. Now it has four different types of entropy. Statistical physicists consider the instantaneous distribution of the current state $\mathbf{p}(t)$, then $S(\mathbf{p}(t))$ and H($\mathbf{p}$(t), $\boldsymbol{\pi}$) are identified as instantaneous Gibbs entropy and the free energy of the canonical system, respectively, where $\boldsymbol{\pi}$ is the invariant distribution. The free energy describes how far the current state of the system is from its equilibrium state. Since the entropy of the infinitely long sequence is in general $+\infty$, one considers the entropy "produced" per step, which is a different class of quantity. The average Shannon entropy produced per step is called metric entropy. With the comparison, one would ask the connection with the change of instantaneous Gibbs entropy $\Delta S(\mathbf{p}(t))$. In fact, $\Delta S(\mathbf{p}(t))$ is shown less or equal that the Shannon entropy produced per step in Proposition 3.1. The choice of the reference measure in the relative entropy here is the probability measure of the time-reversed process. Then the average relative entropy produced per step is called entropy production rate, which quantifies the irreversibility of the system. So the notions of instantaneous Gibbs entropy, metric entropy and entropy production rate are three distinct concepts and each represents a different aspect of the same stochastic dynamics: The randomness in the system at an instance, the randomness generated in a step, and the randomness with respect to its time reversal [125]. In this work, we are interested in these concepts under the representations of finite i.i.d. RDS.

This dissertation is organized as follows: In Chapter 2, I provide the definition of a general RDS with a medium level of rigor, and provide some simple examples in finite state space. I discuss the linear representations of deterministic maps in an RDS and its corresponding MC. Then I introduce the stochastic Perron-Frobenius operator and stochastic Koopman operator. I show a finite i.i.d. RDS uniquely defines a MC, but a given MC is generally compatible with many possible RDS. In Chapter 3, I first provide a brief, but rather coherent presentation of the theory of entropy and entropy production of MC. Then I

use metric entropy to characterize the different RDS representations of an MC and the upper and lower bounds of the metric entropy of RDS associated with an MC are analyzed. At last, I establish several interesting relationships about entropy production between the MC and its corresponding maximum entropy RDS, and the doubly stochastic MC and its invertible RDS. In Chapter 4, I introduce the synchronization in discrete state setting and show the sufficient and necessary conditions for finite RDS and maximum entropy RDS respectively. I also compare existing results with the help of Lyapunov exponents and give one explicit example. At last, I discuss the Multiplicative ergodic theorem (MET) and the extension of my results to countable states. In Chapter 5, I observe that under certain mild conditions, the forward probability in a hidden Markov model exhibits synchronization, which yields an efficient estimation with subsequences. Here I develop a mini-batch gradient descent algorithm for parameter inference in the hidden Markov model. I first efficiently estimate the rate of synchronization, which was proven as the gap of top Lyapunov exponents, and then fully utilize it to approximate the length of subsequences in the mini-batch algorithm. I theoretically validate the algorithm and numerically demonstrate the effectiveness. In Chapter 6, I study the dynamics of a free-draining polymer, specifically the mean looping time. I reformulate and generalize the heuristic Wilemski-Fixman (WF) method, which was previously only known in the field of polymer physics, into a systematical model reduction method for the narrow escape problem in the multi-dimensional Ornstein-Uhlenbeck process. The asymptotic results of mean looping time are analytically extracted.

Chapter 2

MARKOV CHAIN AND FINITE RANDOM DYNAMICAL SYSTEM

The theory of random dynamical system (RDS) studies the action of random maps, drawn from a collection with prescribed probability, on a state space. Heuristically, the difference between an MC and an RDS is that the randomness in the former arises in a particular dynamics while in the latter is embedded in the "law of motion". In this chapter, I present a brief overview of some essential concepts of RDS that are particularly relevant and follow the construction to rigorously define the RDS [4, 6, 60, 108]. We shall start with general RDS and later specify each term for RDS with finite state space, and focus mainly on i.i.d. case. We establish the connection between the finite i.i.d. RDS and MC. For a given finite i.i.d. RDS, the transition probability on state space can be defined and it further induces a finite-state MC. In the mean time, for a given MC, it is shown that there exists a representation by means of i.i.d. random transformations. This RDS representation of an MC, however, is not unique. Different RDS representations of the same MC yield different behaviors, such as synchronization among trajectories of different initial states. The presentation is pedagogically self-contained.

2.1 RDS set-up

Let $\mathscr{S}$ be the state space, Γ be a family of maps from $\mathscr{S}$ into itself. Let Q be the probability measure on the σ-field of Γ. The set Γ is interpreted as the set of all admissible laws of dynamics.

Definition 2.1. $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ *is a metric*[1] *dynamical system if* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space*

[1]The term metric is often used in the literature for historical reasons.

and $\theta(t) : \Omega \to \Omega, t \in \mathbb{Z}$ *is a family of measure-preserving transformations such that*

1. $\theta(0) = id, \theta(s) \circ \theta(t) = \theta(s+t)$ *for every* $s, t \in \mathbb{Z}$.

2. *The mapping* $(t, \omega) \to \theta(t)\omega$ *is measurable.*

3. $\theta(t)\mathbb{P} = \mathbb{P}$ *for every* $t \in \mathbb{Z}$.

The set of the map $\theta(t)$ forms a commutative group and preserves the measure $\mathbb{P}$. Distinctly different from physicists' notion of dynamics as "step-by-step" motion, in stochastic process, the space Ω contains all the possible paths, and $(\Omega, \mathcal{F}, \mathbb{P}, \theta(t))$ is a stationary process. This two-sided discrete-time dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, \theta(t)), t \in \mathbb{Z}$ is also known as *base flow* of random dynamical system. In many applications, the base flow is usually ergodic. If the property 3 is not fulfilled, then $(\Omega, \mathcal{F}, \mathbb{P}, \theta(t))$ is called *measurable dynamical system*. Non-stationary dynamics belong to the latter, as illustrated next.

Definition 2.2. *A measurable random dynamical system (RDS) on the complete separate metric space* $(\mathscr{S}, d)$ *over a metric dynamical system* $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ *is a map with one-sided time,* $\mathbb{N} \times \Omega \times \mathscr{S} \to \mathscr{S} : (t, \omega, i) \to \varphi(t, \omega)i$*, with the following properties:*

1. *The map* $(t, \omega, i) \to \varphi(t, \omega)i$ *is* $\mathcal{B}(\mathbb{N}) \otimes \mathcal{F} \otimes \mathcal{B}(\mathscr{S}), \mathcal{B}(\mathscr{S})$*-measurable.*

2. *The map* $i \to \varphi(t, \omega)i$ *satisfies the cocycle property:*

$$\varphi(0, \omega) = id, \varphi(s+t, \omega) = \varphi(s, \theta(t)\omega) \circ \varphi(t, \omega) \tag{2.1}$$

for every $s, t \in \mathbb{N}$ *and* $\omega \in \Omega$.

From the definition, the RDS is driven by the base flow and for one particular noise realization ω, one can treat $i \to \varphi(t, \omega)i$ as a non-autonomous dynamical system, which defines one-point motion. The cocycle property is intuitively understood as follows: evolve some initial state i for s steps with particular noise realization ω and then go through t more steps with the same noise from the s steps mark; it gives the same result as evolving the

same initial state i for $t+s$ steps with the same noise realization ω. The map $\varphi(t,\omega)$ may not be invertible, so the RDS is defined one-sided in time. We call an RDS *ergodic* if there exists a probability measure $\boldsymbol{\pi}$ on $\mathscr{S}$, such that for any $i \in \mathscr{S}$, the law of the one-point motion $\varphi(t,\omega)i$ converges to $\boldsymbol{\pi}$. We don't assume this one-point motion is Markovian. It is possible to relax the metric dynamical system to measurable dynamical system, but the limiting behaviors of RDS will be unclear.

2.2 RDS with finite state space

In what follows, while all the mathematical definitions are general enough for a continuous state space $\mathscr{S}$, I will give explicit examples in terms of a finite state space $\mathscr{S} = \{1, 2, \cdots, n\}$. Here RDS with the finite state space is denoted as *finite RDS*. Any $\alpha \in \Gamma$ is called a *deterministic transformation* on $\mathscr{S}$. Note Γ is a monoid with the composition of transformations as the operation. If the finite state space $\mathscr{S}$ has n states, then there are n^n possible deterministic transformations. Therefore the cardinality $\|\Gamma\| = n^n$. The i.i.d RDS can be intuitively described as follows [15]: As a dynamics in the state space, the system starts initially with some state i_0 in $\mathscr{S}$; a map α_1 in Γ is chosen according to the probability measure Q and the system moves to the state $i_1 = \alpha_0(i_0)$ in step 1; again, independently of previous maps, another map α_1 is chosen according to the probability measure Q and the system moves to the state $i_2 = \alpha_1(i_1)$. The procedure repeats. The initial state i_0 can be a fixed state or an $\mathscr{S}$-valued random variable independent of all maps α_t. The stochastic process X_t is constructed by means of composition of independent random maps, $X_t = \alpha_{t-1} \circ \cdots \circ \alpha_0(i_0)$.

In terms of the language of the RDS defined in Sec. 2.1, each term has explicit expression. The space Ω is the full shift, $\Omega = \Gamma^{\mathbb{Z}}$, which is the set of all possible two sided infinitely long sequences of deterministic transformations.

$$\Omega = \Big\{\omega : (\cdots \alpha_{-1}, \alpha_0, \alpha_1, \alpha_2 \cdots, \alpha_k, \cdots) \Big| \alpha_k \in \Gamma \Big\} \tag{2.2}$$

The probability measure is the Bernoulli measure defined on the cylinder set,

$$\mathbb{P}\big([\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k]\big) = Q(\alpha_0)Q(\alpha_1)\ldots Q(\alpha_k) \tag{2.3}$$

From the construction, maps at different steps are chosen independently with the same probability measure Q.

The mappings $\theta(t)$ are the left Bernoulli shift for t elements, i.e, $\theta(t)\omega = \alpha_t$. Define the time-one mapping $\varphi(1,\omega) = \alpha_0$ which is the first element of the sequence of deterministic transformations. Then the map $\varphi(t,\omega)$ is the composition of i.i.d. random maps, $\alpha_{t-1}\circ\cdots\circ\alpha_0$. If it applies to an initial state i, it generates a one-point motion $X_t(\omega) = \varphi(t,\omega)i$. Now I construct finite i.i.d. RDS rigorously. Clearly X_t is a Markov chain (MC) and its transition probability is

$$\Pr(i,G) = Q(\alpha : \alpha(i) \in G) \tag{2.4}$$

for any $i \in \mathscr{S}$ and any measurable set $G \in \mathcal{B}(\mathscr{S})$.

Another example to generate random maps is via a Markov chain. Then the probability measure $\mathbb{P}$ is the Markov measure. The measure of a cylinder set is defined by

$$\mathbb{P}([\alpha_0, \alpha_1, \ldots, \alpha_k]) = \boldsymbol{\pi}_{\alpha_0} p_{\alpha_0\alpha_1} \cdots p_{\alpha_{k-1}\alpha_k} \tag{2.5}$$

where $p_{\alpha_i\alpha_j}$ is the transition probability of the Markov chain from the map α_i to α_j and $\boldsymbol{\pi}$ is the stationary probability of the Markov chain. One can check this Markov measure is still invariant with the Bernoulli shift map. However, the stochastic process induced $X_t = \varphi(t,\omega)i$ may not be a Markov chain in general. Even if one keeps as many steps of memory as possible, the Markov property may not hold any more. Here is the explicit example.

Example 2.1. Let the state space be $\mathscr{S} = \{1,2,3\}$ and the set of deterministic transformations Γ be

$$\Gamma = \left\{ \underbrace{\begin{pmatrix} 1\to 1 \\ 2\to 1 \\ 3\to 1 \end{pmatrix}}_{\alpha_1}, \underbrace{\begin{pmatrix} 1\to 3 \\ 2\to 1 \\ 3\to 1 \end{pmatrix}}_{\alpha_2}, \underbrace{\begin{pmatrix} 1\to 1 \\ 2\to 2 \\ 3\to 1 \end{pmatrix}}_{\alpha_3}, \underbrace{\begin{pmatrix} 1\to 2 \\ 2\to 1 \\ 3\to 1 \end{pmatrix}}_{\alpha_4} \right\}.$$

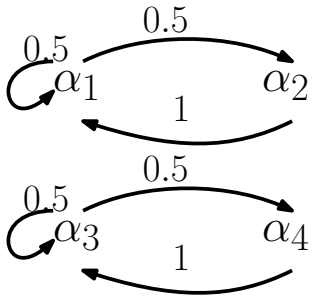

Figure 2.1: Random maps are generated via the Markov chain.

Define the transition probability between different maps as following:

$$P_{\alpha_1\to\alpha_1}=0.5,\ P_{\alpha_1\to\alpha_2}=0.5,\ P_{\alpha_2\to\alpha_1}=1,\ P_{\alpha_3\to\alpha_3}=0.5,\ P_{\alpha_3\to\alpha_4}=0.5,\ P_{\alpha_4\to\alpha_3}=1.$$

It is illustrated by the Figure. 2.1.

Choosing the initiate map with equal probability (0.5) to be α_1 or α_3, i.e, $\mathbf{p}_{\alpha_1}(0) = 0.5$ and $\mathbf{p}_{\alpha_3}(0) = 0.5$. Then possible map sequences can be $\alpha_{\mathbf{1}}\alpha_{\mathbf{1}}\alpha_{\mathbf{1}}\alpha_{\mathbf{2}}\alpha_{\mathbf{1}}\alpha_{\mathbf{1}}\alpha_{\mathbf{2}}\alpha_{\mathbf{1}}\cdots$ or $\alpha_{\mathbf{3}}\alpha_{\mathbf{4}}\alpha_{\mathbf{3}}\alpha_{\mathbf{3}}\alpha_{\mathbf{4}}\alpha_{\mathbf{3}}\alpha_{\mathbf{3}}\alpha_{\mathbf{3}}\cdots$. Consider the RDS trajectory starting from state $X_0 = 1$, then the corresponding state sequences are $111131131\cdots$ and $112112111\cdots$. Notice that 2 and 3 cannot appear in the same sequence. Assume it is a Markov chain with respect to the memory of previous t elements. Then we find

$$\begin{aligned} 0.5 &= \Pr\left\{X_{s+1}=2|X_{s-t}=2, X_{s-t+1}=\cdots=X_s=1\right\} \\ &= \Pr\left\{X_{s+1}=2|X_{s-t+1}=\cdots=X_s=1\right\} \\ &= \Pr\left\{X_{s+1}=2|X_{s-t}=3, X_{s-t+1}=\cdots=X_s=1\right\} \ = \ 0, \end{aligned}$$

which is a contradiction. Thus the stochastic process X_t is not a Markov chain with any finite length of memory.

It is also possible to generate random maps via an independent but not identical process. Then the measure is defined by

$$\mathbb{P}([\alpha_0,\alpha_1,\alpha_2,\ldots,\alpha_k]) = Q_0(\alpha_0)Q_1(\alpha_1)\ldots Q_k(\alpha_k) \tag{2.6}$$

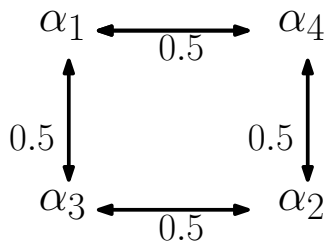

Figure 2.2: Random maps are generated via this Markov chain. This is the illustration of state transition diagram for the Markov chain.

where $Q_0, Q_1, \ldots$ might be different measures. Shift maps $\theta(t)$ in general doesn't preserve this measure and is no longer stationary. So it is only the measurable dynamical system. However, the stochastic process $X_t = \varphi(t, \omega)i$ is still well-defined and follows a time-inhomogeneous Markov chain with its transition probability at step t

$$\Pr_t(i, G) = Q_t(\alpha : \alpha(i) \in G) \tag{2.7}$$

Except for some special cases, different probability measures Q_t will result in different transition probability P_t.

From these three examples, it seems that independence of random maps may be crucial to the Markov property of the stochastic process X_t. It turns out that's not true. Here is the counter-example.

Example 2.2. Let the state space be $\mathscr{S} = \{1, 2\}$ and the set of deterministic transformations Γ be

$$\Gamma = \left\{ \underbrace{\begin{pmatrix} 1 \to 1 \\ 2 \to 2 \end{pmatrix}}_{\alpha_1}, \underbrace{\begin{pmatrix} 1 \to 2 \\ 2 \to 1 \end{pmatrix}}_{\alpha_2}, \underbrace{\begin{pmatrix} 1 \to 1 \\ 2 \to 1 \end{pmatrix}}_{\alpha_3}, \underbrace{\begin{pmatrix} 1 \to 2 \\ 2 \to 2 \end{pmatrix}}_{\alpha_4} \right\}.$$

Then a MC with the transition matrix

$$M = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$$

can be represented by i.i.d. RDS with probability $Q(\alpha_1) = 0.2, Q(\alpha_2) = 0.2, Q(\alpha_3) = 0.3, Q(\alpha_4) = 0.3$. It is possible to generate the random maps via a Markov chain, but the dynamics on the state space are still Markovian. If the initial distribution of the deterministic maps is $\mathbf{p}_{\alpha_1}(0) = \mathbf{p}_{\alpha_2}(0) = 0.5$ and the state transition diagram is illustrated in Fig. 2.2, then $\mathbf{p}_{\alpha_1}(t) = \mathbf{p}_{\alpha_2}(t), \mathbf{p}_{\alpha_3}(t) = \mathbf{p}_{\alpha_4}(t)$ at any steps. So the RDS induces a Markov chain in the state space with the transition matrix M. On the other hand, if I consider two-point motion $X_0 = 1$, $Y_0 = 2$ and apply the same sequence of maps, it is impossible to have $X_0 = 1, X_1 = 1, X_2 = 2$ and $Y_0 = 2, Y_1 = 2, Y_2 = 1$ since α_1 cannot go to α_2 in the single step.

For a given finite i.i.d. RDS, Eq. 2.4 uniquely defines an *induced Markov chain*. From stochastic dynamics perspective, there is no difference between an i.i.d. RDS and its induced MC if only a single sample trajectory is simulated. In other words, the one-point motion in this RDS is the Markov chain with transition probability $P(i, G)$. The difference between them comes out when one is interested in two simultaneous sample trajectories with different initial states (but same $\{\alpha_n\}$) since two sequences $\{X_n\}$ in the RDS are not independent, as was showed in Example 2.2. This is sometimes called "identical noise realizations" [23] or "two-point motion" [10]. The difference in the two theories has been described as "two cultures" in [6].

2.3 Linear Representation of Deterministic Maps

We first start with the linear representation of deterministic maps. An RDS on a n-dimensional vector space X is called *a linear RDS* if $\varphi(t, \omega) \in \mathcal{L}(X)$ for each $t \in \mathbb{N}, \omega \in \Omega$, where $\mathcal{L}(X)$ is the space of linear operators of X. If state i denotes the standard basis $\mathbf{e}_i$ in n-dimensional vector space $\mathbb{R}^n$, the deterministic transformation $\alpha \in \Gamma$ has a linear representation in the $n \times n$ matrices, called *deterministic transition matrix*.

$$(P)_{ij} \triangleq \begin{cases} 1, & j = \alpha(i), \\ 0, & \text{otherwise}, \end{cases} \qquad i, j \in \mathscr{S} \tag{2.8}$$

The dynamics of the map α applying on the state i is represented by the multiplication $\mathbf{e}_i P_\alpha$. Note that $\mathbf{e}_i$ is a row vector. Moreover, $\mathbf{e}_i$ can be considered as the probability concentrated on state i. Now P_α is a 0-1 matrix and has exactly one entry 1 in each row and 0s otherwise. So P_α is the representation of $\varphi(t,\omega)$ in the space of linear operators of $\mathbb{R}^n$. Among all deterministic transition matrices, there are $n!$ permutation matrices and they correspond to all invertible maps of the finite state space $\mathscr{S}$ to itself; all other $n^n - n!$ matrices correspond to non-invertible maps, which necessarily have at least one column of 0s. Some results on eigenvalues and singular values are discussed on Appendix. A.

Second, composition of transformations is represented by the matrix multiplication, i.e, $P_{\alpha_1} \cdot P_{\alpha_2} = P_{\alpha_1 \circ \alpha_2}$. In addition, this linear RDS $\varphi(t,\omega)$ has the form of random matrices production and it is easy to see the cocycle property (2.1). The stochastic process X_t starting from $X_0 = i$ is $X_t(\omega) = \varphi(t,\omega)i$, and its linear representation is

$$\mathbf{v}(t) = \mathbf{e}_i P_{\alpha_0} \cdot P_{\alpha_1} \cdots P_{\alpha_{t-1}} \tag{2.9}$$

It is defined in the push-forward sense. Define another stochastic process Y_t starting from $Y_0 = i$, $Y_t(\omega) = \varphi(t, \theta(-t)\omega)x_0$ and its linear representation is

$$\mathbf{u}(t) = \mathbf{e}_i P_{\alpha_{-t}} \cdot P_{\alpha_{-2}} \cdots P_{\alpha_{-1}} \tag{2.10}$$

$Y_t(\omega)$ is defined in the pullback sense. If the RDS is i.i.d. and ergodic, $X_t(\omega)$ follows a MC and $Y_t(\omega)$ has the same distribution as $X_t(\omega)$ for each t. But $X_t(\omega)$ and $Y_t(\omega)$ have different behaviors: $X_t(\omega)$ moves ergodically through the state space $\mathscr{S}$ along t; $Y_t(\omega)$ could converge to a limit as $t \to +\infty$. Similar idea was discussed in iterative random functions [33]. Although P_{α_i} is picked randomly for Y_t, it is multiplied on the left hand side which is the beginning of the matrices sequence, and the rest of the matrices remain the same. Roughly speaking, the random matrix multiplication may have the memory decay effect along the time and the last couple matrices which are fixed may determine the vector $\mathbf{u}(t)$. Here is an elementary example to illustrate the significant difference between a push-forward matrix multiplication and a pullback matrix multiplication: Consider 3×3 deterministic transition

matrices and their random products: If the matrix

$$P^* = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

is chosen, then the product of any deterministic transition matrix multiplied on the left of P^* will be invariant. This is not the case if a deterministic transition matrix is multiplied on the right of P^*:

$$P^* \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \; P^* \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The pullback product has a limit, while the push forward product does not.

Finially, from the definition of Perron-Frobenius-Ruelle operator (or transfer operator) for deterministic map α, $F : \mathbb{R}^n \to \mathbb{R}^n$, $(F\mathbf{v})_j = \sum_{i:\alpha(i)=j} \mathbf{v}_i, \mathbf{v} \in \mathbb{R}^n$. So P_α is the representation of Perron-Frobenius operator for the deterministic transformation α, and $\mathbf{v} \to \mathbf{v}P_\alpha$ can also be interpreted as the evolution of probability mass $\mathbf{v}$ corresponding to the mapping α. From the definition of Koopman operator for α, $K : \mathbb{R}^n \to \mathbb{R}^n$, $(K\mathbf{u})_j = \mathbf{u}_{\alpha(j)}, \mathbf{u} \in \mathbb{R}^n$. So P_α^T is the representation of Koopman operator for α. We introduce the stochastic Perron-Frobenius operator family and stochastic Koopman operator family associate with a finite RDS.

Definition 2.3. *The stochastic Perron-Frobenius operator $F_{s,t} : \mathbb{R}^n \to \mathbb{R}^n$ for every $0 \le s \le t$ associate to finite RDS φ is defined by*

$$(F_{s,t}\mathbf{v})_j \triangleq \mathbb{E}^{\mathbb{P}}\Big[\sum_{i:\varphi(t-s,\theta(s)\omega)i=j} \mathbf{v}_i\Big] \tag{2.11}$$

The stochastic Koopman operator $K_{s,t} : \mathbb{R}^n \to \mathbb{R}^n$ for every $0 \le s \le t$ associate finite RDS φ is defined by

$$(K_{s,t}\mathbf{u})_j \triangleq \mathbb{E}^{\mathbb{P}}\Big[\mathbf{u}_{\varphi(t-s,\theta(s)\omega)j}\Big] \tag{2.12}$$

The expectation is taken with respect to probability measure $\mathbb{P}$. We refer the family of operators $F_{s,t}, K_{s,t}$, parametrized by time s and t, as the stochastic Perron-Frobenius operator family and the stochastic Koopman operator family respectively.

If θ is the stationary process, $K_{s,t}$ and $F_{s,t}$ are also stationary, i.e, both family of operators can be expressed by the time difference, K_{t-s} and F_{t-s}. Furthermore, if the stochastic process X_t is a MC, then both family of operators form semigroups, i.e, one-parameter family of linear operators with the properties, $F_0 = id$ and $F_{t+s} = F_t \circ F_s$. Here the semigroups are characterized through their generators, F_1 and K_1, which are corresponding stochastic Perron-Frobenius operator and Koopman operator for time-one random map $\varphi(1, \omega)$. In terms of matrix representation (2.8), the generator M is represented by $M = \mathbb{E}^Q[P_\alpha]$, which is exactly the same as Markov transition matrix for X_t. The operator composition is also represented by matrix multiplication, so the stochastic Perron-Frobenius operator $F_t = M^t$. Then the stochastic Koopman operator is then represented by the adjoint of the matrix M. More importantly, this adjoint property is also true for general case.

Theorem 2.1. *For every* $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$,

$$\langle F_{s,t}\mathbf{v}, \mathbf{u}\rangle = \langle \mathbf{v}, K_{s,t}\mathbf{u}\rangle \tag{2.13}$$

where $\langle \mathbf{v}, \mathbf{u}\rangle = \mathbf{v}\mathbf{u}^*$.

Proof. We first check $\mathbf{v} = \mathbf{e}_i$ and $\mathbf{u} = \mathbf{e}_j$.

$$\langle F_{s,t}\mathbf{e}_i, \mathbf{e}_j\rangle = \mathbb{E}^{\mathbb{P}}\Big[\sum_{i:\varphi(t-s,\theta(s)\omega)i=j} \mathbf{e}_i\Big] = \mathbb{P}\Big[\omega : \varphi(t-s, \theta(s)\omega)i = j\Big] \tag{2.14}$$

$$\langle \mathbf{e}_i, K_{s,t}\mathbf{e}_j\rangle = \mathbb{E}^{\mathbb{P}}\Big[(\mathbf{e}_j)_{\varphi(t-s,\theta(s)\omega)i}\Big] = \mathbb{P}\Big[\omega : \varphi(t-s, \theta(s)\omega)i = j\Big] \tag{2.15}$$

Both operators are linear, so the adjoint property is true for any vector $\mathbf{v}, \mathbf{u} \in \mathbb{R}^n$. □

2.4 Finite i.i.d. RDS Representation of a MC

In the present work, I are interested in the reverse question: Can, and how, a Markov chain be represented by the compositions of i.i.d. random transformations? In the world

of stochastic modeling, this provides a more refined stochastic description of dynamics that is consistent with a Markov model. In more precise mathematical writings: Given the transition probability $P(i,\cdot)$, $i \in \mathscr{S}$, does there exist a probability measure on Γ, such that $Q\{\alpha : \alpha(i) \in G\} = P(i,G)$, for all $i \in \mathscr{S}$ and any measurable set $G \in \mathcal{B}(\mathscr{S})$? There is a proof of the "can" for a general continuous transformations, given in [17] and [60]. Quas also showed the sufficient conditions for the representation of a MC on a manifold by smooth maps [95]. Here I will particularly discuss the finite state situation.

In the finite state situation, it follows the Theorem 2.2, which is an analog of the Birkhoff-von Neumann theorem in the theory of doubly stochastic matrices [16] and [79].

Theorem 2.2. *The set of $n \times n$ Markov transition matrices forms a convex polyhedron with deterministic transition matrices as its vertices.*

The proof is based on a min-max algorithm. Here is a scratch of the algorithm: at each step the weight α is the minimum of maximum entry in each row of the matrix M and the corresponding deterministic matrix P is the index of maximum entry in each row; then redefine $M - \alpha P$ as M and keep this iteration until M becomes a zero matrix.

In more plain words, if matrix M is a Markov transition matrix, then there exists a probability measure Q on σ-field of Γ, such that

$$M = \mathbb{E}^Q[P_\alpha]. \tag{2.16}$$

This implies that there always exists at least one RDS representation for any finite Markov chain with the probability measure Q.

Such representation in general is not unique. This gives rise to the question of "how": Which representation is reasonable under some prior information or requirements. This will be answered in Sec. 3.2 Here is an example that illustrates the existence and proves non-uniqueness.

Example 2.3. If the state space $\mathscr{S} = \{1, 2\}$ and Markov transition probability matrix is

$$M = \begin{pmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}$$

It can be decomposed by min-max algorithm,

$$M = 0.6 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + 0.3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 0.1 \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}. \tag{2.17}$$

In the mean time, another decomposition could be

$$M = 0.18 \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} + 0.28 \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} + 0.42 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + 0.12 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.18}$$

The reason for non-uniqueness is that a transition probability only determines the statistical property of the one-point motion of a possible RDS, while an RDS also describes the simultaneous motion of two or more points. We will elaborate this in detail here.

Here I assume that the n-state MC X_t with the transition probability matrix $M_{ij} = \Pr\{X_{t+1} = j | X_t = i\}$ is irreducible and aperiodic. Then there exists a unique stationary distribution $\boldsymbol{\pi}$ and for any initial distribution $\mathbf{p}(0)$, the MC will converge to the stationary distribution, i.e, $\lim_{t\to+\infty} \mathbf{p}(t) = \lim_{t\to+\infty} \mathbf{p}(0)M^t = \boldsymbol{\pi}$.

For the entire path, a more rigorous construction of MC is to consider a measurable dynamical system $(\Omega', \mathcal{F}', \mathbb{P}', \theta(t))$, where $\Omega' = \mathscr{S}^{\mathbb{N}}$, the set of one-sided infinite long sequences of states, θ is again the shift map. The probability measure $\mathbb{P}'$ is defined on the cylinder set $[i_0, i_1, \ldots, i_t]$,

$$\mathbb{P}'([i_0, i_1, \ldots, i_t]) = \mathbf{p}_{i_0}(0) M_{i_0 i_1} \ldots M_{i_{t-1} i_t} \tag{2.19}$$

Since $\mathbf{p}(0)$ may be not necessarily its stationary distribution, $\mathbb{P}'$ is not θ-invariant. But marginalizing states at previous steps $0, 1, \ldots, t-1$, it gives the probability at step t, i.e, $\sum_{i_0,\ldots,i_{t-1}} \mathbb{P}'([i_0, i_1, \ldots, i_t]) = \mathbf{p}_{i_t}(t)$, where $\mathbf{p}(t) = \mathbf{p}(0)M^t$. The stochastic process X_t is defined as $X_t(\omega) \triangleq \theta(t)\omega = \omega_t$.

It would be interesting to discuss the relationship between measurable dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, \theta(t))$ in finite i.i.d RDS and measurable dynamical system $(\Omega', \mathcal{F}', \mathbb{P}', \theta(t))$ in its induced MC. Each element $\omega_1 = \cdots \alpha_0 \alpha_1 \alpha_2 \cdots \alpha_t \cdots$ in the product space Ω is a sequence of deterministic transformations. If this sequence is applied to n different initial conditions,

it simultaneously produces n different sequences in the product space Ω'. At the same time, multiple elements in Ω apply on the same initial conditions may have exactly the same sequence of states. So this provides a more refined stochastic description of dynamics that is consistent with the MC. This also implies that knowing one sample trajectory in the MC may not be enough to fully determine which transformation was picked in the i.i.d RDS view. We use "may" because in some special situations, this can actually be uniquely determined. We shall call such RDS's having *no common dynamics*.

Another question arises is in how many ways can a given Markov transition probability matrix be expressed in the form of (2.16). In fact, one can ask the following combinatorial question: In the representations of a Markov transition matrix in the form (2.16), what is the least possible number of deterministic transition matrices, $\kappa_*(M)$? Some results of upper bounds for $\kappa_*(M)$ can be obtained. Unlike a stochastic differential equation which has a well-defined deterministic counterpart, a Markov chain doesn't have an unambiguous deterministic reference. The "least", therefore, could be interpreted as the "closest to a deterministic dynamics".

The following theorem gives an upper bound estimation.

Theorem 2.3. *If M is $n \times n$ Markov transition matrix, then*

$$\kappa_*(M) \leq n^2 - n + 1. \tag{2.20}$$

Proof. There are n linear conditions on the row sums of a $n \times n$ Markov transition matrix. Therefore, the dimension is $n^2 - n$. Carathodory theorem shows that every Markov transition matrix is in a convex hull of $n^2 - n + 1$ deterministic transition matrices. Then $\kappa_*(\mathbf{M}) \leq n^2 - n + 1$. □

Notice that in Example 2.3, this upper bound has been reached. For irreducible periodic Markov transition matrices, the upper bound can be further improved by using the period of the matrix [77].

Chapter 3

ENTROPY AND ENTROPY PRODUCTION OF STOCHASTIC DYNAMICS

Entropy and entropy production are two distinctly different key concepts originated in thermodynamics, the study of Newtonian particles in terms of their stochastic motions — called heat. In physicists' theory, entropy is a function of the state of a system. Entropy production, however, is associated with the amount of heat being generated in a process; it is path dependent in general. In fact, the physicists of the earlier time carefully introduced the notations of $\mathrm{d}A$ and $đQ$, where $\mathrm{d}A$ represents a change in a state function A that is path independent, and $đQ$ is associated with the accumulation of heat Q (or work $đW$) that is a function of a path. The celebrated First Law of Thermodynamics states that $đQ + đW = \mathrm{d}E$, where E is called internal energy. We see that if the work $đW = \mathbf{F} \cdot \mathrm{d}\mathbf{x}$ is due to a force $\mathbf{F}$ with a potential, $\mathbf{F} = -\nabla U(\mathbf{x})$, then $đQ = \mathrm{d}\big(E + U\big)$. In terms of the nonlinear stochastic dynamics, therefore, entropy, as a state function, should be a functional of the instantaneous probability distribution $\mathbf{p}(t)$, but entropy production is associated with the transition probability.

The entropy and entropy production introduced above can be rigorously established in the theory of Markov chains. The notions of instantaneous Gibbs entropy, metric entropy [118], and entropy production rate [51] are three distinct concepts, each represents a different aspect of the same stochastic dynamics. In the present work, I am interested in these concepts and the connections with the representation of finite i.i.d. RDS. From now on, I assume the MC is irreducible and aperiodic, unless otherwise stated.

3.1 Overview of Shannon Entropy and Relative Entropy

The Shannon entropy for the probability measure $\mathbf{p}$ is the expectation of the information content,

$$S(\mathbf{p}) \triangleq \mathbb{E}^{\mathbf{p}}[-\log(\mathbf{p}(\omega))]. \tag{3.1}$$

One also can consider the relative entropy or Kullback-Leibler divergence of $\mathbf{p}$ with respect to $\boldsymbol{\mu}$, $H(\mathbf{p}, \boldsymbol{\mu})$,

$$H(\mathbf{p}, \boldsymbol{\mu}) \triangleq \begin{cases} \mathbb{E}^{\mathbf{p}}\left[\log\left(\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}\boldsymbol{\mu}}(\omega)\right)\right] & \mathbf{p} \ll \boldsymbol{\mu} \\ +\infty & \text{Otherwise} \end{cases} \tag{3.2}$$

From the definition, $-\log(\mathbf{p}(\omega))$ and $\log\left(\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}\boldsymbol{\mu}}(\omega)\right)$ are random variables on the probability space $(\Omega, \mathcal{F}, p)$. But there is no natural reference measure μ to define for the relative entropy here. The random variable $-\log(\mathbf{p}(\omega))$ is very special because it directly uses the probability measure to reflect the probability space. But in applied science, it is not always (in fact mostly not) possible to know the explicit expression for the probability space. In most cases, I only have the observable of the probability space and it could be a biased one, which is the random variable $X(\omega)$. But the information content $-\log \Pr(X(\omega))$ gives the best estimate of $-\log \mathbf{p}(\omega)$. In particular, if the map $\omega \to X(\omega)$ is one-to-one, then the Shannon entropy of this random variable is the same as the Shannon entropy of the probability space, i.e, $S(\Pr(X)) = -\sum_{X(\omega)} \Pr(X(\omega)) \log \Pr(X(\omega)) = -\sum_{\omega} \mathbf{p}(\omega) \log \mathbf{p}(\omega) = S(\mathbf{p})$. If the map is surjective, then $S(\Pr(X)) \leq S(\mathbf{p})$ due to the convexity of the log function.

In the MC, I specify the probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ for the stochastic process in Sec. 2.4. The random element $\omega' \in \Omega'$ is the infinite long sequence of the MC. The probability space is equipped with the filtration $\{\mathcal{F}'_t\}_{t\geq 0}$ with $\mathcal{F}'_t \subseteq \mathcal{F}'$ where t is non-negative and $t_1 \leq t_2 \to \mathcal{F}'_{t_1} \subseteq \mathcal{F}_{t_2}$. The process X_t is called to be adapted to the filtration $\{\mathcal{F}'_t\}_{t\geq 0}$ if the random variable $X_t : \Omega' \to \mathscr{S}$ is a $(\mathcal{F}'_t, \mathcal{B}(\mathscr{S}))$-measurable function for each $t \geq 0$.

For the MC, applying (3.1) to the finite time distribution of MC restricted to σ-field ${\mathcal{F}'}_0^t$,

the Shannon entropy $S(\mathbb{P}'_{[0,t]})$, where $\mathcal{F}'^{t}_{0} = \sigma(X_s : 0 \le s \le t)$ [58]

$$S(\mathbb{P}'_{[0,t]}) = -\sum_{i_0,\dots,i_t} \mathbf{p}_{i_0}(0) M_{i_0 i_1} \dots M_{i_{t-1} i_t} \log\Big(\mathbf{p}_{i_0}(0) M_{i_0 i_1} \dots M_{i_{t-1} i_t}\Big) \tag{3.3}$$

So the entropy $S(\mathbb{P}')$ for the probability space, can be calculated as the time goes to $+\infty$,

$$S(\mathbb{P}') = \lim_{t\to+\infty} S(\mathbb{P}'_{[0,t]}) \tag{3.4}$$

But the problem is that this quantity $S(\mathbb{P}')$ is $+\infty$.

Then physicists use the random variable X_t, the current state, as the observable for the probability space. Then the entropy of the instantaneous distribution $\mathbf{p}(t)$ is $S(\mathbf{p}(t)) = -\sum_i \mathbf{p}_i(t) \log \mathbf{p}_i(t)$ which is finite. This quantity $S(\mathbf{p}(t))$ doesn't have monotonicity as $t \to \infty$, instead, the relative entropy $H\big(\mathbf{p}(t), \boldsymbol{\pi}\big)$ has. Unlike previous case, this unique invariant measure $\boldsymbol{\pi}$ is the most important one in the MC so this relative entropy becomes a natural choice. Then I have [94, 117]:

Theorem 3.1. *$H\big(\mathbf{p}(t), \boldsymbol{\pi}\big)$ is a non-increasing function of t.*

Proof. For $t \ge 1$,

$$\begin{aligned}
\Delta H\big(\mathbf{p}(t-1), \boldsymbol{\pi}\big) &= H\big(\mathbf{p}(t), \boldsymbol{\pi}\big) - H\big(\mathbf{p}(t-1), \boldsymbol{\pi}\big) \\
&= \sum_{i\in\mathscr{S}} \mathbf{p}_i(t) \log\left(\frac{\mathbf{p}_i(t)}{\boldsymbol{\pi}_i}\right) - \sum_{i\in\mathscr{S}} \mathbf{p}_i(t-1) \log\left(\frac{\mathbf{p}_i(t-1)}{\boldsymbol{\pi}_i}\right) \\
&= \sum_{i,j\in\mathscr{S}} \left[\mathbf{p}_j(t-1) M_{ji} \log\left(\frac{\mathbf{p}_i(t)}{\boldsymbol{\pi}_i}\right) - \mathbf{p}_i(t-1) M_{ij} \log\left(\frac{\mathbf{p}_i(t-1)}{\boldsymbol{\pi}_i}\right)\right] \\
&= \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t-1) M_{ij} \log\left(\frac{\boldsymbol{\pi}_i \mathbf{p}_j(t)}{\mathbf{p}_i(t-1)\boldsymbol{\pi}_j}\right) \\
&\le \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t-1) M_{ij} \left(\frac{\boldsymbol{\pi}_i \mathbf{p}_j(t)}{\mathbf{p}_i(t-1)\boldsymbol{\pi}_j} - 1\right) \\
&= \sum_{i,j\in\mathscr{S}} \frac{\boldsymbol{\pi}_i M_{ij} \mathbf{p}_j(t)}{\boldsymbol{\pi}_j} - \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t-1) M_{ij} \;=\; 1 - 1 \;=\; 0.
\end{aligned}$$

□

Here $\Delta H\big(\mathbf{p}(t-1), \boldsymbol{\pi}\big)$ is defined as the change of the relative entropy w.r.t the stationary distribution. It signifies the non-stationarity of the MC.

For a doubly stochastic MC, the invariant distribution is the uniform distribution, $\boldsymbol{\pi}_i = 1/n$, $S(\mathbf{p}(t)) = \log(n) - H\big(\mathbf{p}(t), \boldsymbol{\pi}\big)$. Therefore, the above theorem becomes the statement "entropy never decreases". This setting is known as *microcanonical system* in statistical physics.

For any MC, the change of Shannon entropy of the instantaneous distribution $\Delta S(\mathbf{p}(t)) \triangleq S(\mathbf{p}(t+1)) - S(\mathbf{p}(t))$ satisfies

$$\begin{aligned} \Delta S(\mathbf{p}(t)) &= \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t) M_{ij} \log\left(\frac{\mathbf{p}_i(t)}{\mathbf{p}_j(t+1)}\right) \\ &= \underbrace{\sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t) M_{ij} \log\left(\frac{\mathbf{p}_i(t) M_{ij}}{\mathbf{p}_j(t+1) M_{ji}}\right)}_{\text{non-negative}} + \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(t) M_{ij} \log\left(\frac{M_{ji}}{M_{ij}}\right). \end{aligned} \tag{3.5}$$

If the MC is detailed balance, $\boldsymbol{\pi}_i M_{ij} = \boldsymbol{\pi}_j M_{ji}$, then the second term on the right-hand-side can be expressed as $\Delta\overline{E}(\mathbf{p}(t)) = \overline{E}(\mathbf{p}(t+1)) - \overline{E}(\mathbf{p}(t))$, which is defined as

$$\overline{E}(\mathbf{p}(t)) = \sum_{i\in\mathscr{S}} \mathbf{p}_i(t)\Big(-\log \boldsymbol{\pi}_i\Big). \tag{3.6}$$

Note that ΔS and $\Delta\overline{E}$ are changes in functions of state, $S(t)$ and $\overline{E}(t)$. This setting is known as Gibbsian *canonical system* in statistical physics. $\overline{E}$ should be identified with the internal energy; and the relationship (3.6) between internal energy and equilibrium measure is known as the *Boltzmann distribution*.

In statistical physics, S is called Gibbs entropy. Then $\overline{E} - S$ should be identified with the notion of *free energy* there. The Theorem 3.1 thus becomes "free energy of a canonical system never increases; it reaches its maximum when a system is at its equilibrium".

The identification of the non-negative term in (3.5) with the concept of *entropy production rate* in nonequilibrium thermodynamics appeared repeatedly in physics and chemistry literature, see [13, 30, 52, 103]. For MC without detailed balance, the last term in (3.5) cannot be expressed as the change of a state function, but it can be identified with *heat exchange*

rate. Then in the stationary state, when $\Delta S = 0$, there is positive entropy production rate that is balanced with the heat dissipation. Such a state is called a *nonequilibrium steady state* [51].

On the other hand, mathematicians study how the Shannon entropy $S(\mathbb{P}'_{[0,t]})$ restricted to σ-field $\mathcal{F}'^{t}_{0}$ increases with time. Define the Shannon entropy of a step as $đS(\mathbb{P}'_{[0,t]}) \triangleq S(\mathbb{P}'_{[0,t+1]}) - S(\mathbb{P}'_{[0,t]})$. So $đS(\mathbb{P}'_{[0,t]}) = -\sum_{i,j} \mathbf{p}_i(t) M_{ij} \log M_{ij}$ is the Shannon entropy generated at step t. Then the asymptotic limit of $đS(\mathbb{P}'_{[0,t]})$ is $\lim_{t\to+\infty} đS(\mathbb{P}'_{[0,t]}) = -\sum_{i,j} \boldsymbol{\pi}_i M_{ij} \log M_{ij}$. In the meantime, this limit also quantifies the average randomness generated per step in an MC.

$$\lim_{t\to+\infty} \frac{S(\mathbb{P}'_{[0,t]})}{t} = -\sum_{i,j} \boldsymbol{\pi}_i M_{ij} \log M_{ij}. \tag{3.7}$$

This limit is called the *metric entropy* of the MC.

The concept of metric entropy itself was first introduced by Kolmogorov and further improved by Sinai [36, 85, 118]. It has been very successfully used in solving the isomorphism, or conjugacy, of dynamical systems. The metric entropy measures the maximal rate of information production a system is capable of generating [57]. It is a well-developed subject but can be technical in nature. Some mathematical derivations, in a heuristic fashion, can be found in Appendix B.

The Shannon entropy of the instantaneous distribution $S(\mathbf{p}(t))$ and the metric entropy h_{MC} are two different classes of stochastic quantities since the first one is the Shannon entropy and the second one is the Shannon entropy production. What should be compared are the Shannon entropy of a step $đS(\mathbb{P}'_{[0,t]})$ and the difference of Shannon entropy of the instantaneous distribution $\Delta S(\mathbf{p}(t))$ because both quantify the entropy produced at time t. However, the current state X_t as the observable of the probability space, without the information of past history, is biased. In fact, the family of probability distributions of current state $\mathbf{p}(t)$ itself doesn't have the information of the probability of the sequence, i.e, the marginal probability $\mathbf{p}(t)$ cannot uniquely determine the joint probability. So simply knowing $\mathbf{p}(t)$ doesn't guarantee the process is Markovian, even though there exists a matrix

M which fulfills the requirement of transition matrix, such that $\mathbf{p}(t)M = \mathbf{p}(t+1)$ for all $t \geq 0$. On the other hand, if one uses the the sequence up to t as the random variable, then it is one-to-one map from the probability space restricted to the σ-field $\mathcal{F}'^{t}_{0}$. It can accurately reflect the Shannon entropy generated at each step.

A better analogy is: starting with many indistinguishable particles, they follow some unknown stochastic dynamics; taking the snapshots at time 0, time 1 and etc, they give the instantaneous distributions. Then color these particles with different colors, rerun the dynamics and take the snapshots at time 0, time 1 and etc again. This time not only the instantaneous distributions are known, but also the past histories of each particle at the current time. Not surprisingly, the second picture generates more randomness each step than the first picture.

Proposition 3.1.

$$\Delta S(\mathbf{p}(t)) \leq đS(\mathbb{P}'_{[0,t]})$$

The equality is reached when the dynamic is the deterministic transformation.

Proof. We know $\sum_k \mathbf{p}_k(t)M_{kj} = \mathbf{p}_j(t+1) \geq \mathbf{p}_i(t)M_{ij}$ for all i. The equality is reached when $M_{ij} = 1$ for all i. So $\frac{\mathbf{p}_i(t)}{\mathbf{p}_j(t+1)} \leq \frac{1}{M_{ij}}$ if $\mathbf{p}_j(t+1) > 0$.

$$\Delta S(\mathbf{p}(t)) = \sum_{i,j \in \mathscr{S}} \mathbf{p}_i(t)M_{ij} \log\left(\frac{\mathbf{p}_i(t)}{\mathbf{p}_j(t+1)}\right) \leq \mathbf{p}_i(t)M_{ij} \log\left(\frac{1}{M_{ij}}\right) = đS(\mathbb{P}'_{[0,t]})$$

□

3.2 Metric Entropy of MC and Its i.i.d RDS Representations

In the finite i.i.d RDS, there is a i.i.d process to choose the deterministic transformation at each step. This i.i.d process is isomorphic to the 1-sided $(a_1, a_2, \ldots, a_N)$ Bernoulli shift, so the entropy for this i.i.d process is naturally defined as $h = -\sum_{i=1}^{N} a_i \log(a_i)$. Moreover, the deterministic transformation with finite state has zero metric entropy once it is chosen.

Therefore, the randomness of this system is solely generated by the i.i.d process. So the metric entropy for this RDS is $h_{\mathrm{RDS}} = -\sum_{\alpha\in\Gamma} Q(\alpha)\log Q(\alpha)$.

It is worth mentioning that the metric entropy of the RDS in many cases can be infinite. This is mainly because when there are countably infinite transformations in Γ, the infinite sum in the entropy may not converge. A different notion of entropy of i.i.d. RDS that remedies the difficulty is defined as the weighted mean of the metric entropy for all deterministic transformations with probability mass as their weight [60]. We do not need to be concerned with this since our $\mathscr{S}$ is always finite.

We are now in the position to address the question that for a given MC, which RDS representation is reasonable according to certain requirement. In fact, the metric entropy of its corresponding RDS becomes a good characterization for different representations. It is natural to ask the lower bound and upper bound of them for a given MC.

3.2.1 Lower Bound of Metric Entropy

Kifer notified this question and gave the result that $h_{\mathrm{RDS}} \geq h_{\mathrm{MC}}$ [60]. In the finite i.i.d RDS, different sequences of deterministic transformations applies on a same initial condition might induce the same MC trajectory, as I discussed in Sec. 2.4. That means more information is required to determine which deterministic transformation is chosen at each step. In other words, the RDS generates more information than the MC at each step. With this intuition, Kifer's result becomes clear; a general proof can be found in his book [60]. We will provide an elementary proof here for the finite i.i.d RDS and illustrate the condition under which the equality is attained.

Denote a deterministic transition matrix P as $P_{i_1,i_2,\ldots,i_n}$, if the corresponding map $1 \to i_1, 2 \to i_2, \ldots, n \to i_n$ which is denoted as $\alpha_{i_1,i_2,\ldots,i_n}$.

Definition 3.1. *The deterministic transformations α_1 and α_2 have common dynamics if there exists a state s in the space $\mathscr{S}$ such that $\alpha(s) = \alpha_2(s)$.*

By the representation of deterministic transition matrices $P_{i_1,i_2,\ldots,i_n}$ and $P_{j_1,j_2,\ldots,j_n}$, the

definition is equivalent with $P_{i_1,i_2,\ldots,i_n} - P_{j_1,j_2,\ldots,j_n}$ has at least one row being all zeroes. In our notation, that is, $\exists k$ such that $i_k = j_k$. If two deterministic transformations have no common dynamics, then $P_{i_1,i_2,\ldots,i_n} - P_{j_1,j_2,\ldots,j_n}$ has no rows that being all zeroes. This definition can be extended to multiple deterministic transformations. If no pair of two deterministic transformations has common dynamics, I call all these deterministic transformations have no common dynamics.

Theorem 3.2. *In the finite state space $\mathscr{S}$, $h_{RDS} \geq h_{MC}$. The equality is reached if and only if deterministic transformations with positive probability have no common dynamics.*

Proof. The metric entropy of the Markov chain is

$$\begin{aligned}
h_{\mathrm{MC}} &= -\sum_{j,k=1}^{n} \pi_j M_{jk} \log M_{jk} \\
&= -\sum_{j,k=1}^{n} \pi_j \Big(\sum_{i_j=k} Q(\alpha_{i_1,i_2,\ldots,i_n}) \Big) \log \Big(\sum_{i_j=k} Q(\alpha_{i_1,i_2,\ldots,i_n}) \Big) \\
&\leq -\sum_{j,k=1}^{n} \pi_j \sum_{i_j=k} \big(Q(\alpha_{i_1,i_2,\ldots,i_n}) \log Q(\alpha_{i_1,i_2,\ldots,i_n}) \big) \\
&= -\sum_{i_1,i_2,\ldots,i_n} \big(Q(\alpha_{i_1,i_2,\ldots,i_n}) \log Q(\alpha_{i_1,i_2,\ldots,i_n}) \big) \sum_{j} \pi_j \\
&= -\sum_{i_1,i_2,\ldots,i_n} Q(\alpha_{i_1,i_2,\ldots,i_n}) \log Q(\alpha_{i_1,i_2,\ldots,i_n}) \;=\; h_{\mathrm{RDS}}.
\end{aligned}$$

The equality holds if M_{jk} is positive, among all deterministic transition matrices satisfying $i_j = k$, only one has positive probability mass and others are 0. This positive probability of course is M_{jk}. Therefore, all deterministic transformations with positive probability have no common dynamics. □

In the case that the equality is attained, and when simulating a trajectory in the state space $\mathscr{S}$ with such RDS, the deterministic transformation is uniquely identified in each step. The following corollaries illustrate more about the RDS and its induced Markov chain as the equality attained.

Corollary 3.1. *If $h_{RDS} = h_{MC}$, the cardinality of the deterministic matrices with positive probability is no greater than n, i.e, $\|\Gamma\| \leq n$*

Corollary 3.2. *If $h_{RDS} = h_{MC}$, each row of the induced Markov transition matrix is a permutation of the first row.*

In fact, $h_{\mathrm{RDS}} = h_{\mathrm{MC}}$ if and only if I can recover the RDS map sequence from the MC trajectory. For a given Markov transition probability matrix, it is not necessarily true that there exists an RDS representation whose metric entropy is the same as that of the MC. In fact, for most Markov transition matrices, it is not reachable. If exists, such RDS might not be unique as well. Nevertheless, it provides a possible lower bound of the metric entropy. In order to give a reachable lower bound, it usually need to solve a non-convex problem and no desirable solution is available.

3.2.2 Upper Bound of Metric Entropy

It is natural to ask what is the upper bound of the metric entropy of all RDS representations given a finite MC. Such representation is uniquely attainable since a strictly concave function over a convex hull has the unique maximum [98]. It turns out I can find an analytic expression for the decomposition.

Lemma 3.3. $\sum_{i_1,i_2,\ldots,i_n=1}^{n} M_{1i_1} M_{2i_2} \ldots M_{ni_n} = 1.$

So there exists a probability measure Q, that is, $Q(\alpha_{t_1,t_2,\ldots,t_n}) = M_{1t_1} M_{2t_2} \ldots M_{nt_n}$.

Theorem 3.4. *The metric entropy $h_{RDS} \leq -\sum_{j,k} M_{kj} \log M_{kj}$ and the equality holds if and only if $Q(\alpha_{i_1,i_2,\ldots,i_n}) = \Pi_k M_{ki_k}$.*

Proof. We can show

$$\begin{aligned} -\sum_{k,j} M_{kj}\log M_{kj} - h_{\mathrm{RDS}} &= -\sum_{k,j}(\sum_{i_k=j} Q(\alpha_{i_1,i_2,\ldots,i_n}))\log M_{kj} - h_{\mathrm{RDS}} \\ &= -\sum_{i_1,i_2,\ldots,i_n} Q(\alpha_{i_1,i_2,\ldots,i_n})\log(\Pi_k M_{ki_k}) - h_{\mathrm{RDS}} \\ &= \sum_{i_1,i_2,\ldots,i_n} Q(\alpha_{i_1,i_2,\ldots,i_n})\log\frac{Q(\alpha_{i_1,i_2,\ldots,i_n})}{\Pi_k M_{ki_k}} \\ &\geq 0 \end{aligned}$$

The Gibbs inequality is applicable since $Q(\alpha_{i_1,i_2,\ldots,i_n})$ and $\Pi_k M_{ki_k}$ are probability mass functions. The equal sign holds if and only if these two functions are identical, i.e, $Q(\alpha_{i_1,i_2,\ldots,i_n}) = \Pi_k M_{ki_k}$. □

The example in (2.18) is the maximum entropy representation for the given Markov transition matrix M. It is easy to see if all entries of the transition matrix M are positive, then all n^n deterministic matrices will have positive probability. Furthermore, the most probable deterministic transition matrix corresponds to the deterministic transformation that maps to the state with the largest probability given the current state, i.e, its deterministic matrix has entry one in the position that is maximum in each row of the transition matrix M. It is a very insightful result since this is the most reasonable "deterministic counterpart" for a given MC with transition probability matrix M.

3.3 Entropy Production Rate of MC and Its i.i.d RDS Representations

The entropy production rate of MC is another quantity that measures the dynamical asymmetry w.r.t. its time reversal. It is in the same class as the metric entropy which both reflect the entropy production.

3.3.1 Entropy Production Rate of MC

From now on, I consider the transition probability matrix M satisfies the condition $M_{ij} > 0 \leftrightarrow M_{ji} > 0$ for any $i, j \in \mathscr{S}$, then it is possible to define the relative entropy of the

distribution of the process with respect to its time reversal restricted to σ-field ${\mathcal{F}'}_0^t$. The time-reversed process X^- is defined as follows,

$$X_s^-(\omega) = X_{t-s}(\omega),\ \forall s \in [0,t] \tag{3.8}$$

So X^- is ${\mathcal{F}'}_0^t$ measurable. The time-reversed process is also called *adjoint process* of the MC. For the sample sequence $i_0, i_1, \ldots, i_t$, the time-reversed process gives $i_t, i_{t-1}, \ldots, i_0$. Define $\mathbb{P}'^-_{[0,t]}$ as the probability measure for the time-reversed process $X_s^-(\omega)$. The probability measure for the time-reversed process $\mathbb{P}'^-$ on this cylinder set is

$$\mathbb{P}'^-([i_0, i_1, \ldots, i_t]) = \mathbf{p}_{i_t}(0) M_{i_t i_{t-1}} \ldots M_{i_1 i_0} \tag{3.9}$$

Here I assume $\mathbf{p}_i(0) > 0$ for all i.

Proposition 3.2. *The time-reversed process of a Markov chain $(M, \boldsymbol{\pi})$ is Markovian. Moreover, the transition matrix of the time-reversed process is*

$$M_{ij}^- = \frac{\mathbf{p}_j(0) M_{ji}}{\mathbf{p}_i(0)},\ i, j \in \mathscr{S} \tag{3.10}$$

Proof. The Markovian property means $P(X_{t+1}^- = i_{t+1} | {\mathcal{F}'}_0^t) = P(X_{t+1}^- = i_{t+1} | X_t^-)$ for any $t \geq 0$. The left-hand-side is

$$\Pr(X_{t+1}^- = i_{t+1} | {\mathcal{F}'}_0^t) = \frac{\mathbb{P}'^-([i_0, \ldots, i_t, i_{t+1}])}{\mathbb{P}'^-([i_0, \ldots, i_t])} = \frac{\mathbf{p}_{i_{t+1}}(0) M_{i_{t+1} i_t}}{\mathbf{p}_{i_t}(0)}$$

The right-hand-side is

$$\Pr(X_{t+1}^- = i_{t+1} | X_t^-) = \frac{\mathbb{P}'^-([i_t, i_{t+1}])}{\mathbb{P}'^-([i_t])} = \frac{\mathbf{p}_{i_{t+1}}(0) M_{i_{t+1} i_t}}{\mathbf{p}_{i_t}(0)}$$

Moreover

$$\begin{aligned}
\mathbb{P}'^-([i_0, i_1, \ldots, i_t]) &= \mathbf{p}_{i_t}(0) M_{i_t i_{t-1}} \ldots M_{i_1 i_0} \\
&= \mathbf{p}_{i_0}(0) \underbrace{\left(\frac{\mathbf{p}_{i_1}(0) M_{i_1 i_0}}{\mathbf{p}_{i_0}(0)} \right)}_{M_{i_0 i_1}^-} \cdots \underbrace{\left(\frac{\mathbf{p}_{i_t}(0) M_{i_t i_{t-1}}}{\mathbf{p}_{i_{t-1}(0)}} \right)}_{M_{i_{t-1} i_t}^-}
\end{aligned}$$

the probability measure can be rewritten as the Markov measure with the transition matrix $M_{ij}^- = \frac{\mathbf{p}_j(0) M_{ji}}{\mathbf{p}_i(0)}$. □

In particular, for the stationary MC, the transition matrix for the time-reversed process is $M^-_{ij} = \frac{\boldsymbol{\pi}_j M_{ji}}{\boldsymbol{\pi}_i}$. The stationary distribution of the time-reversed process is also $\boldsymbol{\pi}$. Moreover, this time-reversed process has the same metric entropy as the original MC.

Since $\mathbf{p}_i(0) > 0$, $\mathbb{P}'_{[0,t]}$ is absolutely continuously with respect to $\mathbb{P}'^-_{[0,t]}$. Then the relative entropy of the measure of Markov chain w.r.t. the measure of time-reversed process, $H(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]})$ is

$$H\left(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]}\right) = \sum_{i_0,\ldots,i_t \in \mathscr{S}} \mathbf{p}_{i_0}(0) M_{i_0 i_1} \ldots M_{i_{t-1} i_t} \log\left(\frac{\mathbf{p}_{i_0}(0) M_{i_0 i_1} \ldots M_{i_{t-1} i_t}}{\mathbf{p}_{i_t}(0) M_{i_t i_{t-1}} \ldots M_{i_1 i_0}}\right) \tag{3.11}$$

Similarly, it relates to the stationary entropy production rate of MC via

$$e_p = \lim_{t\to+\infty} \frac{1}{t} H\left(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]}\right) \tag{3.12}$$

So the e_p is intuitively understood as the asymptotic average of entropy produced per step with respect to its time-reversed probability. It is also a property for a stationary MC. It has the following explicit expression [51].

Theorem 3.5.

$$e_p = \sum_{i,j\in\mathscr{S}} \boldsymbol{\pi}_i M_{ij} \log\left(\frac{M_{ij}}{M_{ji}}\right) \tag{3.13}$$

Proof. From the Definition 3.12

$$\begin{aligned} e_p &= \lim_{t\to+\infty} \frac{1}{t} H\left(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]}\right) \\ &= \lim_{t\to+\infty} \frac{1}{t} \left\{ \sum_{i\in\mathscr{S}} \big(\mathbf{p}_i(0) - \mathbf{p}_i(t)\big) \log \mathbf{p}_i(0) + \sum_{s=0}^{t-1} \sum_{i,j\in\mathscr{S}} \mathbf{p}_i(s) M_{ij} \log\left(\frac{M_{ij}}{M_{ji}}\right) \right\} \\ &= \sum_{i,j\in\mathscr{S}} \boldsymbol{\pi}_i M_{ij} \log\left(\frac{M_{ij}}{M_{ji}}\right) \end{aligned}$$

□

One in fact has a result stronger than (3.12): The relative entropy in (3.11) of a step is

$$\begin{aligned} \mathrm{đ}H\left(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]}\right) &\equiv H\left(\mathbb{P}'_{[0,t+1]}, \mathbb{P}'^-_{[0,t+1]}\right) - H\left(\mathbb{P}'_{[0,t]}, \mathbb{P}'^-_{[0,t]}\right) \\ &= \sum_{ij} \mathbf{p}_i(t) M_{ij} \log\left(\frac{\mathbf{p}_i(0) M_{ij}}{\mathbf{p}_j(0) M_{ji}}\right). \end{aligned} \tag{3.14}$$

One can show that the asymptotic relative entropy of a step in (3.14) is also the e_p. Note that for a stationary MC, the entropy production rate is exactly the time-averaged relative entropy, i.e, $e_p = \frac{1}{t} H\big(\mathbb{P}'_{[0,t]}, \mathbb{P}'^{-}_{[0,t]}\big)$ for any $t > 0$. There are many other equivalent expressions for the entropy production rate. For instance,

$$e_p = \sum_{i,j} \boldsymbol{\pi}_i M_{ij} \log\Big(\frac{\boldsymbol{\pi}_i M_{ij}}{\boldsymbol{\pi}_j M_{ji}}\Big), \quad \text{or} \quad e_p = \sum_{i,j} \boldsymbol{\pi}_i M_{ij} \log\Big(\frac{M_{ij}}{M^{-}_{ij}}\Big),$$

or

$$e_p = \frac{1}{2} \sum_{i,j} (\boldsymbol{\pi}_i M_{ij} - \boldsymbol{\pi}_j M_{ji}) \log\Big(\frac{\boldsymbol{\pi}_i M_{ij}}{\boldsymbol{\pi}_j M_{ji}}\Big).$$

So the entropy production rate $e_p = 0$ if and only if the MC is detailed balance, $\boldsymbol{\pi}_i M_{ij} = \boldsymbol{\pi}_j M_{ji}$.

As I have discussed earlier in Sec. 3.1, both $đS\big(\mathbb{P}'_{[0,t]}\big)$ and $đH\big(\mathbb{P}'_{[0,t]}, \mathbb{P}'^{-}_{[0,t]}\big)$ belong to the same class of stochastic quantity. As their asymptotic limits, the metric entropy h_{MC} and the entropy production rate e_p characterize the average randomness generated in the dynamic stepping from t to $t+1$ and average dynamic asymmetry with respect to time reversal, respectively.

3.3.2 Cycle Distributions and Maximum Entropy RDS

Besides expressing the entropy production rate in terms of the transition matrix M and its corresponding stationary distribution $\boldsymbol{\pi}$, a different representation can be given in terms of a collection of cycles $\mathcal{C}$ and weights $\{w_c : c \in \mathcal{C}\}$ on these cycles. These are regarded as cycle distributions of the MC [54]. In addition, there is an associated graph-based diagram method to compute the weight w_c which was first discovered by T. L. Hill [49] and proved by Qians [51]. In the setting of maximum entropy RDS, this graphical method can be further formulated as a function on the probability coefficient of the deterministic map with the single attractor. The entropy production rate e_p of the MC can be expressed in terms of the cycle weights as,

$$e_p = \sum_{c \in \mathcal{C}} w_c \log \frac{w_c}{w_{c_-}}. \tag{3.15}$$

Here c_- is denoted the reversed cycle of c. The previous proof given in [51] is quite involved. Here I will give a shorter, combinatorial proof for Eq. (3.15).

We start with discussing a related graphical method to solve the invariant distribution for the MC, i.e. to solve $\boldsymbol{\pi}$ for $\boldsymbol{\pi}M = \boldsymbol{\pi}$. The MC can be viewed as a directed graph with the transition probability as edge weight. Firstly, construct the complete set of spanning directed rooted *trees* which all edges are directed toward the root. A tree has the maximum possible edges without forming any loops. A tree with n nodes has $n-1$ edges and each node, except the root, has exactly one outgoing edge. If one views the directed graph as a discrete map for a dynamical system, the directed rooted tree gives a single fixed point. Second, assign the weight of previous directed rooted tree T as the product of its edge weights, $e(T)$. In fact, this weight connects with the coefficient of the corresponding map α (with root state goes to itself) under maximum entropy RDS as follows, $Q(\alpha) = e(T)M_{ii}$, where i is the root state. Finally, the weight of a set of graphs is the sum of their weights. Then the invariant distribution $\boldsymbol{\pi}$ can be expressed by $e(T)$ [21, 49, 62].

Theorem 3.6. *The invariant distribution for the irreducible and aperiodic MC is given by*

$$\boldsymbol{\pi}_i = \frac{e(\mathcal{T}_i)}{\Sigma},\ i = 1, \ldots, n. \tag{3.16}$$

where $\mathcal{T}_i$ is the set of directed rooted trees whose root is state i and the normalization factor $\Sigma = \sum_{i=1}^{n} e(\mathcal{T}_i)$.

The original proof was based on Cramer's rule but Hill discovered an elegant proof which is included here.

Proof. Consider the equation to solve $\sum_i \boldsymbol{\pi}_i M_{ij} = \boldsymbol{\pi}_j$. It is equivalent with solving the following equation

$$\sum_{\substack{k=1 \\ k\neq j}}^{n} \boldsymbol{\pi}_k M_{kj} = \boldsymbol{\pi}_j - \boldsymbol{\pi}_j M_{jj} = \sum_{\substack{i=1 \\ i\neq j}}^{n} M_{ji}\boldsymbol{\pi}_j \tag{3.17}$$

Let $\mathcal{G}_j$ be the set of directed graphs that have exactly one limit cycle and j is contained in that cycle. If a directed rooted tree $T \in \mathcal{T}_j$, adding the edge $j \to i$ will create an element

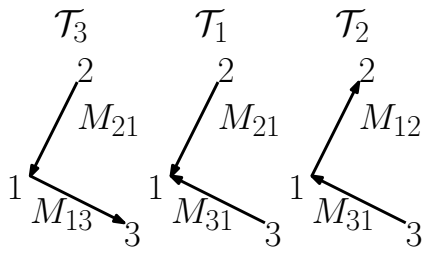

Figure 3.1: $e(\mathcal{T}_1) = M_{21}M_{31}, e(\mathcal{T}_2) = M_{12}M_{31}, e(\mathcal{T}_3) = M_{21}M_{13}$

$G \in \mathcal{G}_j$ and $e(G) = M_{ji}e(T)$. If a directed graph $G \in \mathcal{G}_j$, deleting the edge $j \to i$ will create an element $T \in \mathcal{T}_j$ and $M_{ji}e(T) = e(G)$. So I have $\sum_{\substack{i=1 \\ i\neq j}}^{n} M_{ji}e(\mathcal{T}_j) = e(\mathcal{G}_j)$. Here I am considering the outgoing edge from j and I can consider the incoming edge $k \to j$ as well. Similarly, $\sum_{\substack{k=1 \\ k\neq j}}^{n} M_{kj}e(\mathcal{T}_k) = e(\mathcal{G}_k)$. So $\boldsymbol{\pi}_j \propto e(\mathcal{T}_j)$. After renormalization, the solution (3.16) is the invariant distribution. □

We will give one example for the theorem.

Example 3.1. If the MC has the transition matrix

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & 0 \\ M_{31} & 0 & M_{33} \end{bmatrix},$$

each set $\mathcal{T}_i$ has only one element. The directed rooted trees and their weights are shown in Fig. 3.1. So the invariant distribution is

$$\boldsymbol{\pi} = \left(\frac{M_{21}M_{31}}{\Sigma}, \frac{M_{12}M_{31}}{\Sigma}, \frac{M_{21}M_{13}}{\Sigma} \right) \tag{3.18}$$

where $\Sigma = M_{21}M_{31} + M_{12}M_{31} + M_{21}M_{13}$.

The connection between two proofs is the matrix-tree theorem and is discussed in Appendix C. From the proof, it gives the insight for the following proposition to connect the RDS picture.

We call the attractor of the deterministic map α, $\mathcal{A}(\alpha)$, is *single* if the attractor is either exactly one fixed point or one limit cycle. The size of the single attractor $\|\mathcal{A}(\alpha)\|$ is the period of the limit cycle or just one for the fixed point. If the attractor is not single, denote $\mathcal{A}(\alpha) = \emptyset$ and the size of it is 0.

Let $\mathcal{G}_i$ be the set of directed graphs that has exactly one limit cycle and i is contained in that cycle. Unlike the set of directed rooted trees $\mathcal{T}_i$, for any $G \in \mathcal{G}_i$, $e(G)$ is exactly the probability of the corresponding deterministic map under maximum entropy RDS. So the set of directed graph corresponds to the deterministic maps with single attractor is $\cup_i\{\mathcal{G}_i, \mathcal{T}_i\}$. For convenience, denote $\mathcal{A}(G)$ as the single attractor of this corresponding deterministic map for $G \in \cup_i\{\mathcal{G}_i, \mathcal{T}_i\}$.

Proposition 3.3. *The normalization factor Σ is equal to the expected size of the single attractor of the deterministic map under maximum entropy RDS, i.e,*

$$\sum_{i=1}^{n} e(\mathcal{T}_i) = \mathbb{E}^Q(\|\mathcal{A}(\alpha)\|) \tag{3.19}$$

Proof. From appendix C, I know $\sum_{\substack{j=1 \\ j\neq i}}^{n} M_{ij} e(\mathcal{T}_i) = e(\mathcal{G}_i)$. So

$$\sum_{i=1}^{n} e(\mathcal{T}_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij} e(\mathcal{T}_i) = \sum_{i=1}^{n} \Big(e(\mathcal{G}_i) + M_{ii} e(\mathcal{T}_i) \Big)$$

$\sum_{i=1}^{n} M_{ii} e(\mathcal{T}_i)$ is the probability of the deterministic maps whose attractor is exactly one fixed point. It can be rewritten as $\sum_{T\in\cup_i \mathcal{T}_i} M_{\mathcal{A}(T)\mathcal{A}(T)} e(T)$. Here $\mathcal{A}(T)$ is the root state. For any $G \in \mathcal{G}_i$, $e(G)$ is counted $\|\mathcal{A}(G)\|$ times in $\sum_{i=1}^{n} e(\mathcal{G}_i)$. So it can rewritten as $\sum_{i=1}^{n} e(\mathcal{G}_i) = \sum_{G\in\cup_i \mathcal{G}_i} \|\mathcal{A}(G)\| e(G)$ and can be interpreted as the expected size of the single limit cycle. In total, it is the expected size of the single attractor of the deterministic map under maximum entropy RDS.

□

The MC will generate an infinite sequence of cycles almost surely. The set of cycles $\mathcal{C}$ contains all possible directed cycles along almost all sample paths of X_t and the weight w_c is

t	0	1	2	3	4	5	6	7	8
$X_t(\omega)$	2	1	3	1	3	1	1	2	2
$\eta_t(\omega)$	[2]	[2,1]	[2,1,3]	[2,1]	[2,1,3]	[2,1]	[2,1]	[2]	[2]
cycles				(1,3)		(3,1)	(1)	(1,2)	(2)

Table 3.1: The derived chain η_t and the cycles formed for this sample trajectory X_t [51].

the mean occurrences of the directed cycle c for almost all sample paths. The main reason to introduce this graphic method is it can be extended to find the cycle weight w_c. Here I present the method, the basic idea of the proof and the insight in RDS picture.

Start with the sample sequence ω, decompose cycles along the sequence by discarding the cycles formed at step t and keep the track of the remaining states in the sequence. The remaining sequence is called *derived chain* $\eta_t(\omega)$ and $w_{c,t}(\omega)$ is the *number of occurrences* of the cycle c up to step t. The cycle weight $w_c(\omega)$ is defined as average rate of occurrence $w_c(\omega) = \lim_{t\to+\infty} \frac{w_{c,t}(\omega)}{t}$. The weight also enjoys the ergodicity, i.e. the limit converges almost surely to a constant w_c. The precise definition is given in [51]. Here is one example to illustrate the idea. If I give the trajectory $(2, 1, 3, 1, 3, 1, 1, 2, 2, \dots)$ of $X_t(\omega)$, where the transition matrix is

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}.$$

The derived chain dynamics and cycles are counted as in Table 3.1.

Note the self-loop $2 \to 2$ here is regarded as a degenerated cycle. Without loss of generality, I include such self-loops to $\mathcal{C}$. Cycles are recorded by the ordered sequence $c = (i_1, \dots, i_t)$ with i_s are distinct for $1 \le s \le t$. The reverse of the cycle $(i_1, \dots, i_t)$ is defined as $c_- = (i_t, \dots, i_1)$. So $(1, 3)$ and its reverse $(3, 1)$ are the same cycle, but $(2, 1, 3)$ and its reverse $(3, 1, 2)$ are different. These cycles are the single attractors of deterministic

maps, $\mathcal{C} = \mathcal{A}(\cup_i\{\mathcal{G}_i, \mathcal{T}_i\})$.

In fact, the dynamic of the derived chain itself also deserves a look. η_t follows another MC with state space as all possible distinct ordered sequences. From the example, one can see that the first element in the derived chain state is always the initial state. Here the MC for the derived chain has 15 states. The transition matrix $\widetilde{M}$ is reducible with 3 classes based on the initial state. For the class starting with state 2, the set of all possible states are $[\mathscr{S}]_2 = \{[2], [2,1], [2,3], [2,1,3], [2,3,1]\}$. The submatrix of $\widetilde{M}$ on $[\mathscr{S}]_2$ is

$$\widetilde{M}^{[2]} = \begin{array}{c} \\ [2] \\ [2,1] \\ [2,3] \\ [2,1,3] \\ [2,3,1] \end{array} \begin{array}{c} \begin{array}{ccccc} [2] & [2,1] & [2,3] & [2,1,3] & [2,3,1] \end{array} \\ \begin{pmatrix} M_{22} & M_{21} & M_{23} & & \\ M_{12} & M_{11} & & M_{13} & \\ M_{32} & & M_{33} & & M_{31} \\ M_{32} & M_{31} & & M_{33} & \\ M_{12} & & M_{13} & & M_{11} \end{pmatrix} \end{array} \tag{3.20}$$

The directed graph corresponds to $\widetilde{M}^{[2]}$ is shown in Fig. 3.2. There is another example of derived chain dynamics in Fig. 3.3. The transition matrix of the MC is

$$M = \begin{bmatrix} 0 & M_{12} & 0 & M_{14} \\ M_{21} & 0 & M_{23} & M_{24} \\ 0 & M_{32} & 0 & M_{34} \\ M_{41} & M_{42} & M_{43} & 0 \end{bmatrix} \tag{3.21}$$

Here are the properties of the derived chain dynamics on the connectivity and cycle formation. Reader can refer Fig. 3.2 and Fig. 3.3 for more intuitive ideas. Denote the size of the derived chain state $[i_1, \ldots, i_t]$ as t. Then the largest possible size of the derived chain state is n.

First, the number of states in derived chain is much more than the original MC and not every state is connected in derived chain dynamics even when the original MC is completely connected. The derived chain state with size of t can only go to the state with size of $s \leq \min(t+1, n)$. There is no connection between states with the same size except with

itself. Each state with the last element i_t implies there is a path in the original MC from i_1 to i_t without forming a cycle. So the number of such paths is the same as number of states with the last element i_t. Furthermore, Each states with size n gives a Hamiltonian path in the original MC from i_1 which is the path from i_1 visits each state exactly once. Since the last element of the derived chain state is the current state in the original MC, the weight of each edge $[i_1, \ldots, i_t][i_1, \ldots, i_s]$ if exists, is the transition probability $M_{i_t i_s}$. Therefore the invariant distribution for the derived chain state $\mathbf{\Pi}^{i_1}([i_1, \ldots, i_t])$ satisfies the following equation,

$$\begin{aligned}\mathbf{\Pi}^{i_1}([i_1, \ldots, i_t]) = \mathbf{\Pi}^{i_1}([i_1, \ldots, i_{t-1}])M_{i_{t-1}i_t} + \mathbf{\Pi}^{i_1}([i_1, \ldots, i_t])M_{i_t i_t} \\ + \sum_{j_1, \ldots, j_r} \mathbf{\Pi}^{i_1}([i_1, \ldots, i_t, j_1, \ldots, j_r])M_{j_r i_t}\end{aligned} \tag{3.22}$$

Second, in the original MC the current state cannot tell which cycle could be possibly formed at next step, but the derived chain state can. The derived chain has the past history on the path from i_1 to the current state of the original MC after removing cycles. Each time the derived chain state $[i_1, \ldots, i_t]$ goes to the state with the same or less size $[i_1, \ldots, i_s], s \leq t$, a cycle $(i_s, i_{s+1}, \ldots, i_t)$ in the original MC sequence is formed. In particular, any cycle contained with i_1, i.e, $(i_1, \ldots, i_t)$, can only be formed by the derived chain state $[i_1, i_2, \ldots, i_t]$ going to $[i_1]$.

From these properties, one can pick the suitable class to calculate the weight of cycles. In particular, for the cycle $c = (i_1, i_2, \ldots, i_t)$, it is much convenient to use the class $[\mathscr{S}]_{i_1}$, i.e, the initial state of the original MC is i_1. Hence the average rate of the occurrence for this cycle is same as the average frequency of the pair $[i_1, \ldots, i_t][i_1]$ in the derived chain dynamics. Since the average frequency of state $[i_1, \ldots, i_t]$ goes to $\mathbf{\Pi}^{i_1}([i_1, i_2, \ldots, i_t])$, the average frequency of the pair $[i_1, \ldots, i_t][i_1]$ is $w_c = \mathbf{\Pi}^{i_1}([i_1, i_2, \ldots, i_t]) \cdot M_{i_t i_1}$.

Proposition 3.4. *The directed graph G_2 corresponds to the derived chain on $[\mathscr{S}]_{i_1}$, is homomorphic to the directed graph G_1 to the MC on $\mathscr{S}$.*

Proof. Consider the function from $[\mathscr{S}]_{i_1}$ to $\mathscr{S}$, $f([i_1, \ldots, i_t]) = i_t$. For any edge $[i_1, \ldots, i_t][i_1, \ldots, i_s] \in E(G_2)$, the weight of the edge is $M_{i_t i_s}$ and it means there is an edge from i_t to i_s in G_1. So $f([i_1, \ldots, i_t])f([i_1, \ldots, i_s]) \in E(G_1)$. □

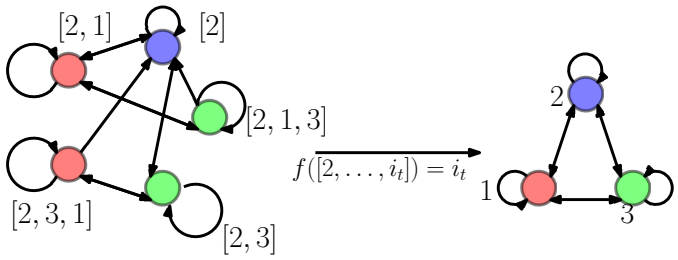

Figure 3.2: A 3-state completely connected MC. The transition matrix of derived chain dynamics is in (3.20).

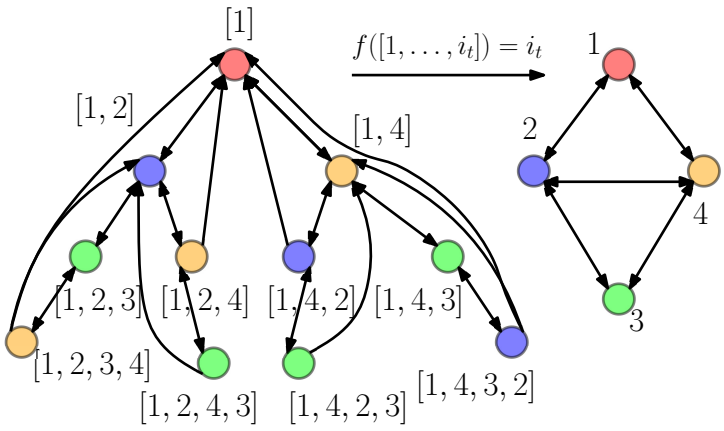

Figure 3.3: A 4-state MC with initial state 1.

The Fig. 3.2 and Fig. 3.3 demonstrate the homomorphism of the directed graph G_2 (left) to G_1 (right). The homomorphism f induces the equivalent class on $[\mathscr{S}]_{i_1}$ by using the pre-image $f^{-1}(i_t)$ for any $i_t \in \mathscr{S}$. These equivalent classes are colored in the same color in these two examples. In fact, the invariant measures on both dynamics are also connected by the pre-image of the homomorphism as shown in the following theorem.

A directed rooted tree is denoted as $T_{[i_1,\ldots,i_t]}$ if it has root i_t and has edges $i_1 i_2, \ldots i_{t-1} i_t$. By convention, if $t = 1$, $T_{[i_1]} = T_{i_1}$. And $\mathcal{T}_{[i_1,\ldots,i_t]}$ is the set of all such directed rooted trees $T_{[i_1,\ldots,i_t]}$.

Theorem 3.7. *The invariant distribution for the derived chain dynamics is given by*

$$\mathbf{\Pi}^{i_1}([i_1,\ldots,i_t]) = \frac{e(\mathcal{T}_{[i_1,\ldots,i_t]})}{\Sigma},\ [i_1,\ldots,i_t] \in [\mathscr{S}]_{i_1}. \tag{3.23}$$

where the normalization factor is $\Sigma = \sum_{i=1}^{n} e(\mathcal{T}_i)$. *Moreover,* $\mathbf{\Pi}^{i_1}(f^{-1}(i_t)) = \boldsymbol{\pi}(i_t)$.

Note:

$$e(\mathcal{T}_{[i_1,\ldots,i_t]}) = M_{i_1 i_2}\ldots M_{i_{t-1}i_t} e(\mathcal{T}_{i_1,\ldots,i_t}) \tag{3.24}$$

where $\mathcal{T}_{i_1,\ldots,i_t}$ is the set of directed forests whose roots are $i_1,\ldots,i_t$.

Proof. The first part is to show $\mathbf{\Pi}^{i_1}([i_1,\ldots,i_t]) \propto e(\mathcal{T}_{[i_1,\ldots,i_t]})$. The technique is similar with the proof of Theorem 3.6. Based on Eq. (3.22), it is equivalent with solving the following equation,

$$\begin{aligned}\sum_{i_s=1}^{i_{t-1}} \mathbf{\Pi}^{i_1}([i_1,\ldots,i_t]) M_{i_t i_s} &= \mathbf{\Pi}^{i_1}([i_1,\ldots,i_{t-1}]) M_{i_{t-1}i_t} \\ &+ \sum_{j_1,\ldots,j_r} \mathbf{\Pi}^{i_1}([i_1,\ldots,i_t,j_1,\ldots,j_r]) M_{j_r i_t}\end{aligned} \tag{3.25}$$

Let $\mathcal{G}_{[i_1,\ldots,i_t]}$ be the set of directed graphs that have exactly one limit cycle with i_t contained and the graphs have edges $i_1 i_2,\ldots,i_{t-1}i_t$. It has the following six steps.

1. $\mathcal{T}_{[i_1,\ldots,i_t]} + i_t i_s \subset G_{[i_1,\ldots,i_t]}$ If a directed rooted tree $T \in \mathcal{T}_{[i_1,\ldots,i_t]}$, adding an edge $i_t i_s$ will create an element $G \in \mathcal{G}_{[i_1,\ldots,i_t]}$, where $i_s = i_1,\ldots,i_{t-1}$.

2. $G_{[i_1,\ldots,i_t]} - i_t i_s \subset \mathcal{T}_{[i_1,\ldots,i_t]}$ If a directed graph $G \in \mathcal{G}_{[i_1,\ldots,i_t]}$, deleting the edge $i_t i_s$ will create an element $T \in \mathcal{T}_{[i_1,\ldots,i_t]}$, where $i_s = i_1,\ldots,i_{t-1}$.

3. $\mathcal{T}_{[i_1,\ldots,i_{t-1}]} + i_{t-1} i_t \subset \mathcal{G}_{[i_1,\ldots,i_t]}$ If a directed rooted tree $T \in \mathcal{T}_{[i_1,\ldots,i_{t-1}]}$, adding an edge $i_{t-1}i_t$ will create an element $G \in \mathcal{G}_{[i_1,\ldots,i_t]}$ because i_t has a path to i_{t-1} in the tree and now with the edge $i_{t-1}i_t$, it must have a limit cycle with i_t contained.

4. $\mathcal{T}_{[i_1,\ldots,i_t,j_1,\ldots,j_r]} + j_r i_t \subset \mathcal{G}_{[i_1,\ldots,i_t]}$ If a directed rooted tree $T \in \mathcal{T}_{[i_1,\ldots,i_t,j_1,\ldots,j_r]}$, adding an edge $j_r i_t$ will create a limit cycle $(i_t, j_1,\ldots,j_r)$, which is an element in $\mathcal{G}_{[i_1,\ldots,i_t]}$.

5. $\mathcal{G}_{[i_1,\ldots,i_t]} - j_r i_t \subset \mathcal{T}_{[i_1,\ldots,i_t,j_1,\ldots,j_r]}, j_r \neq i_{t-1}$ If a directed graph $G \in \mathcal{G}_{[i_1,\ldots,i_t]}$ has the limit cycle $(i_t, j_1, \ldots, j_r)$, $r \geq 1$, where $j_1, \ldots, j_r$ have no common elements with $i_1, \ldots, i_{t-1}$, deleting the edge $j_r i_t$ will create an element $T \in \mathcal{T}_{[i_1,\ldots,i_t,j_1,\ldots,j_r]}$ since the root is j_r and apart from the path from i_1 to i_t, there is another path from i_t to j_r: $i_t j_1, \ldots j_{r-1} j_r$,

6. $\mathcal{G}_{[i_1,\ldots,i_t]} - i_{t-1} i_t \subset \mathcal{T}_{[i_1,\ldots,i_{t-1}]}$ If a directed graph $G \in \mathcal{G}_{[i_1,\ldots,i_t]}$ has the limit cycle $(i_s, \ldots, i_t, j_1, \ldots, j_r)$, $r \geq 1$, where $1 \leq s < t$ and where $j_1, \ldots, j_r$ have no common elements with $i_1, \ldots, i_s$, deleting the edge $i_{t-1} i_t$ will create an element $T \in \mathcal{T}_{[i_1,\ldots,i_{t-1}]}$ because the root is i_{t-1}.

From 1 and 2, the RHS of Eq. (3.25) is

$$\text{RHS} = \sum_{i_s=i_1}^{i_{t-1}} e(\mathcal{T}_{[i_1,\ldots,i_t]}) M_{i_t i_s} = e(\mathcal{G}_{[i_1,\ldots,i_t]}).$$

From 3-6, the LHS of Eq. (3.25) is

$$\text{LHS} = e(\mathcal{T}_{[i_1,\ldots,i_{t-1}]}) M_{i_{t-1} i_t} + \sum_{j_1,\ldots,j_r} e(\mathcal{T}_{[i_1,\ldots,i_t,j_1,\ldots,j_r]}) M_{j_r i_t} = e(\mathcal{G}_{[i_1,\ldots,i_t]}).$$

Now I have proved $\mathbf{\Pi}^{i_1}([i_1, \ldots, i_t]) \propto e(\mathcal{T}_{[i_1,\ldots,i_t]})$.

The second part is to prove the normalization factor is $\Sigma = \sum_{i=1}^{n} e(\mathcal{T}_i) = e(\cup_i \mathcal{T}_i)$, which is the same as the normalization factor in calculating the invariant distribution of the original MC. By the definition, $\Sigma = \sum_{[i_1,\ldots,i_t] \in [S]_{i_1}} e(\mathcal{T}_{[i_1,\ldots,i_t]}) = e(\cup_{i_2,\ldots,i_t} \mathcal{T}_{[i_1,\ldots,i_t]})$. If a directed rooted tree with root i_t, $T \in \mathcal{T}_{i_t}$, there exists a path from i_1 to i_t. If $t = 1$, the path degenerates to a point. So $T \in \cup_{i_2,\ldots,i_{t-1}} \mathcal{T}_{[i_1,i_2,\ldots,i_{t-1},i_t]}$. If a directed rooted tree $T \in \cup_{i_2,\ldots,i_{t-1}} \mathcal{T}_{[i_1,i_2,\ldots,i_{t-1},i_t]}$, then T has root i_t and $T \in \mathcal{T}_{i_t}$. So I have

$$e(\cup_{i_2,\ldots,i_{t-1}} \mathcal{T}_{[i_1,i_2,\ldots,i_{t-1},i_t]}) = e(\mathcal{T}_{i_t}). \tag{3.26}$$

That implies the weights of the union of set on i_t are the same, i.e, $e(\cup_{i_t} \mathcal{T}_{i_t}) = e(\cup_{i_2,\ldots,i_t} \mathcal{T}_{[i_1,\ldots,i_t]})$. This proves the normalization factor is $\Sigma = \sum_{i=1}^{n} e(\mathcal{T}_i)$ and the invariant distribution $\mathbf{\Pi}^{i_1}([i_1, \ldots, i_t])$ is given by Eq. (3.23).

Moreover, if divide Σ on both hand-side of Eq. (3.26), it gives the following relationship on both invariant distributions by using the pre-image of the homomorphism, $\mathbf{\Pi}^{i_1}(f^{-1}(i_t)) = \boldsymbol{\pi}(i_t)$.

□

By using Theorem 3.7, I built a bridge between the cycle coordinates $(\mathcal{C}, w_c)$ and the maximum entropy RDS of the MC.

Corollary 3.3. *The mean number of occurrences of the cycle c per step, w_c is*

$$w_c = \frac{Q(\alpha : \mathcal{A}(\alpha) = c)}{\mathbb{E}^Q(\|\mathcal{A}(\alpha)\|)} \tag{3.27}$$

where $Q(\alpha : \mathcal{A}(\alpha) = c)$ is the probability of the deterministic map with single attractor c under maximum entropy RDS.

Proof. From the previous discussion, if $c = (i_1, \ldots, i_t)$, $w_c = \frac{e(\mathcal{T}_{[i_1,\ldots,i_t]})M_{i_t i_1}}{\Sigma}$. The numerator $e(\mathcal{T}_{[i_1,\ldots,i_t]})M_{i_t i_1}$ is equal to the weight of the set of directed graph with the only one limit cycle $(i_1, \ldots, i_t)$ and it can also be interpreted as the probability of the deterministic map with single attractor c, i.e, $Q(\alpha : \mathcal{A}(\alpha) = c)$. The denominator $\Sigma = \mathbb{E}^Q(\|\mathcal{A}(\alpha)\|)$ is proven in Proposition 3.3. □

Instead of dividing the step size t in finding the cycle weight w_c, divide the total number of cycles formed and I will have the following ergodic theorem. Though the proof is straightforward and elementary, the result in fact is very interesting.

Corollary 3.4. *The frequency of the cycle c along the sample sequence almost surely converges to the probability of the deterministic map with the attractor c given the attractor of the map is single,*

$$\lim_{t \to +\infty} \frac{w_{c,t}(\omega)}{\sum_{c \in \mathcal{C}} w_{c,t}(\omega)} = \mathbf{p}_c, \ \textit{almost surely} \tag{3.28}$$

where $\mathbf{p}_c = Q(\alpha : \mathcal{A}(\alpha) = c \mid \mathcal{A}(\alpha)$ is single).

Proof. We know $\lim_{t\to+\infty} w_{c,t}(\omega)/t = w_c$ almost surely. After using Eq. (3.27) I have

$$\lim_{t\to+\infty} \frac{w_{c,t}(\omega)/t}{\sum_{c\in\mathcal{C}} w_{c,t}(\omega)/t} = \frac{w_c}{\sum_{c\in\mathcal{C}} w_c} = \frac{Q(\alpha : \mathcal{A}(\alpha) = c)}{\sum_{c\in\mathcal{C}} Q(\alpha : \mathcal{A}(\alpha) = c)}.$$

□

So $\sum_{c\in\mathcal{C}} w_c$ is the mean number of occurrences of any possible cycle per step. Then the reciprocal of this quantity, denoted $\lambda = \frac{1}{\sum_{c\in\mathcal{C}} w_c}$, will be the mean steps to generate a cycle and is the "time unit" of forming the cycle.

Corollary 3.5.

$$\lim_{t\to+\infty} \frac{\sum_{c\in\mathcal{C}} w_{c,t}(\omega)\|c\|}{\sum_{c\in\mathcal{C}} w_{c,t}(\omega)} = \lambda \text{ almost surely.} \tag{3.29}$$

Proof. We know $\sum_{c\in\mathcal{C}} w_c\|c\| = 1$.

$$\lim_{t\to+\infty} \frac{\sum_{c\in\mathcal{C}} w_{c,t}(\omega)\|c\|/t}{\sum_{c\in\mathcal{C}} w_{c,t}(\omega)/t} = \frac{\sum_{c\in\mathcal{C}} w_c\|c\|}{\sum_{c\in\mathcal{C}} w_c} = \frac{1}{\sum_{c\in\mathcal{C}} w_c}.$$

□

For each sample sequence $X_t(\omega)$, it induces a sequence of cycles and $w_{c,t}(\omega)$ counts the number of cycles c occurred up to t. Instead of studying $X_t(\omega)$, I try to study the dynamics of cycles. Unfortunately, the dynamics of cycle is not Markovian and seemingly complicated, but it is still ergodic with invariant distribution $\mathbf{p}_c$. The cycle weight w_c can be express by $w_c = \mathbf{p}_c/\lambda$.

With the expression of the cycle weight w_c, now I can connect the cycle coordinates with the edge coordinates M_{ij} and invariant distribution $\boldsymbol{\pi}$.

Corollary 3.6.

$$\sum_{c\in\mathcal{C}} w_c J_c(i) = \boldsymbol{\pi}_i, \quad \sum_{c\in\mathcal{C}} w_c J_c(i,j) = \boldsymbol{\pi}_i M_{ij} \tag{3.30}$$

$$\text{where } J_c(i) = \begin{cases} 1 & \text{if } i \in c \\ 0 & \text{Otherwise} \end{cases}, \; J_c(i,j) = \begin{cases} 1 & \text{if } ij \text{ is an edge of } c \\ 0 & \text{Otherwise} \end{cases}.$$

Proof. The RHS of the first equality is

$$\sum_{c\in\mathcal{C}} w_c J_c(i) = \lim_{t\to+\infty} \frac{\sum_{c\in\mathcal{C}} w_{c,t} J_c(i)}{t}$$

The numerator is the number of occurrences of the state i up to t without counting the derived chain. But the maximum length of the derived chain is n.

$$\frac{\#\text{state } i - n}{t} \leq \frac{\sum_{c\in\mathcal{C}} w_{c,t} J_c(i)}{t} \leq \frac{\#\text{state } i}{t}.$$

Taking the limit $t \to +\infty$, it gives $\sum_{c\in\mathcal{C}} w_c J_c(i) = \boldsymbol{\pi}_i$.

Since ij is an edge of c, I can write it out RHS of the second equality explicitly,

$$\sum_{c\in\mathcal{C}} w_c J_c(i,j) = \sum_{i_1,\ldots,i_t} \frac{e(\mathcal{T}_{[j,i_1,\ldots,i_t,i]})}{\Sigma} M_{ij} = \boldsymbol{\pi}_i M_{ij}$$

From Eq. (3.26), $\sum_{i_1,\ldots,i_t} \frac{e(\mathcal{T}_{[j,i_1,\ldots,i_t,i]})}{\Sigma} = \frac{e(\mathcal{T}_i)}{\Sigma} = \boldsymbol{\pi}_i$ □

Theorem 3.8. *The entropy production rate e_p is represented by the cycle coordinates $(\mathcal{C}, w_c)$ and furthermore, by the invariant distribution of cycle dynamics $\mathbf{p}_c$ and λ*

$$e_p = \sum_{c\in\mathcal{C}} w_c \log \frac{w_c}{w_{c_-}}, \quad e_p = \frac{H(\mathbf{p}_c, \mathbf{p}_{c_-})}{\lambda} \tag{3.31}$$

Proof. Use the second equality of Eq. (3.30) in the expression of entropy production rate in (3.13), assume the cycle $c = (i_1, \ldots, i_t)$. Due to the Eq. (3.24), I have $w_c = M_{i_1 i_2} \ldots M_{i_t i_1} \frac{e(\mathcal{T}_{i_1,\ldots,i_t})}{\Sigma}$ and $w_{c_-} = M_{i_1 i_t} \ldots M_{i_2 i_1} \frac{e(\mathcal{T}_{i_1,\ldots,i_t})}{\Sigma}$.

$$e_p = \sum_{i,j} \Big(\sum_{c\in\mathcal{C}} w_c J_c(i,j) \Big) \log \frac{M_{ij}}{M_{ji}} = \sum_{c\in\mathcal{C}} w_c \log \frac{M_{i_1 i_2} \ldots M_{i_t i_1}}{M_{i_1 i_t} \ldots M_{i_2 i_1}} = \sum_{c\in\mathcal{C}} w_c \log \frac{w_c}{w_{c_-}}.$$

Use the expression I get in Eq. (3.27)

$$\begin{aligned} e_p &= \sum_{c\in\mathcal{C}} w_c \log \frac{w_c}{w_{c_-}} = \sum_{c\in\mathcal{C}} \frac{Q(\alpha : \mathcal{A}(\alpha) = c)}{\mathbb{E}^Q(\|\mathcal{A}(\alpha)\|)} \log \frac{Q(\alpha : \mathcal{A}(\alpha) = c)}{Q(\alpha : \mathcal{A}(\alpha) = c_-)} \\ &= \frac{\sum_{c\in\mathcal{C}} Q(\alpha : \mathcal{A}(\alpha) = c)}{\mathbb{E}^Q(\|\mathcal{A}(\alpha)\|)} \sum_{c\in\mathcal{C}} \mathbf{p}_c \log \frac{\mathbf{p}_c}{\mathbf{p}_{c_-}} = \frac{H(\mathbf{p}_c, \mathbf{p}_{c_-})}{1/(\sum_{c\in\mathcal{C}} w_c)} \end{aligned}$$

□

From the theorem, entropy production rate e_p is proportional to the relative entropy of the invariant distribution of cycle dynamics with respect to its reverse cycle. The extra constant term $1/\lambda$ is the average number of cycle occurrences per step which bridges from the cycle dynamics back to the original MC.

3.3.3 Entropy Production of Doubly Stochastic MC and its invertible RDS

Entropy production characterizes dynamic randomness, which can be divided conceptually as "uncertainties in the past" and "uncertainties in the future". The former is represented by non-invertible, "many-to-one" maps while the latter is best represented by stochastic, "one-to-many" dynamics. In connection to the entropy production in non-invertible dynamics, Ruelle has introduced the notion of *folding entropy* [100].

In Sec. 3.3.2, the entropy production is discussed in terms of maximum entropy representation of the MC. Here I establish a relationship between the discrete state, finite RDS with only invertible transformations, *invertible RDS*, and the entropy production rate of its corresponding MC, which is always doubly stochastic: Both the rows and columns of the MC transition matrix sum to 1. Thus the invariant distribution is uniform, i.e, $\boldsymbol{\pi}_i = 1/n$. From the Birkhoff-Von Neumann theorem, the set of doubly stochastic matrices is the convex hull of the set of $n \times n$ permutation matrices, and the vertices are precisely permutation matrices [127]. In other words, for every doubly stochastic MC, one can assign the probability measure $Q(\alpha)$ on the the set of invertible maps Δ, such that $\sum_{\alpha\in\Delta} Q(\alpha) = 1$ and $Q(\alpha : \alpha(i) = j) = M_{ij}$. It provides an invertible RDS representation for a doubly stochastic MC through an i.i.d. process. Such representation is different from the maximum entropy representation, and it may not be unique. For the invertible RDS, each invertible map α has a well-defined inverse map α^{-1}; its matrix representation is the inverse of its permutation matrix, $P_{\alpha^{-1}} = P_\alpha^{-1}$. So the cycle of the inverse map is reversed compared with the original map.

With this set up, one can introduce a *time-reversal dual* probability measure on the set of permutation matrices Δ, $Q^-(\alpha) = Q(\alpha^{-1})$. The probabilities of the invertible map and its

inverse are flipped. Q^- define a dual RDS through an i.i.d. process, and its corresponding MC is exactly the time-reversed process with transition matrix $M^-_{ij} = M_{ji}$.

Theorem 3.9. *The entropy production rate of a doubly stochastic MC has an upper bound in terms of the relative entropy of measure Q with respect to Q^-.*

$$e_p \leq H(Q, Q^-). \tag{3.32}$$

The equality holds if and only if $Q = Q^-$.

Proof. The entropy production rate $e_p = \frac{1}{n}\sum_{ij} M_{ij} \log \frac{M_{ij}}{M_{ji}}$. Based on the log-sum inequality, which states that for two sets of non-negative numbers $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \Big(\sum_{i=1}^{n} a_i\Big) \log \frac{(\sum_{i=1}^{n} a_i)}{(\sum_{i=1}^{n} b_i)}$$

The equality holds if and only if $a_i = b_i$.

$$\begin{aligned}
e_p &= \frac{1}{n}\sum_{ij} \Big(\sum_{\alpha:\alpha(i)=j} Q(\alpha)\Big) \log \frac{\sum_{\alpha:\alpha(i)=j} Q(\alpha)}{\sum_{\alpha:\alpha(j)=i} Q(\alpha)} \\
&= \frac{1}{n}\sum_{ij} \Big(\sum_{\alpha:\alpha(i)=j} Q(\alpha)\Big) \log \frac{\sum_{\alpha:\alpha(i)=j} Q(\alpha)}{\sum_{\alpha:\alpha(i)=j} Q^-(\alpha)} \\
&\leq \frac{1}{n}\sum_{ij} \sum_{\alpha:\alpha(i)=j} \Big(Q(\alpha) \log \frac{Q(\alpha)}{Q^-(\alpha)}\Big) \\
&= \sum_{\alpha} Q(\alpha) \log \frac{Q(\alpha)}{Q^-(\alpha)}.
\end{aligned} \tag{3.33}$$

The last equality (3.33) is because the term $Q(\alpha) \log \frac{Q(\alpha)}{Q^-(\alpha)}$ sums up exactly n times for each α which is the number of edges. The equality holds if and only if $Q(\alpha) = Q^-(\alpha)$. □

The condition that the equality holds is the sufficient condition of the doubly stochastic MC being detailed balance, i.e, $M^- = M$.

Theorem 3.9 should be compared and contrasted with an earlier result from Kifer and Ye *et al.*. [60, 127]: $h_{\mathrm{RDS}} \geq h_{\mathrm{MC}}$, i.e, the metric entropy of MC in (3.7) is upper bounded by the metric entropy of RDS, $h_{\mathrm{RDS}} = -\sum_{\alpha} Q(\alpha) \ln Q(\alpha)$. h_{RDS} exists unique finite upper bound which is maximum entropy RDS, but $H(Q, Q^-)$ here could be $+\infty$.

Chapter 4

SYNCHRONIZATION

Each realization of an RDS is a non-autonomous dynamical systems [4]. Synchronization is a well-developed concept in the theory of non-autonomous ODEs [71, 73]. Through studying synchronization in an RDS, one appreciates the fact that RDS formulation is a more refined model of stochastic dynamics than an MC. The same phenomenon has also been terms as *random sink* [10, 73] in neuron dynamics applications. It has a deep relation to the concept of *coupling* in the theory of MC [72, 113].

4.1 Synchronization in Finite i.i.d. RDS

In finite i.i.d RDS defined in Sec. 2.4, if we apply the sequence of deterministic transformations on multiple initial conditions simultaneously, then it induces multiple sequences of states. These sequences are not independent; actually once they collide at some instance, they will be together forever. This phenomenon is called *synchronization*. In fact, it is easy to see that if any pair of sequences synchronize in the finite step almost surely, any multiple sequences will synchronize to one single sequence almost surely. So we can reduce this to the study of two-point motion [10].

We shall first give a mathematical definition for synchronization: remind $\varphi(1,\omega) = \alpha_0$ for $\omega = (\dots, \alpha_{-1}, \alpha_0, \alpha_1, \dots)$ and the map $\varphi(t,\omega) = \alpha_{t-1} \circ \alpha_{t-2} \dots \alpha_1 \circ \alpha_0$.

Definition 4.1. *The finite i.i.d. RDS on the state space $\mathscr{S}$ over the metric dynamical system $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$ synchronizes if for any two different initial states i_0 and j_0 in $\mathscr{S}$, i.e, $i_0 \neq j_0$,*

$$\lim_{t \to +\infty} \Pr\Big\{\omega : \varphi(t,\omega) i_0 = \varphi(t,\omega) j_0, t \in \mathbb{N}\Big\} = 1 \tag{4.1}$$

There are always some sequences that they will never synchronize but we claim the probability of these sequences is 0. So they are insignificant.

We note, however, that not every RDS will possess such property. Given a Markov chain with the transition probability matrix M, we would like to have a survey on the synchronization of its RDS representations. We are particular interested in the maximum metric entropy representation, i.e, $Q(\alpha_{i_1,i_2,\ldots,i_n}) = M_{1i_1}M_{2i_2}\ldots M_{ni_n}$.

In studying synchronization, one needs simultaneous construction of two infinite sequences by applying the sequence ω on the two initial states as a pair (x_0, y_0), so the product space will be $\Omega' \otimes \Omega'$ and the shift map can be induced from θ in the RDS. It describes a two-point motion.

In terms of this RDS, it becomes a new Markov chain (X, Y) whose state space is $\mathscr{S} \times \mathscr{S}$. Its transition probability matrix W is defined as

$$W_{(s,j)\to(k,\ell)} = \begin{cases} \displaystyle\sum_{i_s=k, i_j=\ell} \alpha_{i_1,i_2,\ldots,i_n} & \text{when } s \neq j, \\ M_{sk} & \text{when } s = j \text{ and } k = \ell, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

W is an $n^2 \times n^2$ stochastic matrix. Each component of this new Markov chain is the one-point motion in the RDS, namely the original Markov chain, since summing out the probability of the other component will degenerate the transition probability W into the original transition probability matrix M.

The synchronization of the RDS means that only the states $\{(1,1),(2,2),\ldots,(n,n)\}$ are possibly recurrent, while all other states are transient. Moreover, the transition probability among these recurrent states are exactly the same as the original transition probability because the two-point motion collapses down to a copy of the one-point motion. In the

maximum entropy RDS, the previous expression can be further simplified as

$$W_{(s,j)\to(k,\ell)} = \begin{cases} M_{sk}M_{j\ell} & \text{when } s \neq j, \\ M_{sk} & \text{when } s = j \text{ and } k = \ell, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

In probability theory, Eq. 4.3 is the transition probability of a Markov chain, as the most basic example of coupling of two Markov chains, first used by Wolfgang Doeblin. This Markov chain (X, Y) behaves as follows: if $X_n \neq Y_n$, then the two components make independent movements according to the transition matrix $\mathbf{M}$ at each step; if $X_n = Y_n$, they make the same movements. Thus, $\{(1,1),(2,2),\ldots,(n,n)\}$ is absorbing.

Actually, in terms of Eq. 4.4 below, it is very clear why Doeblin's coupling is different from two independent Markov chains: The transition probability matrix W' for the two-point motion of two independent Markov chains is

$$W'_{(s,j)\to(k,\ell)} = M_{sk}M_{j\ell}. \tag{4.4}$$

W is the same as W' except at these n rows of $(1,1),(2,2),\ldots,(n,n)$.

With the introduction of W, we can find the precise condition under which an RDS synchronizes. We have the following theorem.

Theorem 4.1. *For a Markov chain, its corresponding RDS with the maximum-metric-entropy representation synchronizes if and only if this MC has unique absorbing communicating class, and the MC restricted on this communicating class is aperiodic.*

Remark 1. From this theorem, it is easy to see that if a Markov chain is irreducible and aperiodic, then the maximum entropy RDS synchronizes.

Proof. $\Rightarrow$: Remember that once a trajectory falls into an absorbing communicating class, it will never get out again. If the Markov chain has at least two absorbing communicating classes, then states from these two classes will never synchronize.

If the MC restricted on an absorbing communicating class is k-periodic, then this communicating class can be divided into k subclasses. Each time the MC jumps from one subclass to the next. States from different subclasses will never synchronize.

$\Leftarrow$: Denote the absorbing communicating class by $\mathscr{C}$. Since the absorbing communicating class is unique, there exists a positive integer n_1 such that for any state $i \in \mathscr{S} \setminus \mathscr{C}$, there exists a state $j \in \mathscr{C}$ to make $p^{n_1}(i,j) > 0$. Since the Markov chain restricted on $\mathscr{C}$ is irreducible and aperiodic, there exists a positive integer n_2 such that for any two states $i, j \in \mathscr{C}$, $p^{n_2}(i,j) > 0$. Set $n = n_1 + n_2$, then for any two states $i \in \mathscr{S}$, $j \in \mathscr{C}$, $p^n(i,j) > 0$.

Now for any two initial states $i_0, j_0 \in \mathscr{S}$, we can find admissible sequences to reach any state $k_0 \in \mathscr{C}$ after nth step:

$$\underbrace{i_0, i_1, i_2, \ldots i_{n-1}, i_n}_{n \text{ steps}} = k_0 \quad \text{and} \quad \underbrace{j_0, j_1, j_2, \ldots j_{n-1}, j_n}_{n \text{ steps}} = k_0.$$

If this two sequence collide before the step n, i.e, $m = \min\{k : i_k = j_k\} < n$, we truncate the sequence up to m steps. Now there are two admissible sequences which reach the same state at step m in the first time. So in the intermediate states, $i_k \neq j_k$, for $1 \leq k \leq m-1$. The probabilities of both sequences in MC are strictly positive.

Now we want to calculate the probability for both sequences with initial states i_0 and j_0, $p(i_0, j_0)$, in the maximum-metric-entropy representation RDS.

$$\begin{aligned} p(i_0, j_0) &= \Pr\left\{\omega : \varphi(m,\omega)i_0 = i_k, \varphi(m,\omega)j_0 = j_k, 1 \leq k \leq m\right\} \\ &= (M_{i_0 i_1} M_{j_0 j_1} (M_{i_1 i_2} M_{j_1 j_2}) \ldots (M_{i_{m-1} i_m} M_{j_{m-1} j_m}) \\ &= \Pr\left\{i_0, i_1, i_2, \ldots i_{m-1}, i_m\right\} \Pr\left\{j_0, j_1, j_2, \ldots j_{m-1}, j_m\right\}. \end{aligned}$$

This probability is strictly positive. $p(i_0, j_0)$ gives a lower bound of the probability of sequences starting with i_0 and j_0 synchronize within n steps. For convention, if $i_0 = j_0$, $p(i_0, j_0) = 1$. Among all different pairs of initial states, define $p = \min_{i,j} p(i,j)$, which is the lower bound of the probability for any two initial states to synchronize within n steps. The

probability that any two sequences don't synchronize within t steps is $P_{\text{non-sync}} = \Pr\{\omega : \varphi(t,\omega)i_0 \neq \varphi(t,\omega)j_0\} \leq (1-p)$. So the probability for any two sequences do not synchronize within kt steps is $P^k_{\text{non-sync}} < (1-p)^k$. It implies any two sequences do not synchronize in finite steps has probability 0. Thus, the finite i.i.d RDS synchronizes by definition.

□

We shall show several examples of synchronization through numerical simulations. We consider a Markov chain with 4×4 transition matrix, thus $\|\Gamma\| = 256$. We will simulate four sample trajectories in the RDS setting and these trajectories start from four different initial conditions and stop once all four collide into one trajectory; count down steps required. We are interested in two questions: first, given a Markov transition matrix, what kind of RDS representation will synchronize and what kind will not?; second, if it synchronizes, what is the probability distribution of the synchronization steps (or coupling time)? A more rigorous definition for the random variable N_s, the synchronization steps is

$$N_s(\omega) = \min\left\{t : \varphi(t,\omega)s_1 = \varphi(t,\omega)s_2 = \varphi(t,\omega)s_3 = \varphi(t,\omega)s_4, t \in \mathbb{N}\right\} \tag{4.5}$$

where s_1, s_2, s_3, s_4 are four different initial conditions.

For the first question, we have shown the sufficient and necessary condition for the maximum-metric-entropy representation. It is possible that an irreducible and aperiodic MC in other RDS representations will not synchronize. A trivial example is an irreducible and aperiodic doubly stochastic Markov matrix that is decomposed into a convex combination of permutation matrices. However, if Γ contains non-permutation matrices, it is still possible that RDS will not synchronize.

Example 4.1. Consider the 4×4 Markov transition matrix M which is decomposed into

the following way,

$$\begin{pmatrix} 0.4 & 0.3 & 0.1 & 0.2 \\ 0.3 & 0.4 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.4 & 0.3 \\ 0.2 & 0.1 & 0.6 & 0.1 \end{pmatrix} = 0.3 \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} + 0.1 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$
$$+0.2 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + 0.1 \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} + 0.3 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.6}$$

The first four deterministic matrices are permutation matrices and the last one is not. In such RDS representation, it is easy to see that two trajectories starting from state 1 (or state 2) and state 3 (or state 4) will not synchronize into one, even though the Markov transition matrix is irreducible and aperiodic.

For the second question, it is very hard to calculate the distribution of synchronization steps due to deterministic transition matrices being non-commutative. We simulate various examples of Markov transition matrices in maximum metric entropy representation and min-max representation and histogram the frequency of synchronization steps. In Figure. 4.1, we give two examples of Markov transition matrices M_1 and M_2. They are

$$M_1 = \begin{pmatrix} 0 & 0.38 & 0.41 & 0.21 \\ 0 & 0 & 0 & 1 \\ 0.18 & 0.30 & 0.44 & 0.08 \\ 0.39 & 0.11 & 0 & 0.5 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 0 & 0.86 & 0.14 \\ 0.14 & 0.51 & 0 & 0.35 \\ 0.09 & 0 & 0.38 & 0.53 \\ 0 & 0.47 & 0.46 & 0.07 \end{pmatrix}. \tag{4.7}$$

Both matrices are irreducible and aperiodic so maximum-metric-entropy representation will synchronize. We also verified min-max representations synchronize through finding the absorbing states after constructing the 16×16 matrixes W in (4.2). From the Figure. 4.1, the maximum-metric-entropy representation synchronizes faster than the min-max representation on average for the first Markov transition matrix; however, the min-max one is faster

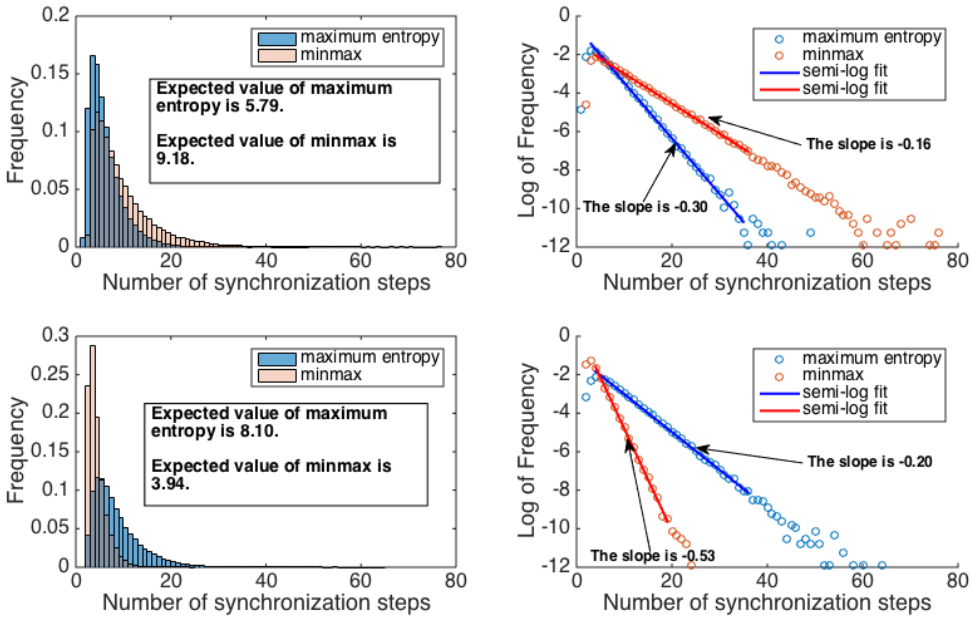

Figure 4.1: The top left and bottom left figures are the histograms of the synchronization steps for Markov transition matrices M_1 and M_2 after 150,000 independent simulations in each representation. The top right and bottom right figures are the logarithm of the frequency scatter plots of the histograms with the linear fits.

on average for the second. The tail frequency for both representations shows an exponential decaying, as illustrated by the semi-log linear fits of the tail probability $p(x) \propto a \exp(-bx)$. The slope b is related to the second largest eigenvalue of the transition matrix of the two-point motion, W.

This synchronization has found an application in sampling. Here is the short overview of perfect sampling. In Eq. 2.10, we discuss the pullback product may have the limit, however, the push forward doesn't. Consider $Y_t(\omega)$ starting with the whole state space $\mathscr{S}$, $Y_t(\omega) = \varphi(t, \theta(-t)\omega)\mathscr{S}$, which are n simultaneous sequences starting with state $1, \ldots, n$, $\mathrm{supp}(Y_t(\omega))$ is non-increasing. Once multiple sequences collide at some instance, they will be together forever. For a fixed ω, $\lim_{t\to+\infty} \mathrm{supp}(Y_t(\omega))$ may be smaller than $\mathscr{S}$, even can be a

singleton (with some assumptions). This means these n simultaneous sequences synchronize into one sequence. If the support of $Y_t(\omega)$ as $t \to +\infty$ is almost surely a singleton, it is equivalent with the RDS synchronization, i.e, for any different initial states x_1 and x_2, i.e, $x_1 \neq x_2$, $\lim_{t\to+\infty} \Pr(\omega : \varphi(t,\omega)x_1 = \varphi(t,\omega)x_2) = 1$. Then the limit $Y_\infty(\omega) = \lim_{t\to+\infty} Y_t(\omega)$ exists almost surely and is ω-dependent. Moreover, $Y_\infty(\omega)$ follows the invariant distribution $\boldsymbol{\pi}$ since $X_t(\omega)$ follows the invariant distribution $\boldsymbol{\pi}$ as $t \to +\infty$, but the limit doesn't exist. This is exactly the idea of *the coupling from the past* or *perfect sampling* [90]. The algorithm works because $Y_\infty(\omega)$ can be sampled in finite time. There exists some finite $t_0(\omega)$ such that $\varphi(t, \theta(-t)\omega)\mathscr{S}$ is singleton for all $t \geq t_0(\omega)$ almost surely. Then this singleton is $Y_\infty(\omega)$ for this given ω and exactly has the law of $\boldsymbol{\pi}$. So this method is also called perfect sampling. However, not every finite RDS have such properties and one sufficient condition is the RDS is monotone and ergodic [101].

4.2 Synchronization in Continuous RDS

Here I give an example on one-dimensional continuous space RDS and illustrate the connection between the synchronization and the Lyapunov exponents.

Example 4.2. Let $f : \mathbb{S}^1 \to \mathbb{S}^1$ be a smooth map and $F : \mathbb{R} \to \mathbb{R}$ a lift of f. Here $F(x) = x + a\sin(2\pi x)$ with $a > 0$ and identify $\mathbb{S}^1$ with $\mathbb{R}\backslash\mathbb{Z}$. x here is identified as the angle of rotation which has the domain $[0, 1)$. We define the family of the map $f_\phi(x) = F(x+\phi)-\phi \mod 1$ with $\phi \in [0, 1)$. It can be considered as the conjugation of the map F by the rotation ϕ, i.e, rotating ϕ angle, applying the map F and rotating $-\phi$ angle back. Define the i.i.d RDS $\varphi(t,\omega) : \mathbb{S}^1 \to \mathbb{S}^1$ as follows,

$$\varphi(t,\omega) = f_{\phi_{t-1}} \circ \cdots \circ f_{\phi_0} \tag{4.8}$$

where $\omega = (\ldots, \phi_0, \phi_1, \ldots, \phi_{t-1}, \ldots)$ and the probability measure is $\mathbb{P} = \lambda^{\mathbb{Z}}$ such that these angles are sampled in the i.i.d. manner with the uniform distribution. It is clear that the Lebesgue measure λ on $\mathbb{S}^1$ is a stationary distribution for the Markov transition probabilities

associate to RDS $\varphi(t,\omega)$. In fact, as far as F is not the rational rotation, the Lebesgue measure is the only stationary distribution. The Lyapunov exponent is defined as

$$\lambda(\omega) = \lim_{t\to+\infty} \frac{1}{t} \log \|\mathrm{d}\varphi(t,\omega)x_0\| \tag{4.9}$$

In fact,

$$\frac{1}{t} \log \|\mathrm{d}\varphi(t,\omega)x_0\| = \frac{1}{t} \log \|f'_{\phi_{t-1}}(x_{t-1}) f'_{\phi_{t-2}}(x_{t-2}) \dots f'_{\phi_0}(x_0)\| \tag{4.10}$$

$$= \frac{1}{t} \sum_{i=0}^{t-1} \log |f'_{\phi_i}(x_i)| = \frac{1}{t} \sum_{i=0}^{t-1} \log |F'(x_i+\phi_i)| \tag{4.11}$$

So, by ergodic theorem, the time average is equal to the space average for almost every ω,

$$\lim_{t\to+\infty} \frac{1}{t} \sum_{i=0}^{t-1} \log |f'_{\phi_i}(x_i)| = \int_0^1 \log(|F'(x)|)\mathrm{d}x, \text{ a.s.} \tag{4.12}$$

So the Lyapunov exponent $\lambda(\omega)$ is the constant $\int_0^1 \log(|F'(x)|)\mathrm{d}x$ almost surely.

Here in this example, $|F'(x)| = |1 + 2a\pi\cos(2\pi x)|$. This deterministic map $F(x)$ has different properties with different a value, as summarized as follows,

1. $0 < a < \frac{1}{2\pi}$, $F'(x) > 0$ so the map F is diffeomorphism. The fixed point $x = 1/2$ is stable since $|F'(x)| < 1$.

2. $\frac{1}{2\pi} < a < \frac{1}{\pi}$, $F'(x)$ could be negative, so the map is an injective map. The fixed point $x = 1/2$ is still stable since $|F'(x)| < 1$.

3. $\frac{1}{\pi} < a < a^*$, the map F has the period doubling property, i.e, the value will oscillations among 2^n values.

4. $a^* < a$, the dynamics with most values of a beyond this region will exhibit chaotic behavior. Some of the value will have positive Lyapunov exponents.

In Fig. 4.2, it plots the Lyapunov exponents for this deterministic map with the parameter a.

On the other hand, the Lyapunov exponent $\lambda = \int_0^1 \log(|F'(x)|)\mathrm{d}x$, is negative when $a < \frac{1}{\pi}$ and is positive when $a > \frac{1}{\pi}$. So $a = \frac{1}{\pi}$ is the critical point for the noise-induced synchronization and when the parameter a is beyond this point, the synchronization phenomenon disappears. Because after a little bit calculus drill, you can show $\int_0^1 \log|1+2\cos(2\pi x)|\mathrm{d}x = 0$ and $\int_0^1 \log|1+\cos(2\pi x)|\mathrm{d}x = -\log 2$. Fig. 4.3 shows the Lyapunov exponents for RDS with the parameter a. Although this deterministic map is considered as one sample sequence in RDS, Lyapunov exponent for the deterministic map is not relevant with the one for RDS. Because this sequence is not typical in RDS. In Fig. 4.4, $a = 0.3$ is in the synchronization region and $a = 0.37$ is not. Even though for both a, the fixed point or the period-2 cycle are in the contracting region which is stable, the weighted expanding region in the right figure is more than the weight contracting region. The left figure is not. The random angle rotation ϕ in RDS will take the state out of the stable region again and again. In fact, since the invariant measure is Lesbegue, each expanding or contracting region will be visited with the equal probability. Therefore, during the competition of expanding and contracting in the RDS, more contracting than expanding will result in noise-induced synchronization. In Fig. 4.5 and Fig. 4.6, we sampled two sample trajectories $X_1(t)$ and $X_2(t)$ driven by the same sequence of maps, but starting from different initial points. For $a = 0.3$, the distance of these two trajectories decreases asymptotically, while at some steps it may increases. In the end, the contraction in the dynamics dominates the expansion. On the other hand, for $a = 0.37$, the expansion in the dynamics dominates the contraction and it shows the non-synchronization behavior.

One may wonder that the connection between the bifurcation point from fixed point to period-2 cycle is also $a = \frac{1}{\pi}$. In fact, it is purely coincident. I can give another example, such that the attractor is the fixed point but doesn't have synchronization behavior. In Fig. 4.7, the slopes of this piecewise linear function are 13/5, -3/5 and 13/5. The Lyapunov exponent of RDS is positive and RDS doesnt synchronize, however, the deterministic map has the single stable fixed point.

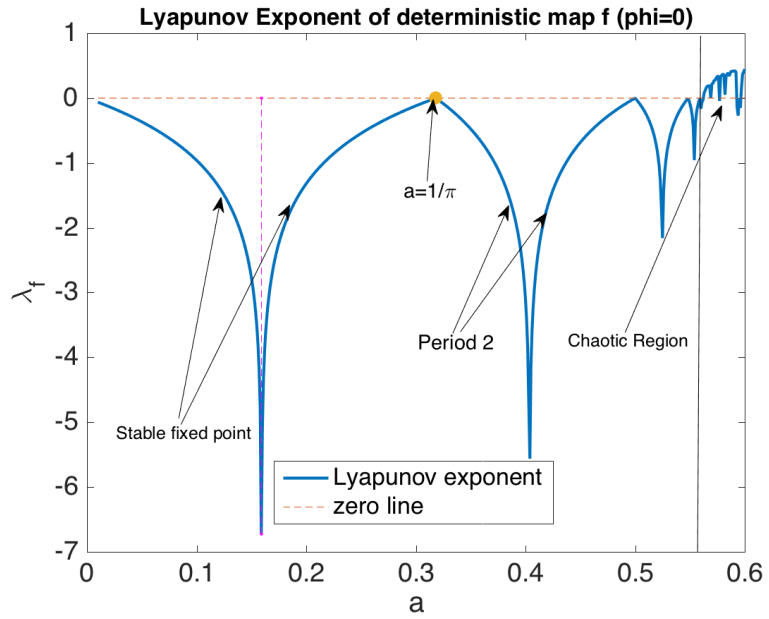

Figure 4.2: The Lyapunov exponent of deterministic map $F(x)$

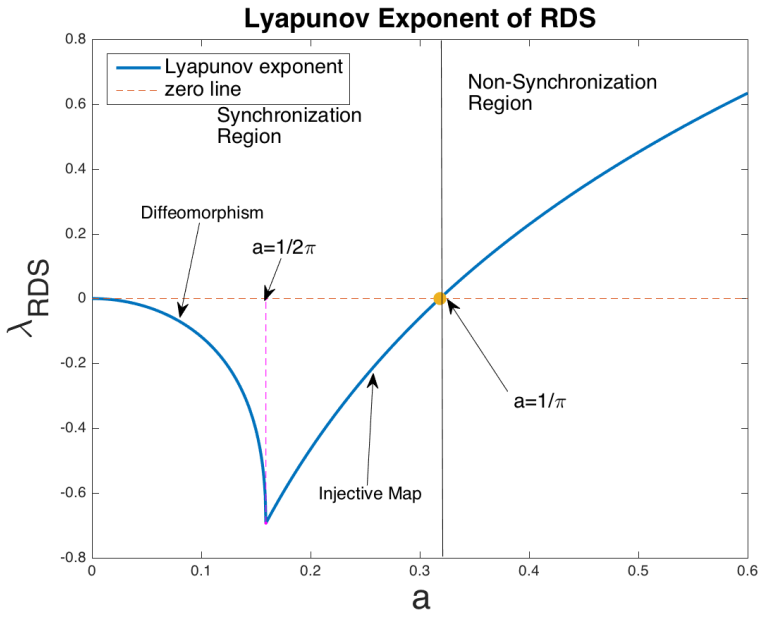

Figure 4.3: The Lyapunov exponent of RDS

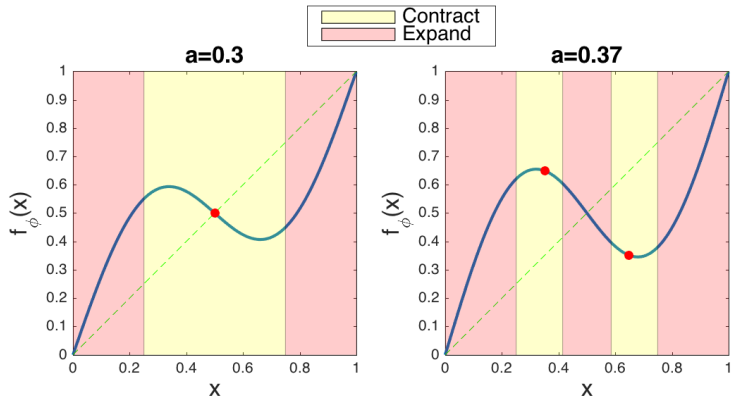

Figure 4.4: The expanding and contracting regions for $a = 0.3$ and $a = 0.37$

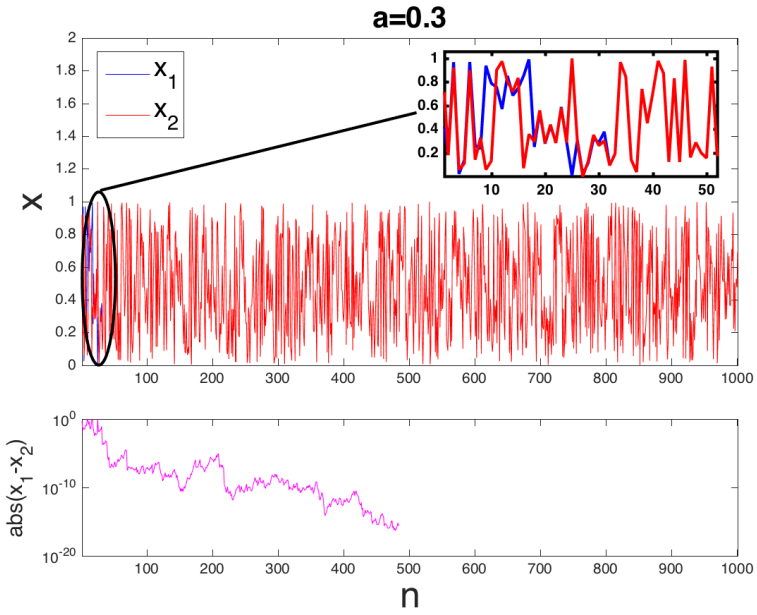

Figure 4.5: The noise-induced synchronization phenomenon for $a = 0.3$.

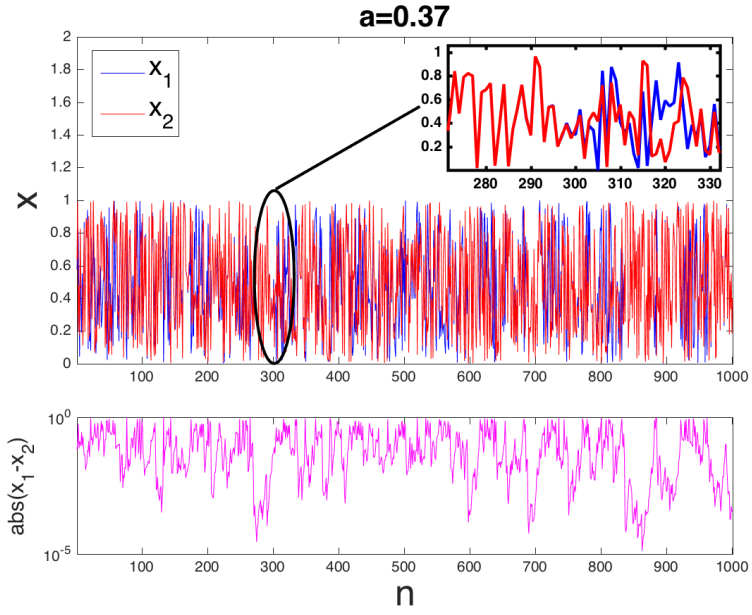

Figure 4.6: The non-synchronization phenomenon for $a = 0.37$.

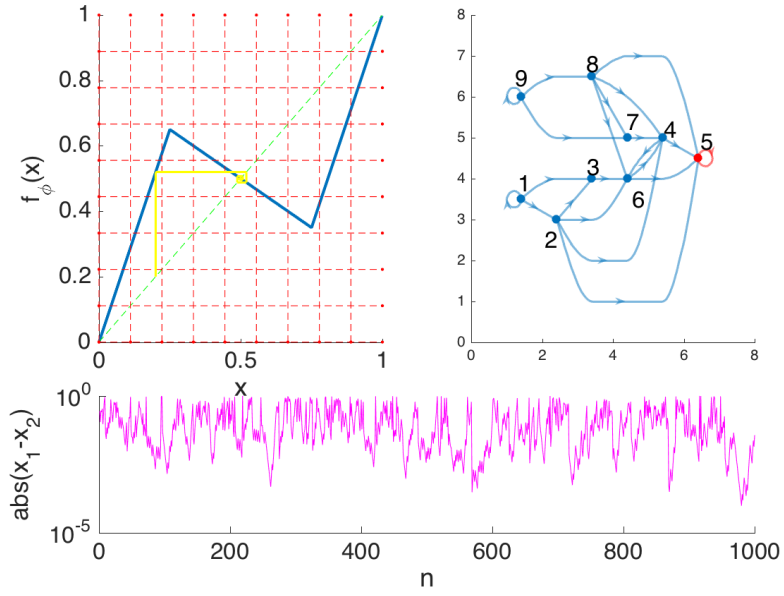

Figure 4.7: Another example.

4.3 Multiplicative Ergodic Theorem

In Sec. 4.2, I discussed the one-dimensional RDS and the Lyapunov exponent is constant almost surely due to ergodic theorem. In particular, in Eq. 4.10, $f'_{\phi_i}(x_i)$ is a scalar and I can rearrange the order. After taking the logarithm, it turns the multiplication into addition such that ergodic theorem can be applied. However, in multiple-dimension case, the ergodic theorem is not enough. It is important to study the asymptotic properties of products of a large number of random matrices [31].

The degenerate case is to consider the product of deterministic matrix, i.e, A^t, then, one can show $\lim_{t\to+\infty}\frac{1}{t}\log\|A^t\| = \max_i \log(|\sigma_i|)$, where σ is the spectrum of M. Here the matrix norm is generic. It implies that the largest (in modulus) eigenvalue determines the growth of the vector with nonzero projection on the eigenvector corresponding to the eigenvalue. Another simple case is given by the family of commutative diagonalizable matrices, M_i, with probability p_i, i.e., $A_iA_j = A_jA_i$. Each $A_s = S\Sigma(s)S^{-1}$ performs a simultaneous diagonalization, where Σ_i is the diagonal matrix. The products of i.i.d. random matrices has the following expression, $\Pi_{s=1}^t A_s = S(\Pi_{s=1}^t\Sigma(s))S^{-1}$. Due to law of large number, $\lim_{t\to+\infty}\frac{1}{t}\log\|\Pi_{s=1}^t M_s\| = \max_i\sum_s \log|\sigma_i(s)|p_s$ where $\sigma_i(s)$ is the i-th entry of the diagonal matrix.

The surprising fact is this can be generalized to the case of independent random matrices, even with noncommuting random matrices A_i. The ordering becomes important and the arguments for scalar don't apply any more. First, I need to define it more rigorously.

Consider the state space $\mathscr{S} = \mathbb{R}^n$, Γ is the family of maps from $\mathbb{R}^n$ to $\mathbb{R}^n$. The cardinality of Γ can be finite, countable or continuum. Let Q be the probability measure on the σ-field of Γ. The map $\varphi(t,\omega)\mathbf{u}_0 = \alpha_{t-1}\ldots\alpha_0(\mathbf{u}_0)$ with $\mathbf{u}_0 \in \mathbb{R}^n$. The derivative of the map is $\mathrm{d}\varphi(t,\omega)\mathbf{u}_0$, which is $\frac{\mathrm{d}\alpha_{t-1}}{\mathrm{d}x}(\mathbf{u}_{t-1})\frac{\mathrm{d}\alpha_{t-2}}{\mathrm{d}x}(\mathbf{u}_{t-2})\ldots\frac{\mathrm{d}\alpha_0}{\mathrm{d}x}(\mathbf{u}_0)$ and $\frac{\mathrm{d}\alpha_i}{\mathrm{d}x}(\mathbf{u}_i)$ is the Jacobian matrix of the map α_i evaluated at $\mathbf{u}_i$. In the simplest case, these maps are linear maps and the probability measure $\mathbb{P}$ is Bernoulli measure, $\mathbb{P} = Q^{\mathbb{Z}}$, then $\frac{\mathrm{d}\alpha_i}{\mathrm{d}x}(\mathbf{u}_i)$ is the constant matrix. Therefore, the derivative of the map $\mathrm{d}\varphi(t,\omega)\mathbf{u}_0$ is the production of i.i.d. random matrices.

Define these matrices $A : \Omega \to M_{n\times n}(\mathbb{R})$, $\mathrm{d}\varphi(t,\omega)\mathbf{u}_0 \triangleq P_t(\omega) = \Pi_{i=1}^{t} A_i$ where A_i are i.i.d. random matrices with probability measure Q. Starting with the nonzero test vector $\mathbf{v}$, the Lyapunov exponent defined as

$$\lambda_1 = \lim_{t\to+\infty} \frac{1}{t} \log \|P_t(\omega)\mathbf{v}\| \tag{4.13}$$

exists and is a nonrandom quantity almost everywhere. The quantity is called maximum Lyapunov exponent. This is the called Furstenberg-Kesten theorem [41, 42]. The only condition needed here turns out to be $\log^+ \|A_\omega\|$ is integrable, i.e, $\int \log^+ \|A_\omega\| \mathrm{d}\mathbb{P} < \infty$, where $\log^+ \|A\| = \max\{\log \|A\|, 0\}$. Roughly speaking, there is a subclass Ω^* of typical sequences which has full measure over Ω and a generic vector $\mathbf{v}$ grows exponentially with rate λ_1. Although there exist very improbable sequences in Ω which lead to a different asymptotic limit of the growth rate, they don't change the almost sure convergence in (4.13) [31]. Later in this section, I will discuss their relevance for other growth rate in multiplicative ergodic theorem. But in general, it is difficult to determine a particular sequence ω belongs to or not to Ω^* without knowing the explicit value of its asymptotic growth rate. Moreover, in most cases, the analytical results are out of reach. To get an idea of the difficulties, here is an example.

Example 4.3. Consider the random matrix A by setting $A = R_{\pi/2} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ with probability $p \leq 1$, and $A = H_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$ with probability $1-p$. Let $\{A_t\}$ be an i.i.d. sequence with the same law as A. It turns out the maximum Lyapunov exponent for $p > 0$ is 0.

To understand this, we will apply a test vector $\mathbf{v}_0 = (0,1)^T$ and let $\mathbf{v}_t = P_t\mathbf{v}_0$. The nature of $R_{\pi/2}$ and H_2 leads to one of entries of $\mathbf{v}_t$ is 0. Denote $\mathbf{v}_t$ in state 1 if the first entry is non-zero and $\mathbf{v}_t$ in state 2 if the second entry is non-zero. This two states dynamics is a Markov Chain. The Markovian property is given as follows: The matrix $R_{\pi/2}$ takes one state to another and the matrix H_2 remains the same state. The diagram above illustrates

this Markov Chain. The transition matrix M is

$$M = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \tag{4.14}$$

In each sequence in the Markov chain, the pair of 11 doubles the vector norm $\|\mathbf{v}\|$ due to the first entry of H_2 and the pair of 22 cuts the vector norm by half. So for each ω',

$$\lim_{t\to+\infty} \frac{1}{t} \log \|\mathbf{v}_t\| = \lim_{t\to+\infty} \frac{1}{t}(\#(11)\log(2) - \#(22)\log(2)) \tag{4.15}$$

Moreover,

$$\lim_{t\to+\infty} \frac{1}{t}\#(11)\log(2) \quad \to \quad \pi_1 M_{11}\log(2) \tag{4.16}$$

$$\lim_{t\to+\infty} \frac{1}{t}\#(22)\log(2) \quad \to \quad \pi_2 M_{22}\log(2) \tag{4.17}$$

where π_1 and π_2 are stationary density of state 1 and state 2. This limit is because MC is irreducible and aperiodic and the time average is replaced by the space average. So $\lim_{t\to+\infty} \frac{1}{t} \log \|\mathbf{v}_t\| = \pi_1 p_{11}\log(2) - \pi_2 p_{22}\log(2) = 0$.

One can get the same result for another test vector $\mathbf{v}_0' = (1,0)^T$. In addition,

$$\lim_{t\to+\infty} \frac{1}{t} \log \|A_{n-1}\ldots A_0\| \leq \lim_{t\to+\infty} \max\{\frac{2}{t}\log\|P_t\mathbf{v}_0\|, \frac{2}{t}\log\|P_t\mathbf{v}_0'\|\} = 0 \tag{4.18}$$

So we can proof the maximum Lyapunov exponent is 0. On the other hand, when $p = 1$, the Lyapunov exponent is $\lambda = \log 2$. So the Lyapunov exponent is a discontinuous function of p. In fact, more recent results show that in i.i.d. case, the Lyapunov exponent is in general continuous with respect to p and the matrix A as far as p doesn't vanish [18, 116].

In most cases, the top Lyapunov exponent will be positive and there are very few cases that have zero Lyapunov exponent. But the eigenvalues of these matrices are irrelevant with the Lyapunov exponents [31]. Here is the example to illustrate this fact.

Example 4.4. Consider the random matrix A by setting $M_1 = \begin{pmatrix} x & -1 \\ 1 & 0 \end{pmatrix}$ with probability $1/2$, and $M_2 = \begin{pmatrix} y & -1 \\ 1 & 0 \end{pmatrix}$ with probability $1/2$ with $|x| < 2$ and $|y| < 2$. Let $\{A_t\}$ be an

i.i.d. sequence with the same law as A. The eigenvalues of A are complex with modulus 1, so $\mathbf{v}_t = M_1^t \mathbf{v}_0$ and $\mathbf{v}'_t = M_2^t \mathbf{v}_0$ don't increase exponentially with t. But from Furstenberg's result [42], the Lyapunov exponent λ_1 is positive. The basic ideas are both matrices are in the group of $SL(2,\mathbb{R})$, i.e, $\det(M_1) = \det(M_2) = 1$. The criteria for positive Lyapunov exponents here are:

1. There is no $C \in \mathrm{GL}(2,\mathbb{R})$ such that CMC^{-1} is an orthogonal matrix for M_1, M_2.

2. There is no finite set L in the unit circle, such that $M(L) = L$ for M_1, M_2. In the previous example, there exists a set $L = \{0, \pi/2, \pi, 3\pi/2\}$ and $A(L) = L$ for all A.

Now I will present the idea of Furstenberg's theorem for product of i.i.d. random 2×2 matrices. The effect of the vector multiplying the matrix $\mathbf{v}_1 = A\mathbf{v}_0$ can be decomposed into two part: the stretch of the vector norm $\|\mathbf{v}_0\| \to \|\mathbf{v}_1\|$ and the rotation of the argument $\Theta_0 \to \Theta_1$. It is well expressed in the polar coordinate representation of $\mathbf{v}$. We can consider $\mathbf{v}$ and $-\mathbf{v}$ are in the same equivalent class since the sign will not matter with the matrix norm. That means $\Theta = \Theta + \pi$. First, I am interested in the dynamics of the rotation of the argument on the *projective space* $P^1 = [0, \pi)$ of $\mathbb{R}^2$. The projective space here is simply the half unit circle. The dynamics of argument as follows,

$$\overline{\mathbf{u}}_{\Theta_{t+1}} = \overline{A_t \mathbf{u}_{\Theta_t}} \tag{4.19}$$

Here I write $\mathbf{u}_\Theta = (\cos(\Theta), \sin(\Theta))$ and for $\mathbf{v} \in \mathbb{R}^2 \backslash \{0\}$, I write $\overline{\mathbf{v}} \in P^1$ for the equivalence class of $\mathbf{v}$. Then Θ_t follows the Markov chain which is referred as the *projective Markov chain* for the matrix product P_t.

The transition probability $\Pr(\Theta, \cdot)$ for the projective Markov chain with $\Theta \in P^1$ is given by

$$\Pr(\Theta, G) = Q(A : \overline{A\mathbf{u}_\Theta} \in G) \tag{4.20}$$

where $G \subset P^1$. The invariant measure μ for the projective Markov chain is defined as follows,

$$\int_{P^1} P(\Theta, G) \mathrm{d}\mu(\Theta) = \mu(G) \tag{4.21}$$

If the projective Markov chain is ergodic, then the the law of argument Θ_t will converge to this invariant measure μ. For the given matrix sequence ω, the stretch of the vector norm $\log \frac{\|A_0(\omega)\mathbf{v}\|}{\|\mathbf{v}\|} = \log \|A_0(\omega)\mathbf{u}_\Theta\|$ is a function of the argument of $\mathbf{v}$, which is Θ, denoted as $f(\omega, \Theta)$. Then the asymptotic growth of the vector norm is,

$$\frac{1}{t}\log \|A_{t-1}\dots A_0\mathbf{u}_{\Theta_0}\| = \frac{1}{t}(f(\omega,\Theta_0) + f(\theta(1)\omega,\Theta_1) + \cdots + f(\theta(t-1)\omega,\Theta_{t-1})) \quad (4.22)$$

where Θ_t is given by Eq. 4.19 and θ is the left shift operator. By ergodic theorem, the time-averaging is equal to the space-averaging.

$$\lim_{t\to+\infty}\frac{1}{t}\sum_{i=0}^{t-1} f(\theta(i)\omega,\Theta_i) = \int\int f(\omega,\Theta)\mathrm{d}\mathbb{P}\mathrm{d}\mu = \int_{P^1}\int_A \log\|A\mathbf{u}_\Theta\|\mathrm{d}Q(A)\mathrm{d}\mu(\Theta) = \lambda \quad (4.23)$$

$\int_A \log\|A\mathbf{u}_\Theta\|\mathrm{d}Q(A)$ is understood as the average stretch of vector norm at the fixed argument Θ. This averaging is taken on the matrices weighted with the probability. Then the second integral is the averaging with respect to the invariant measure of the argument. The Eq. 4.23 demonstrates the limit exists and is independent with ω almost surely.

But the Furstenberg-Kesten theorem neglects the finer structure given by the lower growth rate connected to the eigenvalues of $(P_t^*P_t)^{1/2}$, which is the singular values of P_t, different from the largest one. In the degenerate case, that is A is a fixed matrix and assume the spectrum of the matrix σ is not degenerate, the eigenvalues of $(P_t^*P_t)^{1/2}$ tend to be $|\sigma|^t$, so Eq. 4.13 will give n different values. If a nonzero vector $\mathbf{v}$ orthogonal to the eigenvector corresponding to the largest eigenvalue σ_1, then the limit in Eq. 4.13 will be $\leq \log|\sigma_2|$. The equal sign holds only if $\mathbf{v}$ has nonzero component on the second eigenvector. In general, if $\mathbf{v}$ has zero component on the eigenvectors corresponding to first $k-1$ eigenvalues, and nonzero on the others, then the limit is given by $\log|\sigma_k|$. Oseledec shows that such a hierarchy of the exponential growth of the vector norm can be extended to the random matrices production, which is called Lyapunov spectrum. Of course, the largest one is λ_1 in the Furstenberg-Kesten theorem.

The Oseledets multiplicative ergodic theorem has many versions: invertible, non-invertible [4] and semi-invertible [44].

Theorem 4.2. *(Invertible version) Let θ be an invertible ergodic measure-preserving transformation of a probability space $(\Omega, \mathbb{P})$. Let $A : \Omega \to GL(d, \mathbb{R})$ satisfy $\int \log \|A_\omega^{\pm 1}\| \mathrm{d}\mathbb{P} \leq \infty$. Then there exist $\lambda_1 > \lambda_2 > \cdots > \lambda_k > -\infty$ and subspaces $V_1(\omega), \ldots, V_k(\omega)$ such that:*

1. *V_i is a decomposition of $\mathbb{R}^d$: $V_1(\omega) \oplus \ldots V_k(\omega) = \mathbb{R}^d$.*

2. *$A(\omega)V_i(\omega) = V_i(\theta\omega)$, $\mathbb{P}$-a.s.*

3. *$\lim_{t\to+\infty} \frac{1}{t} \log \|P_t(\omega)\mathbf{v}\| = \lambda_i$, $\mathbb{P}$-a.s and all $\mathbf{v} \in V_i(\omega)\backslash\{0\}$, where $P_t(\omega) = A(\theta(t-1)\omega) \ldots A(\omega)$.*

Theorem 4.3. *(Non-invertible version) Let θ be a not-necessarily invertible ergodic measure-preserving transformation of a probability space $(\Omega, \mathbb{P})$. Let $A : \Omega \to M(d, \mathbb{R})$ satisfy $\int \log \|A_\omega\| \mathrm{d}\mathbb{P} \leq \infty$. Then there exist $\lambda_1 > \lambda_2 > \cdots > \lambda_k \geq -\infty$ and subspaces $U_1(\omega), \ldots, U_k(\omega)$ such that:*

1. *U_i is a flag: $\mathbb{R}^d = U_1(\omega) \supset U_2(\omega) \supset U_k(\omega) = \{0\}$.*

2. *$A(\omega)U_i(\omega) \subset U_i(\theta\omega)$, $\mathbb{P}$-a.s.*

3. *$\lim_{t\to+\infty} \frac{1}{t} \log \|P_t(\omega)\mathbf{v}\| = \lambda_i$, $\mathbb{P}$-a.s and all $\mathbf{v} \in U_i(\omega)\backslash U_{i+1}(\omega)$, where $P_t(\omega) = A(\theta(t-1)\omega) \ldots A(\omega)$.*

Sometimes these random matrices are singular but θ is still invertible. So here is the semi-invertible version, whose result is as strong as the invertible one.

Theorem 4.4. *(Semi-invertible version) Let θ be an invertible ergodic measure-preserving transformation of a probability space $(\Omega, \mathbb{P})$. Let $A : \Omega \to M(d, \mathbb{R})$ satisfy $\int \log \|A_\omega\| \mathrm{d}\mathbb{P} \leq \infty$. Then there exist $\lambda_1 > \lambda_2 > \cdots > \lambda_k \geq -\infty$ and subspaces $V_1(\omega), \ldots, V_k(\omega)$ such that:*

1. *V_i is a decomposition of $\mathbb{R}^d$: $V_1(\omega) \oplus \ldots V_k(\omega) = \mathbb{R}^d$.*

2. *$A(\omega)V_i(\omega) = V_i(\theta\omega)$, $\mathbb{P}$-a.s.*

3. $\lim_{t\to+\infty}\frac{1}{t}\log\|P_t(\omega)\mathbf{v}\| = \lambda_i$, $\mathbb{P}$-*a.s and all* $\mathbf{v}\in V_i(\omega)\backslash\{0\}$, *where* $P_t(\omega) = A(\theta(t-1)\omega)\dots A(\omega)$.

In terms of finite RDS, these random matrices are deterministic transition matrices, which in general singular. So we have more refined result,

Proposition 4.1. *Let P be the induced linear cocycle over $(\Omega,\mathcal{F},\mathbb{P},\theta)$ in finite RDS acting on the state space $\mathbb{R}^n$ equipped with the standard Euclidean norm $\|\cdot\|$. Then for any $\omega\in\Omega$ and any $\mathbf{v}\in\mathbb{R}^n$,*

$$\lambda(\omega,\mathbf{v}) := \lim_{t\to+\infty}\frac{1}{t}\log\|P_t(\omega)\mathbf{v}\|$$

exists and equals either 0 or $-\infty$, and consequently, the Lyapunov exponents of P take at most two values $\lambda_1 = 0$ or $\lambda_2 = -\infty$. Moreover, the corresponding multiplicities of λ_1,λ_2 satisfy inequalities

$$m_1(\omega)\geq 1,\ m_2(\omega)\leq k-1,\ \mathbb{P}-a.e.\ \omega\in\Omega, \tag{4.24}$$

i.e., the Lyapunov exponent $\lambda_1 = 0$ is always attained.

Proof. We note that with n,ω varying, there are only a finite number of choices of $P_t(\omega)$. It follows that for any $\mathbf{v}\in\mathbb{R}^n$, $\|P_t(\omega)\mathbf{v}\|$ only take a finite number of different values for all $t\in\mathbb{N},\ \omega\in\Omega$. Now we let $\omega\in\Omega$ be fixed.

If $P_t(\omega)\mathbf{v}\neq\mathbf{0}$ for all $n\in\mathbb{N}$, then $\|P_t(\omega)\mathbf{v}\|$ have uniform positive upper and lower bounds, i.e., there exist constants $0<\kappa_1<\kappa_2<+\infty$ independent of t,ω such that $\kappa_1\leq\|P_t(\omega)\mathbf{v}\|\leq\kappa_2$ for all $t\in\mathbb{N}$. Thus, $\lim_{t\to+\infty}\frac{1}{t}\log\|P_t(\omega)\mathbf{v}\|$ exists and equals 0.

If there exists $t\in\mathbb{N}$ such that $P_t(\omega)\mathbf{v}=\mathbf{0}$, then $P_s(\omega)\mathbf{v}=\mathbf{0}$ for any $s\geq t$. Thus $\lim_{t\to+\infty}\frac{1}{t}\log\|P_t(\omega)\mathbf{v}\| = -\infty$.

We now argue that $m_1(\omega)\geq 1$ for $\mathbb{P}$-a.e. $\omega\in\Omega$, i.e., the Lyapunov exponent $\lambda_1=0$ is always attained. For otherwise, there exists an $t\in\mathbb{N}$ such that $P_t(\omega)\mathbf{v}=0$ for all $\mathbf{v}\in\mathbb{R}^n$. This is impossible since $P_t(\omega)$ is not a zero matrix. Since $m_1(\omega)\geq 1$ and $m_1(\omega)+m_2(\omega)=n$, we obtain the other inequality in (4.24).

□

Similar results are being extended to countable state RDS as well. These are still in preparation. The Lyapunov spectrum can be used to characterize the sufficient and necessary condition for synchronization in finite RDS.

Theorem 4.5. *The finite RDS synchronizes if and only if P admits precisely two Lyapunov exponents $\lambda_1 = 0$, $\lambda_2 = -\infty$ with respective multiplicities*

$$m_1(\omega) = 1, \quad m_2(\omega) = n - 1, \ \mathbb{P} - a.e.\ \omega \in \Omega. \tag{4.25}$$

if λ_2 exists, then for $\mathbb{P}$-a.e. $\omega \in \Omega$, the spectral subspace

$$V(\omega) =: \{\mathbf{v} \in \mathbb{R}^n : \lambda(\omega, \mathbf{v}) = -\infty\} \tag{4.26}$$

exists and $m_2(\omega) = \dim V(\omega)$. To prove this theorem, we need to start with the following lemma,

Lemma 4.6. *If λ_2 exists, then for $\mathbb{P}$-a.e. $\omega \in \Omega$,*

$$V(\omega) \subseteq E_0 =: \mathbf{v} = (v_1, \cdots, v_n)^\top : \sum_{i=1}^{n} v_i = 0\}. \tag{4.27}$$

Proof. Let $\mathbf{v} = (v_1, \cdots, v_n)^\top \in V(\omega)$. Then there exists $t \in \mathbb{N}$ such that $P_t(\omega)\mathbf{v} = \mathbf{0}$. If we denote $P_t(\omega)\mathbf{v} = \mathbf{u} = (u_1, ..., u_n)^\top$, then $\sum \lim_{i=1}^{n} u_i = \sum_{i=1}^{n} v_i$. Hence $\mathbf{v} \in E_0$. □

Now it is ready to prove Thm. 4.5.

Proof. Suppose the finite RDS synchronizes. Then for $\mathbb{P}$-a.e. $\omega \in \Omega$, there are integers $t_\omega \in \mathbb{N}$, $\ell_\omega \in \{1, \cdots, n\}$ such that for all $s \geq t_\omega$,

$$P_t(\omega)s_i = s_{\ell_\omega}, \ \ i = 1, \cdots, n.$$

For given ω and $s \geq t_\omega$, it follows that the matrix $P_s(\omega)$ has every entries on the ℓ_ω-th row being 1 and all other entries being 0. Let $\mathbf{v} \in E_0$, where E_0 is the co-dimension-1 hyperplane defined in (4.27). We then have $P_s(\omega)\mathbf{v} = \mathbf{0}$, i.e., $\mathbf{v} \in V(\omega)$, where $V(\omega)$ is the spectral subspace defined in (4.26). Hence $E_0 \subseteq V(\omega)$, and by Lemma 4.6, we actually have

$$V(\omega) = E_0. \tag{4.28}$$

It follows that $m_2(\omega) = \dim V(\omega) = \dim E_0 = n-1$ and $\lambda_2 = -\infty$ is attained. By Proposition 4.1, the Lyapunov exponent $\lambda_1 = 0$ is always attained. Since $m_2(\omega) = n-1$, we have $m_1(\omega) = 1$.

Now suppose (4.25) holds. Then $\dim V(\omega) = n-1$, $\mathbb{P}$-a.e. $\omega \in \Omega$. It follows from Lemma 4.6 that (4.28) holds. Let $s_i,\ s_j$ be any two distinct elements of $\mathscr{S}$ and denote by e_i, respectively e_j, the i-th, respectively the j-th, standard unit vector in $\mathbb{R}^n$. Since $e_i - e_j \in E_0$, we have by (4.28) that $e_i - e_j \in V(\omega)$ for $\mathbb{P}$-a.e. $\omega \in \Omega$. It follows from (4.26) that, for a fixed such ω, there exists $t(\omega)$ sufficiently large such that for all $t \geq t(\omega)$,

$$P_t(\omega)(e_i - e_j) = \mathbf{0},$$

i.e.,

$$P_t(\omega)e_i = P_t(\omega)e_j.$$

This show that any pair of elements in $\mathscr{S}$ synchronizes, hence the finite RDS synchronizes.

□

Similar results could be extended to countable state RDS as well.

Chapter 5

APPLICATION OF RDS: STATISTICAL INFERENCE IN HIDDEN MARKOV MODEL

Hidden Markov model (HMM) and its variants have seen wide applications in time series data analysis. It is assumed in the model that the observation variable Y probabilistically depends on the latent variables X with *emission distribution* $p(y_t|x_t)$ at each time t. The underlying probability of the discrete random variables X follows a Markov chain with *transition probability* $p(x_t|x_{t-1})$ [96]. HMM is the simplest dynamic Bayesian network and has proven a powerful model in many applied fields including speech recognition [53, 96], computational biology [64, 65, 106], machine translation [83, 84], cryptanalysis [55] and finance [14]. Model parameters and hidden variables are inferred for prediction or classification tasks.

Traditionally, model parameters and hidden variables are estimated iteratively for the HMMs through the celebrated Baum-Welch algorithm [81]. For this maximum likelihood estimation, a forward-backward procedure is used which computes the posterior marginals of all hidden state variables given a sequence of observations. Later, Bayesian algorithms are also developed through forward filtering backward sampling algorithm and variational Bayes method which handles conjugate emission models on the natural parameter space in a similar vein as the Baum-Welch algorithm.

In all the aforementioned approaches for inference in HMMs, marginalization over hidden variables is involved. This step is the crux of the computation burden. For long observation sequences, this step causes problems of scalability, computation error, and even numerical stability in inference for HMMs [40, 59, 81]. Hence an important question is: can one only use part of the data to approximate marginal likelihood over hidden variables of the entire chain, so that stochastic algorithms can be developed with controllable error?

To economize on computational cost at each iteration, we will take advantage of the memory loss property for the filtered state probability. The key idea is that successive blocks of sufficiently long subsequence observations can be considered almost independent of each other. In this paper, we make use of this memory loss property to approximate the predictive distribution of hidden states $p(x_t|y_{1:t})$ by only using part of the observation sequence $p(x_t|y_{t-B+1:t})$. This is achieved by formulating $p(x_t|y_{1:t-1})$ as a long sequence of heterogeneous matrices (comprised of emission probabilities and the transition probability) applied successively on an initial probability vector.

However, a critical question that needs to be answered is how long should the subsequence be? Though previous theory exists to quantify the length, the resulting lengths are often longer than the entire sequence which is practically not useful. So one needs to evaluate the rate of memory loss accurately and efficiently to control the length of the subsequence. If we recall the process of calculating filtered state probability, it can be considered as independent and identically distributed random matrix production if we treat observations as random events. We thereby make use of the random dynamical system (RDS) theory and describe the long time behavior of random matrices production with multiplicative ergodic theorem (MET), or Oseledec's theorem [4]. In particular, there exists the Lyapunov spectrum. Previous results showed the rate of memory loss is upper bounded by the gap of the top two Lyapunov exponents, $\lambda_2 - \lambda_1$ and is in fact realized almost surely [7, 8, 28]. In particular, the memory loss property requires the Markov chain to be irreducible and aperiodic and the emission distribution to be positive, such that the gap is strictly negative [68, 69]. In this work, we develop an algorithm to accurately and efficiently calculate this gap and the length of subsequence.

5.1 Overview of Hidden Markov Model

Hidden Markov models (HMM) are a class of discrete-time stochastic process $\{X_t, Y_t, t \geq 0\}$: $\{X_t\}$ is a latent discrete valued state sequence generated by a Markov chain, with values in the finite set $\{1, 2, \ldots, n\}$; $\{Y_t\}$ is corresponding observations generated from distributions

determined by the latent states X_t. Here it assumes Y_t taking values in $\mathbb{R}^d$, but it can easily extended to discrete states.

We can use the forward algorithm to compute the joint distribution $p(x_t, y_{1:t})$ by marginalizing over all other state sequences $x_{1:t-1}$. Y_t is conditionally independent of everything but X_t and X_t is conditionally independent of everything but X_{t-1}, i.e, $p(y_t|x_{1:t}, y_{1:t-1}) = p(y_t|x_t)$ and $p(x_t|x_{t-1}, y_{1:t-1}) = p(x_t|x_{t-1})$. The algorithm takes advantage of the conditional independence rules of HMM to perform the calculation recursively. With Bayes's rule, it follows,

$$p(x_t, y_{1:t}) = \sum_{x_{t-1}=1}^{n} p(y_t|x_t)p(x_t|x_{t-1})p(x_{t-1}, y_{1:t-1}) \tag{5.1}$$

In Eq. 5.1, $p(y_t|x_t)$ is called emission distribution with emission parameter $\{\phi_i\}_{i=1}^n$, $p(x_t|x_{t-1})$ is the transition probability of the Markov chain which is represented by a transition matrix M. In most cases, we assume M is primitive, i.e, the corresponding Markov chain is irreducible and aperiodic. We denote the parameter of interest as $\theta = \{M, \phi\}$. If the emission distribution is Gaussian distribution, then the emission parameters are the mean μ and the covariance σ. By using notation of vectors and matrix operations, this joint distribution can be represented by a row vector $\mathbf{p}_t = p(x_t, y_{1:t}|\theta)$ with jth component is $p(x_t = j, y_{1:t}|\theta)$. The forward algorithm can be calculated by the following non-homogenous matrix product.

$$\mathbf{p}_t = \mathbf{p}_0 M D_1 M D_2 \dots M D_t \tag{5.2}$$

where $\mathbf{p}_0$ is the initial state distribution. D_t is a diagonal matrix with jth entry as $D_{jj}(y_t) = p(y_t|x_t = j, \phi_j)$ which is the emission distribution assuming the current state is at j. Moreover, if one treats the observation y_t as random events, then $D(y_t)$ are random matrices sampled independently at each step. If one starts with invariant distribution of the Markov chain initially, $\boldsymbol{\pi}$, then these matrices are sampled in i.i.d. manner with probability density distribution $f(y)$.

$$f(y) = \sum_j \boldsymbol{\pi}_j p(y|x_t = j, \phi_j) \tag{5.3}$$

If the initial distribution is not $\boldsymbol{\pi}$, after an appropriate number of time steps, the distribution follows the invariant distribution and one can assume these matrices are sampled

in i.i.d. manner anyway. Now it is turned into a product of random matrices problem and these diagonal matrices randomly rescale the columns of M. $\mathbf{p}_t$ is called forward probability.

If we normalize the vector $\mathbf{p}_t$, it obtains the *filtered state probability*, $\boldsymbol{\rho}_t = p(x_t|y_{1:t}, \theta)$, which is not the invariant distribution of the Markov chain,

$$\boldsymbol{\rho}_t = \frac{p(x_t, y_{1:t}|\theta)}{p(y_{1:t}|\theta)} = \frac{\mathbf{p}_t}{\mathbf{p}_t \cdot \mathbb{1}} \tag{5.4}$$

This process is called filtering. The normalization constant $(\mathbf{p}_t \cdot \mathbb{1})$ gives the total probability for observing the given sequence up to step n irrespective of the final states, which is also called marginal likelihood $p(y_{1:t}|\theta)$. Not only this process ensures the numerical stability of random matrices production, but also $\boldsymbol{\rho}_t$ provides the scaled probability vector of being each state at step t. Note the probability vector $\boldsymbol{\rho}_t$ lives in a simplex, S^{n-1}, which is also called projective space in dynamical system, or space of measure in probability theory. Instead, the joint probability $\mathbf{p}_t$ is in $\mathbb{R}^{n+}$.

Another joint probability column vector $\mathbf{b}_i = p(y_{i+1:t}|x_i, \theta)$ is the probability of observing all future events starting with a particular state x_i. It can be computed by the backward algorithm similarly and it is called backward probability. We begin with $\mathbf{b}_t = \mathbb{1}$, and it gives

$$\mathbf{b}_i = MD_{i+1} \dots MD_t \mathbb{1} \tag{5.5}$$

It is again a product of random matrices. One can similarly renormalize the backward probability vector for better numerical stability, $\boldsymbol{\beta}_i = \mathbf{b}_i/(\mathbf{b}_i \cdot \mathbb{1})$ such that $\boldsymbol{\beta}_i \propto p(y_{i+1:t}|x_i, \theta)$. In fact, with forward and backward probability, we can calculate the probability $p(x_i|y_{1:t}, \theta) \propto \boldsymbol{\rho}_i^T \circ \boldsymbol{\beta}_i$ which is the Hadamard product of two vectors. In fact, the highest entry of this probability vector can give rough idea which latent state at step i lies.

5.2 Exponential Forgetting

Heuristically, in this very long heterogenous matrix multiplication (5.2), one observes that the final vector is irrelative to the initial vector and almost determined by the last several matrices multiplications, up to a normalization constant. As a matter of fact, if one is not

interested in the precise value of the final vector, the subsequence of matrices with length B are sufficient to approximate the vector. In more mathematical precise writings: Start any two different initial state probability vector $\mathbf{p}_0$ and $\mathbf{p}'_0$ and after applying exactly the same sequence of matrices, they generate two sequence of filtered state probability $\boldsymbol{\rho}_t$ and $\boldsymbol{\rho}'_t$. The distance of two sequence goes to 0 asymptotically almost surely, i.e,

$$\lim_{t\to+\infty} \|\boldsymbol{\rho}_t - \boldsymbol{\rho}'_t\| = 0 \ a.s. \tag{5.6}$$

This phenomenon is called *exponential forgetting of prediction filter* or *loss of memory of HMM*.

Example 5.1. In the figure 5.1, Markov chain has three state, emission distribution is a one-dimensional Gaussian on each state and the parameter ϕ is

$$M = \begin{bmatrix} 0.005 & 0.99 & 0.005 \\ 0.01 & 0.03 & 0.96 \\ 0.95 & 0.005 & 0.045 \end{bmatrix}, \mu = [0, 0.5, -0.5], \sigma = [1, 1, 1]$$

Starting with every point in the simplex as initial conditions, we apply these points by the same sequence of random matrices. One observes that the triangle consisting all points starts to shrink along n and after 40 steps, the triangle is contained within a small circle with radius ϵ. As n goes to $+\infty$, it will synchronize into a random fixed point, since it is sequence dependent. That implies if one allows error of ϵ, it may only requires the last 40 matrices which is irrelevant with the initial condition. So it significantly simplifies computational complexity.

If the diagonal matrices D_i are homogenous, it degenerates to the corollary of Perron-Frobenius theorem for primitive matrices. Now natural questions to arise are: what are conditions for such phenomenon and under these conditions, what is the rate of convergence. This rate in fact answers the critical question that how long the length B should be for a given ϵ. More questions about the rate are how to estimate the rate numerically or even

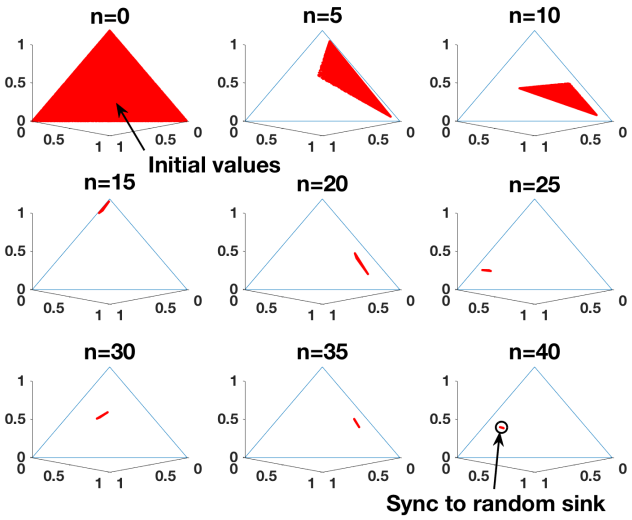

Figure 5.1: Starting with every point in the simplex, apply the same sequence of random matrices, and the triangle is contained within a small circle with radius ϵ after 40 steps.

analytically and does the rate continuously depends on the parameter θ. In fact, the sufficient conditions for this phenomenon are given in Le Gland *et al.* [68, 69],

Theorem 5.1. *If Markov transition matrix is primitive and the emission distribution $p(y_t|x_t)$ is strictly positive, then for any $\mathbf{p}_0, \mathbf{p}_0' \in S^{n-1}$, there exists a strictly negative $-c$ such that*

$$\limsup_{t\to+\infty} \frac{1}{t} \log \|\boldsymbol{\rho}_t - \boldsymbol{\rho}_t'\| \leq -c, \quad \textit{almost surely} \tag{5.7}$$

The theorem implies the filtered state probability forget almost surely their initial conditions exponentially fast and the rate is at least c. The techniques they used are Hilbert metric and Birkhoff contraction coefficient $\tau(M)$, which are extensively applied in non-negative matrix theory [47, 104]. Definitions of both terms are included in the Appendix D where it is also shown that $\tau(M) < 1$ for positive matrix M which is a sub-class of primitive matrix. It is a bit surprising that eigenvalues of each matrix in the heterogenous matrix production have little to do with this asymptotic behavior. In particular, one can construct a matrix sequence that spectrum radius of each is uniformly less 1, but the product doesn't even converge to 0. It is because the spectral radius doesn't process sub-multiplicity property, on

the other hand, this Birkhoff contraction coefficient does. Moreover, $\tau(M) = 0$ if and only if each row of M is a scalar multiple of the first row, which is also called weak ergodicity. At last, this coefficient is invariant with rescaling rows and columns of matrix. From these three properties, one immediately concludes when M is positive, the heterogeneous matrix production in Eq. 5.2 has the weak ergodicity and the exponential forgetting of the prediction filter follows with convergence rate $\log \tau(M)$. To further relax the positive matrix to primitive matrix, the approach is rather technical.

On the other hand, the long time behavior of random matrices production is well studied in *multiplicative ergodic theorem* (MET) through Lyapunov exponent in Chapter 4. It is the heart of RDS. Lyapunov exponent is exactly the generalization of absolute value of eigenvalues in the terms of random matrices production. Atar *et al.* [8] and Collet *et al.* [28] gave the exact convergence rate by Lyapunov exponents,

Theorem 5.2. *The rate of convergence of filtered state probability is the gap of the top two Lyapunov exponents almost surely,*

$$\limsup_{t\to+\infty} \frac{1}{t} \log \|\boldsymbol{\rho}_t - \boldsymbol{\rho}'_t\| = \lambda_2 - \lambda_1, \quad \textit{almost surely} \tag{5.8}$$

So the convergence rate is upper bounded by the gap between the first two Lyapunov exponents of the products of random matrices in (5.2) and in fact realized for almost all realizations. Furthermore, they showed this gap is strictly negative when the transition matrix is primitive and the emission distribution is positive. Then it recovered Le Gland's results. There is a nice connection between two results: Peres [88] proved the gap of the first two Lyapunov exponents, $\lambda_2 - \lambda_1$ in i.i.d random matrices production is upper bounded by $\log \tau(M)$. So for positive matrices, these two results connect with each other naturally.

The asymptotic limit of the rate of growth for the product of independent random matrices, $\lim_{t\to+\infty} \frac{1}{t} \log \frac{\|\mathbf{p}_t\|}{\|\mathbf{p}_0\|}$, is as been studied started at the beginning of the 60s. It has great relevance for development of the ergodic theory of dynamical system. Furstenberg and Kesten [41, 42] showed

Theorem 5.3.

$$\lambda_1 = \lim_{t\to+\infty} \frac{1}{t} \log \frac{\|\mathbf{p}_t\|}{\|\mathbf{p}_0\|}, \quad \textit{almost surely} \tag{5.9}$$

this limit λ_1 exists almost surely, moreover, it is nonrandom quantity and independent of the choice of metric and initial condition.

It is considered as the extension of strong law of large number to i.i.d random matrices [31]. This limit is called maximum Lyapunov exponent. It is rather surprising result since the order of sequence seems not much important even for non-commutative matrix multiplication. However, the Furstenberg-Kesten theorem neglects the finer structure given by the lower growth rates, other than the maximum Lyapunov exponent. Later Oseledets [86] showed there exists Lyapunov spectrum Λ, like eigenvalue spectrum, from the multiplicative ergodic theorem (MET). Similarly Lyapunov spectrum doesn't depend on the choice of sequence almost surely and thus it is a global property for this random matrix multiplication. For a given initial vector, such set of sequences that gives different asymptotic limit of growth rate has zero measure. Analog with eigenvector, it also has Lyapunov vector which describes characteristic expanding and contracting directions, but it depends on the particular ergodic sequence. It is discussed in Chap. 4.

The filtered state probability $\boldsymbol{\rho}_t$ is projected onto the simplex S^{n-1} in (5.4) and the dynamics of it will be an induced RDS. There is a nice theorem connecting Lyapunov spectrum of both RDS. [4]

Theorem 5.4. *Lyapunov spectrum of the induced RDS is that of the corresponding RDS minus the maximum Lyapunov exponent, i.e, $\Lambda' = \Lambda - \lambda_1$.*

Specifically, when the condition in theorem 5.1 is fulfilled, then maximum Lyapunov exponent of the induced RDS, $\lambda'_1 = 0$ with multiplicity 1 and the next one is $\lambda'_2 = \lambda_2 - \lambda_1$ which is what we desire to estimate.

In the framework of RDS, the exponential forgetting property defined above is equivalent with *synchronization by noise* discussed in Chap. 4.

However, we note not every RDS processes this property. Specifically, Newman [82] showed the necessary and sufficient conditions for stable synchronization in continuous state RDS. Crudely speaking, in order to see synchronization, one needs two ingredients: local contraction (negative maximum Lyapunov exponents) so that nearby points approach each other; along with a global irreducibility condition. In discrete state RDS, conditions for synchronization are discussed as well. In HMM, the global irreducibility holds since the transition matrix M is primitive and the local contraction is guaranteed by this gap $\lambda_2 - \lambda_1$. It recovers the results previously obtained. A more intuitive picture will be presented in the next section. So the 2-norm of difference for two nearby trajectories has the following behavior,

$$\|\boldsymbol{\rho}_t - \boldsymbol{\rho}'_t\| \leq C \exp\Big((\lambda_2 - \lambda_1)t\Big) \|\mathbf{p}_0 - \mathbf{p}'_0\| \leq \epsilon \tag{5.10}$$

where C is some constant. If one would like to have the error within the radius of ϵ, then the length of the subsequence should be $B \approx \frac{\ln(\epsilon)}{\lambda_2 - \lambda_1}$. However, from the previous literature [8, 68, 69], the explicit analytical estimate of the gap $\lambda_2 - \lambda_1$ for a given parameter is either too loose or still difficult to find. So the numerical estimation is needed.

5.3 Algorithm

In fact, one could sample two sequences of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ with the same matrices sequence and monitor the maximum length needed to achieve ϵ error. However, it suffers numerical instability and lack of robustness, such that some rare events can ruin the estimate. Or one use QR decomposition directly to find the Lyapunov spectrum for (5.2) which takes about $O(n^3)$ order of multiplications per each iteration. But what needed is merely the second largest one instead of the whole spectrum. Then it is possible to have a more efficient algorithm and may provide some new insight as to why this gap governs the exponential forgetting rate. In a realistic case, one may have access to the forward probability $\mathbf{p}_t$ or the filtered state probability $\boldsymbol{\rho}_t$, or at least some portion of them. We would like to take advantage of this information without redoing this time-consuming filtering process.

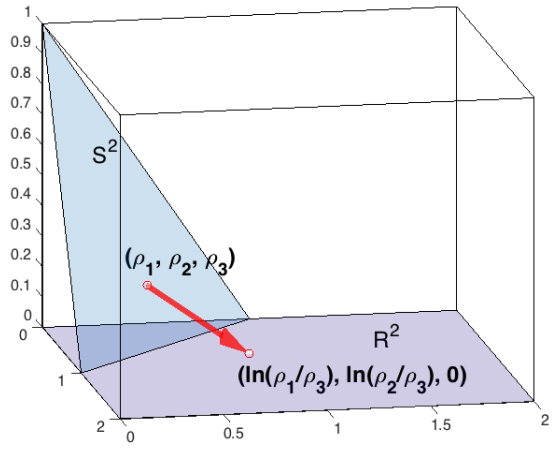

Figure 5.2: Diagram of the projection from a point in the simplex S^2 to $\mathbb{R}^2$.

If $\boldsymbol{\rho} = [a_1, a_2, \ldots, a_n]$, define a projection Π from simplex S^{n-1} to $\mathbb{R}^{n-1}$ as the log ratio relative to the last component. The projection is illustrated for the example in figure 5.2.

$$\Pi : \boldsymbol{\rho} = \Big[a_1, a_2, \ldots, a_n\Big] \to \mathbf{r} = \Big[\underbrace{\log(\frac{a_1}{a_n})}_{r_1}, \underbrace{\log(\frac{a_2}{a_n})}_{r_2}, \ldots, \underbrace{\log(\frac{a_{n-1}}{a_n})}_{r_{n-1}}, 0\Big] \tag{5.11}$$

Denote $r_n = 0$ as convention, such that $\mathbf{r}$ is embedded in $\mathbb{R}^n$. Since M is primitive, a_n cannot be 0 except for at most n initial steps. In the mean time, $\boldsymbol{\rho}$ will be in the interior of the simplex. Such projection from compact space to non-constraint space, illustrated in figure 5.2, is relatively common in numerical optimization which is called soft-max parametrization. It directly implies the constraint condition $\sum_i a_i = 1$ and $a_i > 0$. The inverse of the projection, Π^{-1} is

$$\Pi^{-1} : \mathbf{r} = \Big[r_1, r_2, \ldots, r_{n-1}, 0\Big] \to \boldsymbol{\rho} = \Big[\underbrace{\frac{\exp(r_1)}{\sum_i \exp(r_i)}}_{a_1}, \underbrace{\frac{\exp(r_2)}{\sum_i \exp(r_i)}}_{a_2}, \ldots, \underbrace{\frac{\exp(r_n)}{\sum_i \exp(r_i)}}_{a_n}\Big] \tag{5.12}$$

The index of the summation is from 1 to n. This projection naturally defines an induced RDS for the dynamics of $\mathbf{r}$. Furthermore,

Theorem 5.5. *If the coordinate transformation is one-to-one, the derivative and its inverse exists, then Lyapunov spectrum is invariant under such coordinate transformation.*

Then the projection preserves the Lyapunov spectrum. It also means the synchronization with the variable $\mathbf{r}$ implies the synchronization with $\boldsymbol{\rho}$ and vice versa. Heuristically understanding, $\lambda_1' = 0$ is due to the constraint condition and after the parametrization, the condition is inherited in the last component $r_n = 0$. If we only study the dynamics for the first $n-1$ unconstrained coordinates, it removes this particular Lyapunov exponent of the induced RDS but keeps the rest of the spectrum the same. Now the maximum Lyapunov exponent is the desired difference $\lambda_2 - \lambda_1$.

In addition, the dynamics of $\mathbf{r}$ has the following nice property. The random map for $\mathbf{r}$ has the form as

$$\mathbf{r}_{t+1} = \mathbf{d}(y_{t+1}) + F(\mathbf{r}_t),\ t \geq 0 \tag{5.13}$$

It is composed of a random translation $\mathbf{d}$ and a deterministic map $F(\mathbf{r})$. Each component of the map F is explicitly given as

$$\begin{aligned} F_i(\mathbf{r}) &= \log\Big(\frac{\sum_{j=1}^n \exp(r_j) M_{ji}}{\sum_{j=1}^n \exp(r_j) M_{jn}}\Big),\ 1 \leq i \leq n-1 & (5.14)\\ &= \log\Big(\frac{\exp(\mathbf{r}) \cdot \mathbf{m}_i}{\exp(\mathbf{r}) \cdot \mathbf{m}_n}\Big),\ 1 \leq i \leq n-1 & (5.15) \end{aligned}$$

If we denote $\mathbf{m}_i$ as the i-th column of the transition matrix M and $\exp(\mathbf{r})$ as component-wise exponent, then Eq. 5.14 is written by the inner product form Eq. 5.15.

The random translation is similarly defined as the log ratio of diagonal of D relative to the last component,

$$\mathbf{d}(y_t) = \left[\log\left(\tfrac{p(y_t|x_t=1)}{p(y_t|x_t=n)}\right), \ldots, \log\left(\tfrac{p(y_t|x_t=n-1)}{p(y_t|x_t=n)}\right)\right]. \tag{5.16}$$

Since the emission distribution is positive, the log ratio is well defined. The random map is the translation of the deterministic smooth map F by the i.i.d. random variable $\mathbf{d}(y_t)$ and F is solely dependent on the transition matrix M. It is even more interesting to notice the

Jacobian of this random map is independent with $\mathbf{d}$, it is $J(\mathbf{r}) = \nabla F(\mathbf{r})$ since the random translation will not affect the local contraction or expansion.

The $(n-1)$-by-$(n-1)$ Jacobian matrix $J(\mathbf{r})$ can be explicitly expressed as follows,

$$J_{ij}(\mathbf{r}) = \frac{\exp(\mathbf{r}_j)M_{ji}}{\exp(\mathbf{r}) \cdot \mathbf{m}_i} - \frac{\exp(\mathbf{r}_j)M_{jn}}{\exp(\mathbf{r}) \cdot \mathbf{m}_n},\ 1 \leq i, j \leq n-1 \tag{5.17}$$

Then we will have the corollary following by Theorem 5.4 and Theorem 5.5

Corollary 5.1. *The top Lyapunov exponent of the random matrix production $J(\mathbf{r}_i)$ is $\lambda_2 - \lambda_1$,*

$$\lambda_2 - \lambda_1 = \limsup_{t \to +\infty} \frac{1}{t} \log \|J(\mathbf{r}_t)J(\mathbf{r}_{t-1}) \cdots J(\mathbf{r}_1))\| \tag{5.18}$$

Now the maximum Lyapunov exponent $\lambda_2 - \lambda_1$ is approximated by the finite time Lyapunov exponent. Instead of using QR decomposition, the maximum Lyapunov exponent can be estimated by averaging finite time approximations which is much faster and easier to implement. One can start with a unit test vector, apply these Jacobian matrices sequentially and renormalize the vector at each step. Averaging all these renormalization constants along the timeline will give the approximation of maximum Lyapunov exponent. It is not a concern that all vectors are alignment along the direction of maximal expansion because we are not interested in finer structure of the spectrum. The order of multiplication needed is $O(n^2)$ per each iteration which is faster than QR decomposition. More importantly, if one already has partial data set of the filtered state probability, they can be projected to $\mathbf{r}$ and estimate the Lyapunov exponent directly without further information on observed sequences.

The maximum Lyapunov exponent for this induced random map $\lambda_2 - \lambda_1$ characterizes the rate of separation of infinitesimally close trajectories in $\mathbb{R}^{n-1}$. If two vectors $\mathbf{r}'$ and $\mathbf{r}$ are close enough, one could use their difference to approximate the 2-norm of the difference of $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$, $\|\boldsymbol{\rho}' - \boldsymbol{\rho}\|_2 \leq \frac{1}{4}\|\mathbf{r}' - \mathbf{r}\|_2$. Then the rate of separation for $\boldsymbol{\rho}$ in fact is upper bounded by the gap $\lambda_2 - \lambda_1$, which is the estimation of exponential forgetting rate. This algorithm provides some new insight for the analytical justification for the gap.

We apply this algorithm to approximate the gap of Lyapunov exponent in the previous example. In the figure 5.3, the estimated gap is $\lambda_{\max} = -0.1944$ with data of 10000 and the

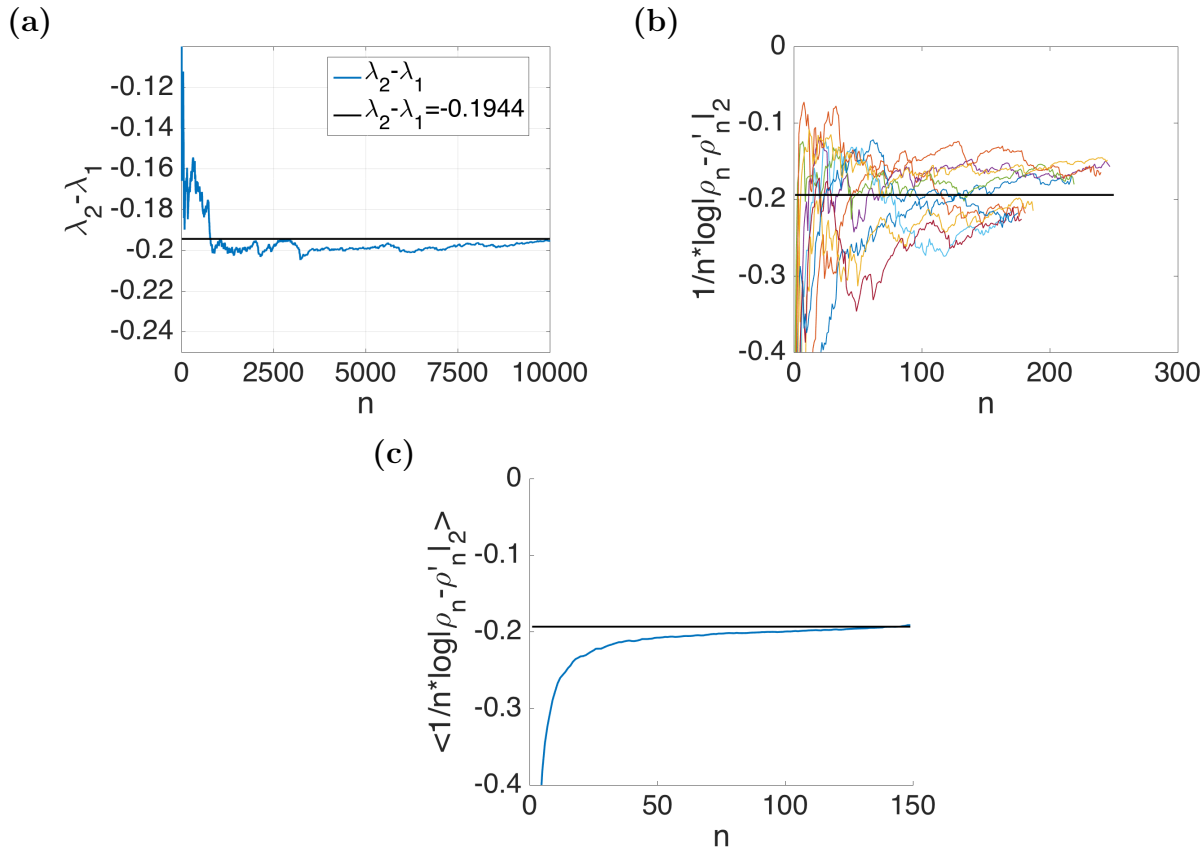

Figure 5.3: (a) We use algorithm 1 to estimate the gap of Lyapunov exponent with the observation sequence with length of 10000. (b) We sample 10 independent sequences for $\frac{1}{t}\log\|\boldsymbol{\rho}_t - \boldsymbol{\rho}'_t\|_2$ and compare with the theoretical limit (black line). (c) We average 500 independent sample sequences and compare with the theoretical limit (black line).

length B needed for $\epsilon = 10^{-15}$ is about 178. On the other hand, one starts with two different initial conditions $\mathbf{p}_0$ and $\mathbf{p}_0'$ and applies the same sequences of random matrices to obtain $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ after normalization. We plot $\frac{1}{t}\log\|\boldsymbol{\rho}_t - \boldsymbol{\rho}_t'\|_2$ along n for ten independent sequences and they roughly converge to the theoretical limit -0.1944. However, as n increases, $\|\boldsymbol{\rho}_t - \boldsymbol{\rho}_t'\|_2$ reaches the machine epsilon and becomes numerically unstable, such that some sequences are cut off beyond $t = 150$. So we are not able to visualize the strong convergence directly. If we average 500 sample sequences, then we can clearly visualize the convergence in mean. With the uniform integrability, convergence in mean is granted by the strong convergence.

Right now, it seems one needs to estimate the gap for each parameter θ. One related result is if matrices are nonsingular and maximum Lyapunov exponents are simple, then it depends continuously on the probability [88]. Bocker and Viana [18] showed Lyapunov spectrum depend continuously on matrices and probability for 2-dimensional case, as far as all probabilities are positive. Moreover, a few of Avila's deepest results [116] that are still in preparation with Eskin and Viana, extend the statement to arbitrary dimension [123]. The book [116] gives a nice introduction on this most recent approach. The direct consequence for this result is it doesn't need to estimate the gap every time and it is safe to reuse the previous estimation for couple steps in parameter inference. The pseudocode of estimating the length B is given in Algorithm 1.

Algorithm 1 Estimate the length B

1: $a \leftarrow 0$, initialize $\mathbf{p}_0$ and e,
2: **for** $i = 0, 1, \ldots, t,$ **Do**
3: $\quad \mathbf{p}_{i+1} \leftarrow \mathbf{p}_i M D_{i+1}$, D_{i+1} is given in (5.2),
4: $\quad \boldsymbol{\rho}_{i+1} \leftarrow \mathbf{p}_{i+1}/(\mathbf{p}_{i+1} \cdot \mathbb{1})$, update $\mathbf{r}_{i+1}$ according to (5.11),
5: $\quad e \leftarrow J(\mathbf{r}_{i+1})e$, $a \leftarrow a + \log\|e\|$, $e \leftarrow e/\|e\|$.
6: **end for**
7: $\lambda \leftarrow a/t, B \leftarrow \log(\epsilon)/\lambda$.

5.4 Statistical Inference

Traditionally, EM, variational inference or MCMC are used to perform inference over θ. These algorithms have found widespread use in statistics and machine learning [40, 81]. However, it is a computational challenge in terms of scalability and numerical stability, to marginalize all hidden state variables given a long sequence of observations. There are many other gradient based algorithms to obtain the maximum likelihood estimator(MLE) or maximum a posteriori (MAP), for instance, stochastic gradient descent method. We must be able to efficiently estimate the gradient of the log-likelihood function or log-posterior function, $\ln p(\theta|y_{1:t})$.

With the prior function $p(\theta)$, the gradient is written as

$$\begin{aligned}\frac{\partial \ln p(\theta|y_{1:t})}{\partial \theta_i} &= \frac{\partial \ln p(y_{1:t}|\theta)}{\partial \theta_i} + \frac{\partial \ln p(\theta)}{\partial \theta_i} \\ &= \sum_{j=1}^{t} \frac{\boldsymbol{\rho}_{j-1} \frac{\partial MD_j}{\partial \theta_i} \boldsymbol{\beta}_j}{\boldsymbol{\rho}_n \cdot \boldsymbol{\beta}_n} + \frac{\partial \ln p(\theta)}{\partial \theta_i}\end{aligned} \tag{5.19}$$

The complexity of matrix multiplication needed to calculate one component of the gradient is $O(t)$ and the space needed is also $O(t)$. So it is prohibitively expensive to compute directly in space and time when t is very large. Moreover, this direct computation is not numerically stable since the numerator and denominator are usually extremely small in such massive matrix multiplication. In fact, there are various algorithms to reduce the complexity, including the following mini-batch gradient descent method, which employs noisy estimates of the gradient based on mini-batch of data [38, 75].

First, instead of summing over all index j from 1 to t, uniformly sample a subset of summand S with cardinality s at each step and use the following estimator for the direction of the full gradient. Here we assume the prior distribution $p(\theta)$ as uniform for the sake of simplicity,

$$\frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta_i} = \frac{t}{s} \sum_{j \in S} \frac{\boldsymbol{\rho}_{j-1} \frac{\partial MD_j}{\partial \theta_i} \boldsymbol{\beta}_j}{\boldsymbol{\rho}_{j-1} MD_j \boldsymbol{\beta}_j} \tag{5.20}$$

Then we expect $\mathbb{E}(\frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta_i}) = \frac{\partial \ln p(\theta|y_{1:t})}{\partial \theta_i}$. This is typically referred to mini-batch gradient descent based techniques and it is very effective in the case of large-scale problems.

Second, instead of computing normalized forward and backward probability $\boldsymbol{\rho}_j$ and $\boldsymbol{\beta}_j$ recursively, we introduce a buffer of length B on left and right ends of the subsequence of random matrices and both vectors are estimated by this much shorter subsequence,

$$\widetilde{\mathbf{p}}_{LB} = \mathbf{p}_0 M D_{j-B} \dots M D_{j-1}, \quad \widetilde{\boldsymbol{\rho}}_{j-1} = \widetilde{\mathbf{p}}_{LB}/(\widetilde{\mathbf{p}}_{LB} \cdot \mathbb{1}) \tag{5.21}$$

$$\widetilde{\mathbf{b}}_{RB} = M D_{j+1} \dots M D_{j+B} \mathbb{1}, \quad \widetilde{\boldsymbol{\beta}}_j = \widetilde{\mathbf{b}}_{RB}/(\widetilde{\mathbf{b}}_{RB} \cdot \mathbb{1}) \tag{5.22}$$

The reason to use the same buffer length for forward and backward probability is that Lyapunov spectrums for forward algorithm and backward algorithm are exactly the same. Therefore, the gap of the top two Lyapunov exponents are the same and the buffer of length is the same.

Theorem 5.6. *Let θ be an invertible ergodic measure-preserving transformation of a probability space $(\Omega, \mathbb{P})$, where $\mathbb{P} = Q^{\mathbb{Z}}$ is the Bernoulli measure. Let $A : \Omega \to M(d, \mathbb{R})$ satisfy $\int \log \|A_\omega^{\pm 1}\| \mathrm{d}\mathbb{P} < \infty$. Let $P_t(\omega)$ be the induced linear cocycle, $P_t(\omega) = A(\theta(t-1)\omega) \dots A(\omega)$. Then the Lyapunov spectrums of these two linear cocycles*

$$\lambda := \lim_{t \to +\infty} \frac{1}{t} \log \|\mathbf{u} P_t(\omega)\|, \ \lambda' := \lim_{t \to +\infty} \frac{1}{t} \log \|P_t(\omega)\mathbf{v}\| \tag{5.23}$$

are equal $\mathbb{P}$-a.s.

The proof is straight forward. The left multiplication and the right multiplication will not change the Lyapunov spectrum but the subspaces $U_i(\omega)$ will be different for two cocycles.

So the gradient (5.20) is approximated as

$$\frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta_i} = \frac{t}{s} \sum_{j \in S} \frac{\widetilde{\boldsymbol{\rho}}_{j-1} \frac{\partial M D_j}{\partial \theta_i} \widetilde{\boldsymbol{\beta}}_j}{\widetilde{\boldsymbol{\rho}}_{j-1} M D_j \widetilde{\boldsymbol{\beta}}_j} \tag{5.24}$$

Note that (5.24), the matrix multiplication required is $O(2Bs)$ after using the buffer and the space needed is $O(s)$. When $2Bs \ll t$, this results in significant computational

speedups over the full batch inference algorithm. This techniques is exactly due to the memory decaying property and the buffer length B is calculated in Algorithm 1. In order to be consistent with this technique, we can uniformly sample the subset of summand in the domain $[B+1, t-B-1]$. Moreover, one can enforce no overlap among the sampled subsequences. In pseudocode, the algorithm is presented in Algorithm 2.

Algorithm 2 Mini-batched based inference for HMM

1: initialize the parameter $\theta = \{M, \phi\}$ and learning rate η
2: **for** $i = 0, 1, \ldots, N_{iter}$, **Do**
3: Periodically estimate the buffer length B according to Algorithm 1
4: i.i.d sample s integers in the subset $[B+1, t-B-1]$ with uniform distribution.
5: Calculate $\frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta}$ according to (5.24).
6: $d = \frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta} / \| \frac{\partial \ln \widetilde{p}(\theta|y_{1:t})}{\partial \theta} \|$
7: $\theta \leftarrow \theta + \eta d$
8: **end for**

This memory decay property not only can be taken advantage of in mini-batched gradient descent based inference on MLE or MAP, but also be used in stochastic gradient-MCMC [74, 75], stochastic variational inference [38], stochastic EM and online learning [59]. There are many more algorithms could be built based on this fundamental property in HMM. No matter what mini-batched based algorithm is used, it is important to estimate the buffer length efficiently and accurately.

Example 5.2. In order to demonstrate the algorithm, we sampled a long observation sequence with length $t = 10^7$ and the parameter given in Example 5.1. We assume we know all parameters except μ_1 and μ_2. In the algorithm, we use the same left and right buffer length $B_2 = B_1 = 200$ and sample size $s = 100$.

The learning rate η starts with 0.05 and decays with the rate of 0.95 along the steps to prevent oscillations. After 25 steps, the algorithm will restart with the latest parameter,

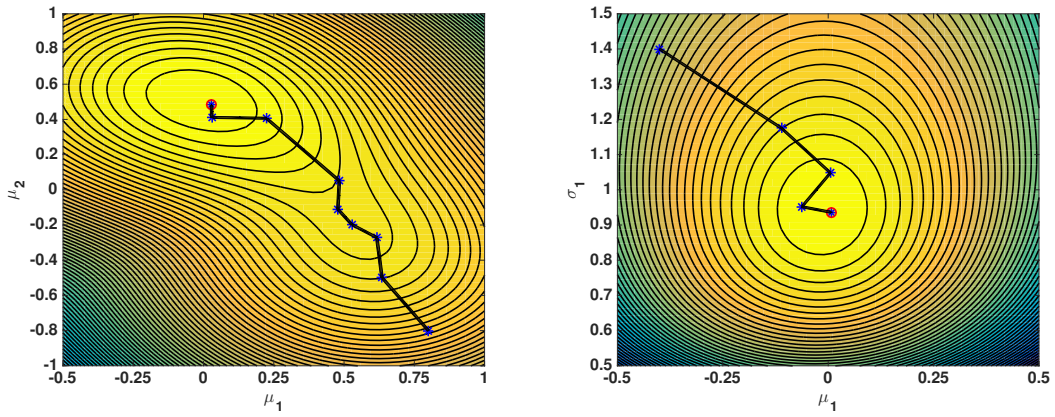

Figure 5.4: Apply the Algorithm 2 to Example 5.1. The background is the contour plot of the log-likelihood function. In the left figure, μ_1 and μ_2 are unknown, the algorithm converges to $(0.03, 0.48)$ starting from $(0.8, -0.8)$. In the right figure, μ_1 and σ_1 are unknown, the algorithm converges to $(0.01, 0.94)$ starting from $(-0.4, 1.4)$.

until the difference of parameter from the previous restart is within the threshold, which in this case is 0.02. From the figure 5.4, the parameter reaches the desire MLE $(0.03, 0.48)$ after 8 starts from the initial guess $(0.8, -0.8)$. Note from the contour plot of the log-likelihood function, there is a region around $(0.5, 0)$ which is the flip of the mean, is very flat and our algorithm is able to escape it due to the stochastic nature. Also, the number of matrix multiplications needed is is about $8 * 10^6$ which is less than the length observation sequence $L = 10^7$. It implies the algorithm has reached MLE even before the filtering procedure is finished in single iteration in EM or gradient descent algorithms. So it significantly speed up the inference procedure.

Although traditionally the EM algorithm has monotonic convergence, ensures parameters constraints implicitly and generally easier to be implemented, the convergence of EM can be very slow in terms of time for each iteration and total iteration steps. Our method can significantly reduce the time for each iteration since we only utilize part of the data by harnessing the memory loss property, but the steps needed is still comparable with EM

algorithm. In fact, one could extend our idea to efficiently estimate the Hessian matrix such that much faster quadratic convergence can be achieved with second-order method. Moreover, one can calculate observed fisher information which is the negative Hessian matrix of the log-likelihood evaluated at MLE. It will give curvature information at MLE and help to decide which MLE may be better without calculating the likelihood explicitly. Another natural extension of our method is to discrete state Kalman filter, which is continuous time version of HMM. Similar exponential forgetting property and the rate being the gap of top two Lyapunov exponents are discussed in [8] for the Wonham filter, but the proof is harder and evolves different techniques, in particular, the naive time discretization may be challenging since one needs to justify the change order of the limit.

In the era of big data, data analysis methods in machine learning and statistics, such as hidden Markov models, play a central role in industry and science. The growth of the web and improvements in data collection technology in science have lead to a rapid increase in the magnitude and complexity of these analysis tasks. This growth is driving the need for the scalable algorithms that can handle the "Big Data". However, we don't need that the whole massive data, instead small portion of data could serve as good as the original. One successful examples are mini-batched based algorithms. Despite the simple chain-based dependence structure, apply such algorithms in HMM are not obvious, since subsequences are not mutually independent. However, with the data set being abundant, we are able to harness the exponential memory decay in filtered state probability and appropriately choose the length of the subchain with the controlled error, to design mini-batched based algorithms. We proposed an efficient algorithm to accurately calculate the gap of the top two Lyapunov exponents, which helps to estimate the length of the subchain. We also prove the validity of the algorithm theoretically and verified it by numerical simulations. In the example, we also proposed the mini-batched gradient descent algorithm for MLE of log-likelihood function and it significantly reduces the computation cost.

Chapter 6

APPLICATION OF STOCHASTIC PROCESS: POLYMER LOOPING

The theory of polymer dynamics is one of the most successful stochastic-process models in chemical science [35,37]. Polymer dynamics in an aqueous solution is naturally stochastic. A polymer molecule dissolved in the so-called theta solvent, known also as free-draining, can be mathematically represented in terms of a multi-dimensional Ornstein-Uhlenbeck (OU) process. While the general theory of an OU process is well developed (see [93] and the references cited within), explicitly analytical results on the kinetics of the formation of a end-to-end loop are still highly sought after in theoretical chemistry and biochemistry. The problem is essentially a perturbation of a linear operator [25, 56, 119].

In the literature, "polymer looping" has been a classic problem with important biological applications. It is an essential step in various intra-cellular processes, including gene regulation, DNA replication, protein and RNA folding [78, 91, 102]. In some biological systems, two reacting molecules may bind distal target sites along a single polymer chain and a biological function requires the polymer to "spontaneously" form a loop structure that brings the two in contact [70]. This polymer-loop-mediated phenomena is due to thermal fluctuations and the loop formation consists of two steps: the "encounter" of the two ends of the polymer defined by a small capture distance, followed by a bond formation between the two ends when they are within the capture radius. When the time scale of the former is much longer than the latter, the loop formation is called diffusion controlled: The diffusive Brownian motion of the chain is then the rate determining step.

As an in-depth study of a multi-dimensional stochastic process, the problem was first discussed by Jacobson and Stockmayer in 1950 even before the dynamic Rouse model [50, 99].

They formulated the free energy cost of looping in terms of an equilibrium distribution. It is quantified as the ratio of respective equilibrium constants for intermolecular and intramolecular synapsis reactions when studying the polymer melts, which is called the J factor. The J factor is directly proportional to the statistical fraction of polymer conformations within the capture radius a, which is the ratio of the rate constants of the cyclization reaction, $k_{\text{coil}\to\text{loop}}/k_{\text{loop}\to\text{coil}}$. The J factor is mainly discussed in the more complicated case where the effect of bending and twisting rigidity is taken into account and so the polymer is no longer flexible [105]. The rate $k_{\text{coil}\to\text{loop}}$ is the inverse of the mean time for the two ends of a polymer to meet within the capture radius a (or simply looping time τ) from some configurations.

The study of the kinetics of loop formation is a classical problem in polymer dynamics. It has led to intense theoretical and numerical research, even for the simplest case, of a Rouse chain. Though seemingly simple, the problem is actually challenging [87]. The mathematical formulation of the kinetics of this looping process involves a non-simple boundary-value problem of a linear PDE in a high dimensional space.

There were two major theoretical approaches in the early stage of the polymer looping investigation; they led to seemingly contradicting results. First, Wilemski and Fixman (WF) [121,122] and Doi [34] estimated that the looping time for long polymer chains scales as $\sim N^2$, where N is the number of monomers. Surprisingly, according to this result, the looping time does not depend on the capture radius a (more details of this method will appear in later sections). On the other hand, Szabo, Schulten and Schulten (SSS) [110] estimated the looping time scales as $\sim N^{3/2}/a$. They approximated the dynamics of the end-to-end distance of the polymer as a single Brownian particle diffusing in a potential of mean force. Recent numerical simulations, however, indicated these time scales are showing up in different time regimes but the prediction from WF is more accurate than that from SSS theory [24, 87]. In fact, more recent theoretical advances show that the looping time may follow a mixed scaling law with $\tau \approx c_1 N\sqrt{N}/a + c_2 N^2$. Toan *et al.* [114] proposed an effective space-dependent diffusion coefficient in SSS theory, such that both theories could be unified into one and the mixed scaling law was revealed. However, it has been difficult to verify whether

this space dependency is really a mathematical approximation of a multi-dimensional OU process, or a different mathematical model all together. Very recently Amitai *et al.* [2] also discovered a similar scaling law, based on the expansion of the eigenvalues of the Fokker-Planck equation in the limit of small capture radius. The first order perturbation in a of the largest eigenvalue matches the SSS result and the next order perturbation roughly scales as N^2. But their theory works well only when N is relatively small ($N < 64$), and the contribution from other eigenvalues is unclear. Guèrin *et al.* [111] and Bènichou *et al.* [12] adapted the renewal equation method on the non-Markovian process of the end-to-end vector and reviewed other previous approaches. The numerical result they obtained is better than previous theories but an analytic scaling law is currently out of reach.

Within the OU process framework and in terms of analytical results, the mean first passage time is a solution to a boundary value problem associated with a high dimensional backward equation. Significant progress has been made in recent years on the asymptotic of passage-time problem associated with three-dimensional diffusion processes with a small exit boundary [20, 25, 27], crossing an unstable limit-cycle barrier [32], and rigorous lower bound on the density of the passage-time of the OU process [112]. The WF theory distinguishes itself from these other works in dealing with a high-dimensional OU process, which is equivalent to a low-dimensional non-Markovian process. With this in mind, the WF theory is currently the only analytically feasible approximation for the non-Markovian process of the end-to-end vector.

Here I discuss the WF theory rigorously and articulate the assumptions behind it. Specifically, I show that the assumption for the Delta sink is weaker than the Heaviside sink, so that the Delta sink gives better approximation results. The integrand appearing in WF theory in the expression for the looping time is not a genuine survival probability. We extract the postulated mixed scaling laws directly from WF theory in the limit of small dimensionless ϵ, the ratio of capture radius a to the Kuhn length b, with the help of perturbation methods. The mixed scaling law I find includes an extra term $N\sqrt{N}$ which will be dominated by N^2 term as $N \to +\infty$. Then I use Monte Carlo simulations to estimate the looping time for various N

and the capture radius ϵ. In general, WF theory overestimates the looping time but agrees very well with the simulation in the small ϵ regime (irrespective of the value of N). We also verify the mixed scaling law numerically through regression. Last but not least, the WF theory naturally gives a direction to the stochastic model reduction of high-dimensional OU process. We will briefly discussion the connection to the celebrated Mori-Zwanzig formalism.

6.1 Overview of Rouse Model

The Rouse model is the simplest flexible polymer model which uses beads and springs to represent the polymer chain in a viscous fluid. It assumes beads have no excluded volume and no hydrodynamic interactions among them. Actually, the notion of a theta solvent is defined as an exact cancellation between these two opposing effects. Two adjacent beads are connected with a harmonic spring with the same spring constant k [35, 63] and in total there are $N+1$ beads and N springs. The potential energy of the entire mechanical system is given by $E(\vec{\mathbf{R}}) = \frac{1}{2}k\sum_{n=1}^{N}(\mathbf{R}_n - \mathbf{R}_{n-1})^2$, where $\mathbf{R}_n$ is the position of nth bead in three-dimensional space. Let b denote the Kuhn segment length or effective bond length, that is defined as the mean square length of the bond. Then, if $k = 3/b^2$ in k_BT units, the Boltzmann distribution of the conformation and the end-to-end vector distribution will be the same as the ones for the Gaussian chain. In addition to the Hookean force from its connecting springs, a bead experiences a frictional force when it moves. Each bead is assumed to have the same friction coefficient ζ. Inertia is negligible since the motion is overdamped. The diffusion coefficient of a monomer is $D_0 = 1/\zeta$ in k_BT units as dictated by the Einstein relation. The dynamics are described by Langevin equations 6.1 with Gaussian white noise [115],

$$\frac{\mathrm{d}\mathbf{R}_i}{\mathrm{d}t} = -D_0\nabla_{\mathbf{R}_i}E(\mathbf{R}) + \mathbf{f}_i \quad i = 0, 1, \ldots, N, \quad \langle \mathbf{f}_n(t)\mathbf{f}_m(t')\rangle = 2D_0\mathbf{I}\delta_{nm}\delta_{tt'} \tag{6.1}$$

Through a normal mode transformation (6.2), the Langevin equations 6.3 for the modes $\mathbf{X}_p$ are decoupled and each mode evolves independently but not identically to the others.

$$\theta_p = \frac{p\pi}{2(N+1)},\ \gamma_p = 12\sin^2(\theta_p),$$
$$\mathbf{X}_p = \frac{1}{N+1}\sum_{n=0}^{N}\mathbf{R}_n\cos\Big((2n+1)\theta_p\Big),\quad \mathbf{R}_n = \mathbf{X}_0 + 2\sum_{p=1}^{N}\mathbf{X}_p\cos\Big((2n+1)\theta_p\Big), \tag{6.2}$$
$$\frac{\mathrm{d}\mathbf{X}_p}{\mathrm{d}t} = -\gamma_p\mathbf{X}_p + \mathbf{F}_p,\ \langle\mathbf{F}_p(t)\mathbf{F}_q(t')\rangle = \frac{1}{N+1}\mathbf{I}\delta_{pq}\delta_{tt'}\ \text{for}\ p+q>0 \tag{6.3}$$

Here γ_p are eigenvalues of the Laplacian matrix of the Rouse chain. The zeroth mode of the chain represents the dynamics of the center of mass for the chain which is Brownian motion. The dynamics is also called a reversible Ornstein-Uhlenbeck process [93] and its equilibrium distribution is the Boltzmann distribution $P_e(\mathbf{X})$ in Eq. 6.4 [92].

$$P_e(\mathbf{X}) = \prod_{p\ \text{odd}}\Big(\frac{(N+1)\gamma_p}{\pi}\Big)^{3/2}\exp(-(N+1)\gamma_p\mathbf{X}_p^2) \tag{6.4}$$

Nondimensionalize the system with the Kuhn length b as the characteristic length and b^2/D_0 as the characteristic time, one can define a dimensionless parameter as the ratio of capture radius to the Kuhn length, $\epsilon = a/b$. It specifies the capture radius in the unit of the Kuhn length so I will call ϵ as capture radius as well. From now on, all quantities and equations are nondimensionalized unless otherwise specified.

The end-to-end vector $\mathbf{R}_{ee}$ can be expressed as the linear combination of the odd order terms of $\mathbf{X}_p$ with coefficient c_p (see Eq. 6.5). All even order modes do not contribute if the chain has a homogenous spring constant. The equilibrium distribution of the end-to-end vector is also a Gaussian with variance N.

$$\mathbf{R}_{ee} = \mathbf{R}_0 - \mathbf{R}_N = \sum_{p\ \text{odd}} c_p\mathbf{X}_p,\ c_p = 4\cos(\theta_p), \tag{6.5}$$
$$P_{\mathbf{r}}(\mathbf{R}_{ee}) = \Big(\frac{3}{2\pi N}\Big)^{3/2}\exp\Big(-\frac{3\mathbf{R}_{ee}^2}{2N}\Big) \tag{6.6}$$

The relaxation time τ_p for the odd modes $\mathbf{X}_p$ is $\tau_p = 1/\gamma_p$. Specifically, τ_1 is the largest, $\tau_1 \approx \frac{(N+1)^2}{3\pi^2}$. So the relaxation time for the end-to-end vector $\mathbf{R}_{ee}$ is dominated by τ_1.

The corresponding Fokker-Planck equation for the Langevin equation 6.3 describes the time evolution of the probability density $P(\mathbf{X},t)$ [43]. This linear operator is denoted as L_F. The Green function of the Fokker-Planck equation is the probability propagator, $G(\mathbf{X},t|\mathbf{X}^0,0)$, and is a Gaussian function.

$$\frac{\partial P(\mathbf{X},t)}{\partial t} = \quad \nabla\cdot\Big(\Gamma\mathbf{X}P(\mathbf{X},t)\Big)+\frac{1}{2(N+1)}\nabla\cdot\Big(\nabla P(\mathbf{X},t)\Big)=L_FP(\mathbf{X},t), \tag{6.7}$$

in which Γ is a diagonal matrix with elements γ_p. But the dynamics of the end-to-end vector $\mathbf{R}_{ee}$, being a projection of a N-dimensional OU process, is *non-Markovian*. One could sample from Boltzmann (canonical) distribution polymer configurations with fixed $\mathbf{R}_{ee}=\mathbf{r}_0$ and study the average end-to-end vector dynamics. These dynamics follow a non-Markovian Gaussian process with the conditional probability $T(\mathbf{r},t|\mathbf{r}_0,0)$. The probability at time t on location $\mathbf{r}$ is also w.r.t. the canonical ensemble. It is determined by the time correlation function of the end-to-end vector, $\phi(t)$, which is the summation of the correlation functions of the odd modes. The conditional probability $T(\mathbf{r},t|\mathbf{r}_0,0)$ looks like a probability propagator, but it doesn't satisfy the Kolmogorov-Chapman equation, i.e, $\int T(\mathbf{r},t|\mathbf{r}',t')T(\mathbf{r}',t'|\mathbf{r}_0,0)\mathrm{d}\mathbf{r}'\neq T(\mathbf{r},t|\mathbf{r}_0,0)$, for $t>t'>0$.

$$\phi(t)=\tfrac{\langle\mathbf{R}_{ee}(t)\mathbf{R}_{ee}(0)\rangle}{\langle\mathbf{R}_{ee}^2\rangle}=\tfrac{2}{N(N+1)}\textstyle\sum_{p\ \text{odd}}\tfrac{1}{\tan^2(\theta_p)}\exp(-\gamma_pt) \tag{6.8}$$

$$T(\mathbf{r},t|\mathbf{r}_0,0)=\Big(\tfrac{3}{2\pi N(1-\phi^2(t))}\Big)^{3/2}\exp\Big(-\tfrac{3}{2N}\tfrac{(\mathbf{r}-\phi(t)\mathbf{r}_0)^2}{1-\phi^2(t)}\Big) \tag{6.9}$$

Once the two ends of the chain get close (within the capture distance ϵ), the reaction may start. So this reactive surface is described as a tubular neighborhood $S_\epsilon=\{\mathbf{X}:\|\mathbf{R}_{ee}\|=\epsilon\}$. In the ideal case, every time the two ends get closer than ϵ, the reaction happens and this corresponds to the absorbing boundary condition $P(\mathbf{X},t)=0$ on the surface S_ϵ. Once the reaction term is imposed, the modes are no longer independent.

In the literature, when addressing the looping problem, it is assumed that the initial distribution is the Boltzmann distribution $P_e(\mathbf{X})$. But the initial condition doesn't fulfill the boundary condition. Physically all configurations that are in a looped state react immediately, so the initial condition becomes the normalized Boltzmann distribution outside

the reactive surface and zero inside the surface. The difference between these two initial conditions are negligible since the probability inside the surface is very small compared with outside. It is worth noting that the polymer chain sampled from the Boltzmann distribution doesn't imply all internal modes of the chain are fully relaxed.

If the PDE 6.7 with the boundary condition specified could be solved, one could integrate the solution $P(\mathbf{X}, t)$ over the configuration space $\mathbf{X}$ and get the probability that the chain hasn't formed a loop at time t, the survival probability $\Sigma(t)$. Then $1 - \Sigma(t)$ is the cumulative probability distribution for the random time of capturing, and the mean looping time τ is the time integral of survival probability through the time domain with some mild assumptions, which can be fully determined. τ is a function of the number of beads N and the dimensionless capture radius ϵ. One intuitively expects that $\tau \to \infty$ as $N \to \infty$, or $\epsilon \to 0$.

6.2 Markovian Approximation

A simple approach is to consider that the two ends of the polymer behave as if they were diffusing apart from the rest of the chain, such that no memory effect is expected. We can project all these internal modes of the polymer into a 1D single variable, the end-to-end distance $r = \|\mathbf{R}_{ee}\|$ (after angular components are averaged). This is the theory first developed by Szabo, Schulten and Schulten [110]. Suppose that the end-to-end distance obeys a reduced Langevin equation of the form with diffusion coefficient D in the unit of D_0,

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -D\frac{\mathrm{d}U(r)}{\mathrm{d}r} + \xi(t),\ \langle \xi(t)\xi(t')\rangle = 2D\delta(t - t'), \tag{6.10}$$

where the effective potential $U(r)$ of the mean field is given by $U(r) = -\ln P_e(r)$, where $P_e(r) = 4\pi r^2 \Big(\frac{3}{2\pi N}\Big)^{3/2} \exp\Big(-\frac{3r^2}{2N}\Big)$. One can calculate the mean first passage time $\tau(r)$ starting from the fixed end-to-end distance r and integrate this space dependent function w.r.t. the equilibrium distribution to get the looping time τ.

The dynamics can be considered those of a single particle diffusing in the effective potential. The looping problem becomes a standard first passage time problem which can be

addressed by solving the adjoint operator equation, $D\Big(\frac{\mathrm{d}^2\tau(r)}{\mathrm{d}r^2} - \frac{\mathrm{d}U}{\mathrm{d}r}\frac{\mathrm{d}}{\mathrm{d}r}\tau(r)\Big) = -1$, with boundary condition $\tau(\epsilon) = 0$. One can assume a reflective boundary condition when the chain is fully stretched, i.e. $\frac{\mathrm{d}}{\mathrm{d}r}\tau(N) = 0$. In fact, the difference with the natural boundary condition here is negligible since the probability of the chain reaching that far is very small for the flexible chain. It is easy to show the solution of the adjoint equation is

$$\tau(r) = \int_\epsilon^r \frac{1}{D\exp(-U(x))}\mathrm{d}x \int_x^\infty \exp(-U(y))\mathrm{d}y \tag{6.11}$$

Averaging over the equilibrium distribution of the end-to-end distance and find,

$$\tau = \int_\epsilon^\infty \tau(r)P_e(r)\mathrm{d}r = \int_\epsilon^\infty \frac{1}{DP_e(x)}\mathrm{d}x\Big(\int_x^\infty P_e(y)\mathrm{d}y\Big)^2 \tag{6.12}$$

Eq. 6.12 is known as the SSS theory of polymer looping. It can be simplified through the Laplace method if I assume that the potential barrier at $r = \epsilon$ is very high. In SSS theory, the diffusion coefficient D is chosen as the relative diffusion coefficient of the two end monomers, $D = 2$. (in Sec. 6.4, it is shown that this choice of diffusion coefficient is related to a short time approximation for the end-to-end correlation function $\phi(t)$). For the Rouse model, the looping time becomes

$$\tau_{\mathrm{SSS}} \approx \frac{1}{8\pi\epsilon P_e(0)} = \frac{\sqrt{\pi}}{6\sqrt{6}}\frac{N\sqrt{N}}{\epsilon} \tag{6.13}$$

Specifically, $P_e(0)$ is exactly proportional to the J factor of the Gaussian chain. Another way to derive the SSS result is to coarse grain the polymer chain to a dumbbell model with the effective spring constant $k_{\mathrm{eff}} = 3/N$.

It turns out that the equilibrium assumption about the chain is not an accurate one. Not all internal modes of the polymer chain have relaxed before the looping process starts. So when $\tau_1 \gg \tau_{\mathrm{SSS}}$, the effective potential is no longer time independent. The motion of all internal modes has be taken into consideration in the problem of looping. If $\epsilon \ll 1/\sqrt{N}$, the looping time τ is longer than the relaxation time of the chain, so all internal modes are fully relaxed before the two ends meet and SSS theory works well.

Amitai *et al.* [2] calculate the perturbation of the spectrum for the Fokker-Planck equation 6.7 and estimate the looping time accordingly. If I assume the natural boundary condition, the system is detailed balanced with the Boltzmann distribution as unique stationary density [93]. Also, in this case, the operator L_F in the linear PDE 6.7 is hermitian in the Hilbert space $L^2(\mathbf{X})$ with the inner product defined as $\langle f, g\rangle = \int_{\mathbb{R}^{3N+3}} \frac{f(\mathbf{X})g(\mathbf{X})}{P_e(\mathbf{X})}\mathrm{d}\mathbf{X}$ [115]. The linear operator can be rewritten as

$$L_F P(\mathbf{X}, t) = \nabla \cdot \Big(P_e(\mathbf{X}) \nabla \Big(\frac{P(\mathbf{X}, t)}{P_e(\mathbf{X})} \Big) \Big) \tag{6.14}$$

Therefore it has a discrete real spectrum $\{-\lambda_i^0\}$ with complete orthogonal eigenfunction basis $\{\phi_i^0\}$. The spectrum consists of nonnegative integer-weighted sums of eigenvalues of the Laplace matrix, $-\gamma_p$, and the eigenfunctions are products of the corresponding Hermite polynomials. The top eigenvalue $\lambda_0^0 = 0$ and the corresponding eigenfunction is the stationary density function. The next two are $\lambda_1^0 = \gamma_1$ and $\lambda_2^0 = \gamma_2$.

But with the absorbing boundary condition, the spectrum $\{-\lambda_i^\epsilon\}$ is shifted due to removing the small tubular neighborhood S_ϵ and there is no longer a stationary density. If ϵ is small enough, the spectrum will be still distinct discrete and real. The time dependent solution can still be expressed as an expansion in the eigenfunctions $\{\phi_i^\epsilon\}$ with coefficients c_i, $P(\mathbf{X}, t) = \sum_{i=0}^{\infty} c_i \phi_i^\epsilon(\mathbf{X}) \exp(-\lambda_i^\epsilon t)$. If the initial condition is the previous stationary density, i.e. the Boltzmann distribution, the eigenfunction ϕ_0^ϵ is close to the Boltzmann distribution so that most of the energy is contributed by this zeroth mode, i.e, $c_0 \approx 1$. The survival probability $\Sigma(t) = \int_{\mathbb{R}^{3N+3}-S_\epsilon} P(\mathbf{X}, t)\mathrm{d}\mathbf{X}$, is in fact the summation of infinitely many exponential distributions. If the survival probability decays to 0 sufficiently fast, i.e, $\lim_{t\to+\infty} t\Sigma(t) = 0$, then the mean first passage time τ is simply the integral of the survival probability from 0 to $+\infty$,

$$\tau = \int_0^\infty \Sigma(t)\mathrm{d}t = \sum_{i=0}^{\infty} \frac{C_i}{\lambda_i^\epsilon} \approx \frac{1}{\lambda_0^\epsilon}. \tag{6.15}$$

Amitai *et al.* [2] discovered that the first passage time roughly follows one single exponential distribution for relatively small N ($N < 64$) by numerical simulation, so the zeroth mode in

Eq. 6.15 is enough. We will discuss the distribution of the first passage time in Sec. 6.5. When ϵ is small enough, the first order perturbation of the zeroth eigenvalue is proportional to the ratio of the partition function of the closed polymer chain to the whole configuration space. This ratio is again exactly the J factor and the proportionality spatial factor is $8\pi\epsilon$.

$$\lambda_0^\epsilon = 0 + 8\pi\epsilon J + O(\epsilon^2) \tag{6.16}$$

The series expansion of the eigenvalues follows results in [25]. The J factor in the scaling of N appears in the expression was discussed in [45]. If one uses Eq. 6.16 in Eq. 6.15, then the first order perturbation result exactly matches the SSS Markovian approximation, i.e, the looping time is exactly Eq. 6.13.

In summary, the SSS theory describes the kinetics of loop formation as a diffusion process in an effective potential of mean force that is derived from the equilibrium distribution for the end-to-end distance $P_e(r)$. It approximates the non-Markovian dynamics of the end-to-end distance r by these simple Markovian dynamics. It assumes that the internal modes have relaxed before the looping process starts (this is called *the local equilibrium assumption*). The condition for this to hold is $\epsilon \ll 1/\sqrt{N}$. As far as the spectrum of the linear operator is concerned, this Markovian approximation corresponds to the first order perturbation of the zeroth mode. In Sec. 6.5, it is shown that the SSS result significantly underestimates the looping time (it can also be proved through a variational principle [89]). Roughly speaking, the Markovian estimate ignores the case that two ends meet each other due to fluctuation, but the center of mass of the chain may be far away from the two ends. Intuitively, this case is more likely to happen for relatively large ϵ or large N. This drawback brings us to a more comprehensive method, Wilemski-Fixman theory.

6.3 WF Theory

One could incorporate a distance-dependent reaction term in the Fokker Planck equation 6.7 with a microscopic rate constant κ and relate the looping time with the time integral of a normalized sink-sink time correlation function. This is the celebrated Wilemski-Fixman

(WF) approximation [121, 122]. Instead of solving a Fokker-Planck equation 6.7 with a complicated boundary condition, the modified equation is of convection-diffusion-reaction type in free space (see Eq. 6.17) and one can express the solution in terms of the Green function $G(\mathbf{X}, t|\mathbf{X}', 0)$ for Eq. 6.7. This method can also be applied to study the Fokker-Planck equation with other difficult boundary conditions, like diffusion-limited catalytically-activated chemical reactions with a special catalytic subvolume [11], facilitated diffusion and looping with a heterogenous Rouse chain. The microscopic rate constant κ, which measures the effectiveness of the reaction, is not the same as the coefficient in the partially absorbing boundary condition in Collins and Kimball's kinetic theory. In fact, Szabo *et al.* [109] shows that the partially absorbing boundary condition is equivalent to the use of a delta sink on the reactive surface in conjunction with a reflecting boundary condition. However, as I will show later, if I let $\kappa \to +\infty$ in the WF approximation, the absorbing boundary condition can be recovered under some assumptions.

$$\frac{\partial W(\mathbf{X}, t)}{\partial t} = L_F W(\mathbf{X}, t) - \kappa S(\mathbf{X}) W(\mathbf{X}, t),\ W(\mathbf{X}, 0) = P_e(\mathbf{X}) \tag{6.17}$$

Two popular choices for the sink function are the Heaviside function and the radial delta function. For a given configuration $\mathbf{X}$, the Heaviside function is $S_1(\mathbf{X}) = \begin{cases} 1, & \text{if } \|\mathbf{R}_{ee}\| \le \epsilon \\ 0, & \text{Otherwise} \end{cases}$. Intuitively, it corresponds to the Brownian particle starting to react once inside the reactive surface S_ϵ. The delta function is $S_2(\mathbf{X}) = \frac{1}{4\pi\epsilon^2}\delta(\|\mathbf{R}_{ee}\| - \epsilon)$. It is also called the Smoluchowski reaction sink. The reactive spherical surface is in three-dimensional space and it separates the whole space into two regions, inside and outside of the reactive surface. But this sink doesn't allow the Brownian particle to pass the surface. Both types of sinks have a three-dimensional representation in terms of the end-to-end vector $\mathbf{r}$.

The solution of Eq. 6.17 starting with the equilibrium distribution $P_e(\mathbf{X})$ can be expressed with the help of Dyson's formula as

$$W(\mathbf{X}, t) = P_e(\mathbf{X}) - \kappa \int_0^t \mathrm{d}t' \int \mathrm{d}\mathbf{X}' G(\mathbf{X}, t|\mathbf{X}', t') S(\mathbf{X}') W(\mathbf{X}', t'). \tag{6.18}$$

The inner space integral gives the probability density at time t in the original Fokker-Planck equation 6.7 if the initial distribution at time t' is $S(\mathbf{X})W(\mathbf{X},t')$. The meaning of Eq. 6.18 is as follows: the probability of a chain to have the configuration $\mathbf{X}$ at time t, is the probability to observe the configuration $\mathbf{X}$ without sink, minus the probability of reaching the configuration $\mathbf{X}$ at time t starting in the sink at some point t' between 0 to t. The advantage of this approach is that the Green function for Eq. 6.7 is known. The survival probability $\Sigma(t)$ is given by $\Sigma(t) = 1 - \kappa \int_0^t \mathrm{d}t' \int \mathrm{d}\mathbf{X}' S(\mathbf{X}')W(\mathbf{X}',t')$.

To solve analytically for $W(\mathbf{X},t)$ is still very difficult. The WF theory takes advantage of the conditional probability T to reduce the dimensionality from $\mathbb{R}^{3N+3}$ to $\mathbb{R}^3$. One can multiply $S(\mathbf{X})$ on both sides of Eq. 6.18 and integrate w.r.t. $\mathbf{X}$ to get Eq. 6.19. If one multiplies with a sink form other than $S(\mathbf{X})$, the result will be an unbalanced sink-sink correlation.

$$\begin{aligned}\int W(\mathbf{X},t)S(\mathbf{X})\mathrm{d}\mathbf{X} = \int P_e(\mathbf{X})S(\mathbf{X})\mathrm{d}\mathbf{X} - \kappa \int_0^t \mathrm{d}t' \int \mathrm{d}\mathbf{X} S(\mathbf{X}) \\ \times \int \mathrm{d}\mathbf{X}' G(\mathbf{X},t|\mathbf{X}',t')S(\mathbf{X}')W(\mathbf{X}',t')\end{aligned} \tag{6.19}$$

If $S(\mathbf{X})W(\mathbf{X},t)$ is the canonical ensemble with the Boltzmann distribution for the end-to-end vector $\mathbf{r}$, one could rewrite Eq. 6.19 in terms of $T(\mathbf{r},t|\mathbf{r}_0,0)$. Specifically, for the Heaviside sink, the space and time dependencies of $W(\mathbf{X},t)$ are separated inside the sink, $W(\mathbf{X},t) = P_e(\mathbf{X})g(t,\kappa)$ when $\|\mathbf{r}\| \leq \epsilon$; for the delta sink, $W(\mathbf{X},t)$ only requires the same space and time separation on the reactive surface, i.e, $W(\mathbf{X},t) = P_e(\mathbf{X})g(t,\kappa)$ when $\|\mathbf{r}\| = \epsilon$. The condition for the delta sink is weaker than for the Heaviside sink, so one would expect that the delta sink performs better than the Heaviside sink. Note that both are weaker than the original WF assumption which requires time and space separation in the whole space. In either case, $W(\mathbf{X},t)$ is not homogenous on the reactive surface. In fact, WF theory allows non-Markovian dynamics for the end-to-end vector of the chain but doesn't capture the full non-Markovian effect.

With the help of the sink-sink correlation function $C(t)$, Eq. 6.19 can be rewritten into

a more compact form Eq. 6.20 and Eq. 6.21, if the separation condition is satisfied.

$$C(t) = \langle S(\mathbf{r},0), S(\mathbf{r},t)\rangle = \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' S(\mathbf{r}) T(\mathbf{r},t|\mathbf{r}',0) S(\mathbf{r}') P_{\mathbf{r}}(\mathbf{r}') \tag{6.20}$$

$$g(t,\kappa)P_0 = P_0 - \kappa \int_0^t g(t',\kappa) C(t-t')\mathrm{d}t', \quad P_0 = \int P_{\mathbf{r}}(\mathbf{r}) S(\mathbf{r}) \mathrm{d}\mathbf{r}. \tag{6.21}$$

One observes that $C(\infty) = (P_0)^2$ for both sinks. For the delta sink, $C(t)$ has a singularity at $t = 0$; for the Heaviside sink, $C(0) = P_0$, the probability of the looped state at equilibrium. In general, there is no closed form expression for $C(t)$.

Assume that on the reactive surface, the time function $g(t,\kappa)$ is asymptotically $q(t)/\kappa + O(1/\kappa^2)$. If I let the microscopic rate constant $\kappa \to +\infty$, then $W(\mathbf{X},t) \to 0$ on the reactive surface asymptotically and the absorbing boundary condition is recovered. Under this assumption, the survival probability will be $\Sigma(t) = 1 - \Big(\int_0^t P_0 q(t')\mathrm{d}t' \Big)$. In this limit, the left hand side of Eq. 6.21 vanishes and I find that

$$P_0 = \int_0^t q(t') C(t-t') \mathrm{d}t'. \tag{6.22}$$

Eq. 6.22 is valid inside the sink for the Heaviside sink and on the reactive surface for the delta sink respectively, and looping time only requires the solution in these regions. This defines a deconvolution problem with the kernel given by $C(t)$. This is an inverse problem and one way to solve it is to use the Laplace transform. The Laplace transform of $q(t)$ and $\Sigma(t)$ are, $\hat{q}(s) = \frac{P_0}{s\hat{C}(s)}$ and $\hat{\Sigma}(s) = \frac{1}{s} - \frac{(P_0)^2}{s^2\hat{C}(s)}$ resepctively. The looping time is given by $\tau = \lim_{s\to 0} \hat{\Sigma}(s)$. Since $C(t)$ is a decreasing function to $C(\infty)$, in the transform variable I have $\hat{C}(s) > C(\infty)/s$. If one approximates $\hat{\Sigma}(s)$ in the denominator by $sC(\infty)$, then $\tau \le \lim_{s\to 0} \frac{\hat{C}(s)}{C(\infty)} - \frac{1}{s}$. In fact, I can show the inequality is an equality as follows.

Define $I(t) = \frac{C(t)}{C(\infty)} - 1$. The improper integral of $I(t)$ from 0 to $+\infty$ is finite, and it is equivalent with $\hat{I}(0) < \infty$ in Laplace transform. The looping time τ is given by

$$\begin{aligned} \tau &= \lim_{s\to 0} \frac{s\hat{C}(s) - (P_0)^2}{s^2\hat{C}(s)} = \lim_{s\to 0} \frac{\hat{I}(s)}{s\hat{I}(s) + 1} \\ &= \hat{I}(0) = \int_0^\infty \mathrm{d}t \Big(\frac{C(t)}{C(\infty)} - 1 \Big) = \lim_{s\to 0} \frac{\hat{C}(s)}{C(\infty)} - \frac{1}{s}. \end{aligned} \tag{6.23}$$

Finally, the integral $\int_0^\infty \mathrm{d}t\Big(\frac{C(t)}{C(\infty)}-1\Big)$ is the famous WF approximation formula which one can evaluate numerically. But the integrand $\frac{C(t)}{C(\infty)}-1$ is clearly not the survival probability $\Sigma(t)$ because this small alteration in the Laplace domain would change the integrand function significantly in the time domain.

We can provide a probabilistic interpretation of the WF formula for the Heaviside sink. $C(0)=P_0$ is the fraction of the polymer that has already formed the loop initially. If one only takes this fraction of the polymers to start the process, the polymer will start to un-loop gradually. We assume that once they are un-looped, these polymers will not form the loop again under the time scale of interest. After sufficiently long time, the distribution of end-to-end vector will be the Gaussian distribution with normalization factor P_0 and the fraction of polymers that is still in the loop state will be $C(\infty)=P_0^2$. So C(t) describes the fraction of the polymer that is still in the loop state at time t if the process starts with the looped polymer. The process of the polymers un-looping themselves is intimately related to the looping problem. We can rewrite the integrand as follows

$$\tau=\frac{1-C(0)}{C(0)}\int_0^\infty \frac{C(t)-C(\infty)}{C(0)-C(\infty)}\mathrm{d}t. \tag{6.24}$$

The expression $\frac{C(t)-C(\infty)}{C(0)-C(\infty)}$ is taken as the approximation of the survival probability for the un-looping process. The integral calculates the expected un-looping time. If one considers this looping process as the dynamical process in a bi-stable system, the ratio $\frac{1-C(0)}{C(0)}$ is the relative stability of the two potential wells, which builds the connection from un-looping time to looping time. A disadvantage with this argument is that it does not work for the delta sink case.

In summary, the approximations made are:

(i) The solution $W(\mathbf{X},t)$ of Eq. 6.17 has the space and time separation inside the sink for the Heaviside sink and on the reactive surface for the delta sink. This approximation is the key step for the construction of this conditional probability w.r.t. canonical ensemble and the reduction in dimension. It is a plausible assumption because the canonical ensemble of polymers will behave as quasi-stationary, at least for the small capture radius case. In order

to go beyond the WF approximation, one has to relax this condition.

(ii) The time function $g(t,\kappa)$ is analytic with respect to the variable $1/\kappa$, such that the absorbing boundary is recovered. Although the assumption is difficult to verify, it is reasonable physically. As the microscopic rate constant κ increases, the reaction is more and more likely to happen once the two ends are within the capture radius. In the limit, the reaction will happen immediately which corresponds to the absorbing condition.

(iii) Instead of solving the deconvolution problem numerically, the WF theory approximates the denominator in the Laplace domain, and it gives a semi-analytical form for the looping time τ directly, but the analytical form for the survival probability is unknown. We will address this issue in Sec. 6.5 and solve numerically for the survival probability $\Sigma(t)$ there.

6.4 Perturbation Method

Although the WF theory provides a good estimate of the looping kinetics, the numerical integration neither provides a reduced model nor gives the scaling law in the two parameters N and ϵ. We are going to apply perturbation techniques on the time integral to extract asymptotic estimates of the looping time.

The sink-sink correlation function $C(t)$ can be expressed as a double integral over two radial variables r and r', after averaging out the azimuthal and polar angles. It uses the fact that $\int \mathrm{d}\mathbf{r}\int \mathrm{d}\mathbf{r}_0 \exp(\mathbf{r}\mathbf{r}') = \int \mathrm{d}r 4\pi r^2 \int \mathrm{d}r' 4\pi r'^2 \frac{\sinh(rr')}{rr'}$.

$$\begin{aligned} C(t) =& \frac{(3/2\pi N)^3}{(1-\phi^2)^{3/2}} \int_0^\infty \mathrm{d}r 4\pi r^2 S(r) \int_0^\infty \mathrm{d}r' 4\pi r'^2 S(r') \exp\Big(-\frac{3}{2N}\frac{r^2+r'^2}{1-\phi^2}\Big) \\ & \times \sinh\Big(\frac{3\phi rr'}{N(1-\phi^2)}\Big) \Big/ \frac{3\phi rr'}{N(1-\phi^2)} \end{aligned} \tag{6.25}$$

We introduce the small dimensionless quantity $x_0 = \frac{3\epsilon^2}{2N} \ll 1$. x_0 is a small quantity since the nature of the looping assumes the capture radius ϵ is much smaller than the average end-to-end distance $\sqrt{N}$. Using x_0 I can find explicit expressions for the integrand $I(t)$.

For the Heaviside sink, the double integral is evaluated by expanding in powers of x_0, $C(t) = \frac{16x_0^3}{9\pi(1-\phi^2)^{3/2}}\Big(1 - \frac{6x_0}{5(1-\phi^2)} + \dots\Big)$. To match the similar form in [34], Pastor *et al.* [87]

proposed the closed form $C(t) \approx \frac{16x_0^3}{9\pi}\Big(1-\phi^2+\frac{4}{5}x_0\Big)^{-3/2}$, which matches the first two order of the expansion in x_0. Then

$$I_H(t) = \frac{C(t)}{C(\infty)} - 1 \approx \Big(\frac{1+\frac{4}{5}x_0}{1-\phi^2(t)+\frac{4}{5}x_0}\Big)^{3/2} - 1 \tag{6.26}$$

The time integral of $I_H(t)$ is roughly approximated in two different time scales. First, in the short time scale, $\phi \approx 1$, the denominator is approximated by $2(1-\phi)+\frac{4}{5}x_0$. The integrand in this time scale is much larger than 1 since x_0 is sufficiently small. The approximation of $I_H(t)$ is, $I_H(t) \approx \Big(\frac{1+\frac{4}{5}x_0}{2(1-\phi(t))+\frac{4}{5}x_0}\Big)^{3/2}$. Second, in the long time scale, $\phi \approx 0$, one could estimate the asymptote as the integrand goes to 0 by exploiting that the quantity $\frac{\phi^2(t)}{1-\phi^2(t)+\frac{4}{5}x_0}$ is small. So the approximation of $I_H(t)$ under this time scale is, $I_H(t) \approx \frac{3}{2}\frac{\phi^2(t)}{1+\frac{4}{5}x_0}$.

For the delta sink, the double integral is evaluated exactly

$$\begin{aligned} C(t) &= \frac{12x_0/N\pi}{\phi\sqrt{1-\phi^2}}\exp\Big(\frac{-2x_0}{1-\phi^2}\Big)\sinh\Big(\frac{2x_0\phi}{1-\phi^2}\Big), C(\infty) = \frac{24x_0^2}{N\pi}\exp(-2x_0) \\ I_{DS}(t) &= \frac{C(t)}{C(\infty)} - 1 = \frac{\exp\Big(2x_0\phi/(1+\phi)\Big) - \exp\Big(-2x_0\phi/(1-\phi)\Big)}{4x_0\phi\sqrt{1-\phi^2}} - 1. \end{aligned} \tag{6.27}$$

Similarly to the Heaviside sink case, it can be approximated in two different time scales. First, in the short time scale, $\phi \approx 1$, so the approximation of $I_{DS}(t)$ is given by $I_{DS}(t) \approx \frac{\exp(x_0)-\exp(-2x_0/(1-\phi))}{4\sqrt{2}x_0\sqrt{1-\phi}}$. Second, in the long time scale, $\phi \approx 0$, one can use a Taylor expansion to estimate the limit of the integrand. So the approximation of $I_{DS}(t)$ is $I_{DS}(t) \approx \frac{3}{2}\phi^2$.

Since the time correlation function for the end-to-end vector $\phi(t)$ doesn't have a closed analytical form, the integral is still not feasible analytically. We will approximate $\phi(t)$ in three different time scales: First, when the time scale is within the relaxation time for the largest mode, $t \le \tau_N = \frac{1}{12}$, take the approximation $\exp(-\gamma_p t) = 1 - \gamma_p t + O(t^2)$. Then $\phi(t)$ can be approximated by $\phi(t) \approx 1 - \frac{6t}{N}$. Second, when time t between two time scales, $\tau_N \ll t \ll \tau_3 = \frac{N^2}{27\pi^2}$, the approximation will be based on $\frac{1}{\tan^2(\theta_p)} \approx \frac{1}{\theta_p^2}$ and $\gamma_p \approx p^2\gamma_1$. Both approximations work well on small p that also contribute the most to the correlation function, $\phi(t) \approx 1 - \frac{8}{\pi^2}\sum_{p \text{ odd}}\frac{1}{p^2}\Big(1-\exp(-p^2\gamma_1 t)\Big)$. One has to further approximate by turning the summation into the integral. Define $x = p\sqrt{\gamma_1 t}$, the coefficient $N^{-1} \ll \sqrt{\gamma_1 t} \ll 1$ is a small

quantity, $\phi(t) \approx 1 - \frac{4}{\pi^2}\sqrt{\gamma_1 t}\int_0^{+\infty}\frac{1}{x^2}(1-\exp(-x^2))\mathrm{d}x = 1 - \frac{4}{N}\sqrt{\frac{3t}{\pi}}$. Last, when the time $t \gg \tau_3$, all other modes are relaxed except the first mode. Then it can be approximated by $\phi(t) \approx \frac{8}{\pi^2}\exp(-\gamma_1 t)$. In summary, the end-to-end vector correlation function $\phi(t)$ has the following analytical approximation

$$\phi(t) \approx \begin{cases} 1 - \frac{6t}{N} & \text{Short Timescale} \\ 1 - \frac{4}{N}\sqrt{\frac{3t}{\pi}} & \text{Median Timescale} \\ \frac{8}{\pi^2}\exp(-\gamma_1 t) & \text{Long Timescale} \end{cases} \tag{6.28}$$

The effective diffusion coefficient D_{eff} for the end-to-end vector is defined by the end-to-end correlation function $\phi(t)$,

$$D_{eff}(t) = \frac{\langle(\mathbf{R}_{ee}(t) - \mathbf{R}_{ee}(0))^2\rangle}{6t} = \frac{N(2-2\phi(t))}{6t} \tag{6.29}$$

In the short timescale, $D_{eff} = 2$ which is time homogeneous. It also verifies the choice of diffusion coefficient in SSS theory. SSS theory doesn't capture the behavior from other timescales, so it cannot reproduce the mixed scaling law.

Numerical simulations show that all three asymptotic results perform very well under the appropriate timescale. Specifically, for $N = 75$, the short timescale approximation works better than the median for $t \le t_1 = \frac{4}{3\pi} \approx 5.1\tau_N$ and the long timescale approximation is better than the median for $t \ge 2.8\tau_3$. Notice when $t \approx \tau_3$, the correlation function ϕ is about 0.76, so one can still treat $\phi \approx 1$ under the median timescale. Then I_H and I_{DS} have analytical approximations $\bar{I}_H$ and $\bar{I}_{DS}$ as follows,

$$\bar{I}_H(t) = \begin{cases} \left(\frac{1+\frac{4}{5}x_0}{\frac{12t}{N}+\frac{4}{5}x_0}\right)^{3/2} & \text{Short Timescale} \\ \left(\frac{1+\frac{4}{5}x_0}{\frac{8}{N}\sqrt{\frac{3t}{\pi}}+\frac{4}{5}x_0}\right)^{3/2} & \text{Median Timescale} \\ \frac{96}{\pi^4(1+\frac{4}{5}x_0)}\exp(-\frac{6\pi^2 t}{(N+1)^2}) & \text{Long Timescale} \end{cases} \tag{6.30}$$

$$\bar{I}_{DS}(t) = \begin{cases} \frac{\exp(x_0)-\exp(-\frac{Nx_0}{3t})}{8\sqrt{3}x_0\sqrt{\frac{t}{N}}} & \text{Short Timescale} \\ \frac{\exp(x_0)-\exp(-\frac{Nx_0}{2}\sqrt{\frac{\pi}{3t}})}{8\sqrt{2}x_0\sqrt{\frac{1}{N}\sqrt{\frac{3t}{\pi}}}} & \text{Median Timescale} \\ \frac{96}{\pi^4}\exp(-\frac{6\pi^2 t}{(N+1)^2}) & \text{Long Timescale} \end{cases} \tag{6.31}$$

Specifically, for $N = 75$ and $\epsilon = 0.75$, the analytical approximations perform very well for both sinks. The time range within which they deviate most is on the boundary of the short timescale and the median timescale. In terms of numerical integration, most of deviation is contributed by the median timescale and its boundaries as expected (figure not shown). The advantage of the approximation is that the integral which provides an estimate of the looping time can be evaluated analytically. Set the two timescale break points as $t_1 = \frac{4}{3\pi}$ and $t_2 = 3\tau_3$. The second breakpoint between the median and long timescales depends on the number of beads, N. We roughly choose $3\tau_3$ as the second breakpoint but this will not change the behavior of scaling law qualitatively.

For the Heaviside sink, one can compute the mixed scaling law explicitly,

$$\begin{aligned} \int_0^\infty \bar{I}_H(t)\mathrm{d}t &= \frac{N\sqrt{N}}{\epsilon}\Big(\frac{1}{6}\sqrt{\frac{5}{6}}(1+\frac{4}{5}x_0)^{3/2}\Big)\Big(1-\sqrt{\frac{6\epsilon^2/5}{16/\pi+6\epsilon^2/5}}\Big) \\ &+ (N+1)^2\frac{16}{\pi^6(1+\frac{4}{5}x_0)}\exp(-\frac{2}{3}\frac{N^2}{(N+1)^2}) + N\sqrt{N}\Big(\frac{\pi}{48}(1+\frac{4}{5}x_0)^{3/2}\Big) \\ &\times\Big(\sqrt{\frac{8N}{\pi}\sqrt{\frac{1}{3\pi}}+\frac{6\epsilon^2}{5}} - \sqrt{16/\pi+6\epsilon^2/5} + \frac{6\epsilon^2/5}{\sqrt{\frac{8N}{\pi}\sqrt{\frac{1}{3\pi}}+\frac{6\epsilon^2}{5}}} - \frac{6\epsilon^2/5}{\sqrt{16/\pi+6\epsilon^2/5}}\Big) \end{aligned} \tag{6.32}$$

When the capture radius $\epsilon \ll 1$, this integral is $\int_0^\infty \bar{I}_H(t)\mathrm{d}t \approx h_1\frac{N\sqrt{N}}{\epsilon} + h_2 N\sqrt{N} + h_3 N^2$ where h_1, h_2 and h_3 are constants. In the limit of $N \to +\infty$, I can ignore the middle term $h_2 N\sqrt{N}$ and recover the mixed scaling law hypothesized before.

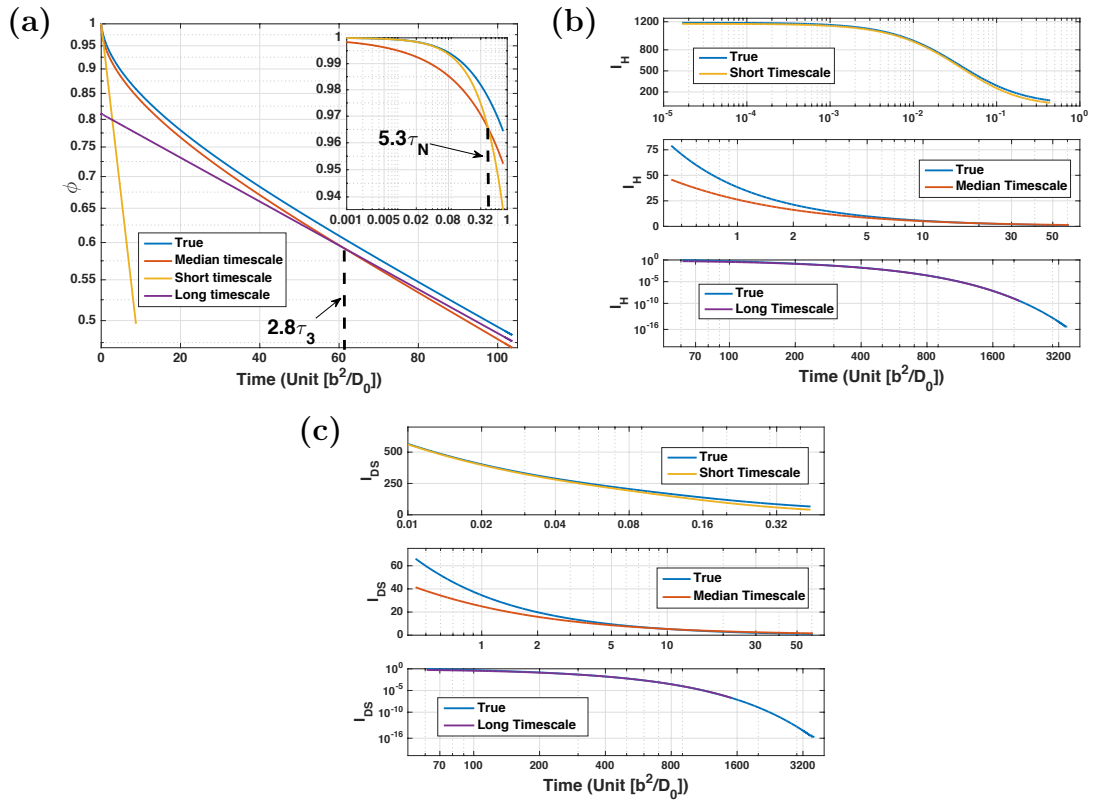

Figure 6.1: (a) Comparison of numerical calculation and analytical approximations for the end-to-end vector correlation function $\phi(t)$ under three different timescales for $N = 75$. (b) The approximated integrand $\bar{I}_H(t)$ from Eq. 6.30 and compared to the numerical evaluation from Eq. 6.26 for $N = 75$ and $\epsilon = 0.75$. (c) The approximated integrand $\bar{I}_{DS}(t)$ from Eq. 6.30 and compared to the numerical evaluation from Eq. 6.27 for $N = 75$ and $\epsilon = 0.75$.

For delta sink, the integral is much more complicated,

$$
\begin{aligned}
I_1 =& \frac{N\sqrt{N}}{\epsilon^2}\frac{1}{9\sqrt{\pi}}\Big(\exp(x_0)-\exp\Big(-\frac{3\pi\epsilon^2}{8}\Big)\Big)+\frac{N\sqrt{N}}{\epsilon}\frac{\sqrt{\pi}}{6\sqrt{6}}\Big(1-\operatorname{erf}\Big(\sqrt{\frac{3\pi\epsilon^2}{8}}\Big)\Big)\\
I_2 =& \frac{N^3}{\epsilon^2}\frac{1}{2^{1/2}3^{15/4}\pi^{5/4}}\Big(\exp(x_0)-\exp\Big(-\frac{\sqrt{3\pi^3}}{2}x_0\Big)\Big)-N\sqrt{N}\frac{\sqrt{\pi}}{18}\exp\Big(-\frac{3\pi\epsilon^2}{8}\Big)\\
&+N^2\frac{\sqrt[4]{\frac{\pi}{3}}}{18\sqrt{2}}\exp\Big(-\frac{\sqrt{3\pi^3}}{2}x_0\Big)-\frac{N\sqrt{N}}{\epsilon^2}\frac{2}{27\sqrt{\pi}}\Big(\exp(x_0)-\exp\Big(-\frac{3\pi\epsilon^2}{8}\Big)\Big)\\
&+N\sqrt{N}\epsilon\frac{\pi^{3/2}}{12\sqrt{6}}\Big(\operatorname{erf}\Big(\frac{\epsilon}{\sqrt{N}}\frac{(3\pi)^{3/4}}{2}\Big)-\operatorname{erf}\Big(\frac{(3\pi)^{1/2}}{2\sqrt{2}}\epsilon\Big)\Big).\\
I_3 =& (N+1)^2\frac{16}{\pi^6}\exp\Big(-\frac{2}{3}\frac{N^2}{(N+1)^2}\Big)
\end{aligned}
$$

$$
\int_0^\infty \bar{I}_{DS}(t)\mathrm{d}t = I_1+I_2+I_3 \tag{6.33}
$$

Similarly when $\epsilon \ll 1$, one can use Taylor expansions for $\exp(x)$ and $\operatorname{erf}(x)$ functions at $x=0$. The integral is roughly $\int_0^\infty \bar{I}_{DS}(t)\mathrm{d}t \approx d_1\frac{N\sqrt{N}}{\epsilon}+d_2N\sqrt{N}+d_3N^2$ as well, where d_1, d_2 and d_3 are constants. In the limit $N\to+\infty$, it recovers the mixed scaling law again. Specifically, d_1 is mostly contributed by the short timescale approximation in I_1 and is about $\frac{\sqrt{\pi}}{6\sqrt{6}}$. It exactly matches the SSS result as predicted in [87]. It implies that the SSS and WF theories give the same asymptotic result $\frac{N\sqrt{N}}{\epsilon}$ in the limit $\epsilon\to 0$ for given N. In practice, this asymptotic result is realized when ϵ is extremely small. d_2 and d_3 are mostly contributed by the median and long timescale approximations. This is also predicted by Doi in [34] and Doi provided a dynamical explanation as well. So both $N\sqrt{N}$ and N^2 term are considered as the next order approximation result when ϵ is still relatively small. One would insightfully rewrite the looping time with the mixed scaling law as $\tau=\frac{N\sqrt{N}}{\epsilon}\Big(d_1+(d_2+d_3\sqrt{N})\epsilon+O(\epsilon^2)\Big)$. Note in the derivation of this scaling law, I assume N as a constant is much larger than 1, i.e, $N\gg 1$. From the scaling law, two seemingly contradicting results from Doi and SSS are the consequence of the non-uniforming convergence of $\epsilon\to 0$ and $N\to+\infty$. Such results are typical in singular perturbation problems.

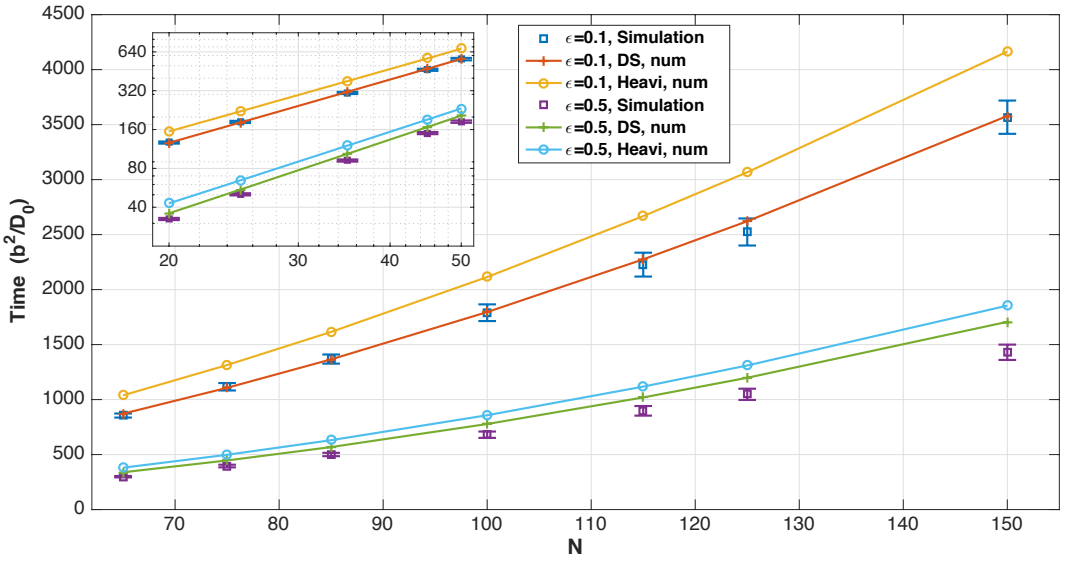

Figure 6.2: Dependence of looping time on N for two capture radius of $\epsilon = 0.1$ and $\epsilon = 0.5$. The looping time is estimated from a Monte Carlo simulation and compared to the numerical integration results in WF theory using Eq. 6.26 and Eq. 6.27.

6.5 Numerical Simulation

The Monte Carlo simulation algorithm is based on [87] but with more sample points and smaller time step. The simulation results are also within the range of their paper. The numerical solver is the Euler-Maruyama method, and the initial condition is sampled from the equilibrium distribution. The trajectory is terminated once the end-to-end distance is within the capture radius and record the passage time. The time step is chosen adaptively: when the end-to-end distance is within 2 Kuhn lengths, much finer time step is picked to prevent overshooting; but a larger time step is permitted when the end-to-end distance is outside of the range. The time step defined is

$$\Delta t = \begin{cases} \Delta_l + \Delta_h \sin\left(\frac{\pi}{2}\left(\frac{\|\mathbf{r}\|-\epsilon}{2-\epsilon}\right)^2\right) & \epsilon \le \|\mathbf{r}\| \le 2 \\ \Delta_l + \Delta_h & \|\mathbf{r}\| > 2 \end{cases} \tag{6.34}$$

with $\Delta_l = 10^{-6}$ and $\Delta_h = 10^{-3}$. This choice of time step is finer than that in [87]. For each parameter set, there are at least $n = 2000$ samples. The passage time is considered to follow a single exponential distribution in [87], so the 95% confidence interval for the mean passage time is $\bar{\tau} \pm \frac{1.96}{\sqrt{n}}\bar{\tau}$. We include this confidence interval for comparison. In fact, one can use the Lilliefors test, which is an improvement of the Kolmogorov-Smirnov test, to return a test decision for the null hypothesis that the passage time comes from an exponential distribution. The result is 1 if the test rejects the null hypothesis at 1% significant level, 0 otherwise [29]. We applied the test to our simulation data for various parameters (I used $N = 20, 25, 35, 45, 50, 65, 75, 85, 100, 115, 125, 150$ and $\epsilon = 0.1, 0.25, 0.5, 0.75, 1, 2$ for a total of 72 parameter combinations). The test shows: when $\epsilon = 0.1$ for all N and $\epsilon = 0.25$ with $N \geq 100$, the test does not reject the null hypothesis at 1% significant level; for other parameter ranges, the test rejects the null hypothesis and that means the data do not fit an exponential distribution. Our result is different from Amitai's result for large N, where they claim that the passage time does not follow a single exponential distribution when $N > 64$. In fact, I can visualize the survival probability for the large N case in Fig. 6.4.

Table 6.1 and Fig. 6.2 show that the WF theory overestimates the mean passage time, as can also be proved by using a variational principle [89]. They also verify our argument that the delta sink should perform better than the Heaviside sink. The analytical results underestimate the numerical integration at various points in the parameter range, by less than 15 percent for Heaviside sink and by less than 10 percent for delta sink. From Table 6.1 and Fig. 6.2, I see that the delta sink results are in remarkable agreement with the simulation for small capture radius ϵ. This implies that the space and time separation approximation relies mostly on the small capture radius assumption and not on the large number of beads. So, it is reasonable to use the WF theory as the simulation results when $\epsilon \ll 1$. At the same time, both analytical results predict that the WF theory has the mixed scaling law with $N\sqrt{N}$, $\frac{N\sqrt{N}}{\epsilon}$ and N^2. We use the WF theory for both sinks under the parameter range $0.1 \leq \epsilon \leq 0.15$ and $100 \leq N \leq 150$ to fit the scaling law. One could use simulation results to fit but the computational cost is enormous. The coefficient d_1 is estimated as 0.1225, very

N, ϵ	Sample points	Simulation	H.n.	H.a.	DS. n.	DS. a.	SSS
50, 0.1	8000	563±12	677	632	564	530	426
50, 0.5	8000	184±4	233	200	205	186	85
50, 1.0	8000	113±2	163	143	141	134	43
75, 0.1	4000	1117±34	1314	1230	1107	1048	783
75, 0.5	4000	396±12	498	437	447	414	157
75, 1.0	4000	261±8	368	330	326	317	783
100, 0.1	2000	1790±76	2114	1984	1796	1710	1206
100, 0.5	2000	681±29	858	762	778	733	241
100, 1.0	2000	466±20	655	595	590	581	121

Table 6.1: Comparison of theoretical results and simulations for selected values of N and the capture radius ϵ. H. n. (Heaviside numerical) and DS. n. (Delta sink numerical) are obtained from the numerical integration in WF theory using Eq. 6.26 and Eq. 6.27. H. a. (Heaviside analytical) and DS. a. (Delta sink analytical) are analytical results from Eq. 6.32 and Eq. 6.33. SSS is the analytical result from Eq. 6.13.

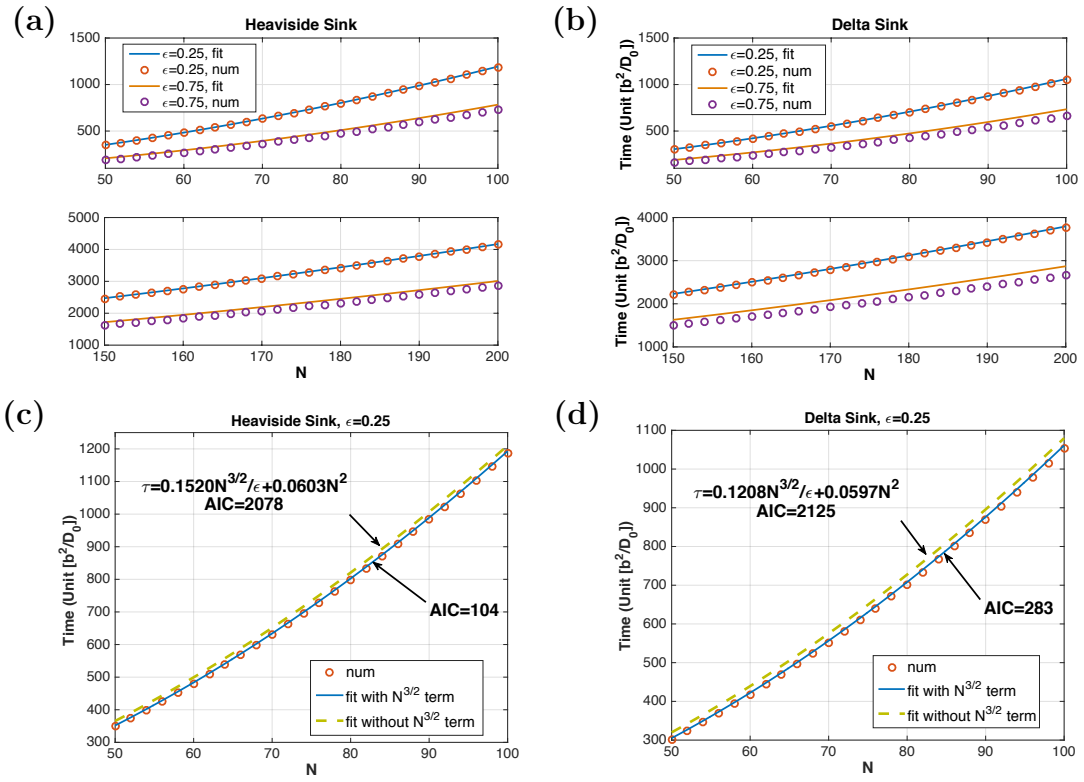

Figure 6.3: (a,b) Comparison of results from regression fit and the WF theory. The fitted lines are plotted according to Eq. 6.35 and Eq. 6.36. The WF results are obtained from numerical integration using Eq. 6.26 and Eq. 6.27. (c,d) Comparison of regression fit with and without $N\sqrt{N}$ term for $\epsilon = 0.25$ and $50 \leq N \leq 100$.

close to the analytically predicted result $\frac{\sqrt{\pi}}{6\sqrt{6}} \approx 0.1206$.

$$\int_0^\infty I_H(t)\mathrm{d}t = 0.1536\ \frac{N\sqrt{N}}{\epsilon} - 0.0982\ N\sqrt{N} + 0.0677\ N^2 + \epsilon_1 \tag{6.35}$$

$$\int_0^\infty I_{DS}(t)\mathrm{d}t = 0.1225\ \frac{N\sqrt{N}}{\epsilon} - 0.1060\ N\sqrt{N} + 0.0677\ N^2 + \epsilon_2 \tag{6.36}$$

With the scaling law at hand, I use the regression equation 6.35 and 6.36 to test other parameter ranges given by $\epsilon = 0.25, 0.75$ and $50 \leq N \leq 100, 150 \leq N \leq 200$ (see Fig. 6.3). When $\epsilon = 0.25$, which is relatively small, the fit agrees remarkably well with the WF approximation for both sinks and in both ranges of N. However, when $\epsilon = 0.75$, the fit

starts to deviate from the WF theory for both sinks and the difference grows with N. Since I know that the WF theory can *overestimate* the looping time and the predictions from the regression fit are larger than those of the WF theory, I conclude that the predictions of the regression fit are not accurate for $\epsilon = 0.75$. For $\epsilon = 0.25$, if one omits the new term $N\sqrt{N}$, Fig. 6.3 shows that for both sinks the regression lines deviate from the numerical integration points in this parameter range and also the values of the Akaike information criterion (AIC) are much higher. The AIC is a measure of the relative quality of a statistical model for a given set of data and and the lower AIC model is better. This indicates that the $N\sqrt{N}$ term needs to be included in the model.

In the previous section, it was mentioned that WF approximates the denominator in the Laplace transform domain to get a semi-analytic form for the looping time and, as a result, the integrand is not a survival probability $\Sigma(t)$. However, the Volterra integral equation 6.22 can be solved numerically by the trapezoidal method. Correspondingly, the survival probability is found numerically and each moment of passage time can be calculated. The most useful moment of passage time, of course, is the mean.

Rewrite Eq. 6.22 to get

$$1 = \int_0^t \Big(P_0 q(t')\Big)\Big(\frac{C(t-t')}{C(\infty)}\Big)\mathrm{d}t' \tag{6.37}$$

The kernel $C(t)/C(\infty)$ is the renormalized sink-sink correlation function and $P_0 q(t)$ is the function to solve for. The survival probability is $\Sigma(t) = 1 - \int_0^t P_0 q(t')\mathrm{d}t'$. There is a difficulty because the kernel has a singular point at $t = 0$ and the numerical integration is very stiff. With the perturbation result in Eq. 6.30, I know the order of the singularity is $1/2$, so careful handling of the kernel function at short time range t is needed.

We plot in Fig. 6.4 the survival probability $\Sigma(t)$ predicted by the WF theory for two different values of N and two different capture radius values ϵ. As expected, the time integral of the survival probability recovers looping time from the WF theory in Eq. 6.23. In addition, I compare with the survival probability from the simulation data. This is computed by creating the histogram for the passage time and calculating the cumulative probability

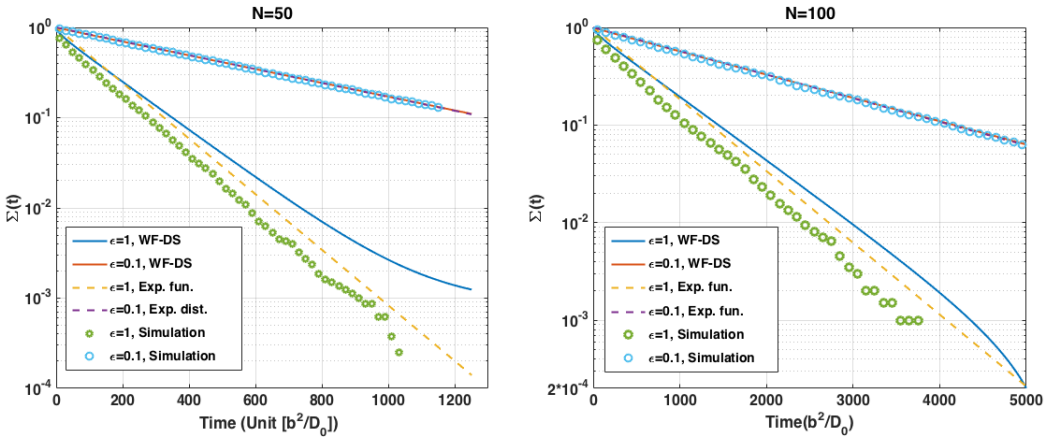

Figure 6.4: Comparison of survival probability from WF theory with delta sink and from simulations for different N and different ϵ. The exponents of the exponential functions are the inverses of the looping time from WF theory in delta sink according to Eq. 6.27.

function for it. Then, the survival probability is one minus the cumulative probability function of the passage time. If the passage time follows a single exponential distribution, the survival probability will be an exponential function with the exponent given by the inverse of the looping time. For small capture radius $\epsilon = 0.1$, the survival probability function from WF agrees with the one from the simulation extremely well. It is interesting to note that the survival probability function obtained through simulation is an exponential function, even for large N. This was also verified previously. The exponent of the exponential function is the inverse of the WF looping time. However, for large capture radius $\epsilon = 1$, the survival probability function from the WF theory is clearly no longer an exponential function and it deviates from the simulation results which also do not show exponential behavior.

6.6 Stochastic Model Reduction

The WF theory as a general method of model reduction deserves further attention. One of the reasons that WF being successful is that it fully utilizes the properties of Gaussian process. In specific, the conditional probability for end-end vector $T(\mathbf{r}, t|\mathbf{r}_0, 0)$ is explicitly given in

Eq. 6.9. But T is not a probability propagator since it doesn't satisfy the Kolmogorov-Chapman equation. This also indicates the reduced model of the end-end vector $\mathbf{r}$ is non-Markovian. Even with the conditional probability T, it doesn't mean the dynamics of the end-end vector is known. Particularly, it will be interesting to study the memory kernel in this non-Markovian dynamics.

One possible description is given by the generalized Langevin equation [39].

$$M\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}(t) = -\int_0^t K(t-s)\mathbf{r}(s)\mathrm{d}s + \mathbf{F}(t) \tag{6.38}$$

where K is the memory kernel. The fluctuating force $\mathbf{F}(t)$ links with the memory kernel K through the fluctuation-dissipation theorem.

$$\langle \mathbf{F}_i(t)\mathbf{F}_j(s)\rangle = 2K(|t-s|)\delta_{ij},\ \langle \mathbf{F}(t)\rangle = 0 \tag{6.39}$$

If $K(t-s) = const\cdot\delta(t-s)$, then it goes back to the Markovian approximation in Sec. 6.2. The probability distribution of this generalized Langevin equation 6.38 follows the Gaussian process. It is the simplest dynamical model of non-Markovian Gaussian process, which plays an important role in understanding the stochastic model reduction.

The stochastic force satisfies $\langle \mathbf{r}(0)\mathbf{F}(t)\rangle = 0$ and it can be interpreted as a manifestation of the causality [107]. Multiply $\mathbf{r}(0)$ and average over the equilibrium ensemble we get

$$M\frac{\mathrm{d}}{\mathrm{d}t}\langle \mathbf{r}(t)\mathbf{r}(0)\rangle = -\int_0^t K(t-s)\langle \mathbf{r}(s)\mathbf{r}(0)\rangle\mathrm{d}s + \langle \mathbf{F}(t)\mathbf{r}(0)\rangle \tag{6.40}$$

Using Laplace transform, one can get

$$\hat{K}(z) = M\frac{1-z\hat{\phi}(z)}{\hat{\phi}(z)} \tag{6.41}$$

The conditional probability $P(\mathbf{r},t|\mathbf{r}_0,0)$,

$$P(\mathbf{r},t|\mathbf{r}_0,0) = \Big(2\pi\frac{1}{M}(1-\phi^2(t))\Big)^{-3/2}\exp\Big(-\frac{1}{2}M\frac{(\mathbf{r}-\phi(t)\mathbf{r}_0)^2}{1-\phi^2(t)}\Big) \tag{6.42}$$

Compare with $T(\mathbf{r}, t|\mathbf{r}_0, 0)$ in Eq. 6.9, I can identify $M = 3/N$, which is the effective spring constant for dumbbell model. So

$$\hat{K}(z) = \frac{3}{N}\frac{1 - z\hat{\phi}(z)}{\hat{\phi}(z)}. \tag{6.43}$$

$$= \frac{3}{N}\frac{\sum_{p \text{ odd}} \frac{1}{\tan^2(\theta_p)}\frac{\gamma_p}{z+\gamma_p}}{\sum_{p \text{ odd}} \frac{1}{\tan^2(\theta_p)}\frac{1}{z+\gamma_p}} = \frac{3}{N} \cdot \Big(\frac{6}{N} + \frac{f(z)}{g(z)}\Big) \tag{6.44}$$

where $f(z), g(z)$ are polynomials and the highest order of $f(z)$ is smaller than the highest order of $g(z)$.

The inverse Laplace transform of Eq. 6.43 is very difficult to find analytically. But $K(t)$ has the following form,

$$K(t) = \frac{3}{N}\Big(\frac{6}{N}\delta(t) - \sum_i a_i \exp(-b_i t)\Big) \tag{6.45}$$

where a_i and b_i are some positive constants determined by θ_p and γ_p. Then the generalized Langevin equation 6.38 is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{r}(t) = -\frac{6}{N}\mathbf{r}(t) + \int_0^t \sum_i a_i \exp(-b_i(t-s))\mathbf{r}(s)\mathrm{d}s + \mathbf{F}(t) \tag{6.46}$$

where $\mathbf{F}$ has the following correlation,

$$\langle \mathbf{F}(t)\mathbf{F}(s)\rangle = -\frac{4N}{3}\sum_i a_i \exp(-b_i|t-s|) + 4\delta(|t-s|). \tag{6.47}$$

Associated with this generalized Langevin equation is the following partial differential equation,

$$\frac{\partial}{\partial t_2}T(\mathbf{r}_2, t_2|\mathbf{r}_1, t_1) = -\frac{\dot{\phi}(t_2-t_1)}{\phi(t_2-t_1)}\nabla \cdot \Big(\mathbf{r}_2 T(\mathbf{r}_2, t_2|\mathbf{r}_1, t_1)\Big) - \frac{N}{3}\frac{\dot{\phi}(t_2-t_1)}{\phi(t_2-t_1)}\nabla^2 T(\mathbf{r}_2, t_2|\mathbf{r}_1, t_1) \tag{6.48}$$

with initial condition $T(\mathbf{r}_2, t_1|\mathbf{r}_1, t_1) = \delta(\mathbf{r}_2 - \mathbf{r}_1)$. This is not a Fokker-Planck equation because of the time-dependent coefficient $\frac{\dot{\phi}(t_2-t_1)}{\phi(t_2-t_1)}$.

The delta kernel corresponds to the Markovian contribution, and the exponential kernel corresponds to the non-Markovian contribution. The Markovian contribution coincides with

the Markovian approximation in Eq. 6.10. Note the effective diffusion coefficient is also recovered as $2D_0$.

Although the analytical result is out of reach at this point, the inverse Laplace transform is possible to calculated numerically. In Fig. 6.5, I plotted $-K(t)$ in the positive t region for $N = 50, 75, 100, 150$. One can tell the kernel is the summation of exponential functions and the tail is governed by the smallest exponential. The kernel memory is quite long for large N, in particular, when $N = 150$, the estimate tail slope is about $-1/83$. So the relaxation time 83 is even longer than $N/3 = 50$. On the other hand, when $N = 50$, the estimated tail slope is about $-1/10$ so the relaxation time is smaller than $N/3$. Therefore, it is necessary to consider this memory term for large N. The form in Eq. 6.46 is very similar with the Mori-Zwanzig formalism [26]. The Mori-Zwanzig formalism was discussed for the deterministic systems, but in some applications, the system might be intrinsically stochastic, like Rouse model. It will be interesting to see how to use this formalism to reduce the stochastic models from the Langevin equation 6.1 for the position of bead $\mathbf{R}_i$ to the generalized Langevin equation 6.46 for the end-end vector $\mathbf{r}$. One possible way is to treat the noise ω as the pre-recorded signal and apply the Mori-Zwanzig formalism for this time-inhomogeneous differential equation with $\mathbf{r}$ as the resolved variable. After averaging the canonical ensemble of all possible sequences, the generalized Langevin equation 6.46 may be deducted directly. This may point to a future direction for stochastic model reduction.

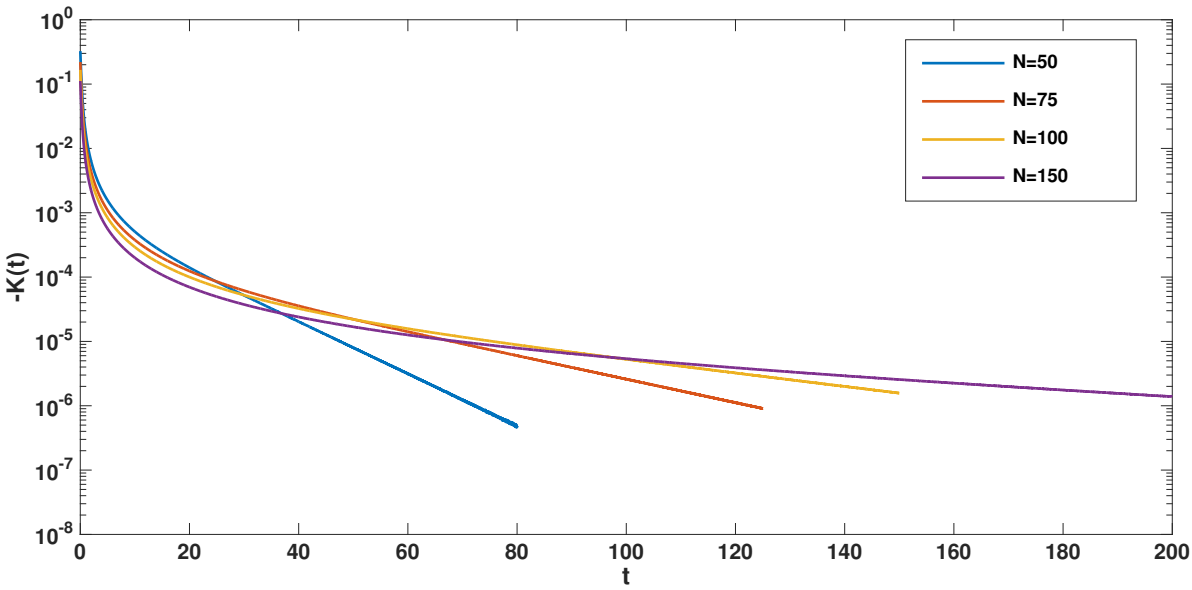

Figure 6.5: The memory kernel $-K(t)$ as the function of t in log scale. Here the delta function part is ignored.

Chapter 7

CONCLUSION AND FUTURE WORK

The RDS theory can be described as the intersection of stochastic processes with dynamical systems. It is an example for the fact that a symbiosis of two mathematical disciplines at the right moment amounts to opening a scientific gold mine, both conceptually and as far as significant applications are concerned [5].

From a dynamical system point of view, RDS deals with measure-preserving dynamical system. It is characteristic feature of the theory of RDS that the problem evolves some ergodic theory and ergodic theorems. The most important one is multiplicative ergodic theorem, which is at the heart of RDS. The Lyapunov spectrum defined from this theorem is the random substitute of the spectrum in linear algebra. From a stochastic process point of view, RDS gives much richer structure than a family of stochastic processes, each evolving for a given initial value. In fact, RDS gives us a family of random transformations. Hence, it makes the study of simultaneous sequences starting from multiple initial values possible. In addition, many quantities previously studied in stochastic process, are also represented in the corresponding RDS.

In the era of BIG DATA, a quote from [23], which has also partially inspired my work, is still very relevant: "[I]n the study of deterministic dynamical systems, environmental noise tends to be suppressed or, at most, plays a secondary role, whereas in the study of statistics the deterministic dynamic kernel of the random generating mechanism tends to give way to the more macroscopic characterization such as the mean functions, the covariance functions, the spectral functions and so on."

In the present work, we have seen that environmental noise which we termed extrinsic noise, actually play very different roles in a stochastic dynamics with the intrinsic noise. It

turn out that RDS is very good modeling framework for extrinsic noise. One interesting property of the extrinsic noise, the noise-induced synchronization, is well studied in the setting of RDS. The possible future direction is to set up the mathematical framework for *random Markov systems* (RMS) which will be the generalization of RDS. At each step, the Markov process is picked instead of the deterministic transformation. A key motivation of this class of models is to distinguish intrinsic noise and extrinsic noise.

One of the particularly attractive features of the discrete state formulation of RDS is the possibility of various in-depth investigations of complex dynamics using a broad spectrum of mathematical tools that are accessible to biologists, chemists, engineers and data scientists. In the present work, the noise-induced synchronization has been found application in efficient inference in the hidden Markov model, which is the discrete state model. The natural extension will be the continuous state version of HMM. This is also known as *nonlinear filters*. In particular, if all latent and observed variables have Gaussian distributions, it is the celebrated Kalman filter. Then it needs the infinite-dimensional version of multiplicative ergodic theorem which is more technical. But such idea of block sampling is still valid. It will be interesting to estimate the length of the block of the observation sequence such that each block is almost independent.

BIBLIOGRAPHY

[1] Linda J. S. Allen, *An introduction to stochastic processes with applications to biology*, Second, CRC Press, Boca Raton, FL, 2011. MR2560499

[2] A. Amitai, I. Kupka, and D. Holcman, *Computation of the mean first-encounter time between the ends of a polymer chain*, Phys. Rev. Lett. **109** (2012Sep), 108302.

[3] Hassan Arbabi and Igor Mezić, *Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator*, SIAM J. Appl. Dyn. Syst. **16** (2017), no. 4, 2096–2126. MR3720364

[4] Ludwig Arnold, *Random dynamical systems*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 1998. MR1723992

[5] ———, *Trends and open problems in the theory of random dynamical systems*, Probability towards 2000 (New York, 1995), 1998, pp. 34–46. MR1632623

[6] Ludwig Arnold and Hans Crauel, *Random dynamical systems*, Lyapunov exponents (Oberwolfach, 1990), 1991, pp. 1–22. MR1178943

[7] R. Atar, *Exponential decay rate of the filter's dependence on the initial distribution*, The Oxford handbook of nonlinear filtering, 2011, pp. 299–318. MR2884600

[8] Rami Atar and Ofer Zeitouni, *Lyapunov exponents for finite state nonlinear filtering*, SIAM J. Control Optim. **35** (1997), no. 1, 36–55. MR1430282

[9] Viviane Baladi, *Positive transfer operators and decay of correlations*, Advanced Series in Nonlinear Dynamics, vol. 16, World Scientific Publishing Co., Inc., River Edge, NJ, 2000. MR1793194

[10] Peter H. Baxendale, *Statistical equilibrium and two-point motion for a stochastic flow of diffeomorphisms*, Spatial stochastic processes, 1991, pp. 189–218. MR1144097

[11] O. Bènichou, M. Coppey, M. Moreau, and G. Oshanin, *Kinetics of diffusion-limited catalytically activated reactions: An extension of the Wilemski-Fixman approach*, The Journal of Chemical Physics **123** (2005), no. 19, 194506.

[12] O. Bénichou, T. Guérin, and R. Voituriez, *Mean first-passage times in confined media: from Markovian to non-Markovian processes*, J. Phys. A **48** (2015), no. 16, 163001, 43. MR3335699

[13] Peter G. Bergmann and Joel L. Lebowitz, *New approach to nonequilibrium processes*, Phys. Rev. (2) **99** (1955), 578–587. MR0074332

[14] R. Bhar and S. Hamori, *Hidden markov models: Applications to financial economics*, Advanced Studies in Theoretical and Applied Econometrics, Springer US, 2006.

[15] Rabi Bhattacharya and Mukul Majumdar, *Random dynamical systems*, Cambridge University Press, Cambridge, 2007. Theory and applications. MR2290421

[16] Garrett Birkhoff, *Three observations on linear algebra*, Univ. Nac. Tucumán. Revista A. **5** (1946), 147–151. MR0020547

[17] R. M. Blumenthal and H. H. Corson, *On continuous collections of measures*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory, 1972, pp. 33–40. MR0397384

[18] Carlos Bocker-Neto and Marcelo Viana, *Continuity of Lyapunov exponents for random two-dimensional matrices*, Ergodic Theory Dynam. Systems **37** (2017), no. 5, 1413–1442. MR3667994

[19] P.B. Bollobas, *Modern graph theory*, Graduate Texts in Mathematics, Springer New York, 1998.

[20] P. C. Bressloff and J. M. Newby, *Stochastic models of intracellular transport*, Rev. Mod. Phys. **85** (2013), 135–195.

[21] S. Roy Caplan and Doron Zeilberger, *T. L. Hill's graphical method for solving linear equations*, Adv. in Appl. Math. **3** (1982), no. 4, 377–383. MR682624

[22] Alexandre N. Carvalho, José A. Langa, and James C. Robinson, *Attractors for infinite-dimensional non-autonomous dynamical systems*, Applied Mathematical Sciences, vol. 182, Springer, New York, 2013. MR2976449

[23] Kung-Sik Chan and Howell Tong, *Chaos: a statistical perspective*, Springer Series in Statistics, Springer-Verlag, New York, 2001. MR1851668

[24] Z. Y. Chen, H.-K. Tsao, and Y.-J. Sheng, *Diffusion-controlled first contact of the ends of a polymer: Crossover between two scaling regimes*, Phys. Rev. E **72** (2005Sep), 031804.

[25] A. F. Cheviakov and M. J. Ward, *Optimizing the principal eigenvalue of the Laplacian in a sphere with interior traps*, Math. Comput. Modelling **53** (2011), no. 7-8, 1394–1409. MR2782819

[26] Alexandre J. Chorin and Ole H. Hald, *Stochastic tools in mathematics and science*, Third, Texts in Applied Mathematics, vol. 58, Springer, New York, 2013. MR3076304

[27] T. Chou and M. R. D'Orsogna, *First passage problems in biology*, First-passage phenomena and their applications, 2014, pp. 306–345. MR3363071

[28] Pierre Collet and Florencia Leonardi, *Loss of memory of hidden Markov models and Lyapunov exponents*, Ann. Appl. Probab. **24** (2014), no. 1, 422–446. MR3161652

[29] W. J. Conover, *Practical nonparametric statistics*, Wiley series in probability and statistics: Applied probability and statistics, Wiley, 1999.

[30] R. T. Cox, *The statistical method of Gibbs in irreversible change*, Rev. Modern Physics **22** (1950), 238–248. MR0038900

[31] A. Crisanti, G. Paladin, and A. Vulpiani, *Products of random matrices in statistical physics*, Springer Series in Solid-State Sciences, vol. 104, Springer-Verlag, Berlin, 1993. With a foreword by Giorgio Parisi. MR1278483

[32] K. Dao Duc, Z. Schuss, and D. Holcman, *Oscillatory survival probability: analytical and numerical study of a non-Poissonian exit time*, Multiscale Model. Simul. **14** (2016), no. 2, 772–798. MR3498507

[33] Persi Diaconis and David Freedman, *Iterated random functions*, SIAM Rev. **41** (1999), no. 1, 45–76. MR1669737

[34] M. Doi, *Diffusion-controlled reaction of polymers*, Chemical Physics **9** (1975), no. 3, 455 –466.

[35] M. Doi and S.F. Edwards, *The theory of polymer dynamics*, International series of monographs on physics, Clarendon Press, 1988.

[36] Tomasz Downarowicz, *Entropy in dynamical systems*, New Mathematical Monographs, vol. 18, Cambridge University Press, Cambridge, 2011. MR2809170

[37] P.J. Flory, *Statistical mechanics of chain molecules*, Hanser Publishers, 1989.

[38] N. Foti, J. Xu, D. Laird, and E. Fox, *Stochastic variational inference for hidden markov models*, Advances in neural information processing systems 27, 2014, pp. 3599–3607.

[39] Ronald Forrest Fox, *Gaussian stochastic processes in physics*, Phys. Rep. **48** (1978), no. 3, 180–283. MR518844

[40] Andrew M. Fraser, *Hidden Markov models and dynamical systems*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. MR2451733

[41] H. Furstenberg and H. Kesten, *Products of random matrices*, Ann. Math. Statist. **31** (1960), 457–469. MR0121828

[42] Harry Furstenberg, *Noncommuting random products*, Trans. Amer. Math. Soc. **108** (1963), 377–428. MR0163345

[43] Crispin Gardiner, *Stochastic methods*, Fourth, Springer Series in Synergetics, Springer-Verlag, Berlin, 2009. A handbook for the natural and social sciences. MR2676235

[44] Cecilia González-Tokman and Anthony Quas, *A semi-invertible operator Oseledets theorem*, Ergodic Theory Dynam. Systems **34** (2014), no. 4, 1230–1272. MR3227155

[45] T. Guèrin, O. Bènichou, and R. Voituriez, *Reactive conformations and non-markovian cyclization kinetics of a rouse polymer*, The Journal of Chemical Physics **138** (2013), no. 9, 094908.

[46] Xiaoying Han and Peter E. Kloeden, *Random ordinary differential equations and their numerical solution*, Probability Theory and Stochastic Modelling, vol. 85, Springer, Singapore, 2017. MR3726872

[47] Darald J. Hartfiel, *Nonhomogeneous matrix products*, World Scientific Publishing Co., Inc., River Edge, NJ, 2002. MR1878339

[48] Boris Hasselblatt and Anatole Katok, *Chapter 1 principal structures*, 2002, pp. 1 –203.

[49] T.L. Hill, *Free energy transduction and biochemical cycle kinetics*, Dover Books on Chemistry, Dover Publications, 2004.

[50] H. Jacobson and W. H. Stockmayer, *Intramolecular reaction in polycondensations. i. the theory of linear systems*, The Journal of Chemical Physics **18** (1950), no. 12, 1600–1606.

[51] Da-Quan Jiang, Min Qian, and Min-Ping Qian, *Mathematical theory of nonequilibrium steady states*, Lecture Notes in Mathematics, vol. 1833, Springer-Verlag, Berlin, 2004. On the frontier of probability and dynamical systems. MR2034774

[52] Luo Jiu-li, C. Van den Broeck, and G. Nicolis, *Stability criteria and fluctuations around nonequilibrium states*, Z. Phys. B **56** (1984), no. 2, 165–170. MR757533

[53] B. H. Juang and L. R. Rabiner, *Hidden markov models for speech recognition*, Technometrics **33** (1991), no. 3, 251–272.

[54] Sophia L. Kalpazidou, *Cycle representations of Markov processes*, Second, Applications of Mathematics (New York), vol. 28, Springer, New York, 2006. Stochastic Modelling and Applied Probability. MR2226353

[55] C. Karlof and D. Wagner, *Hidden markov model cryptanalysis*, Cryptographic hardware and embedded systems - ches 2003: 5th international workshop, cologne, germany, september 8–10, 2003. proceedings, 2003, pp. 17–34.

[56] Tosio Kato, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition. MR1335452

[57] Michael S. Keane, *Ergodic theory and subshifts of finite type*, Ergodic theory, symbolic dynamics, and hyperbolic spaces (Trieste, 1989), 1991, pp. 35–70. MR1130172

[58] A. I. Khinchin, *Mathematical foundations of information theory*, Dover Publications, Inc., New York, N. Y., 1957. Translated by R. A. Silverman and M. D. Friedman. MR0092709

[59] W. Khreich, E. Granger, A. Miri, and R. Sabourin, *A survey of techniques for incremental learning of hmm parameters*, Information Sciences **197** (2012), 105 –130.

[60] Yuri Kifer, *Ergodic theory of random transformations*, Progress in Probability and Statistics, vol. 10, Birkhäuser Boston, Inc., Boston, MA, 1986. MR884892

[61] Yuri Kifer and Pei-Dong Liu, *Chapter 5 - random dynamics*, Handbook of dynamical systems, 2006, pp. 379 –499.

[62] E. L. King and C. Altman, *A schematic method of deriving the rate laws for enzyme-catalyzed reactions*, The Journal of Physical Chemistry **60** (1956), no. 10, 1375–1378, available at `https://doi.org/10.1021/j150544a010`.

[63] I. Klapper and H. Qian, *Remarks on discrete and continuous large-scale models of DNA dynamics*, Biophysical Journal **74** (1998), no. 5, 2504–2514.

[64] Anders Krogh, Michael Brown, I.Saira Mian, Kimmen Sjlander, and David Haussler, *Hidden markov models in computational biology: Applications to protein modeling*, Journal of Molecular Biology **235** (1994), no. 5, 1501 –1531.

[65] Anders Krogh, Björn Larsson, Gunnar von Heijne, and Erik L. L. Sonnhammer, *Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes*, J. MOL. BIOL **305** (2001), 567–580.

[66] G. Lajoie, K. K. Lin, and E. Shea-Brown, *Chaos and reliability in balanced spiking networks with temporal drive*, Phys. Rev. E **87** (2013May), 052901.

[67] Andrzej Lasota and Michael C. Mackey, *Chaos, fractals, and noise*, Second, Applied Mathematical Sciences, vol. 97, Springer-Verlag, New York, 1994. Stochastic aspects of dynamics. MR1244104

[68] François Le Gland and Laurent Mevel, *Basic properties of the projective product with application to products of column-allowable nonnegative matrices*, Math. Control Signals Systems **13** (2000), no. 1, 41–62. MR1742139

[69] ———, *Exponential forgetting and geometric ergodicity in hidden Markov models*, Math. Control Signals Systems **13** (2000), no. 1, 63–93. MR1742140

[70] S. D. Levene, S. M. Giovan, A. Hanke, and M. J. Shoura, *The thermodynamics of DNA loop formation, from j to z*, Biochemical Society Transactions **41** (2013), no. 2, 513–518.

[71] J. Li, F. X.-F. Ye, H. Qian, and S. Huang, *Time Dependent Saddle Node Bifurcation: Breaking Time and the Point of No Return in a Non-Autonomous Model of Critical Transitions*, ArXiv e-prints (November 2016), available at `1611.09542`.

[72] Thomas M. Liggett, *The coupling technique in interacting particle systems*, Doeblin and modern probability (Blaubeuren, 1991), 1993, pp. 73–83. MR1229954

[73] Kevin K. Lin, *Stimulus-response reliability of biological networks*, Nonautonomous dynamical systems in the life sciences, 2013, pp. 135–161. MR3203501

[74] Y.-A Ma, T. Chen, and E. B. Fox, *A complete recipe for stochastic gradient MCMC*, Advances in neural information processing systems 28, 2015, pp. 2899–2907.

[75] Y.-A. Ma, N. J. Foti, and E. B. Fox, *Stochastic Gradient MCMC Methods for Hidden Markov Models*, arXiv.org **stat.ML** (June 2017).

[76] Y.-A. Ma, H. Qian, and F. X.-F. Ye, *Stochastic dynamics: Models for intrinsic and extrinsic noises and their applications (in Chinese)*, Scientia Sinica Mathematica **47** (2017), no. 12, 1693–1702.

[77] Marvin Marcus, Henryk Minc, and Benjamin Moyls, *Some results on non-negative matrices*, J. Res. Nat. Bur. Standards Sect. B **65B** (1961), 205–209. MR0125124

[78] K.S. Matthews, *DNA looping*, Microbiological Reviews **56** (1992), no. 1, 123–136.

[79] Henryk Minc, *Nonnegative matrices*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York, 1988. A Wiley-Interscience Publication. MR932967

[80] David Mumford, *The dawning of the age of stochasticity*, Mathematics: frontiers and perspectives, 2000, pp. 197–218. MR1754778

[81] K.P. Murphy, *Machine learning: A probabilistic perspective*, Adaptive computation and machine learning, MIT Press, 2012.

[82] Julian Newman, *Necessary and sufficient conditions for stable synchronization in random dynamical systems*, Ergodic Theory and Dynamical Systems (2017), 1–19.

[83] F. J. Och and H. Ney, *A comparison of alignment models for statistical machine translation*, Proceedings of the 18th conference on computational linguistics - volume 2, 2000, pp. 1086–1090.

[84] ———, *The alignment template approach to statistical machine translation*, Computational Linguistics **30** (2004), no. 4, 417–449, available at `https://doi.org/10.1162/0891201042544884`.

[85] D. S. Ornstein, *Ergodic theory, randomness, and "chaos"*, Science **243** (1989), no. 4888, 182–187. MR981173

[86] V. I. Oseledec, *A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems*, Trudy Moskov. Mat. Obšč. **19** (1968), 179–210. MR0240280

[87] R. W. Pastor, R. Zwanzig, and A. Szabo, *Diffusion limited first contact of the ends of a polymer: Comparison of theory with simulation*, The Journal of Chemical Physics **105** (1996), no. 9, 3878–3882.

[88] Yuval Peres, *Domains of analytic continuation for the top Lyapunov exponent*, Ann. Inst. H. Poincaré Probab. Statist. **28** (1992), no. 1, 131–148. MR1158741

[89] J. J. Portman, *Non-gaussian dynamics from a simulation of a short peptide: Loop closure rates and effective diffusion coefficients*, The Journal of Chemical Physics **118** (2003), no. 5, 2381–2391.

[90] James Gary Propp and David Bruce Wilson, *Exact sampling with coupled Markov chains and applications to statistical mechanics*, Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995), 1996, pp. 223–252. MR1611693

[91] M. Ptashne, *Gene regulation by proteins acting nearby and at a distance*, Nature **322** (1986/08/21/print), no. 6081, 697–701.

[92] Hong Qian, *A mathematical analysis for the Brownian dynamics of a DNA tether*, J. Math. Biol. **41** (2000), no. 4, 331–340. MR1788985

[93] ———, *Mathematical formalism for isothermal linear irreversibility*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci. **457** (2001), no. 2011, 1645–1655. MR1850931

[94] ———, *A decomposition of irreversible diffusion processes without detailed balance*, J. Math. Phys. **54** (2013), no. 5, 053302, 10. MR3098936

[95] Anthony N. Quas, *On representations of Markov chains by random smooth maps*, Bull. London Math. Soc. **23** (1991), no. 5, 487–492. MR1141021

[96] Lawrence R. Rabiner, *Readings in speech recognition*, 1990, pp. 267–296.

[97] H. Risken, *The Fokker-Planck equation*, Second, Springer Series in Synergetics, vol. 18, Springer-Verlag, Berlin, 1989. Methods of solution and applications. MR987631

[98] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR0274683

[99] P. E. Rouse, *A theory of the linear viscoelastic properties of dilute solutions of coiling polymers*, The Journal of Chemical Physics **21** (1953), 1272–1280.

[100] David Ruelle, *Positivity of entropy production in nonequilibrium statistical mechanics*, J. Statist. Phys. **85** (1996), no. 1-2, 1–23. MR1413234

[101] Michael Scheutzow, *Attractors for ergodic and monotone random dynamical systems*, Seminar on Stochastic Analysis, Random Fields and Applications V, 2008, pp. 331–344. MR2401964

[102] R. Schleif, *DNA looping*, Annual Review of Biochemistry **61** (1992), no. 1, 199–223.

[103] J. Schnakenberg, *Network theory of microscopic and macroscopic behavior of master equation systems*, Rev. Modern Phys. **48** (1976), no. 4, 571–585. MR0443796

[104] E. Seneta, *Non-negative matrices and Markov chains*, Springer Series in Statistics, Springer, New York, 2006. Revised reprint of the second (1981) edition [Springer-Verlag, New York; MR0719544]. MR2209438

[105] D. Shore, J. Langowski, and R. L. Baldwin, *DNA flexibility studied by covalent closure of short fragments into circles*, Proc. Natl. Acad. Sci. U.S.A. **78** (1981), no. 8, 4833–4837.

[106] Erik L. L. Sonnhammer, Gunnar von Heijne, and Anders Krogh, *A hidden markov model for predicting transmembrane helices in protein sequences*, Proceedings of the 6th international conference on intelligent systems for molecular biology, 1998, pp. 175–182.

[107] T. Srokowski and M. Płoszajczak, *Solving the generalized langevin equation with the algebraically correlated noise*, Phys. Rev. E **57** (1998Apr), 3829–3838.

[108] Anatoliy Swishchuk and Shafiqul Islam, *Random dynamical systems in finance*, CRC Press, Boca Raton, FL, 2013. MR3114390

[109] A. Szabo, G. Lamm, and G. H. Weiss, *Localized partial traps in diffusion processes and random walks*, Journal of Statistical Physics **34** (1984), no. 1, 225–238.

[110] A. Szabo, K. Schulten, and Z. Schulten, *First passage time approach to diffusion controlled reactions*, The Journal of Chemical Physics **72** (1980), no. 8, 4350–4357.

[111] O. Bénichou T. Guérin N. Levernier and R. Voituriez, *Mean first-passage times of non-markovian random walkers in confinement*, Nature **534** (2016), no. 7607, 356.

[112] Peter J. Thomas, *A lower bound for the first passage time density of the suprathreshold Ornstein-Uhlenbeck process*, J. Appl. Probab. **48** (2011), no. 2, 420–434. MR2840308

[113] Hermann Thorisson, *Coupling and shift-coupling random sequences*, Doeblin and modern probability (Blaubeuren, 1991), 1993, pp. 85–95. MR1229955

[114] N. M. Toan, G. Morrison, C. Hyeon, and D. Thirumalai, *Kinetics of loop formation in polymer chains*, The Journal of Physical Chemistry B **112** (2008), no. 19, 6094–6106.

[115] N. G. van Kampen, *Stochastic processes in physics and chemistry*, Lecture Notes in Mathematics, vol. 888, North-Holland Publishing Co., Amsterdam-New York, 1981. MR648937

[116] Marcelo Viana, *Lectures on Lyapunov exponents*, Cambridge Studies in Advanced Mathematics, vol. 145, Cambridge University Press, Cambridge, 2014. MR3289050

[117] Jürgen Voigt, *Stochastic operators, information, and entropy*, Comm. Math. Phys. **81** (1981), no. 1, 31–38. MR630330

[118] Peter Walters, *An introduction to ergodic theory*, Graduate Texts in Mathematics, vol. 79, Springer-Verlag, New York-Berlin, 1982. MR648108

[119] Michael J. Ward and Joseph B. Keller, *Strong localized perturbations of eigenvalue problems*, SIAM J. Appl. Math. **53** (1993), no. 3, 770–798. MR1218383

[120] Stephen Wiggins, *Introduction to applied nonlinear dynamical systems and chaos*, Second, Texts in Applied Mathematics, vol. 2, Springer-Verlag, New York, 2003. MR2004534

[121] G. Wilemski and M. Fixman, *Diffusion controlled intrachain reactions of polymers. i theory*, The Journal of Chemical Physics **60** (1974), no. 3, 866–877.

[122] ———, *Diffusion controlled intrachain reactions of polymers. ii results for a pair of terminal reactive groups*, The Journal of Chemical Physics **60** (1974), no. 3, 878–890.

[123] Amie Wilkinson, *What are Lyapunov exponents, and why are they interesting?*, Bull. Amer. Math. Soc. (N.S.) **54** (2017), no. 1, 79–105. MR3584099

[124] F. X.-F. Ye, Y.-A. Ma, and H. Qian, *Estimate exponential memory decay in hidden Markov model and its applications*, ArXiv e-prints (October 2017), available at `1710.06078`.

[125] F. X.-F. Ye and H. Qian, *Stochastic Dynamics II: Finite Random Dynamical Systems, Linear Representation, and Entropy Production*, ArXiv e-prints (April 2018), available at 1804.08174.

[126] Felix X.-F. Ye, Panos Stinis, and Hong Qian, *Dynamic looping of a free-draining polymer*, SIAM J. Appl. Math. **78** (2018), no. 1, 104–123. MR3745006

[127] Felix X.-F. Ye, Yue Wang, and Hong Qian, *Stochastic dynamics: Markov chains and random transformations*, Discrete Contin. Dyn. Syst. Ser. B **21** (2016), no. 7, 2337–2361. MR3543636

Appendix A

EIGENVALUES AND SINGULAR VALUES OF DETERMINISTIC TRANSITION MATRIX

If a deterministic transition matrix M is invertible, and I denote its transpose as M^T, then MM^T is the identity matrix. In fact, $M^T = M^{-1}$ is the inverse of the corresponding one-to-one transformation. Therefore, invertible deterministic transition matrix has all eigenvalues on the unit circle and singular value being 1. If M is a non-invertible deterministic transition matrix, then its eigenvalues are either on the unit circle or zero. It necessarily has at least one column of 0's. Note, however, that both

$$M_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{A.1}$$

have one column of 0's. However, eliminating the transient state 2 yields

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

This indicates that state 3 in M_2 is also a transient state; in fact state 4 is also transient. The deterministic dynamics represented by M_2 is $2 \to 3 \to 4 \to 1\circlearrowleft$. It has a global "attractor" with a single state 1; thus its zero eigenvalue has a multiplicity of 3. In contrast, the deterministic transformation represented by M_1 is $2 \to 3 \to 4 \to 1 \to 3$. Its global "attractor" is a 3-state cycle. This is the dynamical reason why the zero eigenvalue of M_1 has only multiplicity of 1. We can rearrange rows and columns of M_1 to make it block upper triangular, then the above statement is clear. The singular values of a non-invertible

deterministic transition matrix $M_i, i = 1, 2$, provide different kind of information on the dynamics: The square of the singular values are the number of pre-image of a state that are also exactly the eigenvalues of the diagonal matrix, $M_i M_i^T$. For example, both M_1 and M_2 have a same set of singular values $(\sqrt{2}, 1, 1, 0)$.

Appendix B

METRIC ENTROPY AND TOPOLOGICAL ENTROPY

In this section, I review some important concepts in connection with metric entropy. This material can be found in standard textbooks [36, 57, 60, 118]; it is presented for the convenience of the readers.

Let T be a measure preserving transformation of a probability space $(\mathscr{S}, \mathcal{F}, \mu)$. A good motivation for the notion of *metric entropy* h_T, associated with T, is in terms of measurements. A *measurement* is a finite partition of the space $\mathscr{S}$, $P = \{P_1, \cdots, P_k\}$. All these measurable sets P_i are disjoint and their union is $\mathscr{S}$. Now consider a finite portion of an orbit, with length t, generated by T, starting from an initial condition $s \in \mathscr{S}$,

$$s, T(s), T^2(s), \cdots, T^{t-1}(s).$$

Each of the points $T^i(s)$ belongs to exactly one of the sets of the partition P,

$$s \in P_{k_0}, T(s) \in P_{k_1}, \cdots, T^{t-1}(s) \in P_{k_{t-1}}.$$

We say $k = \{k_0, k_1, \cdots, k_{t-1}\}$ the *address* of s with respect with the partition P. It is possible that another orbit will have the same address as k with respect with the partition P so I can collect all initial points such their orbits have the same address k.

$$P^t(k) = \big\{ s \in \mathscr{S} : \text{address of } s = k \big\} \tag{B.1}$$

In fact, $P^t = \{P^t(k) : k \text{ is any address of length } t\}$ is also a partition of $\mathscr{S}$. Moreover, it can be shown that P^t is the join of the partitions $P, T^{-1}P, \cdots, T^{-(t-1)}P$. One likes to quantify the amount of information in an address of an orbit with length t. This is given by the Gibbs-Shannon entropy

$$H(P^t) = -\sum_{j=1}^{m} \mu(P_j^t) \log \mu(P_j^t) \tag{B.2}$$

where m is the number of partitions of P^t.

The amount of information per unit is then $\frac{1}{t}H(P^t)$. Therefore, the information per unit in a measurement P is defined as $H(P,T) = \lim_{t\to\infty}\frac{1}{t}H(P^t)$ and the metric entropy is the supremum of this value over all possible finite measurements, $h_T = \sup_P H(P,T)$. A discovery due to Sinai helps the computation: The supremum is attained for all partitions are *generators*. A generator is a partition P such that two different points of $\mathscr{S}$ have a different address.

In the case of a Bernoulli sequence of only finite states, e.g., "symbols", with probability $(p_1, \cdots, p_m)$, the generator is simply the natural partition $P = (1, 2, \ldots, m)$ since an address of s is $(s_0, s_1, s_2, \ldots)$ which is s itself. For each address of length n, $k = (k_0, k_1, \ldots, k_{t-1})$, I will have $\mu(P^t(k)) = p_{k_0}p_{k_1}\cdots p_{k_{t-1}}$ and hence

$$H(P^t) = -\sum_k \mu(P^t(k))\log\mu(P^t(k)) = -t\left(\sum_{i=1}^m p_i \log p_i\right). \tag{B.3}$$

Therefore the metric entropy is $h(T) = -\sum_{i=1}^m p_i \log p_i$. Bernoulli trial is a sequence of i.i.d. random variables; one can similarly derive for the metric entropy of a Markov chain.

The metric entropy h of a Markov chain with transition probability M_{ij}, $i, j \in \mathscr{S}$, is the asymptotic exponent of the vanishing probability for a single stochastic trajectory $i_0 i_1 i_2 \cdots i_t$ with increasing t, $e^{-h_t t}$ [58]:

$$M_{i_0 i_1} M_{i_1 i_2} \cdots M_{i_k, i_{k+1}} \cdots M_{i_{t-1}, i_t} = \exp\left(\sum_{k=1}^t \log M_{i_{k-1}, i_k}\right) = e^{-h_t t}, \tag{B.4}$$

and

$$h = \lim_{t\to\infty} -\frac{1}{t}\sum_{k=1}^t \log M_{i_{k-1}, i_k} = -\sum_{i,j\in\mathscr{S}} \pi_i M_{ij} \log M_{ij}. \tag{B.5}$$

Here I stipulate that $0 \log 0 = 0$.

The last step to replace time average by expectation from invariant distribution, is based on the ergodicity of the MC: As $t \to \infty$, the frequency of the state i in the sequence goes to π_i and the frequency of the pair ij becomes $\pi_i M_{ij}$.

Note that for a deterministic transformation with probability 1 for transition $i_k \to i_{k+1}$, the probability in (B.4) is a constant. Therefore, its metric entropy $h = 0$.

The topological entropy of a Markov chain η characterizes what is possible and what is not; it is independent of actual values of transition probabilities. Two Markov chains with transition probabilities M_{ij} and M'_{ij}, $i, j \in \mathscr{S}$, has a same topological entropy when $M_{ij} = 0$ if and only if $M'_{ij} = 0$. They both induce a same SFT, and the topological entropy η simply counts the number of possible trajectories generated by the Markov process. Consider an $n \times n$ irreducible binary matrix A, $n = \|\mathscr{S}\|$, satisfying $A_{ij} = 1$ when $M_{ij} > 0$, $A_{ij} = 0$ when $M_{ij} = 0$. We call it the "topological skeletal matrix" of this MC. The number of possible trajectories increases asymptotically with ℓ:

$$\Omega_B(t) \triangleq (1, 0, \cdots, 0) A^t \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \to C_1 e^{(\log \lambda_A) t}, \tag{B.6}$$

in which λ_A is the largest positive eigenvalue of the matrix A. The topological entropy η, following Boltzmann's notion of entropy

$$\eta = \lim_{t \to \infty} \frac{\log \Omega_B(t)}{t} = \log \lambda_A. \tag{B.7}$$

It is easy to show that $\eta \geq h$. For a Markov chain with all $M_{ij} = \frac{1}{n}$, the equality is attained $\eta = h = \log n$.

The definition of topological entropy according to (B.6) also clearly indicates that η is the same for a stochastic process with "increasing number of distinct trajectories" and an endomorphism with "increasing number of distinct pre-image".

For an MC, its topological entropy η is intimately related to how many connected neighbours a state has, or "dimensions" in a lattice system. For an MC with exact ν non-zero elements in each and every row of its transition matrix, $\nu < n$, then the largest eigenvalue of its "topological skeletal matrix" A is $\lambda_A = \nu$. The corresponding right eigenvector is clearly $(1, 1, \cdots, 1)$. This is the "topology" of the MC, which is also the metric entropy

of the Markov chain with $M_{ij} = A_{ij}/\nu$. Its corresponding invariant distribution, e.g. the left-eigenvector is $\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$:

$$\sum_{i=1}^{n} \frac{1}{n} A_{ij} = \frac{1}{n}. \tag{B.8}$$

Then I have the following result which could be very useful for sparse Markov networks:

Proposition B.1. *The metric entropy of a Markov chain*

$$h \leq \sum_{i \in \mathscr{S}} \pi_i \log \nu_i,$$

where ν_i is cardinality of $\{j : M_{ij} > 0\}$, and $\{\pi_i\}$ is the stationary distribution. In particular, if the MC has the maximal number of neighboring states being ν, then $h \leq \log \nu$.

Proof. Let us now consider all the possible Markov chains with state state $\mathscr{S}$ which have the same "topological skeletal matrix" A: $M_{ij} = M_{ij} A_{ij}$. Denote $\nu_i = \sum_j A_{ij}$. Then for each i, the transition probability M_{ij} has lower entropy than the uniform A_{ij}/ν_i whose entropy is simply $\log \nu_i$. The metric entropy for the Markov chain then is:

$$\begin{aligned} h &= -\sum_{i,j \in \mathscr{S}} \pi_i M_{ij} \log M_{ij} \\ &\leq \sum_{i \in \mathscr{S}} \pi_i \log \nu_i. \end{aligned}$$

□

Appendix C

MATRIX-TREE THEOREM

The matrix-tree theorem is a refined formula that gives the complete symbolic series for directed rooted trees with specified roots and more generally for forests with specified roots [19]. We introduce variable M_{ij} for all $i, j \in \mathscr{S}$ and define the monomial x_T for the directed rooted trees T to be the product of the variables M_{ij} for all directed edges $i \to j$ in T. For the example above in Fig. 3.1, they are

$$x_{T_1} = M_{21}M_{31}, x_{T_2} = M_{12}M_{31}, x_{T_3} = M_{21}M_{13} \tag{C.1}$$

Note the weight of the trees I defined before is exactly evaluated in the monomial for given M_{ij}. Moreover, the directed rooted tree are determined by its monomial x_T. Similarly, it can be extended to rooted forests.

Given a subset $I \subset \mathscr{S}$, I define $F_{n,I}$ to be the sum of the monomials for all forests G whose set of roots is I, which is called the generating function for G.

$$F_{n,I} = \sum_{G:\mathrm{roots}(G)=I} x_G. \tag{C.2}$$

The matrix-tree theorem is stated as follows,

Theorem C.1. *In MC, the generating function $F_{n,I}(M)$ for all forests rooted at I, with edges directed towards the roots, is given by the determinant*

$$F_{n,I}(M) = \det D(\{I\}^c) \tag{C.3}$$

where $D = I - M$ and $D(\{I\}^c)$ is the submatrix of the matrix D by deleting the rows and columns with indices $i \in I$.

In the previous example, the matrix D is

$$\begin{bmatrix} M_{12}+M_{13} & -M_{12} & -M_{13} \\ -M_{21} & M_{21} & 0 \\ -M_{31} & 0 & M_{31} \end{bmatrix}.$$

If $I = \{2\}$, then $\det D(\{2\})$ is $(M_{12}+M_{13})M_{31} - M_{31}M_{13} = M_{12}M_{31}$. It agrees with the expression for x_{T_2}.

The generating function $F_{n,\{i\}}(M)$ evaluated at given transition matrix M is equal to $e(\mathcal{T}_i)$ the weight of the set of directed rooted trees whose root is state i. Then the Eq. (3.16) can be rewritten as

$$\boldsymbol{\pi}_i = \frac{D(\{i\}^c)}{\sum_i D(\{i\}^c)}. \tag{C.4}$$

Appendix D

BIRKHOFF CONTRACTION COEFFICIENT

We will introduce Hilbert metric and Birkhoff contraction coefficient on positive matrices, especially on positive stochastic matrices [47, 104].

Let x and y be positive vectors in $\mathbb{R}^n$, the Hilbert metric is defined as $d(x,y) = \ln \frac{\max_i x_i/y_i}{\min_j x_j/y_j}$. But Hilbert metric is not a metric in $\mathbb{R}^n$ since one could check when $x = cy$ for some constant c, $d(x,y) = 0$. Actually, for each positive probability vector in the interior of the simplex S^{K-1}, d determines a metric on them.

The advantage of Hilbert metric for the positive stochastic matrix M is one can show for two different positive probability row vector, x and y, the distance between x and y under M monotonically decreases, $d(xM, yM) < d(x,y)$. This is not guaranteed for other metrics due to the possible non-normal behavior of the matrix. The Birkhoff contraction coefficient $\tau(M)$ is defined as the supreme of the contraction ratio under the matrix M,

$$\tau(M) = \sup \frac{d(xM, yM)}{d(x,y)} \tag{D.1}$$

This coefficient indicates how much x and y are drown together at least after multiplying by M. Actually, there is an explicit formula for computing $\tau(M)$ in terms of the entries of M. Define $\phi(M)$ as

$$\phi(M) = \min_{p,q,r,s} \frac{M_{pq}M_{rs}}{M_{rq}M_{ps}} \tag{D.2}$$

The term $\frac{M_{pq}M_{rs}}{M_{rq}M_{ps}}$ is cross ratios of all 2×2 sub matrices of M and $\phi(M)$ is the minimum amount of them. If there is a row with both zero and positive elements, $\phi(M) = 0$. The formula for $\tau(M)$ is

$$\tau(M) = \frac{1 - \sqrt{\phi(M)}}{1 + \sqrt{\phi(M)}} \tag{D.3}$$

As expected, for positive stochastic matrix M, $\tau(M) < 1$.