TRANSITION PATH THEORY AND TRANSITION STATE

By

Jun Du

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Applied Mathematics—Doctor of Philosophy

2019

ABSTRACT

TRANSITION PATH THEORY AND TRANSITION STATE

By

Jun Du

This thesis will mainly discuss the transition path theory and its extension to transition state. The framework of transition path theory (TPT) is developed in the context of continuous-time Markov chains on discrete state-spaces. Under assumption of ergodicity, Transition path theory will first choose any two subsets (mostly meta stable states) in the finite state-space based on the equilibrium distribution of the transition probability, and then it analyzes the statistical properties of those associated reactive trajectories, for instance, those trajectories by which the random walker transits from one subset to another. Transition path theory gives properties of these trajectories, such as their probability distribution, their probability current and flux, and their rate of occurrence and finally the dominant reaction pathways. In this thesis we will first introduce the framework of transition path theory for Markov chains, and then briefly discuss its relation to electric resistor network theory and Laplacian eigenmaps, and also diffusion maps is discussed as well.

Based on Transition Path Theory (TPT) for Markov jump processes[17, 84], this thesis develop a general approach for identifying and calculating Transition States (TS) of stochastic chemical reacting networks. The thesis first extend the concept of probability current, originally defined on edges connecting different nodes in the configuration space [84], to each sub-network. To locate sub-networks with maximal probability current on the separatrix between reactive and non-reactive events, which will give the Transition States of the reaction, constraint optimization is conducted. The thesis further introduce an alternative scheme to compute the transition pathways by topological sorting, which is shown to be highly efficient

through analysis. Finally, the theory and the algorithms are illustrated in several examples.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Di Liu for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. His advice and support also helped my daily life and career in U.S., I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Guowei Wei, Prof. Dapeng Zhan, and Prof. Peter W. Bates, for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

I would like to thank my fellow doctoral students for their feedback, cooperation and of course friendship.

Last but not the least, I would like to thank my family: my parents and to my sister for supporting me spiritually throughout writing this thesis and my my life in general.

TABLE OF CONTENTS

LIST OF FIGURES . . . . . **vii**

Chapter 1 Introduction **1**
1.1 Background and Motivation 1
1.2 Choice of Modeling 3
1.3 Rare Events in Molecular Dynamics 6
1.4 Transition State Theory 8
1.5 Transition Path Sampling 10
1.6 Transition Path Theory 11

Chapter 2 Stochastic Kinetic Reactions **14**
2.1 Derivation of the chemical master equation 15
2.2 Numerical Methods for the CME 19
2.2.1 Direct Methods for the CME 19
2.2.2 Direct Stochastic Simulation Algorithm 22
2.2.3 Tau-leaping Method and Other Monte Carlo Methods 24
2.2.4 Conclusion 26

Chapter 3 Transition Path Theory for Jump Markov Process **28**
3.1 Probability theory and stochastic process 30
3.2 The Markov Property 32
3.3 Continuous-time Markov process 35
3.4 Main TPT 44
3.4.1 Notations ans assumptions 44
3.4.2 Reactive trajectories 47
3.4.3 Probability distribution of reactive trajectories 50
3.4.4 Probability current of reactive trajectories 55
3.4.5 Transition rate and effective current 58
3.4.6 Relations with electrical resistor networks 61
3.4.7 Dynamical bottlenecks and reaction pathways 62
3.4.8 Relation with Laplacian eigenmaps and diffusion maps 68
3.5 Algorithmic aspects 70
3.5.1 Computation of dynamical bottlenecks and representative dominant reaction pathways 70
3.5.2 Topological Sorting Algorithm for Transition Pathways 72
3.5.3 Representative Transition Pathway Finding 75
3.5.4 Correctness of the algorithm 76
3.5.5 Illustrative example 77

Chapter 4 Extension of TPT to transition state **80**
4.1 Probability current of sub-networks 80
4.2 Time Reversible Case 81
4.3 Non Time Reversible Case 82

Chapter 5 Illustrative example **83**
5.1 Diffusions Processes in Potentials 83
5.2 Toggle Switch Models in 2D and 3D 86
5.3 The Lennard-Jones 13 Cluster 91

Chapter 6 Future Work: Clustering Methods for Directed Graph **96**

Chapter 7 Conclusion **98**

BIBLIOGRAPHY **99**

LIST OF FIGURES

Figure 1: Schematic 48

Figure 2: Schematic representation of the decomposition of $\mathcal{W}_{\mathcal{D}}$. A reaction pathway w (shown in thick black) can be decomposed into two simple pathways $w_{\mathcal{L}}$ and $w_{\mathcal{R}}$. 65

Figure 3: topological sort example 74

Figure 4: network flow with s as source and t as sink, values along the edges are the capacities, corresponding to the effective probability current. 77

Figure 5: Upper: Contour plot of the equilibrium distribution of diffusion in double-well potential. Below: Contour plot of the weighted capacity of each state. The red line in both plot corresponds to the representative pathways. . . 84

Figure 6: Upper: Contour plot of the stationary distribution $\pi_{(x,y)}$ of diffusion in three-hole potential. Results are for $\beta = 1.67$ and a 60×60 mesh discretization. Below: contourf plot of the weighted current at each discretized state. The red line in both plot corresponds to representative pathways. 86

Figure 7: Upper: Contour plot of the Gibbs energy, -log$\pi(x, y)$, of the 2D Toggle switch model on the state-spaces $S = (\mathbb{Z} \times \mathbb{Z}) \cap ([0, 200] \times [0, 60])$. The dark red region in the right upper part of the panel indicates the subset of states with almost vanishing stationary distribution. Results for $a_1 = 156, a_2 = 30, n = 3, m = 1, K_1 = K_2 = 1$, and $\tau_1 = \tau_2 = 1$. Below: Contour plot of the weighted current. Both plots includes the transition path. 87

Figure 8: Contour plot of the stationary distribution $\pi_{(x,y,z)}$ for the 3D Toggle switch model with the state space $S = [0, 63]^3$. States with probability less than machine precision are marked as white colors. 89

Figure 9: Slice plot of the weighted current, the red line is the representative pathways. 92

Figure 10: Disconnectivity graph for the LJ_13 cluster including all the 1510 local minima by David Wales's website database. The global minimum is a Mackay icosahedron, depicted using Xmakemol, while the next-lowest minima correspond to the three distinct capping sites when one atom is removed from the icosahedral shell and placed on the surface. 93

Figure 11: Weighted current of each minima, the red line is the transition path . . . 94

Chapter 1

Introduction

1.1 Background and Motivation

This will give background of chemical reaction networks.

Recent decades have seen a tremendous level of activity in the field of molecular biology, where the use of different technologies enabled researchers to continuously expand the boundaries of knowledge on biological phenomena occurring at the cellular level. Individual cellular components have collected a lot of data, and there's a deep understanding of the interactions, so the emergence of systems biology becomes an interdisciplinary field, and we can treat such biological processes as dynamical networks. As a result, we can build a lot of mathematical models and use simulations to investigate the reaction systems. Comparing to the traditional laboratory methods, this in silicon modeling method can save significant time and cost to model the molecular biology, because it can use quantitative methods under some hypotheses and assumptions about the mechanisms of biological networks. As a result, it's playing more and more significant role in molecular biology.

However, although (the seminal results from [31, 16, 61]) have already assembled an extensive list of the building blocks of living organisms and well studied the internal mechanisms of these living organisms, it is still far from complete to understand how these pieces work together to influence phenotypic heterogeneity. The principal target of systems biology is to

empower the controlling of the sophisticated behavior in living organisms through the robustly propagated changes either upstream or downstream of the bodies adding the location (see [80] for a dramatic argument on this subject). Nevertheless, to achieve such ambitious goals, we need a new study on developing or modifying some mathematical or computational techniques to take care of the complexity of the biological organization.

At the cellular level of biological reaction networks, the stochasticity and discreteness are playing an essential role, and this has increased the awareness as a vital topic for computational systems biology [3]. Many experimental results [48, 14, 15] have supported this argument, as a result, even though the sophistication of dynamics involved makes the study of stochastic fluctuations a very challenging task, many recent reviews [75, 46, 68] have highlighted this as a practically managed work.

As a result, much interest has been recently attracted to topics related to the building of models that makes the necessarily high resolutions to uncover important biological details such as the outcomes of molecular noise. Any scientific research having the goal of investigating a "real" biological reaction process, comprises the issue of how accurate the model built to describe the reality should be, how to more accurately program the problem, how to formulate a model using the existing knowledge so that the unconformity between the model and the real process is not too large and also the model is simple enough to maintain computationally tractable. Once we build such a model, the next issue is whether we can use this mathematical description to investigate processes of legitimate interest, or for functional purposes, we will need an "approximate" formulation. Typically, the last alternative is the only practical choice if the purpose is to move aside from investigating the intercommunications of particular individual molecules and represent preferably the more complicated function of biological systems that comprise many components regulated in biological networks.

In the broader discernment, this thesis concerns with the development, analysis, and application of such "rough" explanations of complicated biological processes. Nevertheless, in order to keep things into the aspect, it is essential to remark that the discrepancy between the "precise" and "rough" models usually is far less than that between the "precise" model and the real process, so picking the proper modeling standard is of up-most concern.

1.2 Choice of Modeling

Ideally, explaining a complicated dynamic system would be achieved by using some deterministic model, which means that given some previous state we can adequately describe the future by applying some fundamental evolution laws that follow the proper dynamics and keep track of the states and velocity of all the particles included, as well as their intercommunications. Such molecular dynamics procedures can be exact, but the sheer complexity of the intercommunications means they usually are too costly from a computational point of view, mainly if the model includes more than single molecules of each type, and the dynamics are to be studied over a longer time interval. The model on such scale is called microscopic and applies Brownian dynamics for the movement of the particles and the Smoluchowski model for their communications. Not confronting the difficulties, improvements in computational methods like the Green's Function Reaction Dynamics (GFRD) algorithm introduced in [76], have allowed the use of such models to some biological reactions, and their application will undoubtedly develop in the future, mainly because of the improvement of hybrid approaches [39]. Nevertheless, biological complexity and the challenge in forming applicable laws that take all consequences into account, currently restrain the application of microscopic models. Countless of the earlier mathematical modeling of cellular processes applied instead of a

macroscopic method, that is, a deterministic model that assumes massive population levels drop the spatial dimension and is used to analyze the usual reaction. Usually, such simplifications can only be done under specific hypotheses, i.e., in addition to having high molecular copy-numbers that discourage the influences of molecular noise, the system is also well-stirred, indicating the molecules are uniformly diffused within a container of fixed size, and the temperature is also fixed. The time evolution of such a system can then be formed through a system of ordinary differential equations (ODEs) representing the concentrations of the molecular states included, known as the reaction rate equations (RRE).

Nevertheless, because biological reactions at the cellular level, such as gene regulatory networks, usually show low copy numbers of engaging units, this means that some of the hypotheses made in this traditional deterministic setting are no longer adequate. In order to achieve an exact model for such systems, which is still reasonably uncomplicated to reproduce notwithstanding the more excellent resolution, randomness should be added into the mathematical model, while maintaining the well-stirred characterization. Hence, a mesoscopic model which lies between the exact but prohibitively expensive microscopic scale and the rough but from a computational point of view quickly convenient macroscopic scale, has risen as the most favorite alternative for modeling stochastic outcomes, as it considers both the stochastic nature of biological reactions and the discreteness of the population amounts. The model is based on the theory that the process driving the progression of the system is memoryless, which means, depends only on the current state of the system and not the whole system history, with the mathematical formulation given by a continuous-time discrete space Markov jump process [35].

In the mesoscopic formulation, the impacts that are either too complicated or too costly to reproduce are checked in terms of random variables. Next, the future can no longer be

unambiguously resolved from the past and is described just in a probabilistic insight. This is fitting for most applications because the issues being pretended are of a quantitative nature, i.e., the time-evolution of the population amounts of the various interacting cellular elements. From the computational point of view, the fulfillment of the Markov jump process can be produced through the Stochastic Simulation Algorithm (SSA), also known as the Gillespie algorithm (see [35]). As a matter of course, each simulation of a given model will produce a different result. However, the probability distribution of the results for a particular time is defined by the underlying mathematical formulation and can be calculated as the solution of the Chemical Master equation (CME). Therefore, the CME gives a "precise" representation of the stochastic model. Nonetheless, the full probability distribution for the state of a biological system over time can only be computed in uncomplicated situations, which restricts the direct use of CME. Numerical approximations of the solution are also not easy to obtain as the CME is affected by the curse of dimensionality: the number of degrees of freedom needed for an accurate approximation grows exponentially with an increase in the number of components of the biological system.

As the degree of freedom present in most problems that deserve study is large, the usual computational method in mesoscopic modeling has been based on Monte Carlo simulations employing the SSA algorithm, either the original modification from [35] or the numerous modifications that have been introduced since (see, e.g., [33, 37, 7, 8]). In theory, the affiliated Monte Carlo error can be made arbitrarily small by improving the number of simulations, but getting a precise estimate of the probability distribution employing stochastic simulations usually is not achievable because any shift in the state of the system demands an update of the state vector. For systems with various time-scales, this can direct to high computational costs.

An alternative is to try to design algorithms to solve CME directly, notwithstanding the difficulties posed by the curse of dimensionality. As both options are computationally costly, the doubt of the advantage of stochastic modeling arises, and whether the acquired computational cost is justified. A comparison between the outcomes acquired using the deterministic and stochastic methods can, therefore, shed light on why, including molecular noise in the model, is crucial, especially in the case of gene regulatory networks.

1.3 Rare Events in Molecular Dynamics

In the traditional depiction of molecular processes, the dynamics of the molecules' microscopic configurations (position and momenta) are mathematically illustrated in terms of the ODE, resulting from formulations of Lagrange and Hamilton. Inside these models, the physical intercommunications of atoms are encoded in the intercommunication potential, which is formed of sums of contributions of various physical origin as the bond arrangement of the molecule and electrostatic intercommunications. However, most biological reactions can just be explained within a thermodynamical context; instead of an individual molecular system as a solution of the standard equations, we are interested in statistical ensembles, since only such ensembles can be the object of experimental research. Throughout this thesis, we will first concentrate on that ensemble view.

Functions of bio-molecules depend on their dynamical characteristics, and mainly on their capability to withstand transitions between long-living states, which is called as conformations. Molecular conformation is any spatial arrangement of the atoms in a molecule that can be interconverted by rotations about formally single bonds. A conformation of a molecule is interpreted as a mean geometric structure of the molecule, which is preserved on a large

time scale compared to the fastest molecular motions where the system is well rotated, oscillated, or fluctuated. From the dynamical point of view, a conformation typically endures for a very long time (comparing to the fastest molecular motions) such that the affiliated subset of microscopic configurations is substantially invariant or metastable [55] concerning the dynamics. Henceforth transitions between separate conformations of a molecule are rare events compared to the fluctuations within each conformation.

A prevalent model to characterize molecular systems, including thermal noise, is the stochastic Langevin dynamics or Smoluchowski dynamics. In physics, the Langevin equation (named after Paul Langevin) is a stochastic differential equation describing the time evolution of a subset of the degrees of freedom. These degrees of freedom typically are collective (macroscopic) variables changing only slowly in comparison to the other (microscopic) variables of the system. The fast (microscopic) variables are responsible for the stochastic nature of the Langevin equation. A Langevin system can be viewed as a mechanical system with additional noise and friction where the noise can be thought of modeling the impact of a heat bath surrounding the molecule, and the friction is chosen such as to counterbalance the energy fluctuations due to the noise [38]. The Smoluchowski dynamics [73] is a Brownian motion that follows from the Langevin dynamics in the high friction limit and acts just on the position space.

Mathematically speaking, the Langevin and Smoluchowski dynamics are time-continuous Markov diffusion processes on a continuous state space. Under weak conditions, both indicate a unique stationary (equilibrium) distribution in configuration space which corresponds to the stationary (canonical) ensemble in experiments under fixed volume and temperature, respectively.

As discussed above, the issue of recognizing conformations amounts to the classification

of metastable sets in configuration space. The characterization of metastability inside the canonical ensemble, therefore, demands the mathematical explanation of the propagation of sub-ensembles. This is achieved by the transfer operator approach; if we define a transition probability from a sub-ensemble C into another sub-ensemble B in time τ, denoted by $p(\tau, C, B)$ then C will be called metastable on a time slice τ if the portion of the systems in that sub-ensemble which stays in C after time τ is nearly one, i.e. $p(\tau, C, C) \approx 1$ [44]. Lastly, the algorithmic approach to decompose the state space into metastable states is based on the spectral characteristics of the transfer operator [79].

1.4 Transition State Theory

Since the 1930s transition state theory (TST) and the evolution thereof based on the reactive flux formalism have produced the main theoretical framework for the information of rare events [28, 81, 82, 40, 6]. Basically, TST was originated in the circumstances of investigating the rate of chemical reactions $R \to P$, where R indicates the reactant and P the product. The concept behind TST is to approximate this reaction rate k by the mean crossing frequency k^{TST} of transitions from R to P by a transition state, which we call the dynamical bottleneck for the reaction. Ordinarily, the transition state can be any dividing surface separating the reactant state R from the product state P . Then the transition state rate, k^{TST} , is proportional to the total flux of reactive trajectories, meaning, trajectories from the reactant to the product side of the separating surface, and can be represented in terms of thermodynamical quantities.

The transition state rate is always an upper boundary of the actual reaction rate because reactive trajectories can recross the transition state back and forth many times during one

reaction. Hence, the actual rate is given by

$$k = \kappa k^{TST}, \tag{1.4.0.1}$$

where κ, the transition coefficient, is a correcting factor estimating for these recrossings. Due to this overestimation, many strategies have been suggested to enhance the transition state rate. For instance, the most initial one is called variational transition state theory [43] and amounts to determine the dividing surface which minimizes the transition state rate constant (see also [17, 18]).

Implementing the computation in practice, nevertheless, may show very challenging, and this difficulty is associated with an insufficiency of the theory. Transition state theory is based on partitioning the system into two, leaving the reactant state on one side of a dividing surface and the product state on the other, and the theory only reveals how this surface is traversed during the reaction. As a result, transition state theory gives minimal information about the mechanism of the transition, which has terrible consequences e.g., if this mechanism is entirely undiscovered a priori. In such a case, it is challenging to choose a suitable dividing surface, and a lousy choice will lead to a very poor estimate of the rate by transition state theory (too many false crossings of the surface that do not correspond to actual reactive events). The TST approximation is then tough to correct. The situation is even worse when the reaction is of diffusive type since, in this case, all surfaces are crossed many times during an individual reactive event, and there is clearly no good transition state dividing surface that exists.

1.5 Transition Path Sampling

How to go from transition state theory and explain rare events whose mechanism is hidden a priori is an ongoing area of research, and various new methods have been improved to undertake these situations. Most well-known among these methods are the transition path sampling (TPS) method of Bolhuis, Chandler, Dellago, and Geissler [57, 53] and the action method of Elber [65, 66] which allow sampling directly the ensemble of reactive trajectories, for instance, the trajectories by which the reaction occurs.

The fundamental idea behind transition path sampling is a generalization of standard Monte Carlo Markov Chain (MCMC) [56, 13] methods on the trajectory space of the considered dynamics. Ordinarily, an MCMC procedure produces a biased random walk on the configuration space such as the number of visits of a configuration x is proportional to its probability $p(x)$. In transition path sampling, a configuration $X(T) = (x_0, x_{\Delta t}, \cdots, x_T)$ is a series of states describing a time discretization of a right dynamical trajectory of settled length $\mathcal{T}$ rather than single states of the dynamics itself. The statistical weight $p(X(\mathcal{T}))$ depends on the initial conditions and on the underlying dynamics. Since we are only focusing on reactive trajectories connecting A and B, TPS finally produces a random walk on the transition path ensemble concerning the reactive path probability

$$p_{AB}(X(\mathcal{T})) = Z_{AB}^{-1}(\mathcal{T})\mathbf{1}_A(x_0)p(X(\mathcal{T}))\mathbf{1}_B(x_{\mathcal{T}}), \tag{1.5.0.1}$$

where Z_{AB} normalizes the distribution of the transition path ensemble and the characteristic $\mathbf{1}_A(x)$ is equal to one if $x \in A$ and 0 otherwise ($\mathbf{1}_B(x)$ is defined analogously).

We want to highlight that reactive trajectories in the transition path ensemble are accurate dynamical trajectories, free of any bias by non-physical forces, limitations, or hypotheses on

the reaction mechanism. The mechanism of the reaction and probably its rate can then be reached a posteriori by examining the ensemble of reactive trajectories. Nevertheless, these procedures are far from trivial. Transition path sampling or the action method does not explain how this study must be done, and a simple investigation of the reactive trajectories may not be enough to explain the mechanism of the reaction. This may sound contradictory at first. However, the issue is that the reactive trajectories may be very complex objects from which it is tough to extricate the quantities of real interest such as the probability density that a reactive trajectory is at a given location in state-space, the probability current of these reactive trajectories, or their rate of occurrence. In a way, this challenge is the same that we would meet, having produced a long trajectory from the law of classical mechanics but neglecting all about statistical mechanics: how to explain this trajectory would then be unclear. Likewise, the statistical framework to describe the reactive trajectories is not given by the trajectories themselves, and further study beyond transition path sampling or the action method is required (for an effort in this direction, see [34]).

1.6 Transition Path Theory

Lately, an analytical framework to illustrate the statistical properties of the reactive trajectories in the circumstances of Markov diffusion processes has been proposed [24, 19]. This framework, called transition path theory (TPT), goes exceeding standard equilibrium statistical mechanics and estimates for the nontrivial bias that the very definition of the reactive trajectories implies – they have to be involved in a reaction.

TPT enables us to learn the statistical properties of the ensemble of all reactive trajectories (not only reactive trajectories concerning a constant length as in TPS) by giving definite

answers to the following questions:

- What is the probability of encountering a reactive trajectory in a given state, meaning, the probability density function of reactive trajectories?
- What is the amount of reactive trajectories going within a given state, meaning, the probability current of reactive trajectories?
- What is the average frequency of transitions between two sets, say A and B, meaning, the rate of reaction?
- What are the mechanisms of transitions, meaning, the transition tubes, or transition pathways?

The essential component in the main objects given by TPT is the committor function $q_{AB}(x) \equiv q(x)$, which is the probability of going rather to the set B than to the set A conditioned on the process has begun in the state x. The committor function $q(x)$ can be seen as an ideal reaction coordinate, because under proper conditions on the dynamics the levels sets of the committor function foliate the state space in sets of equal probability to rather end up in B than A, i.e., it explains the process of reaction from A to B in terms of probabilities.

Lately, an analytical framework to illustrate the statistical properties of the reactive trajectories in the circumstances of Markov diffusion processes has been proposed [24, 19]. This framework, called transition path theory (TPT), goes exceeding standard equilibrium statistical mechanics and estimates for the nontrivial bias that the very definition of the reactive trajectories implies – they have to be involved in a reaction.

TPT enables us to learn the statistical properties of the ensemble of all reactive trajectories

(not only reactive trajectories concerning a constant length as in TPS) by giving definite answers to the following questions:

- What is the probability of encountering a reactive trajectory in a given state, meaning, the probability density function of reactive trajectories?
- What is the amount of reactive trajectories going within a given state, meaning, the probability current of reactive trajectories?
- What is the average frequency of transitions between two sets, say A and B, meaning, the rate of reaction?
- What are the mechanisms of transitions, meaning, the transition tubes, or transition pathways?

The essential component in the main objects given by TPT is the committor function $q_{AB}(x) \equiv q(x)$, which is the probability of going rather to the set B than to the set A conditioned on the process has begun in the state x. The committor function $q(x)$ can be seen as an ideal reaction coordinate, because under proper conditions on the dynamics the levels sets of the committor function foliate the state space in sets of equal probability to rather end up in B than A, i.e., it explains the process of reaction from A to B in terms of probabilities.

Chapter 2

Stochastic Kinetic Reactions

The kinetics of biological reactions can be illustrated using a network of reaction channels $R_1, \cdots, R_M$ that include reactant and product molecules belonging to a set of d different species $S_1, \cdots, S_d$ with d and $M \in \mathbb{N}^+$. For instance, we might understand that when a molecule from the species S_1 encounters a molecule of type S_2 and certain microphysical conditions are met, the two molecules can fuse into a new molecule of type S_3. Such an interaction "law" can be simply specified naturally by using the notation

$$R_1 : S_1 + S_2 \longrightarrow S_3 \tag{2.0.0.1}$$

Although such reaction channels $R_j (j = 1, \cdots, M)$ catch the interactions between the species, they are not adequate by themselves to explain the full dynamics of the biological reaction. This also requires an understanding of the "rates" at which the reaction channels fire and some initial conditions.

Such explanations of biological reactions simply direct to the idea that the mathematical treatment should take into account that any modifications induced by the reaction channels in the copy numbers of species S_i $(i = 1, \cdots, d)$ are discrete. As already briefly addressed in the introduction, this inspiration of using a discrete characterization is, of course, totally correct, as it reproduces the intrinsic discreteness of nature. The scope of this chapter is

to study the mathematical formula that lead to the discrete stochastic approach to reaction kinetics.

2.1 Derivation of the chemical master equation

As the purpose is to discover how the copy numbers of the species $S_1, \cdots, S_d$ grow as time progress, we formally denote the state of the system by

$$X(t) = [X_1(t), X_2(t), \cdots, X_d(t)] \tag{2.1.0.1}$$

and specify the initial condition as $X(t_0) = x_0 \in N_0^d$ (from here on, the boldface notation $\mathbf{x} \equiv [x_1, \cdots, x_d]$ refers to vectors with d components). The components $X_i(t)$ of the state vector above describe random variables that encode the copy numbers x_i of the species S_i, which are existing within the container of volume V at time t. Each time one of the M reaction channels R_j fires, the state $X(t)$ changes. Without knowing the spatial movements of the molecules, the information required to learn the new state is which R_j reaction fired, and when did this event happen. This makes $X(t)$ a stochastic process, as the firing time and the choice of reaction channels are both random events. Therefore, the key to solve the problem is to define the reaction channels R_j in terms of probabilities.

Under the assumptions that the system is *well-stirred* and at *thermal equilibrium*, it has been rigorously shown in [35, 36] that for each reaction channel $R_j (j = 1, \cdots, M)$, there is a function α_j defined such that $\alpha_j(x)dt$ = the probability, given $X(t) = x \in N_0^d$, that a randomly chosen reaction R_j will fire inside the volume V within the infinitesimal time interval $[t, t+dt]$, with $j = 1, \cdots, M$, and a vector describing the corresponding state change,

with components μ_i^j = change in the molecular number of species S_i triggered by the firing of reaction R_j, $i = 1, \cdots, d$ and $j = 1, \cdots, M$.

The function α_j is called *propensity function* and the vector μ_j is usually referred to as the *stoichiometric* vector, and together they completely specify the reaction channel R_j .

For instance, in the case of the bimolecular reaction R_1, the stoichiometric vector encodes the decrease of the molecular numbers for species S_1 and S_2 by one molecule, and the corresponding increase in the copy numbers of S_3 by the same number. Therefore, $X(t)$ changes to $X(t) + \mu^1$, with $\mu^1 = [-1, -1, 1]$, when assuming the whole system contains only three species.

The derivation of the propensity functions is more involved, using probability laws and molecular mechanics arguments and has a solid microphysical foundation. A comprehensive treatment of the subject can be found in, e.g. [36].

Generally speaking, the propensity functions have the following formula

$$\alpha_j(x) = c_j h_j(x) \tag{2.1.0.2}$$

with c_j being a particular reaction rate constant, specified such that $c_j dt$ is the probability that some random combination of fitting R_j reactant molecules will communicate in the next infinitesimal time interval $[t, t + dt)$. We shall now get a closer look at the derivation of the two terms on the right-hand side of the above equation for the case of bimolecular reactions. Assume a fixed volume V contains a well-stirred mixture of D chemical species $S_1, S_2, \cdots, S_D$, reacting through M chemical reaction channels $R_1, R_2, \cdots, R_M$. Well-stirred here has two meanings: the system of molecules is homogeneous, meaning the probability of observing any randomly selected molecule inside any volume ΔV is $\frac{\Delta V}{V}$ and the system of molecules is

in thermal equilibrium, meaning the macroscopic thermal observables do not vary over time. Let the (transition) probability function $p(x,t|x^0,t_0)$ denotes the probability that there will be $x=(x_1,x_2,\cdots,x_D)$ molecules of each species at time t in V , given that the numbers of molecules is x_0 at t_0 . The initial condition x_0 , t_0 is often contained to simplify the notation $(p(x,t))$.

The (chemical) master equation is the time-evolution equation for the grand probability function, applying the Markov property, which states that the conditional probability for the event (x_n,t_n) given the full history of the system satisfies

$$p(x^n,t_n|x^{n-1},t_{n-1};\cdots;x^0,t_0)=p(x^n,t_n|x^{n-1},t_{n-1}), \tag{2.1.0.3}$$

meaning the dependence of the present (x_n,t_n) on past events can be captured only by the dependence on the previous state (x_{n-1},t_{n-1}).

Even though the Markov property is not exactly fulfilled for any given physical/chemical system, it can often be used as an accurate approximation. One important consequence of the Markov assumption is the Chapman-Kolmogorov equation,

$$p(x^2,t_2|x^0,t_0)=\sum_{x_1}p(x^2,t_2|x^1,t_1)p(x^1,t_1|x^0,t_0). \tag{2.1.0.4}$$

The master equation can be obtained instantly from the Chapman-Kolmogorov equation. The time derivative of the grand probability function is defined as

$$\frac{\partial}{\partial t}p(x,t|x^0,t_0)=\lim_{\Delta t\to 0}\frac{p(x,t+\Delta t)-p(x,t)}{\Delta t}. \tag{2.1.0.5}$$

Introduce a dummy variable y using the Chapman-Kolmogorov equation,

$$\frac{\partial}{\partial t}p(x,t|x^0,t_0) = \lim_{\Delta t\to 0}\frac{\sum_y p(x,t+\Delta t|y,t)p(y,t) - p(y,t+\Delta t|x,t)p(x,t)}{\Delta t}. \tag{2.1.0.6}$$

Let $W(x^2|x^1) = \lim_{\Delta t\to 0} p(x^2, t+\Delta t|x^1,t)/\Delta t$ denote the transition probability per unit time from state x_1 to x_2. The above equation for the time derivative can be simplified as

$$\frac{\partial}{\partial t}p(x,t|x^0,t_0) = \sum_y W(x|y)p(y,t) - W(y|x)p(x,t). \tag{2.1.0.7}$$

For chemical systems, $W(x_2|x_1)$ is nonzero if and only if there is chemical reaction connecting x_2 with x_1. The reaction R_j: $x-\mu_j \longrightarrow x$ (μ_j is the j-th column of matrix μ_j) happens with probability $\alpha_j(x-\mu_j)dt$ in interval $[t, t+dt)$, which implies that $W(x|x-\mu_j) = \alpha_j(x-\mu_j)$. Therefore

$$\frac{\partial}{\partial t}p(x,t|x^0,t_0) = \sum_{j=1}^{M}\alpha_j(x-\mu_j)p(x-\mu_j,t) - \alpha_j(x)p(x,t). \tag{2.1.0.8}$$

Another approach to derive the master equation is to write $p(x,t+dt)$ as the sum of the probabilities of the $1+M$ different ways in which the system can reach the state x at time $t+dt$:

$$p(x,t+dt) = p(x,t)\times\left(1-\sum_{j=1}^{M}\alpha_j(x)dt\right) + \sum_{j=1}^{M}p(x-\mu_j,t)\times\alpha_j(x-\mu_j)dt, \tag{2.1.0.9}$$

where the first term describes the probability that no reaction happens during $[t, t+dt)$ and the system remains in state x, while each term in the second summation is the probability that one reaction R_j happens in $[t, t+dt)$ and changes the state $x-\mu_j \longrightarrow x$. Reconstructing

this formula can also establish the master equation.

Taking the sum of all possible states on both sides of the master equation shows that the master equation conserves probability. Next, applying the master equation to calculate the mean value $< x(t) >= \sum xp(x,t)$ yields

$$\frac{d}{dt} < x(t) >= \sum_{j=1}^{M} \mu_j < \alpha_j(x(t)) >= \sum_{j=1}^{M} \mu_j \alpha_j(< x(t) >), \tag{2.1.0.10}$$

which is equivalent to the reaction rate equation if all the propensity functions $\alpha_j(\cdot)$ are linear (the last equality requires the linearity).

The *Chemical Master equation* (CME) is a difference-differential equation that represents the probability flow responsible for producing and ending any given state of the system under the condition of starting state x_0. The first term considers for inflow into state x from neighboring states, while the second term describes the outflow from state x. Solving the CME gives the full picture of the dynamics of the process $X(t)$.

2.2 Numerical Methods for the CME

2.2.1 Direct Methods for the CME

When the state space $\{x_1, x_2, \cdots\}$ (each element x_i is a distinct D-dimensional molecular population vector) is chosen, the chemical master equation can be rewritten into an infinite linear system of ODEs (ordinary differential equation),

$$\frac{dP(t)}{dt} = P(t)A, \tag{2.2.1.1}$$

where $P(t)$ is the complete probability row vector at time t, $P(t) = (p(x^1, t), p(x^2, t), \cdots)$. The time-independent matrix A is defined from the nonnegative propensity functions,

$$A_{ij} = \begin{cases} -\sum_{k=1}^{M} \alpha_k(x^i), & \text{for } i = j, \\ \alpha_k(x^i), & \text{for} j \text{such that} x^j = x^i + \mu_k, \\ 0, & \text{otherwise} \end{cases} \tag{2.2.1.2}$$

It is obvious to see from the definition above that A is remarkably sparse with at most $M+1$ nonzero elements in each row and each row of A sums up to zero.

In theory, the size of the matrix A can be infinite, but in any physical system, the number of molecules of each species is finite. To calculate the solution $P(t)$ numerically, the CME is often truncated to a finite state problem, limiting the state space of the CME to a finite domain, but also large enough to represent the true physical solution. Nevertheless, for most chemical systems, the size of the CME after truncation is still often large, on the order of 10^5 to 10^9, which makes solving the chemical master equation numerically intensive. The CME illustrates "the curse of dimensionality", which states that the computational complexity increases exponentially with the dimension (the number of reacting species here), unless other hypotheses are made.

Numerical approaches solving the CME directly as a linear differential equation can be coarsely classified into two groups, grid-based methods and Galerkin-based ones [74]. The grid-based methods, which explicitly truncate/aggregate the probability distribution into some finite, smaller domains, include the finite state projection (FSP) method [5], adaptive FSP with Krylov-based exponential computation [69], (adaptive) sparse grids methods [41, 42], and others [29, 83].

Grid-based methods first grid the whole domain space into pairwise disjoint subsets, where each subset may contain just one state or tens to hundreds of states. Then, based on some given guidelines, a number of these subsets are selected, and their union forms the computation domain. For example, Zhang et al. [83] use a few runs of SSA simulations to select potential subsets, precisely, the union of those subsets that have been touched by at least one simulation run is used as the computation domain. All states in the identical selected subset are then aggregated into one single state by the aggregation operator, thus reducing the problem size. The solution $P(t)$ can be recovered from the solution to the reduced problem by using the disaggregation operator. In (adaptive) sparse grids methods, a linear combination of some aggregation/disaggregation operator pairs is applied to obtain maximum reduction of the problem size. Nevertheless, in most of these grid-based approaches, disaggregation operators are represented by piece-wise constant/linear polynomial functions. Another numerically oriented method is to approximate the operator A in the master equation by its second order Taylor expansion, which gives the *Fokker-Planck equation*,

$$\frac{\partial}{\partial t}p(x,t) = -\sum_{j=1}^{M} \mu_j^T \Delta_x(\alpha_j(x)p(x,t)) + \frac{1}{2}\mu_j^T \Delta_x(\mu_j^T \Delta_x(\alpha_j(x)p(x,t))). \tag{2.2.1.3}$$

The Fokker-Planck equation is a D-dimensional parabolic partial differential equation (PDE), that can be solved by fully built numerical PDE methods. Computational cost is saved because the spatial discretization for numerical PDE methods can be much coarser than the actual state space. Nonetheless, it is often challenging to decide a priori how well the continuous Fokker-Planck equation approximates the discrete master equation.

2.2.2 Direct Stochastic Simulation Algorithm

The stochastic simulation algorithm (SSA) [22] offers an alternative way to approximate the grand probability function besides solving the CME directly. Define the function $p(\tau, \mu)$ such that $p(\tau, \mu)d\tau$ is the probability that, given the state x at time t, the next reaction in the system will occur in the infinitesimal time interval $[t + \tau, t + \tau + d\tau)$, and will be R_τ. The probability function $p(\tau, \mu)$ can be decomposed as the product of the probability function $p_0(\tau)$, the probability that, given the state x at time t, no reaction will occur in the time interval $[t, t + \tau)$, times $\alpha_\mu d\tau$, the probability that the reaction R_μ will occur in the time interval $[t + \tau, t + \tau + d\tau)$, i.e.,

$$p(\tau, \mu)d\tau = p_0(\tau)\alpha_\mu d\tau \tag{2.2.2.1}$$

Note that the probability that no reaction will occur in the infinitesimal time interval $[t, t+dt)$ is $p_0(dt) = (1 - \sum_j \alpha_j dt)$. Let $\alpha_0 = \sum_j \alpha_j$, then

$$\begin{cases} p_0(\tau + d\tau) = p_0(\tau)(1 - \alpha_0 d\tau) \\ p_0(0) = 1 \end{cases} \Longrightarrow p_0(\tau) = e^{-\alpha_0 \tau} \tag{2.2.2.2}$$

and

$$p(\tau, \mu) = e^{-\alpha_0 \tau}\alpha_\mu = \underbrace{\alpha_0 e^{-\alpha_0 \tau}}_{p_t(\tau)} \cdot \underbrace{\frac{\alpha_\tau}{\alpha_0}}_{p_r(\tau)}, \tag{2.2.2.3}$$

where $p_t(\tau)$ and $p_r(\tau)$ are the probability density/mass functions for the random variables τ and μ, respectively.

The direct stochastic simulation algorithm is stated as follows.

Step 1. Set the time variable $t := 0$ and the state variable x to the initial state.

Step 2. Calculate the propensity functions $\alpha_j(x), 1 \leq j \leq M$, for the current state x and the sum $a_0(x) = \sum_j \alpha_j(x)$.

Step 3. Generate random numbers τ and μ from the distributions with probability density/mass functions $p_t(\tau)$ and $p_r(\tau)$. One way to achieve this is to first generate two uniform $U(0,1)$ random numbers r_1 and r_2 and then choose the next reaction time by

$$\tau = \frac{1}{\alpha_0} \ln \frac{1}{r_1} \tag{2.2.2.4}$$

and the reaction channel k as the integer that satisfies the inequality

$$\sum_{j=1}^{k-1} \alpha_j < r_2 \alpha_0 \leq \sum_{j=1}^{k} \alpha_j. \tag{2.2.2.5}$$

Step 4. Update the system, $t := t + \tau$ and $x := x + \mu_k$. Repeat from Step2 until the final time t_f is reached.

The SSA algorithm stated here is precise, in the sense that the algorithm produces a sample trajectory consistent with the CME. Thus, the grand probability function $p(x, t|x_0, t_0)$ can be measured from a set of trajectories starting from the same initial condition (x_0, t_0). However, Monte Carlo methods like SSA converge very slowly, implying tremendous trajectories are needed to compute statistical parameters and probability distributions accurately. Furthermore, since the SSA is an explicit approach, simulating one trajectory itself may not be easy in some cases.

2.2.3 Tau-leaping Method and Other Monte Carlo Methods

In order to accelerate the original SSA algorithm, improvements have been made by adopting different approximation techniques. One of the most famous and promising approaches is the tau-leaping method [65], which uses the Poisson approximation to "leap-over" many reactions.

Assume that the time step τ is short enough such that the expected change in molecular numbers in the time interval $[t, t+\tau)$ leads to a negligible expected change of the propensity functions, i.e.,

$$\alpha_j(x(t+\tau)) \approx \alpha_j(x(t)), \quad 1 \leq j \leq M. \tag{2.2.3.1}$$

Then the number of firings of reaction R_j in $[t, t+\tau)$ is a Poisson random variable with mean $\alpha_j(x(t))\tau$ and is independent from all other reactions.

The algorithm for the explicit tau-leaping method is similar to the SSA, except that at Step 3, τ is chosen deterministically to satisfy the "leap-condition" and for each reaction channel j, the number of firings k_j is generated as a Poisson random number $P(\alpha_j(x(t))\tau)$ with mean $\alpha_j(x(t))\tau$. At Step 4 the system is updated with $t = t + \tau$ and $x = x + \sum_j k_j \mu_j$.

Another way to express the explicit tau-leaping method is

$$x(t+\tau) = x(t) + \sum_{j=1}^{M} P(\alpha_j(x(t))\tau) v_j. \tag{2.2.3.2}$$

At a coarser scale, suppose that the leap interval τ spans a very large number of firings of each reaction, yet only insignificant changes in each propensity function are induced, which is reasonable for large numbers of molecules. Then, the Poisson random variable $P(\alpha_j\tau)$ is well approximated by the normal random variable $N(\alpha_j\tau, \alpha_j\tau)$, implying the *Langevin*

leaping formula,

$$\begin{aligned} x(t+\tau) &= x(t) + \sum_{j=1}^{M} N(\alpha_j(x(t))\tau, \alpha_j(x(t))\tau)v_j \\ &= x(t) + \sum_{j=1}^{M} \alpha_j(x(t))\tau v_j + \sum_{j=1}^{M} \sqrt{\alpha_j(x(t))\tau} N_j(0,1)v_j. \end{aligned} \tag{2.2.3.3}$$

Note that in this formula $x(t)$ is a real function. The Langevin leaping formula is actually equivalent to the *chemical Langevin equation* (CLE),

$$dx(t) = \sum_{j=1}^{M} \alpha_j(x(t))v_j dt + \sum_{j=1}^{M} \sqrt{\alpha_j(x(t))} v_j dW_j, \tag{2.2.3.4}$$

where W_j is a Wiener process and $dW_j/dt = N_j(0, 1/dt)$ is Gaussian white noise. Note that for very large numbers of molecules, the stochastic term $\sqrt{\alpha_j(x(t))}v_j$ is much smaller than the deterministic term $\alpha_j(x(t))v_j$. In the limit the stochastic fluctuations in the CLE become negligible and the CLE approximates the RRE

$$\frac{dx(t)}{dt} = \sum_{j=1}^{M} \alpha_j(x(t))v_j. \tag{2.2.3.5}$$

The original explicit tau-leaping method is not very efficient when stiffness is present (where some reactions occur much faster than the others, and the fast reactions force τ to be very small). Implicit tau-leaping methods have been proposed to solve this issue [67],

$$x(t+\tau) = x(t) + \sum_{j=1}^{M} \alpha_j(x(t+\tau))\tau v_j + \sum_{j=1}^{M} (P(\alpha_j(x(t))\tau) - \alpha_j(x(t))\tau)v_j. \tag{2.2.3.6}$$

Another approach to deal with the stiffness efficiently and to speed up the original SSA algorithm is to make usage of stochastic versions of the quasi-steady state or partial equilibrium assumptions. In the deterministic case, the quasi-steady-state approximation assumes that at some time scale, instantaneous rates of change for some intermediate species are approximately zero, while the partial equilibrium approximation assumes that some fast reaction channels are continuously in equilibrium. The former one was extended to the stochastic quasi-steady-state approximation (SQSSA) [64] and the latter one was used to generate the slow-scale SSA method [11].

To illustrate the quasi-steady-state approximation (QSSA) and the partial equilibrium assumption, let us consider the system of reactions

$$\begin{aligned} \frac{d[A]}{dt} &= \epsilon^{-1} f([A],[B],[C],\cdots) && fast, \\ \frac{d[B]}{dt} &= g([A],[B],[C],\cdots) && intermediate, \\ \frac{d[C]}{dt} &= \epsilon h([A],[B],[C],\cdots) && slow, \end{aligned}$$

where $0 < \epsilon \ll 1$. For the slow reactant C, assume $d[C]/dt \approx 0$, $[C] \approx$ constant (quasi steady state). For the fast reactant A, assume that $[A]$ changes very rapidly about the mean $\langle [A] \rangle$, which is nearly constant in time (partial equilibrium). Then the system could be simplified to involve only the intermediate reaction for $[B]$, taking $[C]$ constant and $[A] = \langle [A] \rangle$.

2.2.4 Conclusion

Given a specified error boundary ϵ, it is well known that the computational effort for the Monte Carlo simulation methods like SSA is on the order of ϵ^{-2}. Remember that the CME is actually a D-dimensional deterministic discrete partial differential equation (PDE),

where D is the number of species. For classical PDE solvers with k-th order of accuracy $\epsilon = cN^{-k/D}$, where c is a constant and N is the number of unknowns for the numerical method in D dimensions. Assume that the computational effort is a linear function of N, then the computational effort is on the order of $\epsilon^{-D/k}$.

Relatively speaking, calculating the CME directly in full as discrete PDE is not the best way in high dimensions and when only approximate estimates of some statistics are needed, comparing to Monte Carlo methods.

Chapter 3

Transition Path Theory for Jump Markov Process

Continuous-time Markov chains on discrete state-spaces have a tremendous range of applications. In recent years, especially, with the pop of new applications in network science, Markov chains have become the weapon of choice not only to illustrate the dynamics on these networks but also to investigate their topological properties [62, 50]. In these circumstances, there is a need for new techniques to analyze Markov chains on large state-spaces with no particular symmetries, as is suitable for large complicated networks.

A straightforward starting point to analyze a Markov chain is to apply spectral analysis. This is particularly relevant when the chain displays metastability, as was shown in [2, 54] in the context of time-reversible chains. By definition, the generator of a metastable chain possesses one or more clusters of eigenvalues near zero, and the corresponding eigenvectors provide a direct way to partition the chain (and hence the underlying network) into cluster of nodes on which the random walker remains for a very delayed time before finding its way to another such cluster. This method has been used not only in the context of Markov chains originating from statistical physics (such as glassy systems [32, 1], or biomolecules [10]) but also in the meaning of data segmentation and embedding [45, 51, 72, 49, 12, 63, 70]. The issue with the spectral approach, nevertheless, is that not all Markov chains of interest

are time-reversible and metastable, plus, when they are not, the significance of the first few eigenvectors of the generator is less clear.

In this thesis, we will mainly talk about another approach that does not require metastability and applies for non-time-reversible chains as well. The basic concept is to single out two subsets of nodes of interest in the state-space of the chain and ask what the typical mechanism is by which the walker transits from one of these subsets to the other. We can also ask what the rate is at which these transitions occur, and so on. The first object which comes to mind to describe these transitions is the path of maximum likelihood by which they happen. However, this path can again be not very informative if the two states one has singled out are not metastable states. The primary purpose is to prove that we can give a definite meaning to the question of finding common mechanisms and rates of transition, even in chains that are neither metastable nor time-reversible. In so doing, we shall employ the framework of transition path theory (TPT) which has been developed in [78, 19, 27] in the context of diffusions.

TPT addresses questions like

(1) What is the probability distribution of the particles in the transition path ensemble?

(2) What is the transition rate?

(3) What is the probability current of the transition paths?

In a nutshell, given two subsets in state-space, TPT investigates the statistical properties of the corresponding reactive trajectories, i.e., the trajectories by which transition happens between these sets. TPT gives information such as the probability distribution of these trajectories, their probability current and flux, and their rate of appearance. In this thesis, we shall adopt TPT to continuous-time Markov chains and demonstrate the output of the theory via several examples. For the sake of brevity, we will focus just on discrete-

time Markov chains. We choose representative examples driven by molecular dynamics and chemical physics; however, the tools of TPT presented here can also be used for data segmentation and data embedding. In this context, TPT may also provide an option to Laplacian eigenmaps [72, 49] and diffusion map [63], which have become very famous recently in data analysis.

3.1 Probability theory and stochastic process

First, let us quickly compile the relevant theoretical tools from probability and stochastic process theory by adapting some definitions from [[60], Chapter 3].

A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as a triple composed of a sample space of outcomes $\Omega = \{w_1, w_2, \cdots, \}$, a σ-algebra $\mathcal{F}$ over the subsets of Ω and a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$, which satisfies the requirements $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ and

$$\mathbb{P}(\cup_{k=1}^{\infty})A_k = \sum_{k=1}^{\infty} \mathbb{P}(A_k) \tag{3.1.0.1}$$

for all sequences of pairwise disjoint sets $\{A_k\}_{k=1}^{\infty} \in \mathcal{F}$. Further, let $S \neq \emptyset$ be a finite or countable state set and $\mathcal{G}$ a σ-algebra over S, which together define a measurable space $(S, \mathcal{G})$.

Then, a *random variable* $X = X(w)$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ can be defined as a mapping

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{G})$$

between a *sample space* $(\Omega, \mathcal{F})$ and a state space $(S, \mathcal{G})$, both measurable, with the property that the events $\{w \in \Omega : X(w) \in A\} \in \mathcal{F}$ for any $A \in \mathcal{G}$. The *expectation* of the random

variable X is defined by

$$\mathbb{E}X = \int_{\Omega} X(w) d\mathbb{P}(w) \tag{3.1.0.2}$$

as the weighted sum over all the possible outcomes that the random variable can take.

Next, let $\mathcal{B}(U)$ denote the *Borel* σ-algebra of a topological space set U, in other words, the smallest σ-algebra containing all the open sets of U. Every random variable

$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{B}(S))$$

induces then a probability measure on S,

$$\mathbb{P}_X(B) = \mathbb{P}X^{-1}(B) = \mathbb{P}(w \in \Omega; X(w) \in B), \quad B \in \mathcal{B}(S) \tag{3.1.0.3}$$

and we call P_X the distribution of X. For the case of $S = \mathbb{R}^d$, we can write

$$d\mathbb{P}_X(x) = p(x)dx \tag{3.1.0.4}$$

and refer to $p(x)$ as the *probability density function.*

We are now ready to define a *stochastic process* as a collection of random variables $X := \{X(t, w), w \in \Omega, t \in T\}$ with $T = \{t_0 \leq t_1 \leq \cdots\}$ an ordered set of time points. Fixing $w \in \Omega$ we obtain a realization or trajectory $X(t)$ of the process X, and by fixing t we get a random variable $X(w)$.

3.2 The Markov Property

Speaking now in looser terms, we can think about a stochastic process as a system which evolves probabilistically in time, i.e., in which a certain time-dependent random variable exists. We can then measure its values $\{x_0, x_1, \cdots, x_n, \cdots\}$ at certain times $\{t_0 \leq t_1 \leq \cdots \leq t_n \leq \cdots\}$ and assume that a joint probability density

$$p(\cdots; x_n, t_n; x_{n-1}, t_{n-1}; \cdots; x_0, t_0) \tag{3.2.0.1}$$

exists, which describes the dynamics of the system completely [30]. Next, we can use (3.2.0.1) to define the conditional probability density

$$p(\cdots; x_n, t_n; \cdots; x_{j+1}, t_{j+1} | x_j, t_j; \cdots; x_0, t_0) = \frac{p(\cdots; x_n, t_n; x_{n-1}, t_{n-1}; \cdots; x_0, t_0)}{p(x_j, t_j; \cdots; x_0, t_0)} \tag{3.2.0.2}$$

with $0 \leq j < n$.

If all such conditional probabilities (3.2.0.2) would be possible, this would likewise lead to a complete explanation of the dynamics. Nevertheless, such a description would require a complete history of the system and thus be too complicated. A compelling idea to reduce the complexity is the *Markov assumption*. This specifies that the conditional probability is completely determined by the current state and not by the past, i.e.,

$$p(x_n, t_n | x_{n-1}, t_{n-1}; \cdots; x_0, t_0) = p(x_n, t_n | x_{n-1}, t_{n-1}) \tag{3.2.0.3}$$

which is the *Markov property* (notice that in (3.2.0.3) we have used a finite set of measurements to simplify the notation). The Markov property has the important consequence that

we can now express the joint probability density (3.2.0.1) in terms of simple conditional probabilities

$$p(x_n, t_n; \cdots ; x_0, t_0) = p(x_n, t_n|x_{n-1}, t_{n-1})p(x_{n-1}, t_{n-1}|x_{n-2}, t_{n-2}) \cdots p(x_1, t_1|x_0, t_0)p(x_0, t_0) \tag{3.2.0.4}$$

which means that any future state can be described given only an initial condition and the simple transition probability densities $p(x_j, t_j|x_{j-1}, t_{j-1}), 1 \leq j \leq n$, thus simplifying the treatment of processes that exhibit property (3.2.0.3). Such processes are called *Markov processes* and are in effect *memoryless* because the future development of the process depends only on the current state and not on any of the past states.

The Markov property also has another necessary result. Starting from the addition law of probability for mutually exclusive events, and by reducing one of the variables from the joint probability density by taking the sum over that variable, we have

$$p(x_2, t_2|x_0, t_0) = \int p(x_2, t_2; x_1, t_1|x_0, t_0)dx_1 \tag{3.2.0.5}$$

for three measurements taken at $t_0 \leq t_1 \leq t_2$. Using the definition (3.2.0.2) of the conditional probability density and the Markov property (3.2.0.3) we can write (3.2.0.5) as

$$\begin{aligned} p(x_2, t_2|x_0, t_0) &= \int p(x_2, t_2; x_1, t_1|x_0, t_0)dx_1 \\ &= \int p(x_2, t_2|x_1, t_1; x_0, t_0)p(x_1, t_1|x_0, t_0)dx_1 \\ &= \int p(x_2, t_2|x_1, t_1)p(x_1, t_1|x_0, t_0)dx_1 \end{aligned} \tag{3.2.0.6}$$

which is the *Chapman-Kolmogorov equation* (cf. [30]). In the case of discrete variables that take only integer values, the *Chapman-Kolmogorov equation* for discrete state spaces reads

$$\mathbb{P}(X(t_2) = x_2|X(t_0) = x_0) = \sum_{x_1} \mathbb{P}(X(t_2) = x_2|X(t_1) = x_1)\mathbb{P}(X(t_1) = x_1|X(t_0) = x_0). \tag{3.2.0.7}$$

Of course, before applying outcome (3.2.0.4), the question is raised about whether any general process exists that actually perceives the Markov property (3.2.0.3) exactly. If we assume a magnificent time scale for observations, the answer is negative, because, at the very least, we would need the immediate history to predict the probabilistic future. Fortunately, however, processes that have a relatively short memory, meaning that their memory time is far shorter than the timescale used in marking the measurements, are common. Thus, it is logical to assume that a Markov process approximates such systems with sufficient accuracy and the popularity of Markovian models in many fields of science is evidence of this fact.

Another viewpoint of the current discussion about stochastic processes is whether the state space is discrete or continuous and whether the time evolution proceeds discretely or continuously. Considering that the dynamics of biological processes evolve continuously in time, the quantities of interest take integer values, the focus in our case is predictably on the continuous-time Markov process with a discrete state space. In case the state space is finite or countable, and the time evolution discrete, the term Markov chain is sometimes applied. Without loss of generality we shall take the finite state space to be $S = \{1, \cdots, N\} \subset \mathbb{N}$. Let us now present the construction of a continuous-time Markov process.

3.3 Continuous-time Markov process

The starting point for the construction of the continuous-time object is a discrete-time Markov chain which we proceed to define as in [[60], Chapter 3].

Definition 3.3.1. A random sequence $\{X_n\}_{n\geq 0}$ is a discrete-time Markov chain with initial distribution ρ_0 and transition matrix P , if it is a stochastic Markov process on the finite state space S with initial distribution ρ_0 (viewed as a column vector),

$$(\rho_0)_i = \mathbb{P}(X_0 = i), i \in S \tag{3.3.0.1}$$

and transition probability from state i to state j given as

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i), i, j \in S, \tag{3.3.0.2}$$

for every $n \geq 0$ and $\mathbb{P}(X_n = i) > 0$.

If the transition probabilities are independent of n, then the process is said to be homogeneous. The transition probabilities $\{p_{ij}\}_{i,j\in S}$ can be assembled into a transition matrix $P \in \mathbb{R}^{N\times N}$, which satisfies

$$0 \leq p_{ij} \leq 1, \forall i, j \in S \tag{3.3.0.3}$$

$$\sum_{j\in S} p_{ij} = 1. \tag{3.3.0.4}$$

Any matrix that satisfies the above conditions (3.3.0.3) and (3.3.0.4) is called a stochastic matrix.

Further, using the Chapman-Kolmogorov equation for discrete state spaces (3.2.0.7) and induction on n, it can be shown that the n-step transition probability from state i to state j,

denoted by $p_{ij}^n = \mathbb{P}(X_n = j|X_0 = i)$ is equal to $(P^n)_{ij}$, and computing the probability that the Markov chain will be in state j at $n \geq 0$ will reduce to computing the corresponding power of the transition matrix. Consequently, for an initial distribution ρ_0 we have

$$\mathbb{P}(X_n = j) = \sum_{i \in S} \mathbb{P}(X_n = j|X_0 = i) \cdot \mathbb{P}(X_0 = i) = \sum_{i \in S} (\rho_0)_i (P^n)_{ij} = (\rho_0 P^n)_j. \quad (3.3.0.5)$$

Thus, if we know the initial distribution and the transition matrix we can determine the probability distribution at any later time point. Moreover, by using the notation introduced in Definition 3.3.1 for transition probabilities, we can write the general form of the Chapman-Kolmogorov equation as

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} \quad (3.3.0.6)$$

which leads to

$$P^{m+n} = P^m P^n.$$

We turn now to the task of defining a continuous-time Markov process $\{X(t)\}_{t \in \mathbb{R}}$ with the same finite state space S as the discrete-time chain. In addition to observing the Markov property (3.2.0.3), we also want the process to be time-homogenous, i.e. to fulfill

$$\mathbb{P}(X(t) = j|X(s) = i) = \mathbb{P}(X(t-s) = j|X(0) = i) \quad (3.3.0.7)$$

for any states $i, j \in S$ and $s \leq t$. Intuitively, the main difference to the discrete-time setting discussed previously is that transitions can now occur at any time, so we need to establish how long the process will remain in a state $i \in S$ before performing a jump to a new state $j \in S$.

Let T_i denote the waiting time to the next jump while in state i. It can be shown by making use of the Markov property and the time-homogeneity requirement (3.3.0.7) that

$$\mathbb{P}(T_i > s + t | T_i > s) = \mathbb{P}(T_i > t). \tag{3.3.0.8}$$

Thus, T_i satisfies the memoryless requirement, as (3.3.0.8) basically says that the system forgets it has already waited for time s. This leads to the conclusion that T_i is exponentially distributed with a parameter $w(i)$, as the exponential distribution is the only continuous-time distribution that observes the Markov property (cf. [[71], Chapter 9.10]).
We proceed now to study the transition probabilities. First, as $T_i \approx \exp(w(i))$ and satisfies (3.3.0.8), we infer that

$$\mathbb{P}(T_i < dt) = 1 - e^{-w(i)dt} = w(i)dt + \mathcal{O}(dt^2)$$

when $dt \to 0$. Next, using the notation from Definition 3.3.1, we write the probability that the process will jump to state j after leaving state i as

$$p_{ij} = \mathbb{P}(X(T_i) = j | X(0) = i).$$

The transition probability does not depend on the time spent by the process in i, because if it would do so, the Markov property will no longer be observed. By defining

$$w(i, j) = w(i) \cdot p_{ij} \tag{3.3.0.9}$$

as the transition intensity from state i to state j, we can write

$$\begin{aligned}\mathbb{P}(X(t+dt)=j|X(t)=i) &= \mathbb{P}(X(dt)=j|X(0)=i)\\ &= \mathbb{P}(T_i<dt, X(T_i)=j|X(0)=i)\\ &= w(i)\cdot p_{ij}dt+\mathcal{O}(dt^2)\\ &= w(i,j)dt+\mathcal{O}(dt^2)\end{aligned} \tag{3.3.0.10}$$

with $\mathcal{O}(dt^2)$ accounting for the probability of more than one jump in the interval $[t, t+dt)$. Because of the way we have defined the transition intensities (3.3.0.9), we also have for $i \in S$

$$\sum_{j\neq i} w(i,j) = \sum_{j\neq i} w(i)\cdot p_{ij} = w(i)\sum_{j\neq i} p_{ij} = w(i). \tag{3.3.0.11}$$

Taking (3.3.0.11) into account, we can now write the probability that no jump will take place in $[t, t+dt)$ as

$$\begin{aligned}\mathbb{P}(X(t+dt)=i|X(t)=i) &= \mathbb{P}(X(dt)=i|X(0)=i)\\ &= 1-\sum_{j\neq i}\mathbb{P}(X(dt)=j|X(0)=i)\\ &= 1-\sum_{j\neq i} w(i,j)dt+\mathcal{O}(dt^2)\\ &= 1-w(i)dt+\mathcal{O}(dt^2).\end{aligned} \tag{3.3.0.12}$$

We are now ready to give a definition for a time-homogeneous continuous time Markov process with a finite state space S.

Definition 3.3.2. A stochastic process $\{X(t)\}_{t\in\mathbb{R}}$ with a finite state space S is a time-

homogeneous continuous time Markov process, if it satisfies

$$\mathbb{P}(X(t+dt)=j|X(t)=i)=w(i,j)dt+\mathcal{O}(dt^2) \tag{3.3.0.13}$$

$$\mathbb{P}(X(t+dt)=i|X(t)=i)=1-w(i)dt+\mathcal{O}(dt^2) \tag{3.3.0.14}$$

where $j \neq i$ and $w(i)$ is given as above.

A classic (and arguably one of the most important) example of a continuous-time Markov process is the Poisson process, which is an integer valued counting process $N(t)$ of the number of jumps in the time interval $[0,t]$. The Poisson process satisfies

$$\mathbb{P}(N(t+dt)=i+1|N(t)=i)=wdt+\mathcal{O}(dt^2) \tag{3.3.0.15}$$

$$\mathbb{P}(N(t+dt)=i|N(t)=i)=1-wdt+\mathcal{O}(dt^2) \tag{3.3.0.16}$$

with $w>0$ denoting the constant intensity of the process, which no longer depends on the state. Moreover, we have that the independent increments are exponentially distributed,

$$\mathbb{P}(N(t)-N(s)=k)=\frac{e^{-w(t-s)}(w(t-s))^k}{k!} \tag{3.3.0.17}$$

and depend only on $t-s$ making the Poisson process time-homogeneous.

After these preparations, a recipe for the construction of a continuous-time Markov process $\{X(t)\}_{t\in\mathbb{R}}$ can be formulated (see also [PS08, Chapter 5]). The procedure involves two objects, the first ingredient being an independent and identically distributed sequence $\{\tau_n\}_{n\geq 0} \sim \exp(w)$ that will provide the transition times, with the second component represented by a discrete-time Markov chain $\{X_n\}_{n\geq 0}$ with transition matrix P defined as in

(3.3.0.3,3.3.0.4), which provides the values for the states. We remark that $\{X_n\}_{n\geq 0}$ is sometimes called the embedded chain of the stochastic process $\{X(t)\}_{t\in\mathbb{R}}$. From an algorithmic viewpoint, first we set $X(0) = X_0$ and $t_0 = 0$ and let $t_{n+1} = t_n + \tau_n$ be the next jump time. Next, we define $X(t) = X_n$ for any $t \in [tn, t_{n+1})$, $\forall n \geq 0$. The process $X(t)$ thus obtained is called *Markov jump process*, and we note that the algorithm lightly sketched above is another formulation of the SSA algorithm.

Next, we present a matrix characterization for the continuous-time Markov process. Similarly to the discrete case, we can assemble the transition probabilities of a Markov jump process into a matrix $P(t)$ with elements

$$p_{ij}(t) = \mathbb{P}(X(t) = j | X(0) = i). \tag{3.3.0.18}$$

Due to the exponential distribution of the jump times, we also have

$$\mathbb{P}(N(t) = k) = \frac{e^{-wt}(wt)^k}{k!}. \tag{3.3.0.19}$$

Combining the probability with the k-step transition matrix of the embedded Markov chain leads to

$$p_{ij}(t) = \mathbb{P}(N(t) = k) \cdot \mathbb{P}(X_k = j | X_0 = i) = \sum_{k=0}^{\infty} \frac{e^{-wt}(wt)^k}{k!}(P^k)_{ij}. \tag{3.3.0.20}$$

Hence, in matrix form we have

$$P(t) = e^{-wt} \sum_{k=0}^{\infty} \frac{(wt)^k}{k!} P^k = e^{wt(P-I)} = e^{tL} \tag{3.3.0.21}$$

with $L = w(P - I)$ called the generator of the continuous-time Markov jump process. We remark that in case the state space is infinite, handling e^{tL} requires the operator theory of semigroups [[60], Chapter 7.5]. Thus, given an intensity w and the transition matrix P of the embedded chain we can characterize the Markov jump process. Additionally, the generator L satisfies

$$L = \lim_{t \to 0} \frac{P(t) - I}{t} \tag{3.3.0.22}$$

and because P is a stochastic matrix, we have

$$\sum_{j \in S} l_{ij} = 0 \ \forall i \in S, \tag{3.3.0.23}$$

$$l_{ij} \in [0, \infty) \ \forall i, j \in S \text{ with } i \neq j \tag{3.3.0.24}$$

$$\text{and } l_{ii} \leq 0. \tag{3.3.0.25}$$

Summarizing (3.3.0.23), (3.3.0.24) and (3.3.0.25), the rows of L must sum up to zero, the off-diagonal elements are non-negative, while the diagonal elements are non-positive.

We are now eventually in a position to draw the spotlight on the relationship between the time-continuous Markov chain, its generator, and the CME derived in (2.1.0.8). As we have seen, the generator is built using the stochastic matrix $P(t)$ with details defined by (3.3.0.18). The purpose is to conclude a set of differential equations that illustrate the development of the transition probabilities, or in other words, a master equation. Hence, we begin by taking the time derivative

$$\begin{aligned} \frac{d}{dt} p_{ij}(t) &= \lim_{dt \to 0} \frac{p_{ij}(t + dt) - p_{ij}(t)}{dt} \\ &= \lim_{dt \to 0} \frac{1}{dt} \Big((\mathbb{P}(X(t + dt) = j | X(0) = i) - \mathbb{P}(X(t) = j | X(0) = i) \Big). \end{aligned} \tag{3.3.0.26}$$

Using now the Chapman-Kolmogorov equation (3.3.0.6), we introduce a new variable y in (3.3.0.26) and write

$$\frac{d}{dt}p_{ij}(t) = \lim_{dt\to 0}\frac{1}{dt}\Big(\sum_{y\in S}\mathbb{P}(X(t+dt)=j|X(t)=y,X(0)=i)\mathbb{P}(X(t)=y|X(0)=i) \\ -\mathbb{P}(X(t)=j|X(0)=i)\Big). \tag{3.3.0.27}$$

Further, using (3.3.0.9) and (3.3.0.10) to expand the first term in (3.3.0.27), we have that

$$\begin{aligned}
&\sum_{y\in S}\mathbb{P}(X(t+dt)=j|X(t)=y,X(0)=i)\mathbb{P}(X(t)=y|X(0)=i)\\
&\quad=\mathbb{P}(X(t+dt)=j|X(t)=j,X(0)=i)\mathbb{P}(X(t)=j|X(0)=i)\\
&\quad+\sum_{y\neq j}\mathbb{P}(X(t+dt)=j|X(t)=y,X(0)=i)\mathbb{P}(X(t)=y|X(0)=i)\\
&\quad=\mathbb{P}(X(t+dt)=j|X(t)=j)\mathbb{P}(X(t)=j|X(0)=i)\\
&\quad+\sum_{y\neq j}\mathbb{P}(X(t+dt)=j|X(t)=y)\mathbb{P}(X(t)=y|X(0)=i)\\
&\quad=(1-w(j)dt)p_{ij}(t)+\sum_{y\neq j}w(y,j)dt\cdot p_{ij}(t)+\mathcal{O}(dt^2).
\end{aligned} \tag{3.3.0.28}$$

Inserting (3.3.0.28) into (3.3.0.27), rearranging the terms and passing to the limit, yields via

(3.3.0.11)

$$\begin{aligned}\frac{d}{dt}p_{ij}(t) &= \lim_{dt\to 0}\frac{1}{dt}\Big((1-w(j)dt)p_{ij}(t) - p_{ij}(t) + \sum_{y\neq j} w(y,j)dt\cdot p_{iy}(t) + \mathcal{O}(dt^2)\Big) \\ &= -w(j)p_{ij}(t) + \sum_{y\neq j} w(y,j)p_{iy}(t) \\ &= \sum_{y\neq j} w(y,j)p_{iy}(t) - \Big(\sum_{y\neq j} w(j,y)\Big)p_{ij}(t)\end{aligned} \tag{3.3.0.29}$$

which are the forward Kolmogorov equations for the process $X(t)$. Comparing (3.3.0.29) with (2.1.0.8), we observe that the chemical master equation is a special case of the forward Kolmogorov equation, with the inflow and outflow terms readily recognizable.

Equation (3.3.0.29) can also be written in matrix form, by defining the matrix L as

$$L_{ij} = \begin{cases} -w(j), & \text{if} i = j \\ w(i,j), & \text{if} i \neq j. \end{cases} \tag{3.3.0.30}$$

Thus, we obtain

$$\frac{d}{dt}P(t) = P(t)L. \tag{3.3.0.31}$$

When the state space S is finite, and subject to the initial condition $P(0) = I$, equation (3.3.0.31) has the formal solution $P(t) = e^{tL}$. Comparing with (3.3.0.21), it is clear that by defining L as in (3.3.0.30), we have recovered the generator of the Markov jump process.

Besides the forward Kolmogorov equations, we can also obtain another set of differential equations called the backward Kolmogorov equations. Using again Chapman-Kolmogorov equations (3.3.0.6) to expand a transition matrix $Q(t+dt)$ this time as $Q(dt)Q(t)$ and taking

the time derivative of $Q(t)$ at $t = 0$, we have

$$\begin{aligned}\frac{d}{dt}Q(t) &= \lim_{dt\to 0}\frac{Q(t+dt)-Q(t)}{dt}\\ &= \lim_{dt\to 0}\frac{Q(t)Q(dt)-Q(t)}{dt}\\ &= \lim_{dt\to 0}\frac{Q(dt)-I}{dt}Q(t)\\ &= L^{\star}Q(t).\end{aligned}$$

We conclude now this section by referring the readers interested in a more extensive treatment of stochastic processes to the monographs [9],[52]. For a viewpoint closer to the chemical master equation, [77], [30] are recommended.

3.4 Main TPT

3.4.1 Notations ans assumptions

We will consider a Markov jump process on the countable state-space S with infinitesimal generator (or rate matrix) $L = (l_{ij})_{i,j\in S}$:

$$\begin{cases} l_{ij} \geq 0 & \forall i, j \in S, i \neq j, \\ \sum\limits_{j\in S} l_{ij} = 0 & \forall i \in S. \end{cases} \tag{3.4.1.1}$$

Recall that if the process is in state i at time t, then $l_{ij}\Delta t + o(\Delta t)$ for $i \neq j$ gives the probability that the process jumps from state i to state j during the infinitesimal time interval $[t, t + \Delta t]$, and this probability is independent of what happened to the process

before time t. We assume that the Markov jump process is irreducible and ergodic with respect to the unique, strictly positive invariant distribution $\pi = (\pi_i)_{i \in S}$, the solution of

$$0 = \pi^T L. \tag{3.4.1.2}$$

We will denote by $X(t)_{t \in \mathbb{R}}$ an equilibrium sample path (or trajectory) of the Markov jump process, i.e., any path obtained from $X(t)_{t \in [T, \infty)}$ by pushing back the initial condition, $X(T) = x$, to $T = -\infty$. Following standard conventions, we assume that $X(t)_{t \in \mathbb{R}}$ is right-continuous with left limits (*càdlàg*) (i.e., at the times of the jumps the process is assigned to the state it jumps into rather than to the one it jumped from).

We will be intrigued by investigating specific statistical properties of the ensemble of equilibrium paths. In principle, this needs us to create a suitable probability space whose sample space is the ensemble of these equilibrium paths. Such a creation is standard (see, e.g., [47]), and we will not continue on it here since, by the assumption of ergodicity, the statistical properties of the ensemble of equilibrium paths that we are focused on can also be derived from almost any path in this ensemble through suitable time averaging. This is the perspective that we will adopt in this thesis since it gives an operational way to compute expectations from a trajectory generated, e.g., by numerical simulations.

Below, we will also need the process obtained from $X(t)_{t \in \mathbb{R}}$ by time reversal. We will denote this time-reversed process by $\tilde{X}(t)_{t \in \mathbb{R}}$ and define it as

$$\tilde{X}(t) = X^*(-t), \quad \text{where} \quad X^*(t) = \lim_{s \to t-} X(s)$$

By our assumptions of irreducibility and ergodicity, the process $\tilde{X}(t)_{t\in\mathbb{R}}$ is again a *càdlàg* Markov jump process with the same invariant distribution as $X(t)_{t\in\mathbb{R}}$, π, and infinitesimal generator $\tilde{L} = (\tilde{l}_{ij})_{i,j\in S}$ given by

$$\tilde{l}_{ij} = \frac{\pi_j}{\pi_i} l_{ji}.$$

Finally, recall that if the infinitesimal generator satisfies the detailed balance equations

$$\forall i, j \in S: \quad \pi_i l_{ij} = \pi_j l_{ji},$$

then $\tilde{L} \equiv L$ and, hence, the direct and the time-reversed process are statistically indistinguishable. Such a process is called reversible. We do not assume reversibility in this thesis. For the algorithmic part of this paper, it will be convenient to use the notation and concepts of graph theory. We will mainly consider directed graphs $G = G(S, E)$, where the vertex set S is the set of all states of the Markov jump process and two vertices i and j are connected by a directed edge if $(i, j) \in E \subseteq (S \times S)$.
We also recall the following definition.

Definition 3.4.1. A directed pathway $w = (i_0, i_1, i_2, \cdots, i_n)$, $i_j \in S$, $j = 0, \cdots, n$, in a graph G is a finite sequence of vertices such that $(i_j, i_{j+1}) \in E$, $j = 0, \cdots, n-1$. A directed pathway w is called simple if w does not contain any self-intersections (loops), i.e., $i_j \neq i_k$ for $j, k \in 0, ..., n, j \neq k$.

We will later consider several forms of induced directed graphs.

Definition 3.4.2. Let $E' \subset E$ be a subset of edges of a graph $G = G(S, E)$; then we denote by $G[E'] = G(S', E')$ the induced subgraph, i.e., the graph which consists of all edges in E'

and the vertex set

$$S' = \{i \in S : \exists j \in S \text{ such that } (i,j) \in E' \text{ or } (j,i) \in E'\}.$$

Definition 3.4.3. Whenever a $|S| \times |S|$-matrix $C = (C_{ij})$ with nonnegative entries is given, the weight-induced directed graph is denoted by $G\{C\} = G(S, E)$. In this graph the vertex set S is the set of all states of the Markov jump process, and two vertices i and j are connected by a directed edge $(i, j) \in E \subseteq (S \times S)$ if the corresponding weight C_{ij} is positive.

3.4.2 Reactive trajectories

Let A and B be two nonempty, disjoint subsets of the state-space S. By ergodicity, any equilibrium path $X(t)_{t\in\mathbb{R}}$ oscillates infinitely many times between set A and set B. We are interested in understanding how these oscillations happen (mechanism, rate, etc.). If we view A as a reactant state and B as a product state, then each oscillation from A to B is a reaction event, and so we are asking about the mechanism, rate, etc., of these reaction events. To properly define and characterize the reaction events, we proceed by pruning out of each equilibrium trajectory $X(t)_{t\in\mathbb{R}}$ the pieces during which it makes a transition from A to B (i.e., the reactive pieces), and we ask about various statistical properties of these reactive pieces. The pruning is done as follows (see also Figure 1 for a schematic illustration).

First, given a trajectory $X(t)_{t\in\mathbb{R}}$ we define a set of last-exit-before-entrance and first-entrance-after-exit times $\sigma = \{t_n^A, t_n^B\}_{n\in\mathbb{Z}}$ as follows.

Definition 3.4.4. (exit and entrance times). Given a trajectory $X(t)_{t\in\mathbb{R}}$, the last-exit-before-entrance time t_n^A and the first-entrance-after-exit time t_n^B belong to σ if and only

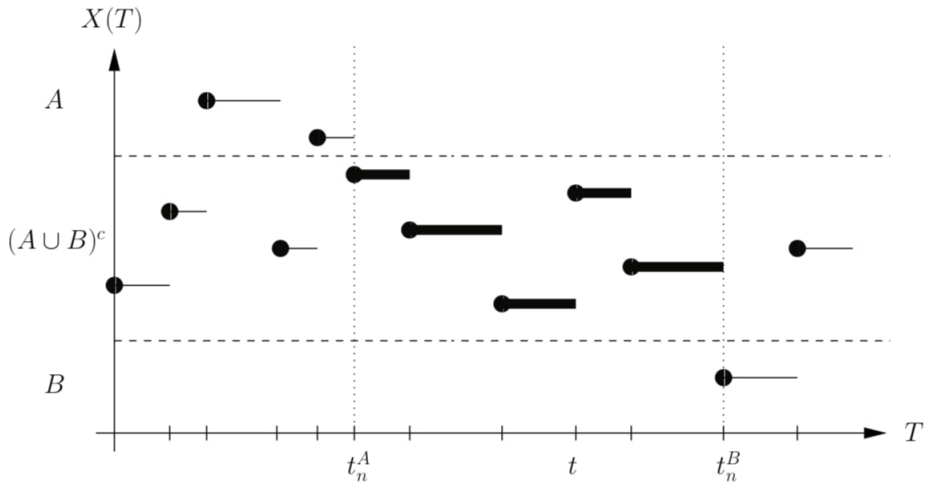


Figure 1: Schematic

if

$$\lim_{t \to t_n^A -} X(t) = x_n^A \in A, \quad X(t_n^B) = x_n^B \in B$$
$$\forall t \in [t_n^A, t_n^B) : X(t) \notin A \cup B. \tag{3.4.2.1}$$

By ergodicity, we know that the cardinality of σ is almost surely (a.s.) infinite. It is also clear that the times t_n^A and t_n^B form an increasing sequence, $t_n^A \leq t_n^B \leq t_{n+1}^A$, for all $n \in \mathbb{Z}$. Notice, however, that we may have $t_n^A = t_n^B$ for some $n \in \mathbb{Z}$ corresponding to events when the trajectory jumps directly from A to B. If, on the other hand, $t_n^A < t_n^B$, then the trajectory visits states outside of A and B when it makes a transition from the former to the latter.

Next, given the set σ, we define the following.

Definition 3.4.5. (reactive times). The set R of reactive times is defined as

$$R = \bigcup_{n \in \mathbb{Z}} (t_n^A, t_n^B) \subset \mathbb{R}. \tag{3.4.2.2}$$

Finally, we denote by $t_1 \equiv t_n^A \leq t_n^2 \leq \cdots \leq t_n^{k_n} \leq t_n^B$the set of all of the successive jumping times of $X(t)$ in $[t_n^A, t_n^B]$, i.e., all of the times in $[t_n^A, t_n^B]$ such that

$$\lim_{t \to t_n^k -} X(t) \neq X(t_n^k) =: x_n^k, \qquad k = 1, \cdots, k_n \in \mathbb{N}, \tag{3.4.2.3}$$

and we define the following.

Definition 3.4.6. (reactive trajectories). The ordered sequence

$$P_n = [x_n^A, x_n^1, x_n^2, \cdots, x_n^{k_n} \equiv x_n^B]$$

consisting of the successive states visited during the nth transition from A to B (including the last state in A, x_n^A, and the first one in B, $x_n^B \equiv x_n^{k_n}$) is called the nth reactive trajectory. The set of all such sequences,

$$P = \bigcup_{n \in \mathbb{Z}} \{P_n\},$$

is called the set of reactive trajectories.

(Note that we have $k_n = 1$ when the trajectory hops directly from A to B at time $t_n^A = t_n^B$, in which case $P_n = [x_n^A, x_n^B]$.)

Since the equilibrium trajectory $\{X(t)\}_{t \in \mathbb{R}}$ used in the construction above is part of a statistical ensemble, the sets R, P_n, and P are also random sets whose statistical properties are induced by those of the ensemble of equilibrium trajectories. In the next sections we obtain explicit expression for various expectations involving these random sets. Using ergodicity, these expectations can be computed a.s. from a single trajectory via time averaging, even though in this case σ, R, P_n, and P are fixed sets. As already explained above, the second viewpoint is the one we will take in this paper since it gives operational definitions to all of

the statistical quantities we are interested in.

3.4.3 Probability distribution of reactive trajectories

A first object relevant to quantify the statistical properties of the reactive trajectories is the following definition.

Definition 3.4.7. The distribution of reactive trajectories $m^R = (m_i^R)_{i \in S}$ is defined so that for any $i \in S$ we have

$$\lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \mathbf{1}_{\{i\}}(X(t)) \mathbf{1}_R(t) dt = m_i^R, \tag{3.4.3.1}$$

where $\mathbf{1}_{C(\cdot)}$ denotes the characteristic function of the set C.

The distribution m_R gives the equilibrium probability that the system is in state i at time t and that it is reactive at that time; i.e., m_i^R can also be expressed as

$$m_i^R = \mathbb{P}(X(t) = i \ \& \ t \in R), \tag{3.4.3.2}$$

where P denotes probability with respect to the ensemble of equilibrium trajectories. To avoid confusion, note that the random objects in (3.4.3.2) are $X(t)$ and R: the time t in this expression is fixed, and m_i^R does not depend on t since we look at equilibrium reactive trajectories.

How can we find an expression for m^R? Suppose we encounter the process $X(t)$ in a state $i \in S$. What is the probability that $X(t)$ is reactive? Intuitively, this is the probability that the process came from A rather than from B times the probability that the process will reach B rather than A in the future. This indicates that the following objects will play an

important role.

Definition 3.4.8. The discrete forward committor $q^+ = (q_i^+)_{i\in S}$ is defined as the probability that the process starting in $i \in S$ will first reach B rather than A. Analogously, we define the discrete backward committor $q^- = (q_i^-)_{i\in S}$ as the probability that the process arriving in state i last came from A rather than B.

The forward and backward committors both satisfy a discrete Dirichlet problem:

$$\begin{cases} \sum\limits_{j\in S} l_{ij} l_j^+ = 0 & \forall i \in (A\cup B)^c, \\ q_i^+ = 0 & \forall i \in A, \\ q_i^+ = 1 & \forall i \in B \end{cases} \tag{3.4.3.3}$$

and

$$\begin{cases} \sum\limits_{j\in S} \tilde{l}_{ij} l_j^- = 0 & \forall i \in (A\cup B)^c, \\ q_i^- = 1 & \forall i \in A, \\ q_i^- = 0 & \forall i \in B \end{cases} \tag{3.4.3.4}$$

Here $L = (l_{ij})_{i,j\in S}$ and $\tilde{L} = (\tilde{l}_{ij})_{i,j} \in S$ denote the infinitesimal generator forward and backward in time, respectively.

Proof. By definition, q_i^+ is the first entrance probability of the process $\{X(t), t \geq 0, X(0) = i\}$ with respect to the set B avoiding the set A. The usual step in dealing with entrance or hitting probabilities with respect to a certain subset of states is the modification of the process such that these states becoming absorbing states. Let $L = (l_{ij})_{i,j\in S}$ be the infinitesimal generator of a Markov jump process and $A \subset S$ be a nonempty subset. Suppose we are

interested in the process resulting from the declaration of the states in A to be absorbing states. Then the infinitesimal generator $\hat{L} = (\hat{l}_{ij})_{i,j\in S}$ of the modified process is given by

$$\hat{l}_{ij} = \begin{cases} l_{ij}, & i \in A^c, j \in S, \\ 0, & i \in A, j \in S. \end{cases} \tag{3.4.3.5}$$

Noe if we make the states in the set A absorbing states, then the discrete forward committor q^+ is the first entrance probability with respect to the set B under the modified process. Thus q^+ satisfies the discrete Dirichlet problem

$$\begin{cases} \sum_{j\in S} \hat{l}_{ij} q_j^+ = 0 & \forall i \in B^c, \\ q_i^+ = 1 & \forall i \in B \end{cases} \tag{3.4.3.6}$$

which is equivalent to the forward committor equation.

Noe observe hat if we substitute the "boundary conditions" into the equations in (3.4.3.3), then we end up with a linear system

$$Uq^+ = v, \tag{3.4.3.7}$$

where the matrix $U = (u_{ij})_{i,j\in(A\cup B)^c}$ is given by

$$u_{ij} = l_i j, \quad i, j \in (A \cup B)^c,$$

and an entry of the vector $v = (v_i)_{i\in(A\cup B)^c}$ on the right-hand side of (3.4.3.7) is defined by $v_i = -\sum_{k\in B} l_i k$ for all $i \in (A \cup B)^c$.

Now let's prove that if the matrix U is irreducible, then the solution of (3.4.3.3) is unique. By the definition of the matrix U there exists at least an index $k \in (A \cup B)^c$ such that

$$|u_{kk}| > \sum_{j \neq k} u_{kj}.$$

But this implies that U is weakly diagonally dominant. Together with its assumed irreducibility, this implies that it is invertible, and this is the end of proof for forward committor equation.

Next, we turn our attention to the discrete backward committor $q_i^-, i \in S$, which is defined as the probability that the process arriving at state i came from A rather than from B. The crucial observation is now that $q^- = (q_i^-)i \in S$ is the discrete forward committor with respect to the reversed time process.

The derivation of (3.4.3.4) is a straightforward generalization of the one of (3.4.3.3). Note that if the Markov jump process is reversible, then the detailed balance condition

$$\pi_i l_{ij} = \pi_j l_{ji} \quad \forall i, j \in S$$

is satisfied and the discrete backward committors solves (3.4.3.4). On one hand, the solution of the discrete Dirichlet problem is unique. On the other hand, a short calculation shows that $1 - q^+$ also satisfies the equation. Consequently, we have $q^- = 1 - q^+$, which ends the proof. □

The committor q_i^+ is related to hitting times with respect to the sets A and B by

$$q_i^+ = \mathbb{P}_i(\tau_B^+ < \tau_A^+). \tag{3.4.3.8}$$

Here $\mathbb{P}_i$ denotes probability conditional on $X(0) = i, \tau_A^+ = \min\{t > 0 : X(t) \in A\}$ denotes the first entrance time of the set A, and $\tau_B^+ = \min\{t > 0 : X(t) \in B\}$ denotes the first entrance time of the set B; q_i^- can be defined similarly using the time-reversed process as

$$q_i^- = \tilde{\mathbb{P}}_i(\tau_B^- > \tau_A^-), \tag{3.4.3.9}$$

where $\tilde{\mathbb{P}}_i$ denotes probability with respect to the time-reversed process conditional on $\tilde{X}(0) = i, \tau_A^- = \inf\{t > 0 : \tilde{X}(t) \in A\}$ denotes the last exit time of the subset A, and $\tau_B^- = \inf\{t > 0 : \tilde{X}(t) \in B\}$ denotes the last exit time of the subset B.

We have the following theorem.

Theorem 3.4.9. *The probability distribution of reactive trajectories defined in (3.4.3.1) is given by*

$$m_i^R = \pi_i q_i^+ q_i^-, \quad i \in S. \tag{3.4.3.10}$$

Proof. Denote by $x_i^{AB,+}(t)$ the first state in $A \cup B$ reached by $X(s)$, $s \geq t$, conditional on $X(t) = i$. Similarly, denote by $x_i^{AB,-}(t)$ the last state in $A \cup B$ left by $X(s)$, $s \leq t$, conditional on $X(t) = i$, or, equivalently, the first state in $A \cup B$ reached by $\tilde{X}(s)$, $s \geq -t$. In terms of these quantities, (3.4.3.1) can be written as

$$m_i^R = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} \mathbf{1}_{\{i\}}(X(t)) \mathbf{1}_A(x_i^{AB,-}(T)) DT.$$

Taking the limit as $T \to \infty$ and using ergodicity together with the strong Markov property, we deduce that

$$m_i^R = \pi_i \mathbb{P}_i(\tau_B^+ < \tau_A^+) \tilde{\mathbb{P}}_i(\tau_B^- > \tau_A^-),$$

which is (3.4.3.10) by definition of q^+ and q^-. $\square$

Notice that $m_i^R = 0$ if $i \in A \cup B$. Notice also that m^R is not a normalized distribution. In fact, from (3.4.3.2)

$$Z_{AB} = \sum_{j \in S} m_j^R = \sum_{j \in S} \pi_j q_j^+ q_j^- < 1 \tag{3.4.3.11}$$

is the probability that the trajectory is reactive at some given instance t in time, i.e.,

$$Z_{AB} = \mathbb{P}(t \in R). \tag{3.4.3.12}$$

The distribution

$$m_i^{AB} = Z_{AB}^{-1} m_i^R = Z_{AB}^{-1} \pi_i q_i^+ q_i^- \tag{3.4.3.13}$$

is then the normalized distribution of reactive trajectories which gives the probability of observing the system in a reactive trajectory and in state i at time t conditional on the trajectory being reactive at time t.

If the Markov process is reversible (i.e., $\pi_i l_{ij} = \pi_j l_{ji}$), then $q_i^+ = 1 - q_i^-$ and the probability distribution of reactive trajectories reduces to

$$m_i^R = \pi_i q_i^+ (1 - q_i^+) \qquad (reversible process). \tag{3.4.3.14}$$

3.4.4 Probability current of reactive trajectories

In this section we are interested in the probability current of reactive trajectories, i.e., the average rate at which they flow from state i to state j. A precise definition amounts to counting how many reactive trajectories jump from state i to state j on average in a time interval of length $s > 0$ and then computing the limit as $s \to 0+$ of the ratio between this

average number and s. In formula, this reads as follows.

Definition 3.4.10. The probability current of reactive trajectories $f^{AB} = (f_{ij}^{AB})_{i,j\in S}$ is defined so that for all pairs of states (i,j), $i,j \in S$, $i \neq j$, we have

$$\lim_{s\to 0+} \frac{1}{s} \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathbf{1}_{\{i\}}(X(t))\mathbf{1}_{\{j\}}(X(t+s)) \times \sum_{n\in\mathbb{Z}} \mathbf{1}_{(-\infty,t_n^B]}(t)\mathbf{1}_{[t_n^A,\infty)}(t+s)dt = f_{ij}^{AB}. \tag{3.4.4.1}$$

In addition, we set $f_{ii}^{AB} = 0$ for all $i \in S$.

In (3.4.4.1), the factor $\sum_{n\in\mathbb{Z}} \mathbf{1}_{(-\infty,t_n^B]}(t)\mathbf{1}_{[t_n^A,\infty)}(t+s)$ is used to prune out of the time average all of the times during which $X(t)$ and $X(t+s)$ are both not reactive. It has this complicated looking form because we want the flux f_{ij}^{AB} to be nonzero even if $i \in A$: for any $i \notin A$ the pruning factor in (3.4.4.1) can be replaced by $\mathbf{1}_R(t)\mathbf{1}_R(t+s)$, but this is not adequate if $i \in A$ because $X(t_A^n) \notin A$ by construction. For $i \notin A$, f_{ij}^{AB} can be also be defined as

$$f_{ij}^{AB} = \lim_{s\to 0+} \frac{1}{s}\mathbb{P}(X(t) = i \;\&\; X(t+s) = j \;\&\; t \in R \& \; t+s \in R). \tag{3.4.4.2}$$

We have the following theorem.

Theorem 3.4.11. *The discrete probability current of reactive trajectories is given by*

$$f_{ij}^{AB} = \begin{cases} \pi_i q_i^- l_{ij} q_j^+ & if i \neq j, \\ 0 & otherwise. \end{cases} \tag{3.4.4.3}$$

Proof. Using the same notation as in the proof of Theorem 3.4.9, (3.4.4.1) can also be

written as

$$f_{ij}^{AB} = \lim_{s\to 0^+} \frac{1}{s} \lim_{T\to\infty} \frac{1}{2T} \mathbf{1}_{\{i\}}(X(t)) \mathbf{1}_{\{j\}}(X(t+s)) \\ \times \mathbf{1}_A(x_i^{AB,-}(t)) \mathbf{1}_B(x_j^{AB,+}(t+s)) dt. \tag{3.4.4.4}$$

Taking the limit $T \to \infty$ and using ergodicity, we deduce that

$$f_{ij}^{AB} = \lim_{s\to 0+} \frac{1}{s} \pi_i q_i^- \mathbb{E}_i[q^+_{X(s)}, \mathbf{1}_{\{j\}}(X(s))],$$

where $\mathbb{E}_i$ denotes the expectation conditional on $X(0) = i$. To take the limit $s \to 0+$ we use

$$\forall \Phi : S \mapsto \mathbb{R} : \lim_{s\to 0+} \frac{1}{s} (\mathbb{E}_i[\Phi(X(s))] - \Phi(i)) = \sum_{j\in S} l_{ij} \Phi(j),$$

and we are done since $i \neq j$. $\square$

This result implies an expected property, namely the conservation of the discrete probability current or flux in each node.

Theorem 3.4.12. *For all $i \in (A \cup B)^c$ the probability current is conserved, i.e.,*

$$\sum_{j\in S} (f_{ij}^{AB} - f_{ji}^{AB}) = 0 \quad \forall i \in (A \cup B)^c. \tag{3.4.4.5}$$

Proof. By the definition of f^{AB} for $i \in (A \cup B)^c$,

$$\begin{aligned}\sum_{j\in S}(f_{ij}^{AB} = f_{ji}^{AB}) &= \pi_i q_i^- \sum_{j\neq i} l_{ij} q_j^+ = \pi_i q_i^+ \sum_{j\neq i} \frac{\pi_j}{\pi_i} l_{ji} q_j^- \\ &= -q_i^- q_i^+ \pi_i l_{ii} + q_i^- q_i^+ \pi_i \tilde{l}_{ii} \\ &= 0,\end{aligned}$$

where we used $\sum_{j\in S} l_{ij} q_j^+ = 0$ if $i \in (A \cup B)^c$ from (3.4.3.3) and $\sum_{j\in S} \tilde{l}_{ij} q_j^- = 0$ if $i \in (A \cup B)^c$ from (3.4.3.4). □

For later use we should also mention that conservation of the current in every state $i \in (A \cup B)^c$ immediately implies the following total conservation of the current:

$$\sum_{i\in A, j\in S} f_{ij}^{AB} = \sum_{j\in S, i\in B} f_{ji}^{AB}, \tag{3.4.4.6}$$

where we used that $f_{ij}^{AB} = 0$ if $i \in S$ and $j \in A$, and $f_{ij}^{AB} = 0$ if $i \in B$ and $j \in S$.

3.4.5 Transition rate and effective current

In this section we derive the average number of transitions from A to B per time unit or, equivalently, the average number of reactive trajectories observed per time unit. More precisely, let $N_T^-, N_T^+ \in \mathbb{Z}$ be such that

$$R \cap [-T, T] = \bigcup_{N_T^- \leq n \leq N_T^+} (t_n^A, t_n^B); \tag{3.4.5.1}$$

that is, $N_T^+ - N_T^-$ is the number of reactive trajectories in the interval $[-T, T]$ in time. Then we have the following definition.

Definition 3.4.13. The transition rate k_{AB} is defined as

$$k_{AB} = \lim_{T\to\infty} \frac{N_T^+ - N_T^-}{2T}. \tag{3.4.5.2}$$

We have the following theorem.

Theorem 3.4.14. *The transition rate is given by*

$$k_{AB} = \sum_{i\in A, j\in S} f_{ij}^{AB} = \sum_{j\in S, k\in B} f_{jk}^{AB}. \tag{3.4.5.3}$$

Proof. From (3.4.4.4) we get

$$\sum_{i\in A, j\in S} f_{ij}^{AB} = \lim_{s\to 0+} \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathbf{1}_A(X(t)) \sum_{j\in S} \mathbf{1}_B(x_j^{AB,+}(t+s))dt. \tag{3.4.5.4}$$

Let us consider the integral; we can always restrict our attention to generic values of T such that there is no $n \in \mathbb{Z}$ for which $T = t_n^A$ or $T = t_n^B$. The integrand in this expression is nonzero if and only if $X(t) \in A$, $X(t+s) \in A^c$ and $t + s \in R$, i.e., if $t_n^A \in (t, t+s)$ for some $n \in \mathbb{Z}$. But this means that the integral of $\mathbf{1}_A(X(t))\mathbf{1}_B(x_j^{AB,+}(t+s))$ on every interval $t \in (t_n^A - s, t_n^A)$ is equal to s and the only contributions to the integral in (3.4.5.4) come from the intervals in $[-T, T] \cap \cup_{n\in\mathbb{Z}}(t_n^A - s, t_n^A)$. But these are exactly $N_T^+ - N_T^-$ intervals such that the whole integral amounts to $(N_T^+ - N_T^-)s$. From (3.4.5.4) and (3.4.5.1), this implies the first identity for the rate k_{AB} . The second identity follows from (3.4.4.6). □

Notice that the rate can also be expressed as

$$k_{AB} = \sum_{i \in A, j \in S} f_{ij}^{+}, \tag{3.4.5.5}$$

where we have the following definition.

Definition 3.4.15. The effective current is defined as

$$f_{ij}^{+} = \max(f_{ij}^{AB} - f_{ji}^{AB}, 0). \tag{3.4.5.6}$$

Identity (3.4.5.5) follows from (3.4.5.3) and the fact that for all $i \in A : f_{ij}^{+} = f_{ij}^{AB}$ since $f_{ji}^{AB} = 0$ and $f_{ij}^{AB} > 0$ if $i \in A$. The effective current gives the net average number of reactive trajectories per time unit making a transition from i to j on their way from A to B. The effective current will be useful to define transition pathways in section 3.4.7.

Theorem 3.4.16. *If the Markov process is reversible, then the effective current reduces to*

$$f_{ij}^{+} = \begin{cases} \pi_i l_{ij}(q_j^{+} - q_i^{+}) & if q_j^{+} > q_i^{+}, \\ 0 & otherwise \end{cases} \quad (reversible\ process), \tag{3.4.5.7}$$

and the reaction rate can be expressed as

$$k_{AB} = \frac{1}{2} \sum_{i,j \in S} \pi_i l_{ij}(q_j^{+} - q_i^{+})^2, \quad (reversible\ process). \tag{3.4.5.8}$$

The last identity can also be written as $k_{AB} = -\sum_{i \in S, j \in B} \pi_i l_{ij} q_i^{+}$ *(for reversible processes!),*

which in turn is identical to the expression that we know from Theorem 3.4.14:

$$k_{AB} = \sum_{i \in S, j \in B, i \neq j} \pi_i l_{ij} (1 - q_i^+), \quad (reversible process).$$

3.4.6 Relations with electrical resistor networks

Before proceeding further, it is interesting to revisit our result in the context of electrical resistor networks [58]. Recall that an electrical resistor network is a directed weighted graph $G(S, E) = G\{C\}$, where $C = (c_{ij})$ is an entrywise nonnegative symmetric matrix (see Definition 3.4.3), called the conductance matrix of G. The reciprocal r_{ij} of the conductance c_{ij} is called the resistance of the edge (i, j). Establishing a voltage $v_a = 0$ and $v_b = 1$ between two vertices a and b induces a voltage $v = (v_i)_{i \in S \setminus \{a,b\}}$ and an electrical current F_{ij} which are related by Ohm's law:

$$F_{ij} = \frac{v_i - v_j}{r_{ij}} = (v_i - v_j) c_{ij}, \quad i, j \in S, i \neq j. \tag{3.4.6.1}$$

Furthermore, Kirchhoff's current law, that is,

$$\sum_{j \in S} F_{ij} = 0 \quad \forall i \in S \setminus \{a, b\}, \tag{3.4.6.2}$$

requires that the voltages have the property

$$v_i = \sum_{j \neq i} \frac{c_{ij}}{c_i} v_j \quad \forall i \in S \setminus \{a, b\}, \tag{3.4.6.3}$$

where $c_i = \sum_{j \neq i} c_{ij}$. A reversible Markov jump process, given by its infinitesimal generator L, can be seen as an electrical resistor network by setting up the conductance matrix C via

$$c_{ij} = \pi_i l_{ij} \quad (j \neq i), \tag{3.4.6.4}$$

where $\pi = (\pi_i)_{i \in S}$ is the unique stationary distribution. Now observe that (3.4.6.3) reduces to

$$0 = \sum_{j \in S} l_{ij} v_j \quad \forall i \in S \setminus \{a, b\}.$$

But this means that the forward committor q^+ with respect to the sets $A = \{a\}$ and $B = \{b\}$ can be interpreted as a voltage (see (3.4.3.3)). Moreover, a short calculation shows that the effective flux, defined in (3.3.0.31), pertains to the electrical current.

3.4.7 Dynamical bottlenecks and reaction pathways

The transition rate k_{AB} is a quantity which is important to describe the global transition behavior. In this section we characterize the local bottlenecks of the ensemble of reactive trajectories which determine the transition rate. In order to get a detailed insight into the local transition behavior we characterize reaction pathways by looking at the amount of reactive trajectories which is conducted from A to B by a sequence of states.

We use the notation of graph theory introduced at the end of section 3.3. Let $G(S, E) = G\{f^+\}$ be the weight induced directed graph associated with the effective current $f^+ = (f_{ij}^+), i_j \in S$. A simple pathway in the graph G, starting in $A \subset S$ and ending in $B \in S$, is the natural choice for representing a specific reaction from A to B because any loop during a transition would be redundant with respect to the progress of the reaction.

Definition 3.4.17. A reaction pathway $w = (i_0, i_1, \cdots, i_n), i_j \in S, j = 0, \cdots, n$ from A to B is a simple pathway such that

$$i_0 \in A, i_n \in B, i_j \in (A \cup B)^c, \quad j = 1, \cdots, n-1.$$

The crucial observation which leads to a characterization of bottlenecks of reaction pathways is that the amount of reactive trajectories which can be conducted by a reaction pathway per time unit is confined by the minimal effective current of a transition involved along the reaction pathway.

Definition 3.4.18. let $w = (i_0, i_1, \cdots, i_n)$ be a reaction pathway in $G\{f^+\}$. We define the min-current of w by

$$c(w) = \min_{e=(i,j)\in w} \{f_{ij}^+\}. \tag{3.4.7.1}$$

The dynamical bottleneck of a reaction pathway is the edge with the minimal effective current

$$(b_1, b_2) = \arg\min_{e=(i,j)\in w} \{f_{ij}^+\}. \tag{3.4.7.2}$$

We call such an edge (b_1, b_2) a bottleneck.

Here and in the following we somewhat misuse our notation by writing $e = (i, j) \in w$ whenever the edge e is involved in the pathway $w = (i_0, i_1, \cdots, i_n)$, i.e., if there is an $m \in \{0, \cdots, n-1\}$ such that $(i, j) = (i_m, i_{m+1})$.

Now it is straightforward to characterize the "best" reaction pathway, that is, the one with the maximal min-current.

Notice that the problem of finding a pathway which maximizes the minimal current is known as the maximum capacity augmenting path problem [115] in the context of solving the

maximal flow problem in a network.

In general, one cannot expect to find a unique "best" reaction pathway because the bottleneck corresponding to the maximal min-current could be the bottleneck of other reaction pathways too.

Definition 3.4.19. Let $\mathbf{W}$ be the set of all reaction pathways and denote the maximal min-current by c_{max}. Then we define the set of the dominant reaction pathways $\mathcal{W}_{\mathcal{D}} \subseteq \mathbf{W}$ by

$$\mathcal{W}_{\mathcal{D}} = \{w \in \mathbf{W} : c(w) = c_{max}\}.$$

To guarantee uniqueness of the bottleneck, we henceforth assume that the nonvanishing effective currents are pairwise different, i.e., $f_e^+ \neq f_{e'}^+$ for all pairs of edges $e = (i, j)$, $e = (i', j')$ with $f_e^+, f_{e'}^+ > 0$. Nevertheless, we are aware that in applications the situation could show up where more than one bottleneck exists because the corresponding currents are more or less equal. This ambiguity is taken into account in an ordered decomposition of the set of all reaction pathways described at the end of this section.

Let $G[\mathcal{W}_{\mathcal{D}}] = G(S_{\mathcal{D}}, E_{\mathcal{D}})$ be the directed graph induced by the set $\mathcal{W}_{\mathcal{D}}$, i.e., the graph whose vertex/edge set is composed of all vertices/edges that appear in at least one of the pathways in $\mathcal{W}_{\mathcal{D}}$. The next lemma shows that the graph $G[\mathcal{W}_{\mathcal{D}}] = G(S_{\mathcal{D}}, E_{\mathcal{D}})$ possesses a special structure which is crucial for the definition of a representative dominant reaction pathway.

Let $b = (b_1, b_2)$ denote the unique bottleneck in $\mathcal{W}_{\mathcal{D}}$. Then the graph $G(S_{\mathcal{D}}, E_{\mathcal{D}} \setminus \{b\})$ decomposes into two disconnected parts $G[\mathcal{L}]$ and $G[\mathcal{R}]$ such that every reaction pathway $w \in \mathcal{W}_{\mathcal{D}}$ can be decomposed into two pathways $w_{\mathcal{L}}$ and $w_{\mathcal{R}}$,

$$w = (\underbrace{i_{l_1}, \cdots, i_{l_n} = b_1}_{=w_{\mathcal{L}}}, \underbrace{b_2 = i_{r_1}, \cdots, i_{r_m}}_{=w_{\mathcal{R}}}), \tag{3.4.7.3}$$

where $w_{\mathcal{L}} \in \mathcal{L}$ is a simple pathway in $G[\mathcal{L}]$ starting in $i_{l_1} \in A$ and ending in $\{b_1\}$ and $w_{\mathcal{R}} \in \mathcal{R}$ is a simple pathway in $G[\mathcal{R}]$ starting in $\{b_2\}$ and ending up in $i_{r_m} \in B$. Whenever we have $\mathcal{L} = \emptyset$, i.e., $(b_1 \in A)$, then $G[\mathcal{L}] = (\{i_{l_1}\}, \emptyset)$; if $\mathcal{R} = \emptyset$, then $G[\mathcal{R}]$ is defined likewise.

Here and in the following we write $w_{\mathcal{L}} \in \mathcal{L}$ (and $w_{\mathcal{R}} \in \mathcal{R}$, respectively) if we want to ex-

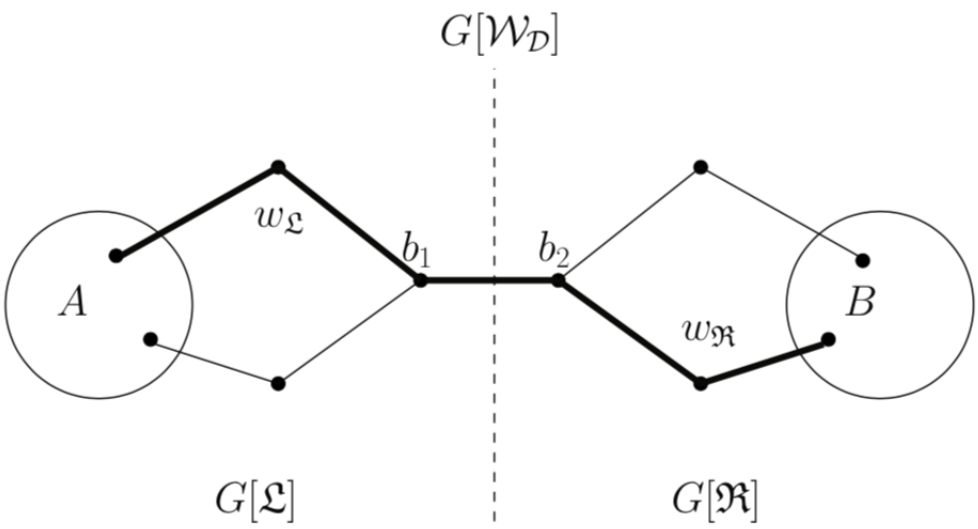


Figure 2: Schematic representation of the decomposition of $\mathcal{W}_\mathcal{D}$. A reaction pathway w (shown in thick black) can be decomposed into two simple pathways $w_\mathcal{L}$ and $w_\mathcal{R}$.

press that for every edge $e \in w_\mathcal{L}$ we have $e \in \mathcal{L}$. The lemma expresses the natural property that the graph $G[\mathcal{W}_\mathcal{D}] = G(S_\mathcal{D}, E_\mathcal{D})$ can be decomposed into two disconnected graphs by removing the bottleneck; see Figure 2 for a schematic illustration.

Proof. It immediately follows from the definition of $\mathcal{W}_\mathcal{D}$ that the bottleneck b is involved in every dominant reaction pathway because otherwise there would exist a pathway $w \in \mathcal{W}_\mathcal{D}$ such that $c(w) > c_{max}$, which leads to a contradiction. By definition, a reaction pathway does not possess any loops. Consequently, the bottleneck b separates $\mathcal{W}_\mathcal{D}$, which proves the assertion. □

According to the lemma, the set of dominant reaction pathways $\mathcal{W}_{\mathcal{D}}$ can be represented as

$$\mathcal{W}_{\mathcal{D}} = \mathcal{L} \times \mathcal{R} := \{(w_{\mathcal{L}}, w_{\mathcal{R}}) : w_{\mathcal{L}} \in \mathcal{L}, w_{\mathcal{R}} \in \mathcal{R}\}. \tag{3.4.7.4}$$

In Figure 2 we give a schematic representation of the decomposition of $\mathcal{W}_{\mathcal{D}}$.
Next, we address the most likely case in applications where more than one dominant reaction pathway exists. By definition, each dominant reaction pathway conducts the same amount of current from A to B, but they could differ, e.g., with respect to the maximal amount of current which they conduct from the set A to the bottleneck, respectively. Now observe that the simple pathways in the set $\mathcal{L}$ could be seen as reaction pathways with respect to the set A and the B-set $\{b_1\}$. Hence, $\mathcal{L}$ itself again possesses a set of dominant reaction pathways $\mathcal{W}_{\mathcal{D}}(\mathcal{L})$, and so on. This motivates the following recursive definition of the representative dominant reaction pathway.

Definition 3.4.20. Let $\mathcal{W}_{\mathcal{D}} = \mathcal{L} \times \mathcal{R}$ and suppose $b = (b_1, b_2)$ is its (unique) bottleneck. Then we define the representative dominant reaction pathway w^* of $\mathcal{W}_{\mathcal{D}}$ by

$$w^* = (w^*_{\mathcal{L}}, w^*_{\mathcal{R}}), \tag{3.4.7.5}$$

where $w^*_{\mathcal{L}}$ is the representative dominant pathway of the set $\mathcal{W}_{\mathcal{D}}\mathcal{L})$ with respect to the set A and the B-set $\{b_1\}$ and $w^*_{\mathcal{R}}$ is the representative of $\mathcal{W}_{\mathcal{D}}\mathcal{R})$ with respect to the A-set $\{b_2\}$ and the set B. If $\mathcal{L} = \emptyset$ and $G[\mathcal{L}] = (\{i\}, \emptyset)$, then $w^*_{\mathcal{L}} = \{i\}$; if $\mathcal{R} = \emptyset$, then $w^*_{\mathcal{R}}$ is defined likewise.

Notice that the representative w^* is unique under the assumption made in Definition

3.4.20. Furthermore, it follows immediately from the recursive definition of w^* that

$$\begin{aligned} w^* &= \arg\max_{w \in \mathcal{W}_{\mathcal{D}}} \min_{e=(i,j)\in w,(i,j)\neq(b_1,b_2)} \{f_{ij}^+\} \\ &= \arg\max_{w \in \mathcal{W}_{\mathcal{D}}} \min_{e=(i,j)\in w,(i,j)\neq(b_1,b_2)} \{f_{ij}^+ - c_{max}\}. \end{aligned} \tag{3.4.7.6}$$

Finally, we turn our attention to the residuum current which results from updating the effective current of each edge along the representative pathway $w_1^* = w^*$ by subtracting the min-current $c_{max}^{(1)} = c_{max}$. That is, the residuum current is defined as

$$f_{ij}^{r,1} = \begin{cases} f_{ij}^+ - c_{max}^{(1)} & if\ (i,j) \in w_1^*, \\ f_{ij}^+ & otherwise. \end{cases} \tag{3.4.7.7}$$

The graph $G_1 = G\{f_{ij}^{r,1}\}$ induced by the residuum current satisfies the current conservation property in analogy to (3.4.4.5). It again possesses a bottleneck, say $\tilde{b}$, a set of dominant pathways, and a representative pathway, say w_2^*. If we denote the min-current of w_2^* with respect to the residuum current by c_{max}, then it should be clear that $c_{max} = c_{max}^{(1)} > c_{max}^{(2)}$ holds. The property (3.4.7.6) of w_1^* guarantees that $c_{max}^{(2)}$ is maximal with respect to all possible residuum currents. We can obviously repeat this procedure by introducing the residuum current $f_{ij}^{r,2}$ by subtracting c_{max} from $f_{ij}^{r,1}$ along the edges belonging to w_2^*, and so on. The resulting iteration terminates when the resulting induced graph $G_{M+1} = G\{f_{ij}^{r,M+1}\}$ no longer contains reaction pathways and leads to an ordered enumeration $(w_1^*, w_2^*, \cdots, w_M^*)$ of

the set $\mathbf{W}$ of all reaction pathways such that

$$c_{max}^{(i)} > c_{max}^{(j)}, \quad 0 \leq i < j \leq M,$$
$$\sum_{i=1}^{M} c_{max}^{(i)} = k_{AB},$$

where the last identity simply follows from the following equation for the rates $k_{AB}(G_i)$ associated with the graphs $G_1, \cdots, G_M$:

$$k_{AB}(G_i) = k_{AB}(G_{i-1}) - c_{max}^{(i)},$$

where G_0 denotes the original graph $\{f_{ij}^{+}\}$, and $k_{AB}(G_{M+1}) = 0$.

The composition of the total rate into fraction coming from currents along reactive pathways is quite a general concept in graph theory. We herein just presented a specification of it. We refer the interested reader to, e.g., [[115], section 3.5].

3.4.8 Relation with Laplacian eigenmaps and diffusion maps

Let us quickly discuss the connection of our results in the context of data analysis (in particular, data segmentation and embedding, i.e., low-dimensional representation). Lately, two classes of methods have been introduced to this aim: Laplacian eigenmaps [45, 51, 72, 49, 12] and diffusion maps [63, 70]. The concept behind these approaches is pretty simple. Given a set of data points, say $S = \{x_1, x_2, \cdots, x_n\}$, one associates a weight induced graph with weight function $w(x, y)$. This graph is constructed locally, e.g., by joining all points with equal weights that are below a cut-off distance from each other. These weights are then renormalized by the degree of each node, which means that $w(x, y)$ can be reinterpreted as

the stochastic matrix of a discrete-time Markov chain. Alternatively, it is also reasonable to interpret the weights as rates and thereby build the generator of a continuous-time Markov chain. In both cases, the properties of the chain are then investigated via spectral analysis of the stochastic matrix or the generator. In particular, the first N eigenvectors with leading eigenvalues, say, $\phi_j(x), j = 1, \cdots, N$, can be used to embed the chain into $\mathbb{R}^N$ via $x \mapsto (\phi_1(x), \cdots, \phi_N(x))$. The eigenvectors can also be applied to segment the original data set into principal components (segmentation).

As explained in the introduction, the spectral approach is particularly important if the Markov chain displays metastability, i.e., if there exist one or more clusters of eigenvalues which are either very close to 1 (in the case of discrete-time Markov chains) or 0 (in the case of continuous-time Markov chains). When the chain is not metastable, however, the significance of the first few eigenvectors is less clear, which makes the spectral approach less appealing. In these situations, TPT may provide an appealing option. For example, if several points (or groups of points) with some specific features can be singled out in the data set, then, by analyzing the reaction between pairs of such groups, one will reveal global information about the data set (for instance, the committor functions between these pairs may be applied for embedding instead of the eigenvectors). The current of reactive trajectories and dominant reaction pathways will also provide supplementary information about the global structure of the data set which is not considered in the spectral approach. In this thesis, we will not, however, develop these ideas any further.

3.5 Algorithmic aspects

Now let's explain the algorithmic details for the computation of the various quantities in TPT. Given the generator L and the two sets A and B, the stationary distribution $\pi = (\pi_i)_{i\in S}$ is computed by solving (3.4.1.2), whereas the discrete forward and discrete backward committors, $q^+ = (q_i^+)_{i\in S}$ and $q^- = (q_i^-)_{i\in S}$, are computed by solving (3.4.3.3) and (3.4.3.4). Solving these equations numerically can be done using any standard linear algebra package. These objects allow one to compute the probability distribution of reactive trajectories $m^R = (m_i^R)_{i\in S}$ in (3.4.3.10), its normalized version $m^{AB} = (m_i^{AB})_{i\in S}$ in (3.4.3.13), the probability current of reactive trajectories $f^{AB} = (f_{ij}^{AB})_{i,j\in S}$ in (3.4.4.3), and the effective current $f^+ = (f_{ij}^+)_{i,j\in S}$ in (3.4.5.6). This also gives the reaction rate k_{AB} via (3.3.0.28) or (3.3.0.30).

Next, we focus on the computation of the bottlenecks and representative dominant reaction pathways which is less standard.

3.5.1 Computation of dynamical bottlenecks and representative dominant reaction pathways

From the definition in (3.4.7.2) of the bottleneck $b = (b_1, b_2)$ associated with the set of dominant reaction pathways $\mathcal{W}_\mathcal{D}$, it follows that

$$f_c^+ > f_b^+ \quad \forall c \in E_\mathcal{D},\ c \neq b,$$

where $f^+ = (f_{ij}^+)_{i,j\in S}$ is the effective current and $E_\mathcal{D}$ is the edge set of the induced graph $G = G[\mathcal{W}_\mathcal{D}]$. This observation leads to a characterization of the bottleneck which is algo-

Algorithm 1: Computation of the bottleneck

Input: Graph $G = G\{f^+\}$.

Step-1: Sort edges of G according to their weights in ascending order
$\Rightarrow E_{sort} = (e_!, e_2, \cdots, e_{|E|})$
Step-2: IF the edge $e_{|E|}$ connects A and B THEN RETURN bottleneck $b := e_{|E|}$.
Step-3: Initialize $l := 1, r := |E|$.
Step-4: WHILE $r - l > 1$
Set $n := \lfloor \frac{r+l}{2} \rfloor$, $E'(m) := \{e_m, \cdots, e_{|E|}\}$.
IF there exists a reaction pathway in $G(S, E'(m))$
THEN $l := m$ ELSE $r := m$.
END WHILE
Step-5: RETURN bottleneck $b := e_l$.
Output: Bottleneck $b = (b_1, b_2)$.

rithmically more convenient. Let $E_{sort} = (e_1, e_2, ..., e_{|E|})$ be an enumeration of the set of edges of $G = G\{f^+\}$ sorted in ascending order according to their effective current. Then the edge $b = e_m$ in E_{sort} is the bottleneck if and only if the graph $G(S, \{e_m, \cdots, e_{|E|}\})$ contains a reaction pathway but the graph $G(S, \{e_{m+1}, \cdots, e_{|E|}\})$ does not. The bisection algorithm stated in Algorithm 1 is a direct consequence of this alternative characterization of the bottleneck and is related to the capacity scaling algorithm [[115], section 7.3] for solving the maximum flow algorithm. For an alternative algorithm in the context of distributed computing which is based on a modified Dijkstra algorithm; see [4].

We also have the following lemma.

Lemma 3.5.1. The computational cost of Algorithm 1 in the worst case is $\mathcal{O}(n \log n)$, where $n = |E|$ denotes the number of edges of the graph $G = G\{f^+\}$.

Proof. Assume that $n = 2^k$, $k > 1$. First, notice that the sorting of the edges of $G = G\{f^+\}$ can be performed in $\mathcal{O}(n \log n)$. In the worst case scenario, the edge $e_1 \in E_{Sort}$ is the bottleneck. When this is the case, the number of edges in the jth repetition of the while-

loop would be

$$\frac{n}{2^j},$$

and we would have $k-1$ repetitions. The cheapest way to determine whether there exists a reactive trajectory is to perform a breadth-first search starting in A; the computational cost of that step depends only linearly on the number of edges to be considered, such that we deduce for the worst case effort $T(n)$ of the entire procedure

$$\begin{aligned} T(n) &= \mathcal{O}(kn) + \mathcal{O}\Big(\frac{n}{2}\Big) + \mathcal{O}\Big(\frac{n}{4}\Big) + \cdots + \mathcal{O}\Big(\frac{n}{2^{k-1}}\Big) \\ &= \mathcal{O}\Big(kn + n\Big(\frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{k-1}}\Big)\Big) \\ &= \mathcal{O}(kn), \end{aligned}$$

which by noting that $k = \log(n)$ ends the proof. □

The algorithm for computing the unique representative pathway w^* of the set of dominant reaction pathways is a direct implementation of the recursive definition of w^* given in (3.4.7.5). Recalling that $\mathcal{W}_{\mathcal{D}}$ can be decomposed as stated in (3.4.7.4) and assuming that f^+ takes different values for every edge (i, j), we end up with Algorithm 2. A rough estimation of the computational cost of this algorithm is $\mathcal{O}(mn \log n)$, where m is the number of edges of the resulting representative pathway w^* and $n = |E|$.

3.5.2 Topological Sorting Algorithm for Transition Pathways

There are two steps to find the path. The first is to sort the edges and compute the bottleneck, with the computational cost of $\mathcal{O}(n \log n)$, where $n = |E|$ denotes the number of edges of

Algorithm 2: Representative pathways

Input: Graph $G = G\{f^+\}$.

Step-1: Determine bottleneck $b = (b_1, b_2)$ in G via Algorithm 1.
Step-2: Determine decomposition $\mathcal{W}_{\mathcal{D}}(G) = \mathcal{L} \times \mathcal{R}$.
Step-3:

$$\text{Set } w^*_{\mathcal{L}} := \begin{cases} b_1 & \text{if } b_1 \in A, \\ \text{result of the recursion with } (G[\mathcal{L}], A, \{b_1\}) & \text{if } b_1 \notin A. \end{cases}$$

Step-4:

$$\text{Set } w^*_{\mathcal{R}} := \begin{cases} b_2 & \text{if } b_2 \in B, \\ \text{result of the recursion with } (G[\mathcal{R}], \{b_2\}, B) & \text{if } b_2 \notin B. \end{cases}$$

Step-5: RETURN bottleneck $(w^*_{\mathcal{L}}, w^*_{\mathcal{R}})$.
Output: Representative $w^* = (w^*_{\mathcal{L}}, w^*_{\mathcal{R}})$ of $\mathcal{W}_{\mathcal{D}}(G)$.

the graph $G = G\{f^+\}$. The second step is to compute the representative pathways, and the computational cost is $\mathcal{O}(mn \log n)$, where m is the number of edges of the resulting representative pathway.

An alternative approach to find dominant pathways is using topological sort. Topological sort of a directed acyclic graph $G = (V, E)$ is a linear ordering of its nodes such that if G has edge (u, v), then u appears before v in the ordering, it can be viewed as an ordering of its nodes along a horizontal line so that all directed edges go from left to right.

For example, the topological sorting for the following graph is (5,4,2,3,1,0). There can be more than one topological sorting for a graph. For example, another topological sorting of the following graph is (4 5 2 3 1 0). The first vertex in topological sorting is always a vertex with in-degree as 0 (a vertex with no in-coming edges). A topological sorting can be done in different ways, one way is similar to deep first search (DFS) with a stack to store the nodes, another way is called Kahn's algorithm, this approach is based on the below fact :

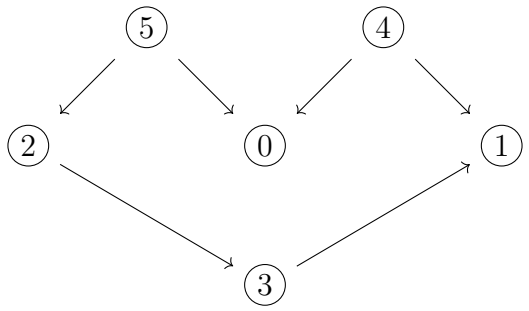


Figure 3: topological sort example

A DAG G has at least one vertex with in-degree 0 and one vertex with out-degree 0.

Proof: Since DAG does not contain a cycle, then all paths will be finite length, now let S be the longest path from s(source) to t(sink), since S is the longest path, then there is no incoming edge to s and outgoing edge from t, otherwise it contradict the truth that S is the longest path, hence we proved that $indegree(s) = 0$ and $outdegree(t) = 0$.

Algorithm 3: Topological sorting

Input: Directed Acyclic Graph (DGA)

Step-1: Compute in-degree (number of incoming edges) for each of the vertex present in the DAG and initialize an empty list L to store the nodes, in-degree computing can be done by traversing the array of edges and simply increase the counter of the destination node by 1.
Step-2: Pick all the vertices with in-degree as 0 and add them into a queue (Enqueue operation), in our effective probability current situation is just the source states.
Step-3: Remove a vertex from the queue (Dequeue operation) and then.

1 Append this vertex to L

2 Decrease in-degree by 1 for all its neighboring nodes.

3 If in-degree of a neighboring nodes is reduced to zero, then add it to the queue.

Step-4: Repeat Step 3 until the queue is empty.
Output: L: horizontal representation of the nodes

The topological sorting algorithm basically traverse the nodes and edges linearly, so the overall time complexity is $\mathcal{O}(|E| + |V|)$.

3.5.3 Representative Transition Pathway Finding

After finding topological order, the algorithm process all vertices and for every vertex, it runs a loop for all adjacent vertices. The outer for-loop runs $|V|$ iterations, but regardless of outer for-loop, the inner for-loop runs a total of $|E|$ iterations since each edge is "traversed" from left to right exactly once, and each iteration on inner loop takes $\mathcal{O}(1)$ time, total adjacent vertices in a graph is $\mathcal{O}(E)$. So the nested loop runs $\mathcal{O}(V + E)$ times. Therefore, overall time complexity of this algorithm is $\mathcal{O}(V + E)$. In the chemical reaction networks case, the number of edges relates to the number of nodes and the dimensions(number of reaction equations), actually usually $|E| = \mathcal{O}(|V|)$, so comparing to $\mathcal{O}(mn \log n)$ introduced in the paper, this algorithm improved the computation to $\mathcal{O}(n)$. For high dimension and large space, it will help the computation of the transition pathways in less time complexity.

Algorithm 4: Representative transition pathways finding

Input: Topological sorting of the DGA

Step-1: Initialize the value(maxflow) at all nodes except source nodes (states A) to be zero;Initial path at all nodes to be empty;

Step-2: For every node u in topological order:
For every adjacent node v of u:
If maxflow(v)< min(maxflow(u),f_{uv}^{+}):
maxflow(v)=min(maxflow(u),f_{uv}^{+}) # update value
path(v)=path(u)+v #update path

Output: path(t), the representative reaction pathway

3.5.4 Correctness of the algorithm

Now we will give the proof of the correctness of the algorithm, we should prove if by induction.

Define:

$$\delta(s,v) = \begin{cases} \max\{\min_{(i,j)\in P} f^+_{i,j} | s \xrightarrow{P} v\} & \text{if path of s to v exists} \\ 0 & otherwise. \end{cases}$$

Let $P = v_1 \to v_2 \to v_3 \to \cdots \to v_m$ be a path, the nodes in $\{v_1, v_2, \cdots, v_{j-1}\}$ are called the predecessors of v_j in P.

Define the dominant path P from s to v such that all predecessors of v are in S as the dominant path from s to v respect to S.

Let $(v_1, v_2, \cdots, v_n)$ be the list of nodes in topological sorted order with $n = |V|$, and $v_1 = s$, $v_n = t$. Right after $j-th$ iteration of the outer for-loop of DAG-maximum-path, $d[v_k], k > j$ is the value of node k gives the current of the maximum path from s to v_k respect to S. Furthermore, $d[v_{j+1}] = \delta(s, v_{j+1})$.

Proof:

At the beginning, $S = \{s\}$, and $d[v_k] = f^+_{s,k}$, clearly it's true.

Now suppose above statement is true for $j = m < n - 1$. Consider $j = m + 1$, in the $(m+1) - th$ iteration, v_{m+1} is included into S and update was done to every adjacent node u of v_{m+1}, which is to the right of v_{m+1}.

If v_{m+1} is not reachable from s, then u is also not reachable from s, so $d[u]$ remains 0.

If v_{m+1} is reachable from s, if u is also reachable from s, then $d[u]$ is updated or not, base on whether $d[u]$ is smaller than $min(d[v_{m+1}], f^+_{v_{m+1},u})$. Thus this new $d[u]$ is the current of the maximum path from s to u with respect to new $S = \{v_1, v_2, \cdots, v_{m+1}\}$.

Furthermore, if $u = v_{m+2}$, then $d[u] = \delta(s, u)$, because there is no other node in $\{v_{m+3}, v_{m+4}, \cdots, v_n\}$

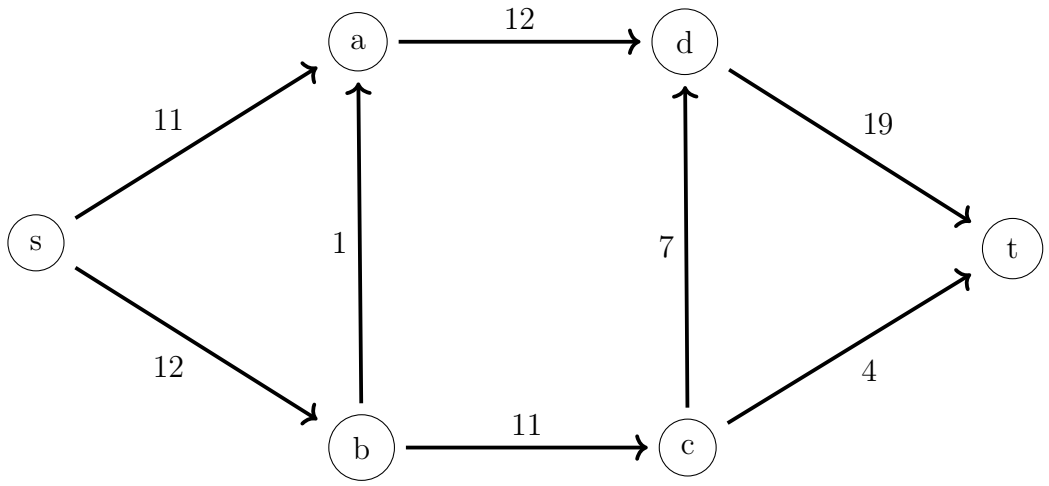


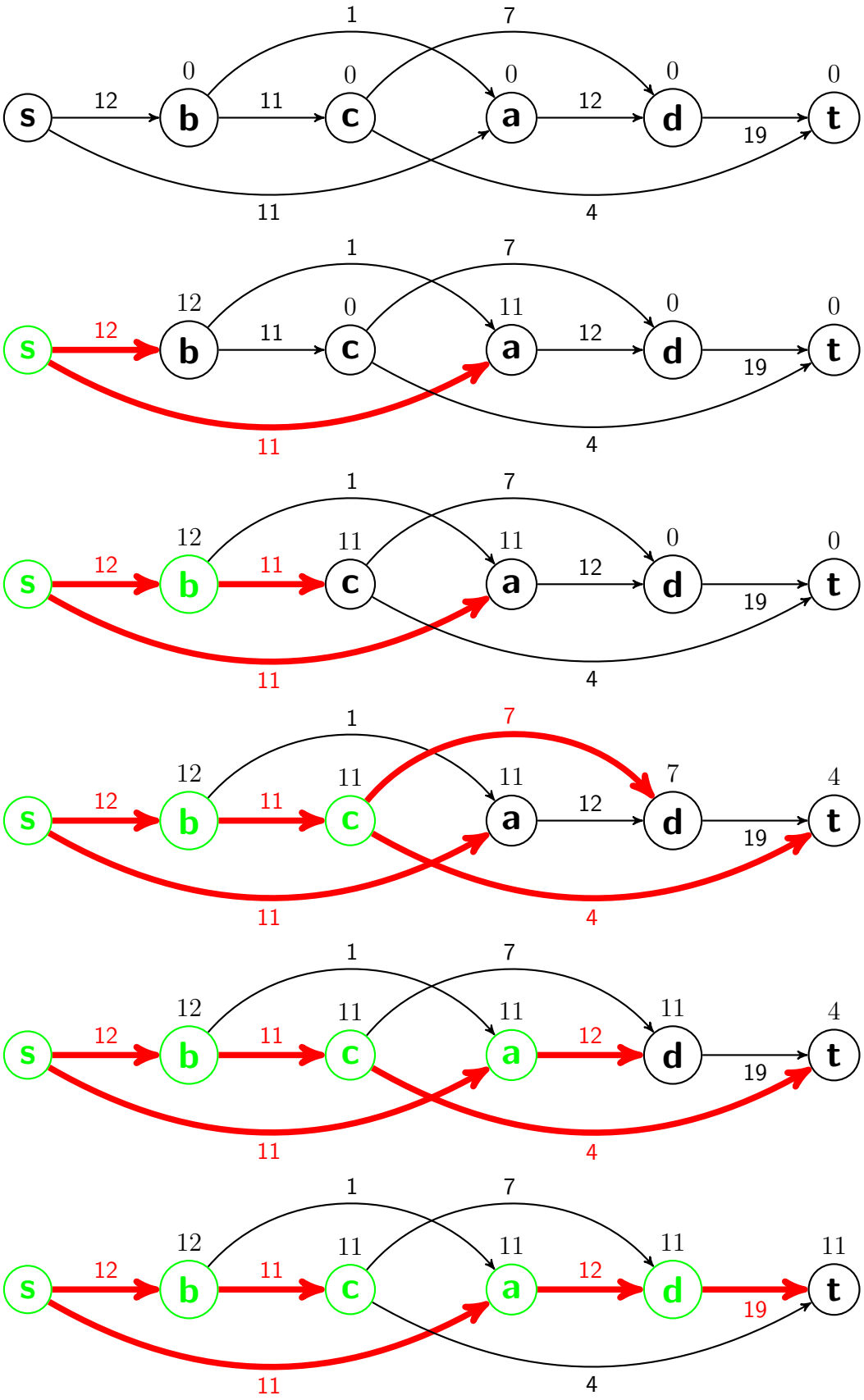
Figure 4: network flow with s as source and t as sink, values along the edges are the capacities, corresponding to the effective probability current.

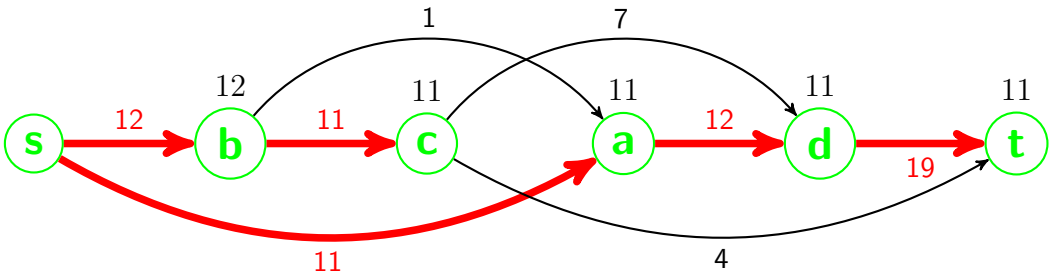
from which v_{m+2} can be reached (by the topological order of nodes).

Hence, the it is true for $j = m + 1$. After $n - 1$ outer iterations, $d[u] = \delta(s, u)$ for all nodes, also include the last node $v_n = t$, so actually the $n - th$ iteration is redundant.

3.5.5 Illustrative example

Below is a simplest example to illustrate the process to find the transition pathway. The artificial example contains source as node s, sink as node t, and the other four intermediate nodes. It is clear to see that path $s \to a \to d \to t$ is the maximum path with current 11, now let's use our algorithm to find this. First step is using algorithm 2 to transfer the network flow into topological sort graph, the following steps go through the nodes from the left to right (visited nodes are marked green), each time we update the maximum value at the adjacent nodes (values above each node) and their corresponding maximum path (marked as red bolder arrows), when we finally reach the sink node t, we also found the representative pathways.


s
b
c
a
d
t
12
11
1
7
12
19
11
4
0
0
0
0
0
0
12
11
1
7
12
19
11
4
12
0
11
0
0
0
12
11
1
7
12
19
11
4
12
11
11
0
0
0
12
11
1
7
12
19
11
4
12
11
11
7
4
0
12
11
1
7
12
19
11
4
12
11
11
11
4
0
12
11
1
7
12
19
11
4
12
11
11
11
4
11
12
11
1
7
12
19
11
4
12
11
11
11
11
11



From above example, we can see, this algorithm doesn't assume the uniqueness of the bottleneck, in case that even if the bottleneck(dominant path) is not unique, in step-2, if maxflow(v)=min(maxflow(u),f^+_{uv}), we can keep both paths in path(v), then we found all representative pathways no matter the uniqueness of the bottleneck. We can also generate the algorithm if people want to find the top k dominate transition pathways, the modification could be done by store k tuples with the k largest values and their corresponding paths at each node.

Chapter 4

Extension of TPT to transition state

In this section we will introduce the definition of probability currents on sub-networks of the system, which will give the impact of a sub-network on the reaction pathways, and can be used to identify the Transition States (TS).

4.1 Probability current of sub-networks

We will first define the probability current on a node instead of an edge, and then generalize it to a sub-network. To this end, we consider the directed network $G(S, E)$ as defined in section 3.4.7 through the effective probability current 3.4.5.6. By the definition of probability current on edges 3.4.4.1, it's the average rate at which the reactive trajectories flow from state i to state j. And given any node in the graph, it is guaranteed that the total amount inflow is equal to total amount of outflow:

$$\sum_{k:(k,i)\in E} f_{ki}^{+} = \sum_{j:(i,j)\in E} f_{ij}^{+}$$

Based on this observation, we can define the current at each node by the amount of the flow pass through it:

Definition 4.1.1. The effective current for a node i in the state space S is defined as

$$C_i^+ = \sum_{j:(i,j)\in E} f_{ij}^+. \tag{4.1.0.1}$$

Notice that in most common situations, like an n dimensional transition path problem on a lattice, every node except the reactant state A and product state B can have an inflow or an equal outflow since we defined the effective probability current that cancels the redundant backward current.

Now let's generalize the definition to any sub-network of the system, based on the observation that the Transition States can be a set of nodes, instead of a single node.

Definition 4.1.2. The current on a given connected set Ω is defined to be

$$C^+(\Omega) = \sum_{i\in\partial\Omega} C_i^+. \tag{4.1.0.2}$$

We ignored in the above definition nodes in the interior Ω° because flows in and out of an interior state should be in balance.

By considering fluxs between different nodes, we are able to characterize Transition States consisting of multiple nodes and in the form of sub-networks with complex topological structures, which also facilitate future investigations on entropy effects.

4.2 Time Reversible Case

First of all, we can not apply probability current to define Transition State directly because it maximum values will be around reactive and product states (A and B) since almost all the trajectories have to pass the states around them. We adopt the strategy reminiscent of

constraint optimization by adding weights to the probability current such that we are finding the transition states around isosurfaces of $\{x|q^+(x) = .5\}$ and $\{x|q^-(x) = .5\}$, where q^+ and q^- are the the forward committors and backward committors. In time reversible case, these are the same since $q^+ = 0.5 \iff q^- = 0.5$, so the optimization function can be defined as

Definition 4.2.1. The transition states of a chemical reaction (time reversible) is defined as

$$TS = \lim_{\sigma \to 0} \arg\max_{\Omega} \{C^+(\Omega) \cdot e^{-(q^+ - 0.5)^2/\sigma^2}\}. \tag{4.2.0.1}$$

4.3 Non Time Reversible Case

In non time reversible case, $q^+ + q^- \neq 1$, if we still just use forward committor as the optimization constraints, the numerical computation shows it's biased from the true saddle points, so we should also add the backward committor. The new optimization function tries to find trade off between these two targets.

Definition 4.3.1. The transition states of a chemical reaction (non time reversible) is defined as

$$TS = \lim_{\sigma \to 0} \arg\max_{\Omega} \{C^+(\Omega) \cdot e^{-((q^+ - 0.5)^2 + (q^- - 0.5)^2)/\sigma^2}\}. \tag{4.3.0.1}$$

Chapter 5

Illustrative example

In this section we illustrate the definition of Transition State in several examples. The first one is time reversible process in double-well and three-hole potentials, the second example is a non-reversible Markov process arising from the modeling of a genetic toggle switch in chemical kinetics, the third example is Lennard-Jones 13 Cluster.

5.1 Diffusions Processes in Potentials

Let us first consider a particle whose dynamics is governed by the stochastic differential equation:

$$\begin{cases} dx(t) = -\dfrac{\partial V(x(t), y(t))}{\partial x}dt + \sqrt{2\beta^{-1}}dW_x(t), \\ dy(t) = -\dfrac{\partial V(x(t), y(t))}{\partial y}dt + \sqrt{2\beta^{-1}}dW_y(t), \end{cases} \tag{5.1.0.1}$$

where $(x(t), y(t)) \in \mathbb{R}^2$ denotes the position of the particles, $\beta > 0$ is a parameter referred to as the inverse temperature. $V(x, y)$ is the potential, and is chosen to be

$$V(x, y) = \frac{5}{2}(x^2 - 1)^2 + 5y^2.$$

The local minima at (-1,0) and (1,0) are separated by a saddle point at (0,0), we choose the inverse temperature($\beta = 1$) such that the process spends most of time within the two wells. The reactant and product states, A and B, are the two local minima points of the

potential V, and correspond to the two maxima in the equilibrium distribution. We consider the domain $\Omega = [-1.5, 1.5] \times [-1, 1]$, which is large enough so that the potential is high at the boundaries and hence the Boltzmann-Gibbs probability density is very small.

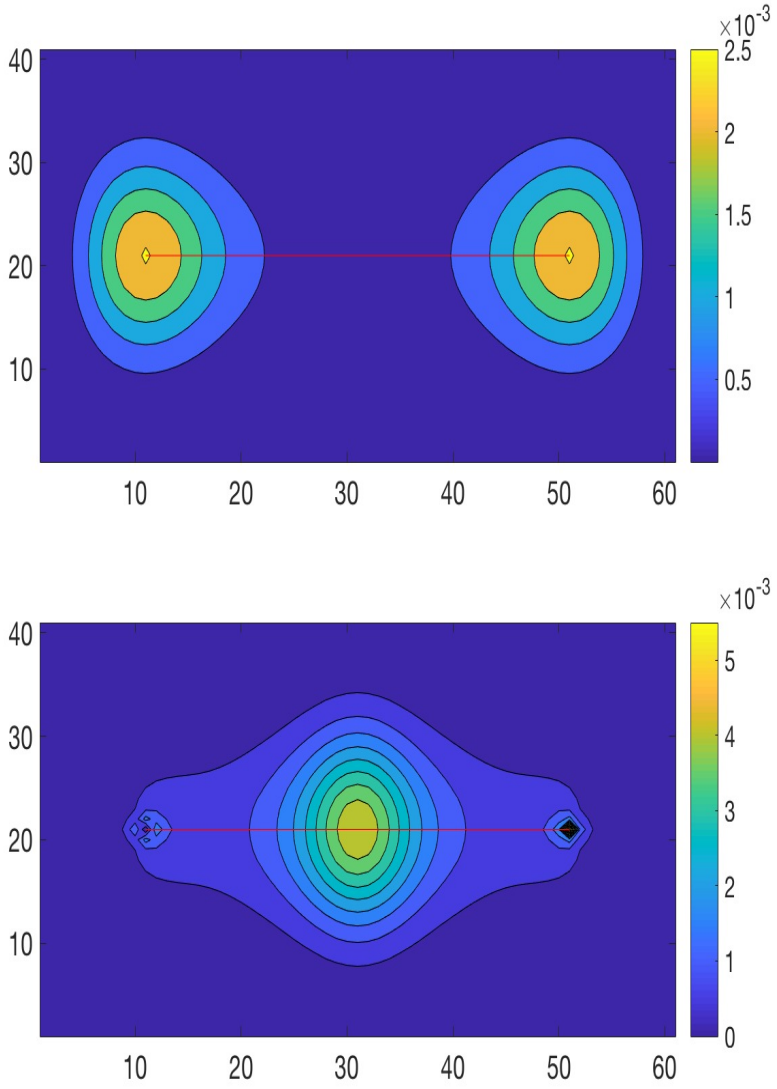


Figure 5: Upper: Contour plot of the equilibrium distribution of diffusion in double-well potential. Below: Contour plot of the weighted capacity of each state. The red line in both plot corresponds to the representative pathways.

To discretize Ω, we use an uniform grid consisting of 500×500 grids. In order to find

the transition states that are saddle points in energy landscape, we maximize the weighted probability currents as discussed in part IV such that it is maximized at $q^+ = 0.5$. The numerical parameter is chosen to be $\sigma^2 = 0.1$. Figure 5 gives the stationary distribution and the contour plot of the weighted probability currents, the red line corresponds to the transition path (representative pathways). We can see that on the lower plot, the weighted current gets its maximum at (0,0), which corresponds to the Transition State(saddle point in upper contour plot).

We next consider a more complex situation. In [93], TPT for diffusion process was illustrated through the example of a particle whose dynamics is governed by the stochastic differential equation (5.1.0.1) with the potrntial $V(x, y)$ chosen to be

$$\begin{aligned} V(x,y) =& 3e^{-x^2-(y-\frac{1}{3})^2} - 3e^{-x^2-(y-\frac{5}{3})^2} \\ &+ 5e^{-(x-1)^2-y^2} - 5e^{-(x+1)^2-y^2} \\ &+ \frac{2}{10}x^4 + \frac{2}{10}(y-\frac{1}{3})^4. \end{aligned} \tag{5.1.0.2}$$

which has also been already considered in [94, 95]. We can see in Figure 6(a) that the potential has two deep minima approximately at ($\pm$1,0), a shallow minimum approximately at (0,1.5), three saddle points approximately at ($\pm$0.6,1.1),(0,-0.4), and a maximum at (0.0.5). The second plot in figure 6 shows the transition states are at the saddle point between two meta-stable states, because here we are choosing high temperature $\beta = 1.67$, so the reaction happens mainly via the lower channel.

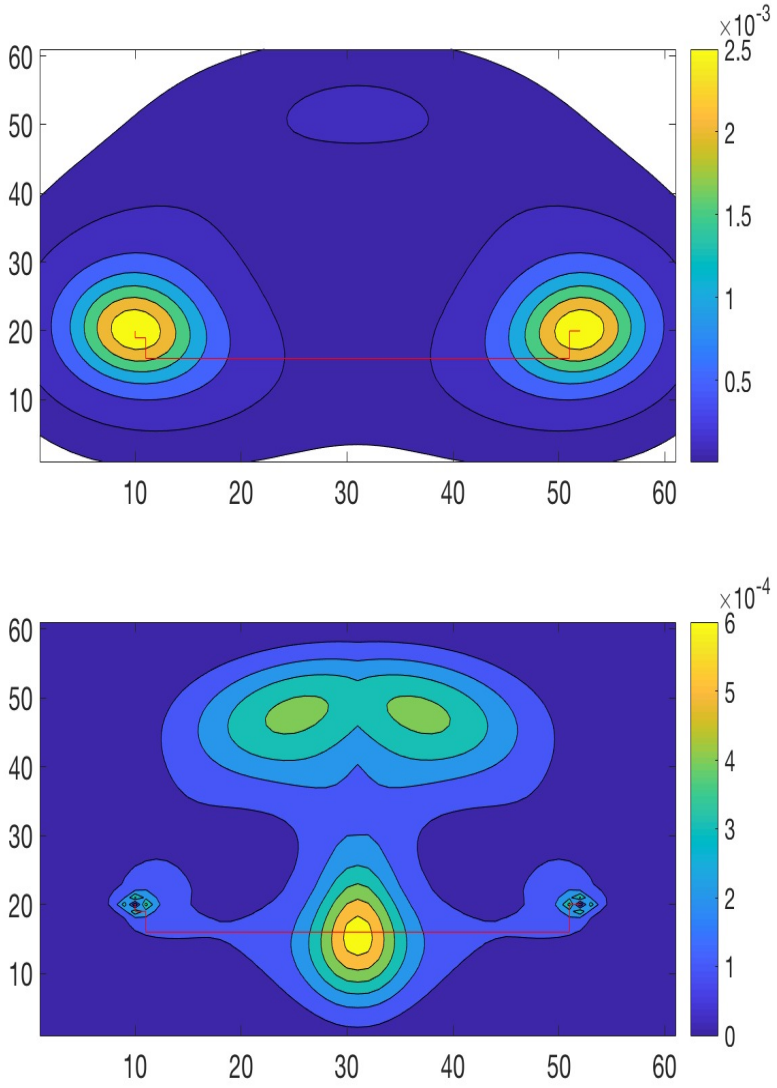


Figure 6: Upper: Contour plot of the stationary distribution $\pi_{(x,y)}$ of diffusion in three-hole potential. Results are for $\beta = 1.67$ and a 60×60 mesh discretization. Below: contourf plot of the weighted current at each discretized state. The red line in both plot corresponds to representative pathways.

5.2 Toggle Switch Models in 2D and 3D

Now consider a Markov jump process which arises as a stochastic model of a genetic toggle switch consisting of two genes that repress each other's expression[91]. The expression of each of the two respective genes results in the production of a specific type of protein; gene

G_A produces protein P_A and gene G_B produces protein P_B.

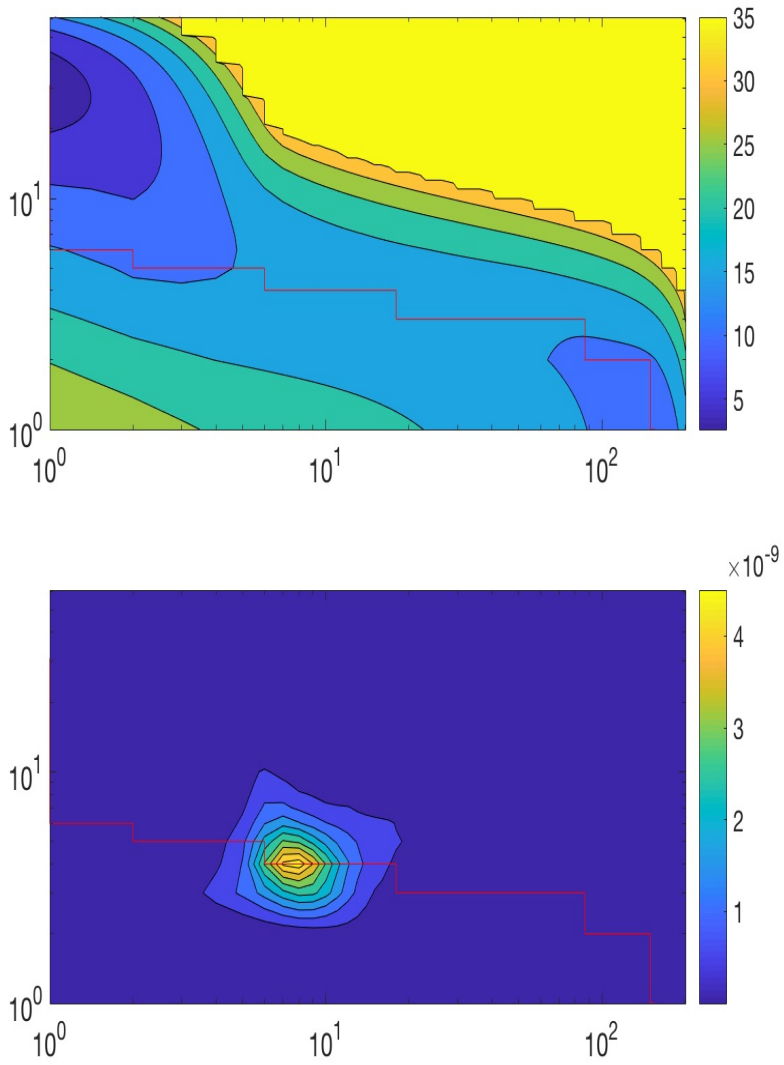


Figure 7: Upper: Contour plot of the Gibbs energy, -log$\pi(x, y)$, of the 2D Toggle switch model on the state-spaces $S = (\mathbb{Z} \times \mathbb{Z}) \cap ([0, 200] \times [0, 60])$. The dark red region in the right upper part of the panel indicates the subset of states with almost vanishing stationary distribution. Results for $a_1 = 156, a_2 = 30, n = 3, m = 1, K_1 = K_2 = 1$, and $\tau_1 = \tau_2 = 1$. Below: Contour plot of the weighted current. Both plots includes the transition path.

Denoting the number of available proteins of type P_A by x and type of P_B by y, the model for the toggle switch proposed is a birth-death process on the discrete state-space

$S = (\mathbb{Z} \times \mathbb{Z}) \cap ([0, d_1] \times [0, d_2])$. Here we choose $d_1 = 200, d_2 = 60$ to illustrate our method. By the governing equation

$$\begin{aligned} \dot{x}_1 &= \frac{a_1}{1 + (x_2/K_2)^n} - \frac{x_1}{\tau}, \\ \dot{x}_1 &= \frac{a_2}{1 + (x_1/K_1)^m} - \frac{x_2}{\tau}, \end{aligned} \tag{5.2.0.1}$$

the generator of this process is given in terms of its action on a test function f as

$$\begin{aligned} (Lf)(x,y) =& c_1(x+1,y)(f(x+1,y) - f(x,y)) \\ &+ \frac{x}{\tau_1}(f(x-1,y) - f(x,y)) \\ &+ c_2(x,y+1)(f(x,y+1) - f(x,y)) \\ &+ \frac{y}{\tau_2}(f(x,y-1) - f(x,y)), \end{aligned} \tag{5.2.0.2}$$

where

$$c_1(x+1,y) = \begin{cases} \frac{a_1}{1+(y/K_2)^n} & \text{if } x \in [0, d_1), \\ 0 & \text{if } x = d_1, \end{cases}$$

$$c_2(x,y+1) = \begin{cases} \frac{a_2}{1+(x/K_1)^m} & \text{if } y \in [0, d_2), \\ 0 & \text{if } y = d_2. \end{cases}$$

For our numerical experiments, we used the parameters $a_1 = 156, a_2 = 30, n = 3, m = 1, K_1 = K_2 = 1$, and $\tau_1 = \tau_2 = 1$.[91]

We show in Figure 7 (Upper) the Gibbs energy, $-\log \pi$, of the birth-death process instead of its stationary distribution π itself with a log-log scaling. Moreover, we neglected all states with almost vanishing stationary distribution. The color scheme is chosen such that the darker the color of a region, the more probable it is to find the process there. One

can clearly see that the process spends most of its time near the two metastable core sets $(x, y) \in \{(155, 0), (155, 1)\}$ and $(x, y) \in \{(0, 30), (1, 30)\}$, as we are interested in the reaction from set A towards set B.

One difference from the diffusion process is that this is not a time reversible process, so $q^+ = 0.5$ doesn't grantee that $q^- = 0.5$. We therefore adopt the weighted probability current (4.3.0.1), which will combine both q^+ and q^- for the optimization function. From Figure 7, we can see that the transition states can be defined as $\{(8, 6)\}$ for these two states, which is located right on the transition path.

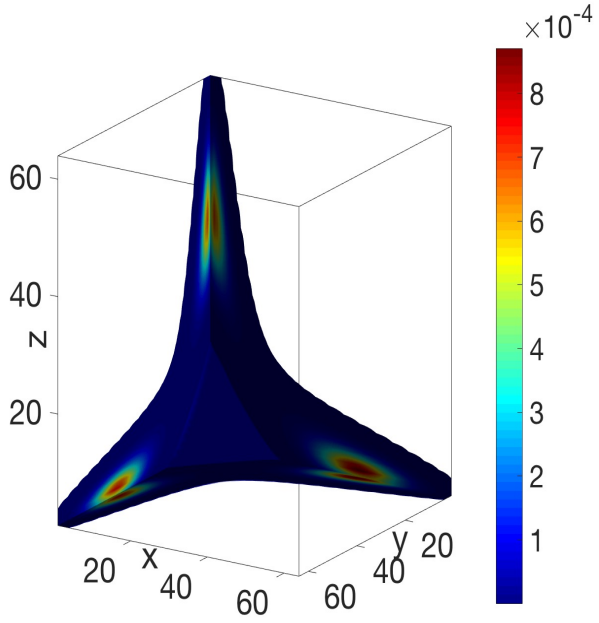


Figure 8: Contour plot of the stationary distribution $\pi_{(x,y,z)}$ for the 3D Toggle switch model with the state space $S = [0, 63]^3$. States with probability less than machine precision are marked as white colors.

Now we apply the method to a more challenging problem with larger state space, by extending it to the 3D genetic toggle switch model, built around three mutually repressing gene products, denoted by S_1, S_2 and S_3. The inhibition of each of the species by the other two components of the system is modeled by using non-standard propensities (reactions R_1

through R_3), while ramaining channels are standard degradation reactions. The reaction channels and corresponding parameter set are summarized below

$$\star \underset{\alpha_4}{\overset{\alpha_1}{\rightleftharpoons}} S_1, \quad \star \underset{\alpha_5}{\overset{\alpha_2}{\rightleftharpoons}} S_2, \quad \star \underset{\alpha_4}{\overset{\alpha_1}{\rightleftharpoons}} S_1$$

where

$$\begin{aligned}
\alpha_1 &= \frac{c_{11}}{(c_{12}+x_2^2)(c_{13}+x_3^2)}, \\
\alpha_2 &= \frac{c_{12}}{(c_{11}+x_1^2)(c_{13}+x_3^2)}, \\
\alpha_3 &= \frac{c_{13}}{(c_{12}+x_2^2)(c_{11}+x_1^2)}, \\
\alpha_4 &= c_4 x_1, \\
\alpha_5 &= c_5 x_2, \\
\alpha_6 &= c_6 x_3.
\end{aligned}$$

The parameters are $c_{11} = 2112.5$, $c_{12} = 845$, $c_{13} = 4225$, $c_{i2} = c_{i3} = 65\{i = 1, 2, 3\}$, $c_4 = 0.0125$, $c_5 = 0.005$ and $c_6 = 0.025$. We first show a 3D visualization of the stationary distribution π in Figure 8, where we neglect the value less than the machine precision.

By solving the ODE system:

$$\begin{aligned}
\dot{x} &= \frac{c_{11}}{(c_{12}+y^2)(c_{13}+z^2)} - c_4 x, \\
\dot{y} &= \frac{c_{12}}{(c_{11}+x^2)(c_{13}+z^2)} - c_5 y, \\
\dot{z} &= \frac{c_{13}}{(c_{11}+x^2)(c_{12}+y^2)} - c_6 z,
\end{aligned} \tag{5.2.0.3}$$

we can locate three metastable sets A, B and C as

$$A = \{\mathbf{x} \in S | 36 \leq x_1 \leq 37, 1 \leq x_2 \leq 2, 1 \leq x_3 \leq 2\},$$

$$B = \{\mathbf{x} \in S | 1 \leq x_1 \leq 2, 36 \leq x_2 \leq 37, 1 \leq x_3 \leq 2\},$$

$$C = \{\mathbf{x} \in S | 1 \leq x_1 \leq 2, 1 \leq x_2 \leq 2, 36 \leq x_3 \leq 37\},$$

which correspond to the three red sets on the 3D contour plot in Figure 8.

To illustrate our proposed approach, we are using state A and B as the initial state and targeting state. We next compute the weighted probability current of each state. Note that we are only interested in the local process for transitions directly between states A and states B, compared with reaction process of $A \to C \to B$. Actually most of the probability current flows using the direct route between the two states A and B, with few of the pathways also make a detour towards the other metastable state states C, before continuing to the set B. However, these pathways have a lower effective current. Figure 9 shows, in the 3D configuration space, the Transition States highlighted with color red.

5.3 The Lennard-Jones 13 Cluster

A Lennard-Jones cluster is made of atoms interacting via the Lennard-Jones pairwise potential

$$V(r) = 4a \sum_{i<j} \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right],$$

where a is the pair well depth, $2^{1/6}\sigma$ is the equilibrium pair separation, and $r = \{r_j\}_{j=1}^{N} \in R^{3N}$ denotes the positions of the N particles in the cluster, r_{ij} is the distance between par-

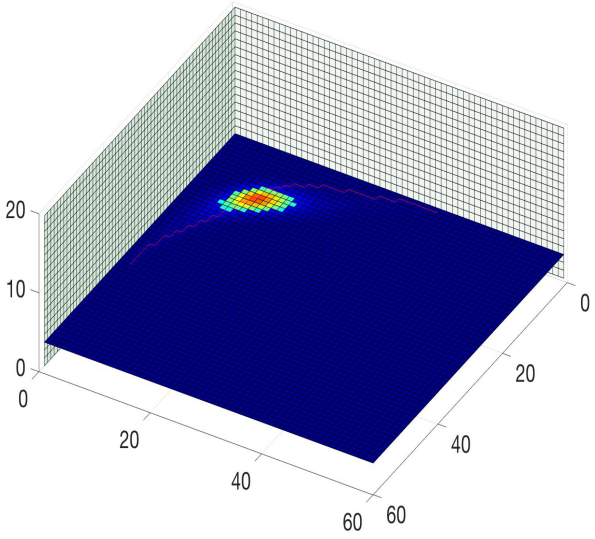


Figure 9: Slice plot of the weighted current, the red line is the representative pathways.

ticles i and j. Clusters of N Lennard-Jones atoms, denoted LJ_N, display a rich variety of behavior, and serve as benchmarks for global optimization [96], and analysis of thermodynamics [97] and dynamics in finite systems [98, 99].

David Wales and collaborators undertook an ambitious program aiming at mapping the evolution of LJ_{13} onto a network with the local minima of the energy being the nodes, therefore reducing the analysis of the dynamics to a Markov chain model that follows a basin hoping mechanism. Two such nodes are connected by an edge if the system can transit from one minimum to the another by crossing a single barrier, and the rate/weight of the directed edge is given via Arhennius formula in terms of the height of the energy barriers. This construction led to a network for LJ_{13} that contains a single connected component with 1510 nodes associated with the lowest local minima on the landscape and 29007 edges. This information is publicly available from the Lennard-Jones Cluster database on Wales's website[100]. The database also contains the information about the generator,

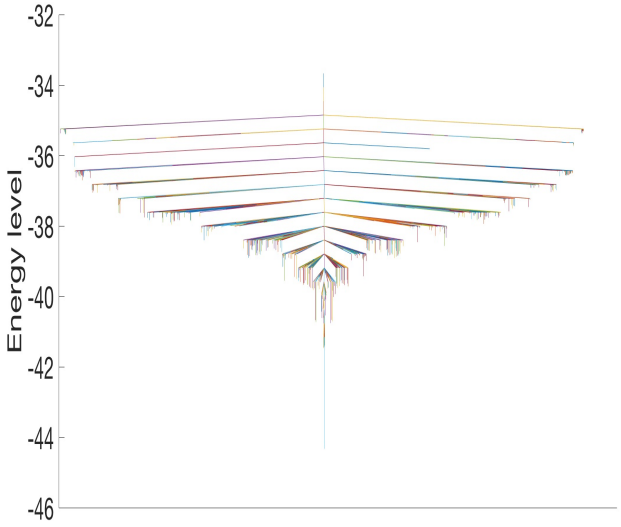


Figure 10: Disconnectivity graph for the $LJ_1 3$ cluster including all the 1510 local minima by David Wales's website database. The global minimum is a Mackay icosahedron, depicted using Xmakemol, while the next-lowest minima correspond to the three distinct capping sites when one atom is removed from the icosahedral shell and placed on the surface.

whose off-diagonal entries are in a form:

$$L_{i,j} = \sum_k \frac{O_i \bar{v}_i^{\kappa}}{O_{i,j}^k (\bar{v}_{i,j}^k)^{\kappa-1}} e^{-\beta(V_{i,j}^k - V_i)}.$$

Here $\beta = 1/(k_B T)$ is inverse proportional to the system's temperature T; O_i, V_i, and $\bar{v}_i$ are respectively the point group order, the value of the potential energy, and the geometric mean vibrational frequency for the local minimum associated with node i. For the transition state k,

$$O_{i,j}^k = O_{j,i}^k, \quad V_{i,j}^k = V_{j,i}^k$$

and

$$\bar{v}_{i,j}^k = \bar{v}_{j,i}^k$$

are the same numbers connecting the local minima i and j (there may be more than one

of them for every pair (i, j) adjacent on the network); and $\kappa = 3 \times 13 - 6 = 33$ is the number of vibrational degrees of freedom. If there is no minimum energy path connecting the minima with index i and j via a single saddle point, we set $k = 1$ and $V_{i,j} = \infty$, which means that $L_{i,j} = L_{j,i} = 0$. Note that, by construction, the generator defined above satisfies detailed-balance with represent to the following Boltzmann-Gibbs equilibrium distribution:

$$\pi_i = \frac{1}{Z(\beta)} \frac{e^{-\beta V_i}}{O_i \bar{v}_i^{\kappa}}, \qquad Z(\beta) = \sum_{i \in S} \frac{e^{-\beta V_i}}{O_i \bar{v}_i^{\kappa}}.$$

All these relations can be easily seen by organizing the states of the chain on a disconnectivity graph [101, 102], that is, a downward facing tree in which each node $i \in S$ lies at the end of a branch at a depth equal to its energy V_i, and branches in the tree are connected at the lowest energy barrier V_k, j that connects all the nodes on one side of the tree to those on the other side. In the corresponding disconnectivity graph, the potential energy increases on the vertical axis, while the horizontal axis is usually arbitrary, although it can be used to reflect properties of the minima. The disconnectivity graph of LJ_{13} cluster is show in figure 10.

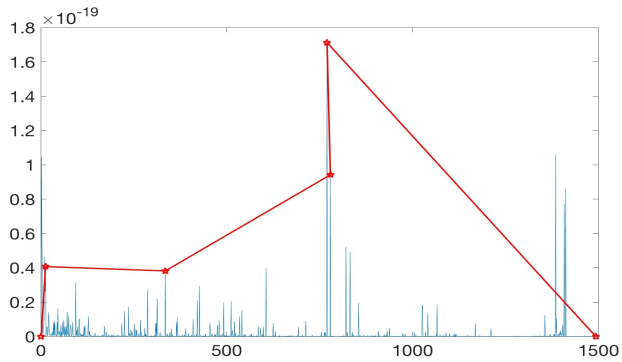


Figure 11: Weighted current of each minima, the red line is the transition path

Because there's only one obvious global minimum, and the next lowest minima connects to

global minimum directly via a saddle point, we only consider the path from global minimum to its global maximum. From the disconnectivity tree, we see that the transition is not accomplished in simple steps, so it is reasonable to apply the transition path theory to find Transition States (among local mimina) along the transition path. The majority of local minima/nodes listed in the original database are not named specifically, so in this sequel we will simply refer the minima by their indices in the database. It is easy to find the initial and final states $A = 1$ (global minimum) and $B = 1493$ (highest local minima). The temperature is measured in unites of a/k_B, in our example, we take $T = 0.15$. We plot the probability currents in Figure 11, with the red line corresponding to the representative pathway: $1 \to 13 \to 335 \to 779 \to 770 \to 1493$. The Transition State therefore should be the highest point at 770.

Chapter 6

Future Work: Clustering Methods for Directed Graph

In [19], the author stated that if several points (or groups of points) with some specific characteristics can be singled out in the data set, then, by analyzing the reaction between pairs of such groups, one will disclose global information about the data set. Furthermore, since the current of reactive trajectories is actually a directed graph as we discussed above, we can try to use the clustering methods in directed graph to find out the essential transition pathways and transition states.

Because the networks present interesting patterns and properties conveying that their internal structure is not governed by randomness. The degree distribution is skewed, following a power-law distribution [108],[109], the mean range between nodes in the network is short (the so-called small-world phenomenon [110],[111],[112]), the relations between entities do not always represent reciprocal relations forming directed networks with nonsymmetric links [107], while edge distribution is inhomogeneously occurring in node groups with high internal edges density and low density between them [107],[113]. The last feature is referred to as clustering or community structure and is of great importance in various fields and real-world applications.

Informally, a cluster or community can be considered as a set of entities that are closer to

each other, compared to the rest of the other entities in the dataset. The notion of closeness is based on a similarity pattern, which is usually defined over the set of entities. In the areas of machine learning and data mining, the task of clustering is also referred as "unsupervised learning" where the aim is to group (cluster) together with similar objects without any prior knowledge about the clusters (e.g., see Ref. [114]).

In the case of networks, the clustering (or community detection) problem refers to grouping nodes into clusters according to their similarity, which usually considers either topological features (e.g., features extracted from the graph), or other features related to the nodes and edges of the graph (e.g., additional information that may be associated with the nodes and edges), or both of them. In other words, the clusters typically match groups of nodes sharing common properties and characteristics. Although there are various definitions for the graph clustering problem, the most popular one states that a cluster corresponds to a set of nodes with more edges inside the set than to the rest of the graph.

Chapter 7

Conclusion

Based on the framework of transition path theory (TPT), we extended the probability current of reactive trajectories to nodes and sub-networks, which allows to identify Transition State with max maximum cur- rents. We also gave alternative approach to compute the transition paths. Future work involved chemical reactions with higher dimensions and complex network dynamics, by combining topological sorting algorithms with clustering methods based on this distance matrix to find the special groups like transition states.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability in stochastic dynamics of disordered mean-field models*, Probab. Theory Related Fields, 119 (2001), pp. 99-161.

[2] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability and low lying spectra in reversible Markov chains*, Comm. Math. Phys., 228 (2002), pp. 219-255.

[3] S. S. Andrews, T. Dinh, and A. P. Arkin, *Stochastic Models of Biological Processes*, Encyclopedia of Complexity and System Science (Robert Meyers, ed.), vol. 9, Springer New-York, 2009, pp. 8730-8749.

[4] A. Gupta, M. Zangril, A. Sundararaj, P. A. Dinda, and B. B. Lowekamp, Free network measurements for adaptive virtualized distributed computing, in Proceedings of the 20th IEEE International Parallel and Distributed Processing Symposium, 2006.

[5] Munsky, B., Khammash, M., *The finite state projection algorithm for the solution of the chemical master equation*, J. Chem. Phys., 124 (2006) 044104.

[6] D. Chandler, *Statistical mechanics of isomerization dynamics in liquids and the transition state approximation*, J. Chem. Phys., 68(6), 1978.

[7] Y. Cao, D. T. Gillespie, and L. R. Petzold, *The slow-scale stochastic simulation algorithm*, J. Chem. Phys. 122 (2005), no. 1, 014116.

[8] Y. Cao, D. T. Gillespie, and L. R. Petzold, *The adaptive explicit-implicit tau-leaping method with automatic tau selection*, J. Chem. Phys. 126 (2007), no. 22, 224101.

[9] D.R. Cox and H.D. Miller, The theory of stochastic processes, Wiley publications in statistics, Wiley, 1965.

[10] Ch. Schutte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys., 151 (1999), pp. 146-168.

[11] Cao, Y., Gillespie, D. T., and Petzold, L. R., *The slow-scale stochastic simulation algorithm*, J. Chem. Phys., 122 (2005) 014116.

[12] D. L. Donoho and C. Grimes, *Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 5591-5596.

[13] D. P. Landau and K. Binder, *A guide to monte carlo simulation in statistical physics*, Cambridge University Press, 2000.

[14] J. Elf, O. G. Berg, and M. Ehrenberg, *Comparison of repressor and transcriptional attenuator systems for control of amino acid biosynthetic operons*, J. Mol. Biol. 313 (2001), 941-954.

[15] M. B. Elowitz and S. Leibler, *A synthetic oscillatory network of transcriptional regulators*, Nature 403 (2000), no. 6767, 335-338.

[16] M. B. Elowitz, E. D. Siggia, P. S. Swain, and A. J. Levine, *Stochastic gene expression in a single cell*, Science 297 (2002), 1183-1186.

[17] F. Tal and E. Vanden-Eijnden, *Transition state theory and dynamical corrections in ergodic systems*, Nonlinearity, 19(2), 2006.

[18] E. Vanden-Eijnden and F. Tal, *Transition state theory: Variational formulation, dynamical corrections, and error estimates*, J. Chem. Phys., 123(18), 2005.

[19] E. Vanden-Eijnden, *Transition path theory*, In M. Ferrario, G. Ciccotti, and K. Binder, editors, Computer Simulations in Condensed Matter: from Materials to Chemical Biology, volume 2 of 703. Springer Verlag, 2006.

[20] W. E, W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events*, Phys. Rev. B, 66, 2002.

[21] W. E, W. Ren, and E. Vanden-Eijnden, *Energy landscape and thermally activated switching of submicron-sized ferromagnetic elements*, J. App. Phys., 93, 2003.

[22] W. E, W. Ren, and E. Vanden-Eijnden, *Finite-temperature string method for the study of rare events*, J. Phys. Chem. B, 109, 2005.

[23] W. E, W. Ren, and E. Vanden-Eijnden, *Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes*, Chem. Phys. Lett., 413, 2005.

[24] W. E and E. Vanden-Eijnden, *Towards a theory of transition paths*, J. Statist. Phys., 123(3):503-523, 2006.

[25] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *String method in collective variables: Minimum free energy paths and committor surfaces*, J. Chem. Phys, 125, 2006.

[26] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, *Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide*, J. Chem. Phys., 123, 2005.

[27] P. Metzner, Ch. Schutte, and E. Vanden-Eijnden, *Illustration of transition path theory on a collection of simple examples*, J. Chem. Phys., 125 (2006), 084110.

[28] H. Eyring, *The activated complex in chemical reactions*, J. Chem. Phys., 3(2), 1935.

[29] Ferm, L. and Lotstedt, P., *Adaptive solution of the master equation in low dimensions*, Applied Numerical Mathematics, 59 (2009) 187-204.

[30] C. W. Gardiner, *Handbook of stochastic methods*, 4th revised and augmented ed., Springer Series in Synergetics, Springer, Berlin, 2009.

[31] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature 403 (2000), no. 6767, 339-342.

[32] G. Ben Arous, A. Bovier, and V. Gayrard, *Aging in the random energy model under Glauber dynamics*, Phys. Rev. Lett., 88 (2002), 087201.

[33] M. A. Gibson and J. Bruck, *Efficient exact stochastic simulation of chemical systems with many species and many channels*, J. Phys. Chem. A 104 (2000), no. 9, 1876-1889.

[34] G. Hummer, *From transition paths to transition states and rate coefficients*, J. Chem. Phys., 120(2), 2004.

[35] D. T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, J. Comput. Phys. 22 (1976), 403-434.

[36] D. T. Gillespie, *A rigorous derivation of the chemical master equation*, Physica A 188 (1992), 404-425.

[37] D. T. Gillespie, *Approximate accelerated stochastic simulation of chemically reacting systems*, J. Chem. Phys. 115 (2001), no. 4, 1716-1733.

[38] C. Hartmann, *Model reduction in classical molecular dynamics.*, PhD thesis, Free University Berlin, 2007.

[39] A. Hellander, S. Hellander, and P. Lotstedt, *Coupled mesoscopic and microscopic simulation of stochastic reaction-diffusion processes in mixed dimensions*, Tech. Re- port 2011-005, Department of Information Technology, Uppsala University, 2011.

[40] C. H. Bennett, *Molecular dynamics and transition state theory: the simulation of infrequent events*, In A. S. Nowick and J. J. Burton, editors, Algorithms for Chemical Computation, 46, pages 63-97. ACS Symposium Series, 1977.

[41] Hegland, M., Burden, C., Santoso, L., MacNamara, S., and Booth H., *A solver for the stochastic master equation applied to gene regulatory networks*, J. Comput. Appl. Math., 205 (2005) 708-724.

[42] Hegland, M., Hellander, A., and Lotstedt, P., *Sparse grids and hybrid methods for the chemical master equation*, BIT Numerical Mathematics, 48 (2008) 265-283.

[43] J. Horiuti, *On the statistical mechanical treatment of the absolute rate of chemical reaction*, Bull. Chem. Soc. Jpn., 13(1), 1938.

[44] W. Huisinga, *Metastability of Markovian Systems: A transfer operator approach in application to molecular dynamics*, PhD thesis, Free University Berlin, 2001.

[45] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888-905.

[46] M. Karn, T. C. Elston, W. J. Blake, and J. J. Collins, *Stochasticity in gene expression: from theories to phenotypes*, Nature Reviews Genetics 6 (2005), no. 6, 451-464.

[47] L. Breiman, Probability, Classics Appl. Math. 7, SIAM, Philadelphia, 1992.

[48] H. H. McAdams and A. P. Arkin, *It's a noisy business! Genetic regulation at the nanomolar scale*, Trends in Genetics 15 (1999), no. 2, 65-69.

[49] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 6 (2003), pp. 1373-1396.

[50] M. E. J. Newman, *The structure and function of complex networks*, SIAM Rev., 45 (2003), pp. 167-256.

[51] M. Meila and J. Shi, *A random walk's view of spectral segmentation*, in Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics, 2001.

[52] J. R. Norris, Markov chains, Cambridge University Press, 1997.

[53] C. Dellago, P. G. Bolhuis, and P. L. Geissler, *Transition path sampling*, Ad- vances in Chemical Physics, 123, 2002.

[54] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schutte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Linear Algebra Appl., 315 (2000), pp. 39-59.

[55] Ch. Schutte, W. Huisinga, and P. Deuflhard, *Transfer operator approach to conformational dynamics in biomolecular systems*, In B. Fiedler, editor, Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, pages 191-223. Springer Verlag, 2001.

[56] D. Frenkel and B. Smit, *Understanding molecular simulation: From Algorithms to applications* Academic Press, 1996.

[57] C. Dellago P. G. Bolhuis, D. Chandler and P. Geissler, *Transition path sampling: throwing ropes over dark mountain passes*, Ann. Rev. Phys. Chem., 53, 2002.

[58] P. G. Doyle and J. L. Snell, Random Walks and Electric Networks, Mathematical Association of America, Washington, D.C., 2000.

[59] P. Metzner, Ch. Schutte, and E. Vanden-Eijnden, *Transition path theory for Markov jump processes*, Multiscale Modeling and Simulation, 2007.

[60] G. A. Pavliotis and A. M. Stuart, *Multiscale methods: Averaging and homogenization*, Springer, 2008.

[61] M. Ptashne, *A genetic switch: Phage lambda revisited*, Cold Spring Harbor Lab- oratory Press, 2004.

[62] R. Albert and A.-L. Barabasi, *Statistical mechanics of complex networks*, Rev. Modern Phys., 74 (2002), pp. 48-97.

[63] R. C. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5-30.

[64] Rao, C. V., Arkin, A. P., *Stochastic chemical kinetics and quasi-steady-state assumption: Application to the Gillespie algorithm*, J. Chem. Phys., 118 (2003) 4999-5010.

[65] R. Elber, A. Ghosh, and A. Cardenas, *Long time dynamics of complex systems*, Account of Chemical Research, 35, 2002.

[66] R. Elber, A. Ghosh, A. Cardenas, and H. Stern, *Bridging the gap between reaction pathways, long time dynamics and calculation of rates*, Advances in Chemical Physics, 126, 2003.

[67] Rathinam, M., Petzold, L., Cao, Y., and Gillespie, D., *Stiffness in stochastic chemically reacting systems: the implicit tau-leaping method*, J. Chem. Phys., 119 (2003) 12784-12794.

[68] J. M. Raser and E. K. O'Shea, *Control of stochasticity in eukaryotic gene expression*, Science 304 (2004), 1811-1814.

[69] Macnamara, S., Burrage, K., and Sidje, R.B., *Multiscale modeling of chemical kinetics via the master equation*, Multiscale Model. Simul., 6 (2008) 1146-1168.

[70] S. Lafon and A. B. Lee, *Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization*, IEEE Trans. Pattern Anal. Mach. Intell., 28 (2006), pp. 1393-1403.

[71] W.J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*, Princeton University Press, 2009.

[72] S. T. Roweis and L. K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), pp. 2323-2326.

[73] M. V. Smoluchowski, *Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen*, Ann. Phys. (Leipzig), 4(21), 1906.

[74] Jahnke, T., Huisinga, W., *Solving the chemical master equation for monomolecular reaction systems analytically*, J. Math. Biol., 54 (2007) 1-26.

[75] T. E. Turner, S. Schnell, and K. Burrage, *Stochastic approaches for modelling in vivo reactions*, Comput. Biol. Chem. 28 (2004), 165-178.

[76] J. S. van Zon and P. R. ten Wolde, *imulating biochemical networks at the particle level and in time and space: Green's Function Reaction Dynamics*, Phys. Rev. Lett. 94 (2005), 8103.

[77] N. G. van Kampen, Stochastic processes in physics and chemistry, 3rd ed., North-Holland Personal Library, Amsterdam: North-Holland, 2001.

[78] W. E and E. Vanden-Eijnden, *Towards a theory of transition paths*, J. Stat. Phys., 123 (2006), pp. 503-523.

[79] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schutte, *Identification of almost invariant aggregates in reversible nearly uncoupled markov chains*, Lin. Alg. Appl., 315(1-3):39-59, 2000.

[80] Darren J. Wilkinson, *Stochastic modelling for quantitative description of heterogeneous biological systems*, Nature Reviews Genetics 10 (2009), no. 2, 122-133.

[81] E. Wigner, *The transition state method*, Trans. Faraday Soc., 34, 1938.

[82] T. Yamamoto, *Quantum statistical mechanical theory of the rate of exchange chemical reactions in the gas phase*, J. Chem. Phys., 33(1), 1960.

[83] Zhang, J.W., Watson, L.T., and Cao, Y., *Adaptive aggregation method for the chemical master equation*, Int. J. Computational Biology and Drug Design, 2 (2009) 134-148.

[84] P. Metzner, C Schutte, and E. Vanden-Eijnden, Transition Path Theory For Markov Jump Process, SIAM Multiscale Modeling and Simulation, 7(3), 1192-1219,2009.

[85] R.J. Allen, P.B. Warren, and P.R. ten Wolde, Sampling Rare Switching Events in Biochemical Networks, Physical Review Letters, 94, 018104, 2005.

[86] H. Eyring, The activated complex and the absolute rate of chemical reactions, Chemical Reviews, 17(1), 65-77, 1935.

[87] J.Horiuti, On the statistical mechanical treatment of the absolute rate of chemical reaction, Bull. Chem. Soc. Japan, 13(1),210-216, 1938.

[88] H.D. Jong, Modeling and Simulation of Genetic Regulatory Systems: A Literature Review, Journal of Computational Biology, 9(1), 67-103, 2002.

[89] P. Bolhuis, D. Chandler, C. Dellago, and P. Geissler, Transition path sampling: Throwing ropes over rough mountain passes, in the dark, Ann. Rev. Phys. Chem., 53(1),291-318, 2002.

[90] P. Bolhuis, Transition-path sampling of β-hairpin folding, Proc. Natl. Acad. Sci. USA, 100(21),12129-12134, 2003.

[91] D.M. Roma, R. O'Flanagan, A. Ruckenstein, A.M. Sengupta, and R.Mukhopadhyay, Optimal Path in Epigenetic Switching, Phys. Rev. E, 71, 011902, 2005.

[92] E. Wigner, The transition state method, Transactions of the Faraday Society, 34, 29-41, 1938.

[93] P. Metzner, C. Schutte, and E. Vanden-Eijnden, Illustration of transition path theory on a collection of simple examples, J. Chem. Phys., 125, 084110, 2006.

[94] S. Park, M. K. Sener, D. Lu, and K. Schulten, Reaction paths based on mean first-passage times, J. Chem. Phys., 119, 1313-1319,2003.

[95] P. Deuflhard and C. Schutte, Molecular conformational dynamics and computational drug design, Applied Mathematics Entering the 21st Century, J. M. Hill and R. Moore, eds.,vol. 116, SIAM, 91-119, 2004.

[96] D. J. Wales and H. A. Scheraga, Global Optimization of Clusters, Crystals, and Biomolecules, Science 285, 1368-1372, 1999.

[97] R. S. Berry, T. L. Beck, H. L. Davis and J. Jellinek, Solid-liquid phase behavior in microclusters, Adv. Chem. Phys. 70B, 75-138, 1988.

[98] C. Dellago, P. G. Bolhuis and D. Chandler, Transition path sampling and the calculation of rate constants, J. Chem. Phys., 108, 1964-1977, 1998.

[99] W. E, W. Ren and E. Vanden-Eijnden, String method for the study of rare events, Phys. Rev. B, 66, 052301, 2002.

[100] http://www-wales.ch.cam.ac.uk/CCD.html

[101] R. Czerminski and R. Elber, Reaction path study of conformational transitions in flexible systems: applications to peptides, J. Chem. Phys. 92(9), 5580-5601, 1990.

[102] O. M. Becker and M. Karplus, The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics, J. Chem. Phys. 106, 1495-1517, 1997.

[103] Kahn, Arthur B., Topological sorting of large networks, Communications of the ACM, 5 (11): 558-562, 1962.

[104] Banisch R, Vanden-Eijnden E., Direct generation of loop-erased transition paths in non-equilibrium reactions, Faraday Discuss., 195, 443-468, 2016.

[105] Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford, Introduction to Algorithms (2nd ed.), MIT Press and McGraw-Hill, pp. 549-552, ISBN 0-262-03293-7,2001.

[106] Tarjan, Robert E, Edge-disjoint spanning trees and depth-first search, Acta Informatica, 6 (2): 171-185,1976.

[107] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167-256.

[108] A.L. Barabasi, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509-512.

[109] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, in: Proceedings of the Conference on Applications,Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM'99, 1999, pp. 25-262.

[110] S. Milgram, The small-world problem, Psychol. Today 1 (1) (1967) 61-67.

[111] R. Albert, H. Jeong, A.-L. Barabasi, The diameter of the world wide web, Nature 401 (1999) 130-131.

[112] J. Leskovec, E. Horvitz, Planetary-scale views on a large instant-messaging network, in: Proceeding of the 17th International Conference on World Wide Web, WWW'08, 2008, pp. 915-924.

[113] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821-7826.

[114] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264-323.

[115] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, Network Flows, Prentice-Hall, Englewood Cliffs, NJ, 1993.