©Copyright 2018
Yue Wang

Some Problems in Stochastic Dynamics and Statistical Analysis of Single-Cell Biology of Cancer

Yue Wang

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Hong Qian, Chair

Matthew Lorig

Krzysztof Burdzy

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

Some Problems in Stochastic Dynamics and Statistical Analysis of Single-Cell Biology of Cancer

Yue Wang

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

With recent development of instrumentations, data collection capabilities, and computational softwares, in cancer biology now one has high-throughput data on individual cells in a population: its heterogeneous genomic makeups and/or phenotypic molecular markers. At this level of description, stochasticity is a significant component of the *dynamics of single cells*. Using purely deterministic representations and mathematical models is no longer realistic nor desirable. In addition, most traditional deterministic models are based on differential equations and dynamical systems with real numbers; single cell data are too discrete to fit a continuous state space.

This dissertation consists of three parts: Motivated by laboratory measurements, carried out at Institute for Systems Biology, on the stochastic population growth of Leukemia cells (HL60), Part I, Chapter 2 reports statistical analysis of the data and develops dynamic models in terms of birth-death processes. It is shown that even in the very earlier stage of ~ 10 cells, there exists already multiple phenotypes in the population: This result invalidates the naive assumption of statistically identical individuals in exponential cell growth. Then in Chapter 3, branching processes are introduced to study the equilibrium of heterogeneous cell population with multiple phenotype switching. A law of large numbers is proven. A proof of mathematical relationship between multitype branching process and multi-dimensional birth-death process is given in Section 3.8.2.

Part II addresses a central issue in data statistics: How to quantify the logic causal effect among observations of interdependent random variables. In terms of the notion of *Markov boundary*, we

provide a rather coherent quantification of a class of causal inference. We prove that in certain special cases, quantifying causal effect is not possible. Computational algorithms for determining such scenarios are proposed and implemented.

Stochastic processes employed in cancer biology and in statistical physics have a fundamental difference: The latter as models for inanimate matters require detailed balance while the former, as models of living organisms, have positive stationary entropy productions. In Part III (Chapter 5), we investigate a theoretical issue of irreversible Markov processes without detailed balance: How to represent the entropy production in a finite system through a lifting of this stochastic process. We prove that the entropy production rate in the finite system can be represented in terms of a potential function in the lifted infinite system with detailed balance. We propose to use this mathematical result to unify the two different statements, due to Clausius and Lord Kelvin, of the Second Law of Thermodynamics.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The author would like to thank Ivana Bozic, Krzysztof Burdzy, Dayue Chen, Siqi He, Sui Huang, Svante Janson, Da-Quan Jiang, May Khalil, Yang Lan, Linyuan Liu, Yifei Liu, Matthew Lorig, Georg Luebeck, Daniel Malinsky, Soumik Pal, Edoardo Pedrini, Hong Qian, Min-Ping Qian, Thomas Richardson, Irit Rubin, Jifan Shi, Jiajun Tong, Xin Tong, Linbo Wang, Weili Wang, Yuting Wei, Daxin Xu, Boyu Zhang, Mingda Zhang, Ruixiang Zhang, Qingyuan Zhao, Zhongkai Zhao, Da Zhou, Joseph Xu Zhou, Xiao-Hua Zhou, Lingxue Zhu for helpful advice and discussions directly related to this dissertation.

The author would also like to thank Yu-Chen Cheng, Qi Guo, Daiwei He, Siqi He, Jing Li, Yi-An Ma, Jiajun Tong, Zikun Wang, Yihao Xuan, Felix Ye, Boyu Zhang, Jize Zhang, Mingda Zhang, Yumeng Zhang, Peng Zheng, Zeyu Zheng, Tianji Zhou and many others for indirect supports to this dissertation.

The author would like to thank his family for their love and supports.

DEDICATION

To “trouble-making” students of PKU

Chapter 1

INTRODUCTION

With recent development of instrumentations, data collection capabilities, and computational softwares, cell biology now has high-throughput data on individual cells in a population: its heterogeneous genomic makeups, biochemical compositions, and phenotypic molecular markers. To many biomedical scientists, *Single-Cell Biology* holds the promise for revolutionizing biology and medicine [117]. At this level of description, stochasticity is a significant component of both the characterization of single-cell data and dynamic, mechanistic representations of the biology. Purely deterministic representations and mathematical models based on differential equations, widely used in mathematical biology [99], are no longer legitimate: They are neither realistic nor desirable for the new biology. In addition, most traditional models are based on real-valued variables; single cell data are too discrete to fit a continuous state space.

To develop the mathematical biology of single cells, we seek inspiration from the field of single-molecule biophysics which had emerged in the early 1990s and risen to prominence in the new millennium. The development of single-molecule biophysics led to the significant discovery of stochastic gene expressions in single cells [129, 115]; that has provided a foundation for understanding cell differentiations and non-genetic heterogeneities in cell populations, and paved the way for the current single-cell biology. Single-molecule biophysics has also inspired new developments in statistical research [127] and new results in probability [70].

We expect a fundamental difference between the stochastic processes employed in single-molecule biophysics and in single-cell biology: The models for macromolecules in biophysics, as inanimate matters, require detailed balance: The long-time limit of a model is a stationary stochastic process that necessarily has time-reversal symmetry, a fundamental property of a part of

a *thermodynamic equilibrium with fluctuations*. Single cells, however, are living organisms. In a conceptualized homeostatic state, they continuously metabolize chemicals and dissipate heat. The stationary stochastic processes that represent cells have positive stationary **entropy production**: A mathematical concept developed in recent years precisely motivated by a need to differentiate nonequilibrium living processes from equilibrium inanimate processes [71].

1.1 Data-driven stochastic cell growth

Stochastic processes in modeling cell growth via clonal expansion has a long history [167]. One early work is the Luria-Delbrück model, which assumes cells grow deterministically, with wildtype cells mutate and become cells with a different phenotype randomly [90]. Since then there have been many further developments which incorporate more stochastic elements into the model, such as Lea and Coulson [84], Koch [81], Moolgavkar and Luebeck [88], Dewanji et al. [25]. We can find various stochastic processes: Poisson processes [12], Markov chains [57], and branching processes [72], or even random sums of birth-death processes [25], all playing key roles in the mathematical theories of cellular clonal growth and evolution.

In applied mathematics, after one establishes a mathematical model in terms of differential equations or stochastic processes, the next obligatory question is how to obtain the behavior of its dynamics. For stochastic processes, there are two complementary perspectives: the stochastic trajectory and its probability distribution as a function of time. The former usually is expressed in terms of a standard, simple process: Brownian motion for diffusion and Poisson process for continuous time Markov chain. The latter can be obtained as a solution to the Chapman-Kolmogorov equation. For simple models, one can directly obtain analytical solutions. For relatively complex processes, one can use generating functions to turn a stochastic problem into a differential equation problem, and then analytically or approximately solve such equations [96], obtaining the probability distribution in a transformed form. There are of course a large repertoire of mathematical methods for obtaining various properties of a stochastic-process model. A particular one used in this dissertation is the martingale method for proving the asymptotic behaviors of branching processes [67].

There is a significant difference between the traditional stochastic dynamics of cell populations with genetic mutations and the current interest of non-genetic heterogeneity: The phenotype changes in the former are practically never reversible, while epi-genetic phenotype switchings are much more likely to be reversible, at least in principle. For a given cell, there are enormous possible genetic mutations in terms of changes in DNA than relatively small number of possible stable non-genetic biochemical variations.

When one compares stochastic-processes models with real data, statistics enters our study. One first needs statistical methods to estimate the value of parameters, as well as to derive the confidence intervals. After the parameters being determined, one needs to evaluate how the model fits the data. When there exist multiple models, one needs to compare their fitnesses on the data to select the best one [77].

In data science, even without a complete model, one can test different assumptions with statistics to provide useful information on the mechanism behind the data. For example, we can test whether two groups have the same mean and variance, which informs us whether we should treat them separately. We can compare and contrast one sample, of cell population growth, containing four initial cells and four samples each with one initial cell, and test whether they behave statistically identical. Studies as such provide implications on whether the descendants of different cells are independent.

At the Institute for Systems Biology, in the laboratory of Dr. Sui Huang, a strain of Leukemia cancer cells (HL60) was cultured starting with a few single cells *in vitro*. The growth of cell population exhibits several phases: The first one is a lag phase during which it is assumed that the cells become accustomed to the new environment and recover from manipulations in the "seeding process". The cell population then enters the exponential growth phase during which the average per capita growth rate is essentially a constant: The total population therefore grows exponentially in this phase. It is a widely held belief that the growth rates for all cells are independent and identically distributed. With the current technology from single-cell biology, one can experimentally check this long-hold hypothesis: We shall show, by using statistics, that even during the exponentially growing phase, individual cells are not independent. The data suggested that the cells should

have at least three states that have different distributions on growth rates.

1.2 Mechanism-based stochastic cell growth

Equally important as data, biologists usually approach to biological problems with a plausible *mechanism* in mind. In applied mathematics, this means developing mathematical models to explain certain biological phenomena. The first step here is to abstract biological mechanisms into mathematical descriptions. The next step then is to study the properties of the mathematical model. This includes finding solutions and proving theorems concerning the essential characteristics of the model. The last step is to translate the mathematical results back into biological setting, in terms of the biological language, to explain existing biological phenomena and predict new ones. Sometimes the mathematics itself becomes so interesting and important that the biology background will be put aside for a while. Of the field of mathematical biology, there are important results even from this mode of research. Two examples from population biology are the Galton-Watson process of the extinction of family names [155] and the Lotka-Volterra system of equations for predator-prey populations [87]. In physiology and biophysics, the Hodgkin-Huxley model for membrane action potentials in neural communication [64] is one example of a work that combines mechanism with data and produces significant new biology as well as pushes the frontier of new applied mathematics.

It has become increasingly clear that at the level of individual cells, one genotype may correspond to multiple phenotypes, which is usually defined by different biochemical compositions, or molecular markers, of a cell. These different cellular phenotypes, same as genotypes, are also relatively stable, and even can be transmitted to "descendants" through cell divisions [3]. More interestingly, individual cells can switch among different phenotypes [29]. One well documented fact is the asymmetric division of stem cells, where one stem cell divides into one stem cell and one non-stem cell. Some cancer non-stem cells can even transform back to cancer stem cells [169]. In such cancer cell populations, there are now ample experiments showing that the proportions of different phenotypes will always converge to the same invariant distribution, regardless of initial proportions [57]. This phenomenon has been named "phenotypic equilibrium".

Branching processes can be used as the proper representations of such population dynamics [66]. In branching processes, there is a series of laws of large numbers under different assumptions, which is quite similar to the "phenotypic equilibrium" phenomenon. Since 1960s, probabilists have already proved that for a continuous-time multitype branching process, the proportions of different types converge to a constant vector under different conditions. In [7], [8] and [139], the proofs require that the coefficient matrix is irreducible, and the largest eigenvalue is positive. In [78], it is required that the branching process is discrete in time. In [159, 160], it is assumed that the initial population tends to infinity. The assumptions required in [67] is the weakest. Still one of the assumptions is not always met in biology: The type with largest growth rate could transform into any other types. In this dissertation, I aim to prove the same result without this assumption.

As a continuous-time Markov process, the waiting time of the branching process studied is exponential. However, the time distribution of cell division in reality deviate significantly from exponential [52]. In our treatment, we apply a special technique that introduces auxiliary types such that the process is always Markovian, but the observed cell-division time is close to reality.

1.3 Data statistics and causality

Applied mathematics in the past was centered around the idea of "dynamics" – in fact, it is not an overstatement that all major areas of modern applied mathematics arouse as a need to understand dynamics, or more narrowly mechanics. The notion of dynamics is unique in science: it is the only formal concept that offers a definitive causality (e.g. a mechanism for a phenomenon). Causality is the foundation for modern science and knowledge; it is certain to say that causality will never stop to be relevant, no matter what new paradigm arises in scientific discovery.

In recent years, with the development of high-throughput measurements, we can acquire large amount of biological data, such as the single cell gene expression level for thousands of cancer cells. Thus it might be feasible now to apply the mechanistic model free, data-driven paradigm, in which causality determination has an essential role to play. In general, purely statistical models can be classified into two groups: Those deal with discrete and/or continuous random variables, and those deal with events and states. For the former data type, linear regression is widely considered

as a power tool for extracting correlations; and for the latter, mutual information has risen to prominence through information theory. To address causal relationship, several methods have appeared in the literature: partial Granger causality [56], conditional mutual information [27], and transfer entropy [135], to name a few.

However, data-driven paradigm might not be able to totally replace mechanism-based paradigm. For example, when one only has partial knowledge for some variables, how should one apply such knowledge into data-based methods? In such cases, model-driven method is natural to emerge. Another problem is, whether the data method is really suitable for some data. In [162], the authors treat ecosystems as nonlinear dynamical systems and only utilize data to produce forecasting. However, with measuring errors and natural stochasticity, the long time behavior of highly nonlinear dynamical systems cannot be predicted even in theory. There is also a possibility that the data is not reliable enough as claimed.

In 1956, Wiener proposed that one can detect whether the prediction of one time series could be improved by incorporating knowledge of the other, so as to measure the causal influence of one time series on another [157]. Later Granger formalized this idea with linear regressions to quantify the prediction accuracy, namely the Granger causality [54]. Lately, the Granger causality is modified to deal with multiple time series [47, 56]. There are also mutual information based quantities. Schreiber proposed transfer entropy in 2000 [135], followed by several improvements later [73, 40]. Another approach is utilizing directed acyclic graphs, mainly by theoretical statisticians [152, 113, 5, 130]. There are many applications of causal inference methods to data in biological research [9, 53, 106].

1.4 Stochastic, nonequilibrium physics of living cells

From the viewpoint of statistical mechanics, life is an open system which exchanges materials (nutrients in the growth medium for cell culture) and energy with the environment. It needs to keep a stationary condition so as to function normally. However, such a stationary state is far from equilibrium, otherwise it is not “living”. A living cell is in a nonequilibrium steady state (NESS) [103]. In recent years, the mathematical theory of NESS, in terms of the notion of entropy

production, has been extensively developed [71].

The idea of stochastic process approach for statistical mechanics can be traced back to Einstein's work on Brownian motion [35]. Onsager exploited stochastic processes to discuss irreversible thermodynamics for systems close to equilibrium [105]. Prigogine started to study irreversible systems far from equilibrium [48]. In order to describe a macroscopic system which is composed by enormous number ($\sim 10^{23}$) of microscopic elements, we need to statistically study the rapidly varying behavior of elements, and then the complex dynamics is projected to several macroscopic variables [71].

For an irreducible continuous time Markov chain, any initial distribution will converge to the unique stationary distribution exponentially fast. Starting from the stationary distribution, and reversing the time, one has another stationary process. Only in special cases, this process is reversible, meaning that the reversed process has the same joint distribution (and the same transition rates) with the original process. In this case, observing this process forwards or backwards in time has no difference. For a reversible process at stationary distribution, the flux between two states should cancel out. This also means that for any cycle, the net flux is zero: The probability of $A \to B \to C \to A$ is the same as $A \to C \to B \to A$. For a Markov chain, we can define its entropy production rate at stationary, which describes the reversibility of this process. It is a nonnegative quantity, and equals zero if and only if the Markov chain is reversible. The cycle flux has essential relationship with reversibility. Biochemical reactions are obviously irreversible with respect to time (how can you imagine a chicken turns back to an egg?). One can even say that irreversibility is one of the chief characteristics of life activities.

For a steady state, the income flux must equal the outcome flux (Kirchhoff's law), thus all fluxes should be in the form of cycles. Therefore we have the circulation decomposition theorem: The net flux on an edge can be decomposed into the flux of cycles passing this edge. In practice, some cycles are not linearly independent, therefore we can determine a basis, and decompose the flux solely in cycles in the basis. The next question is to study the behavior of cycles in the basis.

For a trajectory of a Markov chain, we can define its entropy gain. If we calculate the entropy gain symbolically, then two trajectories with the same starting point have the same entropy gain if

and only if they have the same net winding number for each cycle in the basis. Here the order of cycles finished does not matter. This reminds us that the fundamental group of n dimensional torus has similar properties. We will lift the finite Markov chain into an infinite Markov chain with global potential, through the help of n dimensional torus. We will also prove that the entropy production rate is invariant after the lifting. This approach explains the source of entropy in NESS system, and could be used in unifying the two different statements, due to Clausius and Lord Kelvin, of the Second Law of Thermodynamics.

———— * ———— * ———— * ————

We have discussed a wide array of approaches to cancer biology from the standpoint(s) of applied mathematics in a broad sense. With the rise of data science and the dawning of the age of stochasticity [98] impacting on the traditional applied mathematics, in my Ph.D. research, I have strived to study different aspects of *cancer biology* with various new perspectives. The goal is to have a more comprehensive outlook at the emergent "single-cell cancer biology", from its fundamentals to the the practicals.

The dissertation is organized as follows. In Chapter 2, we study the cancer cell growth dynamics based on experimental data using statistics. We develop models in terms of birth-death processes, and show that there exist multiple phenotypes among cells. In Chapter 3, we prove a law of large numbers in branching processes. The result provides an explanation for the phenotypic-equilibrium phenomenon in cancer cell population. In Chapter 4, we prove that in some special cases, quantifying causal effect is impossible. We have also designed and implemented algorithms to determine such cases. In Chapter 5, we consider the lifting of irreversible stochastic processes and prove the entropy production rates are invariant under the lifting. Chapter 6 discusses some of the possible future works.

Even though I am primarily responsible for all the results in this document, many were the outcomes of collaborative efforts. To acknowledge the contributions from others: Chapter 2, Edoardo Pedrini, Joseph Xu Zhou, Irit Rubin, May Khalil and Sui Huang at Institute for Systems Biology

and Hong Qian; Chapter 3, Da-Quan Jiang at Peking University and Da Zhou at Xiamen University. Chapter 4, Linbo Wang at Harvard University; Chapter 5, Hong Qian. Furthermore, Section 2.5.2 is a collaborative result with Yifei Liu at University of Wisconsin–Madison and Section 3.8.1 is with Zhongkai Zhao at Shandong Normal University.

Chapter 2

STOCHASTIC GROWTH OF LEUKEMIA CELLS REVEALS THE HETEROGENEOUS DIVERSITY

2.1 Introduction

Cancer has been known as a genetic disease caused by oncogenic mutations in somatic cells. According to the clonal evolution theory, cancer cells result from the accumulation of random genetic mutations, which enable the emerging cell clones with novel genotype(s) to gradually out-grow normal cells, break down the tissue homeostasis and gain other cancer hallmarks [59]. In this view, a few genetically distinct clones of homogeneous cells dominate the cancer cell population since they have higher growth fitness than normal cells. With the advance of single-cell technology and the consequent experiments, however, it had been found that nongenetic cell heterogeneity universally exists in cancer cell population [118] - distinct cell subpopulations, transitions and the interactions among them influence cancer cell growth, migration and drug resistance etc. [147]. Thus an contradistinctive view is that cancer is more like an evolving ecosystem [41] in which cells, as biological individuals, diverge into various functionally distinct subpopulations with the persistent cellular heterogeneity [32, 140]. Furthermore, cancer cell "ecology" differs fundamentally from the classic ecological system: Its subpopulations transitions to each other in a reversible way while the species in an animal kingdom do not transition between each other [19]. This enable cancer cells to have remarkable heterogeneity and plasticity, giving rise to the concept of *quasispecies* [33], which play important roles in their growth and later development of resistance to therapeutic treatments [94].

Many new questions arise following the broad hypothesis that heterogeneity and transitions are important factors for cancer. What is the dominant mechanisms behind the onset of cancer? Is it due to the fastest growing clone begins to take off, or the cell population reaches the critical

mass of positive feedback of subpopulations then take off? Should one target the fastest growing subpopulations, or targeting the interactions and interconversions of cancer cells? Therefore, it is important to develop a cancer growth model that fully considers the influence of heterogeneity and inter-conversion of cancer cells on their growth pattern as well as the interplay between these two modalities. The deepen understandings will help us to develop new successful therapeutic strategies. Classically human tumors have been described to follow an exponential growth law [91] but subsequently, to better fit experimental data, two different models have developed, the Gompertz model and the West law model [163]. While there is no one specific model that can adequately describe any one tumor, each model does highlight certain aspects of macroscopic tumor kinetics (mainly the maximum size and the changing in growth rate at different stages). Assuming cancer cells nongenetic heterogeneity with transitions dynamically, the population behavior is influenced by many intrinsic and extrinsic factors which are both variable and unpredictable at single-cell level. Thus, tumor growth cannot be adequately captured by a deterministic model, rather a stochastic population kinetic model is more realistic. Stochastic processes in modeling cell growth via clonal expansion has a long history [167]. One early work is the Luria-Delbrück model, which assumes cells grow deterministically, with wildtype cells mutate and become cells with a different phenotype randomly [90]. Since then there have been many further developments which incorporate more stochastic elements into the model, such as Lea and Coulson [84], Koch [81], Moolgavkar and Luebeck [88], Dewanji et al. [25]. We can find various stochastic processes: Poisson processes [12], Markov chains [57], and branching processes [72], or even random sums of birth-death processes [25], all playing key roles in the mathematical theories of cellular clonal growth and evolution. These models have been applied to clinical data on lung cancer [101], breast cancer [141], and treatment of cancer [142].

This paper focuses on the stochastic growth with heterogeneity of HL60 leukemia cells with near single-cell sensitivities. We note fundamentally at the level of single cells, each cell as an individual organism behaves differently from another; classification and grouping are necessary. Classification requires cell markers whose legitimacy is difficult to determine *a priori*. Still, in the context of cancer biology, the most legitimate maker is cell growth potential. Therefore, we

observed longitudinally the growth of cancer cell populations seeded with very small number of cells (1, 4 or 10 cells). We found the clear evidence that the clonal HL60 leukemia cells exhibit diverse growth patterns. Based on quantitative statistical data analysis, we propose the existence of three distinctive cell states and used branching processes to model the population kinetics with distinctive cell subpopulations and the transition between them. Our results show that the initial cancer cell growth is mainly driven by the fast-growing cell subpopulation in HL60 cancer cell population. This observation opens a new door to develop novel cancer therapy to specifically target the fast-growing cells and the transition to them from other cell subpopulations.

2.2 Results

2.2.1 Experiment summary of the different cell population growth rates from distinct initial seeding cells number

To find how initial cell seeding number influences HL60 cell growth patterns, we chose three experiment settings: each has 1, 4, 10 seeding cells per well and each has 80 wells as replicates. Cells with different seeding numbers were sorted into wells of a 384-well plate filled with growth medium and grew in the same conditions for 23 days. To avoid background noise in cell counting, we took photos and estimated the area occupied by cells every 24 hours for each well from day 4. Based on our benchmark experiment to count cell number and estimate the cell occupied area at the same time, we found that one area unit has approximately 500 cells.

The experiment results of cell growth patterns with various initial seeding numbers were summarized in Figure 2.1. In the left panel of Figure 2.1A the daily cell occupied areas were plotted in linear scale while in the left panel of Figure 2.1B the same data were plotted in $\log_{10}$ scale. Almost all growth curves were straight lines in log scale between day 5 and day 10 - suggesting that the cells were mostly in exponential growth phase during this period. It was clear that cell growth patterns varied even though they came from the same clone. The lower were the cell seeding number, the more significantly cell growth patterns varied. As shown in the left panels of Figure 2.1A and 2.1B, the 10-seeding-cell populations (green curves) were all in exponential phase from day

7 to day 13; almost all 4-seeding-cell populations (blue curves) were in exponential phase from day 8 to day 15, except three wells which lagged behind around 7 days; 1-seeding-cell populations (red curves) significantly varied from each other. They diverged into two distinct groups: growing group and non-growing group. 18 one-seeding-cell wells never took off during the experiment while the exponential growth onset in the remaining wells varied from day 10 to day 21 even they all came from the same clone.

The wells with less initial cells numbers grow slower than other wells even when they were at the moment with the same cell number, as shown in Figure 2.1C left panel. The same happened to the population average growth, as shown in 2.1D. why cell population growth rates were influenced by the initial cell numbers? It was plausible that the growth heterogeneity exists in the HL60 cell line, which was revealed by the randomly sampling the population with small number of cells. More detailed explanation demanded us to construct a quantitative dynamical model to recapitulate the growth dynamics differences from cell populations with distinct initial cell numbers.

2.2.2 Experiment of inheritance of the growth patterns of HL60 cell subpopulations

To test if the HL60 cells can maintain the distinct growth rates across generation, we did further experiment to re-seed the cells with different growth rates. We kept the cells in wells growing for almost 3 weeks. At this point it was obvious that some wells are confluent (full of cells) with fast growing cells; some wells are half full with moderately growing cells; some wells never took off. We chose two wells of cells with different growth rates (one well is full while another is half full), reseeded them into 32 wells separately (about 80 cell per well), and cultivated them for another 20 days. To qualitatively compare their growth rates after reseeding, we measured the time when the cells in each well exceeded the half area of a well (see results in Table 2.2). We found that for fast-growing wells, they exceeded half well during day 11–15, while moderately-growing wells started to exceed half well on day 14, and most of wells exceeded half well at around day 20. 5 wells with moderately-growing did not exceed half well at the end of the experiment. As shown in the analysis in Table 2.2, the fast cells do keep growing fast across generation in the short term. i.e. the heterogeneity of the growth rates are stable with 15–20 generations in our re-seeding experiment.

2.2.3 Statistical tests of cells growth independence during exponential growth phase

After the exponential growth phase, the cell growth slows down or even stops (as shown in Figure 2.1A and B), which are due to limited space, nutrients or cumulated waste. These factors can be attributed to the indirect cell interactions due to their competitions for the limited resources in the environment. Since this carrying capacity effects has been well known, we want to know if cells influence each other's growth before the cell-environment interaction kicks in. However, it is difficult to measure the existence of extracellular signals directly. We would like to induce its existence indirectly: if there exist no strong extracellular interactions during exponential growth phase, the behavior of cells should grow almost independently. Although we only knew the population growth, i.e. the cell area, without detailed information of each cell, we can still statistically test if cells grow independently. In the following sections, we would show that cells did not grow independently, therefore there exists strong extracellular interactions during the exponential growth phase. This implied that leukemia cells might affect each other's growth via some unknown cytokine or chemokine signals, which will be our research in the future.

Statistical analysis of independency of the overall per capital growth rates of the cell subpopulations

We tested whether there are differences of the per area growth rates for three groups with different initial seeding cell numbers. First, we picked out all data points with value between 5 and 50 When cell growth is almost exponential. For each cell area a_i at day i, we calculated the growth rate $a_{i+1}/a_i - 1$, where a_{i+1} is the value of the same well at the next day. If cells are independent, then the growth rates of the wells with different initial cell numbers should be the same. The distributions of growth rates and data points numbers for three groups were shown in Figure 2.1 D. The population average of the growth rate of 10-seeding cell group was the highest while the one of 1-seeding cell was the lowest.

It was not proper to just average over all growth rates, since they have different areas (weights). See Table 2.3 for weighted mean and variance of per area growth rates for different seeding cell

numbers.

To determine whether these mean per area growth rates are statistical significantly different, we applied weighted Welch's t-test [50]. See Table 2.4 for results. We can see that for significance level 0.001, the growth rates for any two groups are different. This implies that even with the same area range, the grow rates of different groups are different. See Figure 2.2 C for plot of growth rate vs. area. We can see that the average growth rates of different groups are different.

Statistical analysis of independency of the daily cell numbers of different cell subpopulations

We tested whether the normalized daily cell numbers of different groups are the same using various statistical tests. Here the cell numbers are measured by the area occupied by HL60 cells in each well. Between day 8 and day 13, the cell growth of all groups were almost exponential. If there were no interactions among cells, we should expect that the ratio of the cell numbers for three groups should be the same as their initial cell numbers, namely 10:4:1. We picked the cell numbers if 10-seeding-cells and 4-seeding-cells group between day 8 and day 13 and normalized them based on their initial cell numbers. Then we applied Welch's t-test [156] to check if they have the same cell numbers for each day. Since we conducted multiple comparison, we could use the false discovery rate to evaluate these p-values. The total cell numbers in different days are positively related, therefore we can use Benjamini-Hochberg-Yekutieli procedure [14]. Here we calculated the p-values from Welch's t-test, and the reference values which help to determine which hypotheses should be rejected at significance level 0.05. See Table 2.5 for results. We can see that at significance level 0.05, the normalized cell number for 1 and 10 seeding cells groups are different in day 10 to day 13. This implies that there might be interaction among cells. See Figure 2.2 A for visualized test results.

We could also use Bonferroni correction [30], which produces a single p-value for multiple comparison. We could reject the null hypothesis that the mean areas are the same for all days between day 8 and day 13 for 4- and 10-seeding-cells groups (significance level 0.00002) and 1- and 10-seeding-cells groups (significant level 0.0009).

Statistical analysis of independency of the variances of daily growth rates of different cell subpopulation

We tested whether the variances of daily cell numbers of different groups are the same using the Brown-Forsythe test [18]. We use the case of 4-seeding-cells group to demonstrate the principle of our statistical test. Assuming that X_1, X_2, X_3, X_4 are the cell numbers at day n from each of initial four cells, if there are no interactions between cells, X_1, X_2, X_3, X_4 can be considered as four independent, identically distributed random variables. We know that the variance of the sum of independent variables is the same as the sum of variance of the variables, i.e. $\text{var}(X_1 + X_2 + X_3 + X_4) = \text{var}(X_1) + \text{var}(X_2) + \text{var}(X_3) + \text{var}(X_4)$. On the other hand, for 1-seeding-cell group, let Y be the area at day n. Since they both came from the same clone of leukemia HL60 cells, if we assume that no interactions exist between the cells, then Y and X_1 are identically distributed. So we have $\text{var}(X_1 + X_2 + X_3 + X_4) = 4\text{var}(Y)$. It means that we can divide the variance of each group by their initial seeding cell number, then compare whether they are the same to determine if cells were growing independently. We applied the Brown-Forsythe test [18] for the variances of daily growth rates normalized by their initial seeding cell number in the exponential growth phase between day 8 and day 13. We calculated the p-values and set the threshold of significance to be 0.05 (see Table 2.6 for details). We found that the variances of 1- and 10-seeding-cell groups are significantly different from day 10 to day 13 (see Figure 2.2 B for detailed results).

From Bonferroni correction, we could reject the null hypothesis that the variances are the same for all days between day 5 and day 10 for 4- and 10-seeding-cells groups (significance level 0.0003), and 1- and 10-seeding-cells groups (significance 0.0002).

2.2.4 Statistical analysis of the inheritance of the subpopulation growth rates of HL60 cells

We separately reseeded 32 samples with 70 cells each into 96-well plate from each group of fast and moderately growing cells. Fast growing cells occupied the whole well while the moderately growing cells occupied about the half well after each of them grew for 2 weeks from one-seeding cell. The total number of possible group combinations (32 vs. 32) is $\binom{64}{32}$, among which there were

6 combinations to have the same mean time for fast wells, and 1 division to have smaller mean time for fast wells, therefore the probability for the random division mean time to be no larger than sample mean time is $7/\binom{64}{32}$. We executed the permutation test [60], and found that the time needed to exceed one half for fast wells is significantly shorter than that of moderate wells, with p-value $2 \times 7/\binom{64}{32} = 7.6 \times 10^{-18}$.

2.2.5 Branching processes model

From Figure 2.1 A, it was obvious that cell growth rates from different wells varied a lot, especially for 4- and 1-seeding cell groups. We constructed a stochastic model with branching processes to capture the large variations of cell growth patterns and performed the simulations to compare with experiment results. Based on the experiment results, we assumed that there existed three phenotypes: fast, moderately and slowly growing cells. Fast growing cells are common in exponential phase wells of 4- and 10-seeding-cells group, and moderately growing cells and slowly growing cells are only common in 1-seeding-cell group, in which slowly growing cells were minority. We also assumed that there existed the transitions between different phenotypes based on our experiment data. For example, the transition clearly happened from the slow growing cells to the moderately growing ones between day 10 and day 15 for one-seeding-cells group, as shown in 2.4. It was interesting to notice that these 4 wells started to take off after a considerably delay compared with the average of the same group and they did not slow down after taking off. In order to fully describe the population behavior of three phenotypes, multi-type discrete-time branching process is applied in our model. As shown in 2.4.11, three phenotypes of fast, moderately and slowly growing cells have their growth rates g_F, g_M, g_S and death rates d_F, d_M, d_S. They transition to each other with constant transition rates.

Denote the population of fast, moderate and slow phenotypes at day n by $F(n), M(n), S(n)$. Then the population at day $n+1$ is:

$$(F(n+1), M(n+1), S(n+1)) = \sum_{i=1}^{F(n)} A_i + \sum_{j=1}^{M(n)} B_j + \sum_{j=1}^{S(n)} C_k,$$

where A_i, B_j, C_k are independent for all i, j, k.

A_i represents the descendants of a fast cell i after one day. It equals $(2, 0, 0)$ with probability g_F, $(0, 1, 0)$ with probability r_{FM}, $(0, 0, 1)$ with probability r_{FS}, $(0, 0, 0)$ with probability d_F, $(1, 0, 0)$ with probability $1 - r_{FM} - r_{FS} - g_F - d_F$.

B_j represents the descendants of a moderate cell j after one day. It equals $(0, 2, 0)$ with probability g_M, $(1, 0, 0)$ with probability r_{MF}, $(0, 0, 1)$ with probability r_{MS}, $(0, 0, 0)$ with probability d_M, $(0, 1, 0)$ with probability $1 - r_{MF} - r_{MS} - g_M - d_M$.

C_k represents the descendants of a slow cell k after one day. It equals $(0, 0, 2)$ with probability g_S, $(0, 1, 0)$ with probability r_{SM}, $(1, 0, 0)$ with probability r_{SF}, $(0, 0, 0)$ with probability d_S, $(0, 0, 1)$ with probability $1 - r_{SM} - r_{SF} - g_S - d_S$.

In Methods section, We explained why branching process was the most suitable for modeling this experiment and why there should be multiple phenotypes with transitions among them. We also did parameter scanning to show that our model was quite robust to perturbations on model parameters (see details in Table 2.7).

The simulation results were shown at the right panels of Figure 2.1 A, B and C in comparison with the experiment data at the left panels. Our model qualitatively captured the growth dynamics patterns of the groups with different initial seeding cell numbers. For example, in Figure 2.1 A, all wells in 10-seeding-cells group in our model grew quickly until they reached carrying capacity. Similar to the experiments, some wells in 1-seeding-cell group in our model never grew while others began to take off after 14 days. In Figure 2.1 C, the growth rates of all cell populations decreased with total cell number, which simply represent that cells reached the carrying capacity of the wells. When wells are less than half full (cell number ¡ 20000), most wells in 10-seeding-cells group grew faster than 1-seeding-cell group even when they had the same cell number. The reason is that, for 10-seeding-cells group, the initial 10 cells are very likely to contain at least one fast cell. This cell will dominate, and the growth rate will be the same with fast cells. For 1-seeding cell group, the initial single cell is possible to be moderate or slow, therefore its descendants are likely to keep the slower growth rate.

2.3 Discussion

As many recent single-cell level data have shown, cancer has many distinct subpopulations with transitions and interactions among them, which influence cancer cell growth, death and migration, as well as many other features. Investigating the "non-genetic heterogeneity hypothesis of cancer cells" is of paramount importance to understand cancers biology. In this paper we showed that the Leukemia HL60 cell line has a cell growth heterogeneity, and the initial cell number influences the overall growth rate of the HL60 cell population. This result might be explained by the multiple distinct cell states and the transitions between them, therefore we introduced multiple phenotypes in our stochastic model with the different growth rates and mutual transitions. In this model we assumed the existence of three phenotypes: fast, moderately and slowly growing cells. The model we built is able to recapitulate the key features in the experimental data.

We also found that a well starting with 1 cell is essentially different from a well starting with 10 cells even when they both reach the same amount of cell number. It seemed that the difference (memory) of cell growth dynamics was long lasting. One possible explanation of the difference in growth rates might be due to the presence/absence in the environment (media) of some soluble factors released by the cells (quorum sensing).

The initial statistical tests on cell growth independence revealed that there must exist cell interactions. However, we found that simple interactions could not explain why the wells from different seeding cell numbers have different growth rates even when they reached the same cell number. Thus we developed a mathematical model with three phenotypes. Currently we do not have direct experimental evidence for it. But this might be tested by performing a single cell RNASeq over the population and compare the results with the composition of the single-cell derived population.

Another alternative would be to build a model with only two phenotypes and some complicated interactions which might lead to the long memory of growth rates on initial cell numbers. We interpreted the difference of growth rates of wells with the same cell areas as the difference of proportions of phenotypes. There are also some alternatives, such as true memory of history in cells, or alteration of the composition of the media. To better understand which of the alternatives

is more likely, we performed a resampling of the cells after having assessed the phenotype (fast or moderately growing). All the wells derived from the fast phenotype are significantly faster then the wells derived from the moderate phenotype, suggesting that the hypotheses of a different proportion of phenotype seems more likely.

To conclude, we have carried out an experimental measurements and mathematical modeling analysis on heterogeneity of cancer cell stochastic growth. We used HL60 as cellular model. The percentage of cancer stem cells in a cancer cell line is generally known/reported; it would be interesting to see whether the percentage of cancer stem cell correlates with the proportion of slow wells from single cell population. A comparative study of the fast and moderate growing single-cell-derived population kinetics in term of heterogeneity will provide further insights into the nature of cancer tissue as a living organism.

2.4 Methods

2.4.1 Cell and sorting

HL60 were maintained in IMDM wGln, 20% FBS(heat inactivated), 1% P/S at a confluence between $0.3 * 10^6$ and $2.5 * 10^6$ cells/ml (GIBCO). Cells were always handled and maintained under sterile conditions (tissue culture hood; 37°C, 5% CO_2, humidified incubator). The day of the experiment cells are collected, washed two times in PBS and stained for vitality (Trypan blue GIBCO). The population of cells is first gated for morphology and then for vitality staining. Only Trypan negative cells are sorted (BD FACSAria II). The cells are sorted in a 384 well plate; IMDM wGln, 20% FBS(heat inactivated), 1% P/S (GIBCO).

2.4.2 Monitoring of the cell growth

Cell growth was monitored using a Leica microscope (heated environmental chamber and CO_2 levels control) with motorized tray. Two days after the sorting (first data-point), each day (at the same time), the 384 well plate is placed inside the environmental chamber. The images are acquired everyday.

2.4.3 Setup of fast and moderately growing cells experiment

HL60 leukemia cells were cultivated for two weeks, and then we chose one full well and one half full well. We supposed the full well was dominated by fast type cells, and the half full well was dominated by moderately growing type cells, which had smaller growth rates. We reseeded cells from these two wells and cultivated them in two 96-well (row A-H, column 1-12) plates. In each plate, B2-B11, D2-D11, F2-F11 wells started with 78 fast cells, C2-C11, E2-E11, G2-G11 wells started with 78 moderate cells. Row A, H, column 1, 12 had no cells, no media, and we found that wells in row B, G, column 2, 11, which were the outmost non-empty wells, evaporated much faster than inner wells. Therefore, the growth of cells in those wells were much slower than inner wells. Hence we only considered inner wells, where D3-D10, F3-F10 started with fast cells, C3-C10, E3-E10 started with moderate cells, totally 32 fast wells, 32 moderate wells. During the experiment, no media was added. Each day, we observed those wells to check whether their areas exceeded one half of the whole well. The experiment was terminated after 20 days.

2.4.4 Welch's t-test

Welch's t-test, or unequal variances t-test, is a two-sample location test which is used to test the hypothesis that two populations have equal means without the assumption that two populations have the same variance [156].

First, define statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

where $\bar{X}_i$, s_i^2, N_i are the sample mean, sample variance and sample size of group i, $i = 1, 2$.

The degree of freedom is

$$\nu = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{s_1^4}{N_1^2\nu_1} + \frac{s_2^4}{N_2^2\nu_2}}$$

where $\nu_1 = N_1 - 1$, $\nu_2 = N_2 - 1$.

Then t satisfies the t-distribution with degree of freedom ν.

2.4.5 *Weighted Welch's t-test*

Now assume for group i ($i = 1, 2$), the sample size is N_i. However, the jth sample is the average of c_i^j i.i.d. variables. Let X_i^j be the observed average for the jth sample. Weighted Welch's t-test is used to test the hypothesis that two populations have equal means, while sample values have different weights [50]. Define

$$\bar{X_i}^W = (\sum_{j=1}^{N_i} X_i^j c_j)/(\sum_{j=1}^{N_i}) c_j$$

$$s_{i,W}^2 = \frac{N_i[\sum_{j=1}^{N_i}(X_i^j)^2 c_j]/(\sum_{j=1}^{N_i} c_i^j) - N_i(\bar{X_i}^W)^2}{N_i - 1}$$

Then substitute $\bar{X}_i$ by $\bar{X_i}^W$, and s_i^2 by $s_{i,W}^2$ in normal Welch's t-test to get t and ν. Now t satisfies the t-distribution with degree of freedom ν.

2.4.6 *Brown-Forsythe test*

Brown-Forsythe test is a multi-sample test which is used to test the hypothesis that two populations have equal variances [18]. Denote the observation of group i by y_{ij}. Let $z_{ij} = |y_{ij} - \tilde{y}_j|$, where $\tilde{y}_j$ is the median of group j. Let us define:

$$F = \frac{(N-p)}{(p-1)} \frac{\sum_{j=1}^{p} n_j(\bar{z}_{\cdot j} - \bar{z}_{\cdot\cdot})^2}{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\bar{z}_{ij} - \bar{z}_{\cdot j})^2},$$

where p is the number of groups, n_j is the number of observations in group j, and N is the total number of observations. Also, $\bar{z}_{\cdot j}$ are the group means of the z_{ij} and $\bar{z}_{\cdot\cdot}$ is the overall mean of the z_{ij}. Then F satisfies F-distribution $F(k-1, N-k)$.

2.4.7 *Benjamini-Hochberg-Yekutieli procedure*

When we have several hypothesis tests and corresponding p-values, Benjamini-Hochberg-Yekutieli procedure can determine which hypotheses should be rejected. The main idea is to control the false discovery rate [14]. Assume we have m null hypotheses $H_1, \cdots, H_m$, and m corresponding

p-values $p_1, \cdots, p_m$. Sort the p-values in ascending order $p_{(1)}, \cdots, p_{(m)}$. Next, find the largest k such that

$$p_{(k)} \leq \frac{k}{mc(m)}\alpha,$$

where $c(m) = 1$ for positively correlated tests. Then at significance level α, we reject the null hypotheses $H_{(i)}$ for $i = 1, 2, \cdots, k$.

2.4.8 Bonferroni correction

When we have several hypothesis tests and corresponding p-values, we can also use Bonferroni correction [30]. Assume we have m null hypotheses $H_1, \cdots, H_m$, and m corresponding p-values $p_1, \cdots, p_m$. Pick out the smallest p-value p_k, then at significance level $\alpha = p_k \cdot m$, we can reject at least one hypothesis of $H_1, \cdots, H_m$.

2.4.9 Permutation test

Permutation test is a non-parametric method to test whether two samples have the same mean [60]. Assume we have two samples $\{x_1, \cdots, x_m\}$, $\{y_1, \cdots, y_n\}$. Consider the null hypothesis: the mean of x and y are the same. For these samples, we calculate a feature, for example the mean difference: $\mu_0 = \frac{1}{m}\sum x_i - \frac{1}{n}y_j$. Then we randomly divide these $m + n$ samples into two groups with size m and n: $\{x'_1, \cdots, x'_m\}$, $\{y'_1, \cdots, y'_n\}$, such that each permutation has equal probability. For these new samples, we can calculate the same feature, namely the mean difference: $\mu'_0 = \frac{1}{m}\sum x'_i - \frac{1}{n}y'_j$. Then the two-sided p-value is defined as

$$p = 2\min\{\mathbb{P}(\mu_0 \leq \mu'_0), 1 - \mathbb{P}(\mu_0 \leq \mu'_0)\}.$$

If μ_0 is an extreme value in the distribution of μ'_0, then the two sample means are different. If the sample size is large, generally the above probability is difficult to calculate precisely, unless the two samples are extremely different.

2.4.10 Model building

The traditional method of population dynamics is ordinary differential equations (ODE), which is deterministic and has continuous variables. In the early stage of our experiments, the cell population is relatively small, therefore it is not appropriate to treat population as continuously changing variables. Also, with such small population, the stochasticity in cell activity cannot be averaged out, it is better for us not to use deterministic model. On single cell level, each cell is not exactly the same with another cell. We cannot consider every nuance among cells and apply different deterministic models for each cell, especially when we only have the information of initial seeding cell numbers. For this purpose, stochastic processes are more suitable, since they can treat the unobservable nuance as stochasticity while being capable to capture the dynamical behavior of population.

Since the population is not large enough in our experiments, our model should be purely stochastic, without any deterministic growth. Since we only consider finite (three) phenotypes, and concern more on concrete population size, Poisson processes are not suitable. Markov chains can partially describe the proportions under some conditions [72], but we can measure population size, not just their ratios, therefore Markov chains are not necessary. Branching processes can describe the population size of multiple phenotypes with symmetric and asymmetric division, transition and death. Also, the parameters can be set to temporal and spatial inhomogeneous, which is convenient. Therefore we utilized branching processes in our model. Now we need to analyze the possible mechanism of leukemia cell growth.

From Table 2.1, we can see that 14 of 80 wells in 1 seeding cell group never grow and all of them are empty at day 20. There are 4 wells which contain more than 10 cells at day 9, but the populations are always under 600. There are another 4 wells which contain more than 10 cells at day 9 and populations at day 19 are under 1,000. For these wells, the populations at day 23 are all larger than 5,000. In comparison, a typical well in 1 seeding cell group has more than 100 cells at day 9, and almost full at day 19 and 23. Besides, no well in 4 and 10 seeding cells groups never grow. Such contrast illustrates that, in some wells with 1-seeding-cell, when they grow into 10

cells, they are still not the same as the wells starting with 10 cells. This implies that there should be heterogeneity in growth rates among cells.

To explain the late growth and no growth, we can divide cells into at least two phenotypes, fast-growing and slow-growing. For 4- and 10-seeding-cells group, since they start with more initial cells, they are very likely to contain at least one fast cell. Such cell leads to fast growth of whole population. For 1-seeding-cell group, if the only initial cell is in slow phenotype, then the population will remain at a very low level or even die out. After a long waiting time, a slow-growing cell can transform into a fast-growing cell, which leads to the late fast growth.

With such two-phenotypes model, we can explain the late growth and never growth in 1-seeding-cell group. However, in such model, once the population starts to grow exponentially fast, it implies that the population is dominated by fast-growing cells. Therefore, when the cell area is large enough (> 5), almost all the cells are in the same phenotype, and the behavior of whole population should only rely on current area. However, as we can see in Figure 2.2 C, at the same area (5–35), wells with 10 initial cells have significantly larger growth rates than wells with 1 initial cell.

To explain this difference, we can introduce another phenotype, moderately-growing. The growth rate of this phenotype is between fast-growing type and slow-growing type, such that starting with this phenotype, the population will grow into exponential phase without waiting too long, but still be slower than the fast-growing type. For 1-seeding-cell group, the only initial cell is likely to be in moderately-growing phenotype, so that at areas around 5-35, the growth rate is around 0.4–0.6. For 4- and 10-seeding-cells groups, they are likely to contain at least one fast-growing cell, therefore the growth rate at areas around 5–35 is around 0.6–1. For area larger than 50, the population is too crowded, and the growth rates of fast-growing and moderately-growing cells tend to be the same.

Another possible explanation for such difference in growth rates is that the population will be 10 small colonies when starting from 10 initial cells. while starting from 1 initial cell, the population will be 1 large colony. With the same area, the total perimeter of 10 small colonies should be much larger than that of 1 large colony. Since total perimeter describes the space of

growth, it should have positive relationship with growth rate. However, this explanation is wrong. We carefully checked the photos, and found that almost all wells will produce 1 large colony with nearly the same shape, and there is no significant relationship between the perimeter and the cell population growth rate. We assumed that the growth rate of each phenotype is a decreasing function of total cell number regardless of colony shape. Three phenotypes can transform between each other. Otherwise the fast-growing phenotype will dominate, and the other two phenotypes will be extinct.

Since the only data we have is cell area, we are not able to verify most assumptions in our model, not to mention determining the values of parameters. We chose the values of parameters ad hoc and illustrated that our model can reproduce experimental phenomena, especially Figure 2.1 A and Figure 2.1 C (See Model Details for parameter scan results) .

2.4.11 Model details

The simulation time interval is half day, but we will only utilize the results in full days. For each initial cell, the probability of being fast, moderately and slowly growing phenotypes are 0.3, 0.5, 0.2. Such proportions might validated from FACS sorting experiment if we have cell surface biomarkers to distinguish them.

Each half day, for a fast-growing cell, it has probability 0.001 to die, probability s_1 to divide. The division produces two fast cells. Denote the total cell number of previous day as N, then $s_1 = (1 - N^2/C^2)/2 + \delta$, where $C = 40000$ is the carrying capacity, δ is random, satisfies the uniform distribution on $[-0.1, 0.1]$, and it is the same for all cells in the same well. If $s_1 > 1$, set $s_1 = 1$. If $s_1 < 0.001$, set $s_1 = 0.001$.

Each half day, for a moderately-growing cell, it has probability 0.001 to die, probability s_2 to divide. The division produces two moderate cells. $s_2 = s_1/1.5$. Each half day, for a slow-growing cell, it has probability 0.001 to transform into fast-growing type, probability 0.001 to die, probability s_3 to divide. The division produces two slow-growing cells. $s_3 = s_1/3$.

Since we only have cell area data, it is difficult to estimate every parameter in our model. Therefore, we ran a parameter scan, namely checking the performance of our model with different

Table 2.1: Population of some wells in 1 seeding cell group, and well 200 is for comparison. Some numbers are approximated. These wells illustrated different behaviors from those wells starting with 10 or 4 cells. Such difference implies that cells from wells with different initial cell numbers are essentially different. Therefore, we had to consider cell type heterogeneity.

Well number	162,167,170, 176,177,179, 182,183,186, 201,234,236, 239,240	165	166	178	211	163	181	192	204	Comparison (200)
Initial number	1	1	1	1	1	1	1	1	1	1
Day 9	<10	89	36	43	16	12	44	25	21	130
Day 19	0	350	120	170	200	300	550	800	600	40000
Day 23	Empty	500	150	200	400	5000	5500	9000	6000	40000

Table 2.2: The distribution of time needed for each well to reach one half area. We reseeded equal numbers of fast cells and moderate cells, and cultivated them under same environment, to compare their inherent growth rates. The results show that fast cells, even reseeded into new media, still grew faster.

Time to reach one half area	11	12	13	14	15	20	>20
Fast wells	26	2	1	2	1	0	0
Moderate wells	0	0	0	1	1	25	5

parameters. See Table 2.7 for results. We can see that within a large range of parameters, our model can replicate the experiment results. Especially, the simulation results are not sensitive with transition rates, therefore we set most transition rates to 0.

Table 2.3: Weighted mean and variance of per area growth rates. The growth rate is positively related to seeding cell number.

Seeding cell number	10	4	1
Weighted mean of per area growth rate	1.6181	1.5418	1.4875
Weighted variance of per area growth rate	0.6666	0.6104	0.6103

Table 2.4: Welch's t-test on the mean per area growth rates during exponential phase. Groups with different seeding cell numbers have significant difference.

Seeding cell numbers	10 vs. 4	10 vs. 1	4 vs. 1
Welch's t-test p-value	2.12×10^{-8}	$< 10^{-12}$	5.35×10^{-5}

Table 2.5: Welch's t-test for same mean. For each day between Day 8 and Day 13, we used Welch's t-test to check whether the mean areas of wells from different initial cells are proportional to initial cell numbers. The test results show that the differences are significant, which implies that cells are not independent.

4 and 10 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.5648	0.9758	0.1618	0.0203	0.0030	0.0000
Reference	0.0417	0.0500	0.0333	0.0250	0.0167	0.0083
Significant				Yes	Yes	Yes
1 and 10 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.3792	0.1232	0.0271	0.0055	0.0013	0.0001
Reference	0.0500	0.0417	0.0333	0.0250	0.0167	0.0083
Significant			Yes	Yes	Yes	Yes
1 and 4 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.2809	0.1458	0.1220	0.0655	0.0269	0.0167
Reference	0.0500	0.0417	0.0333	0.0250	0.0167	0.0083
Significant						

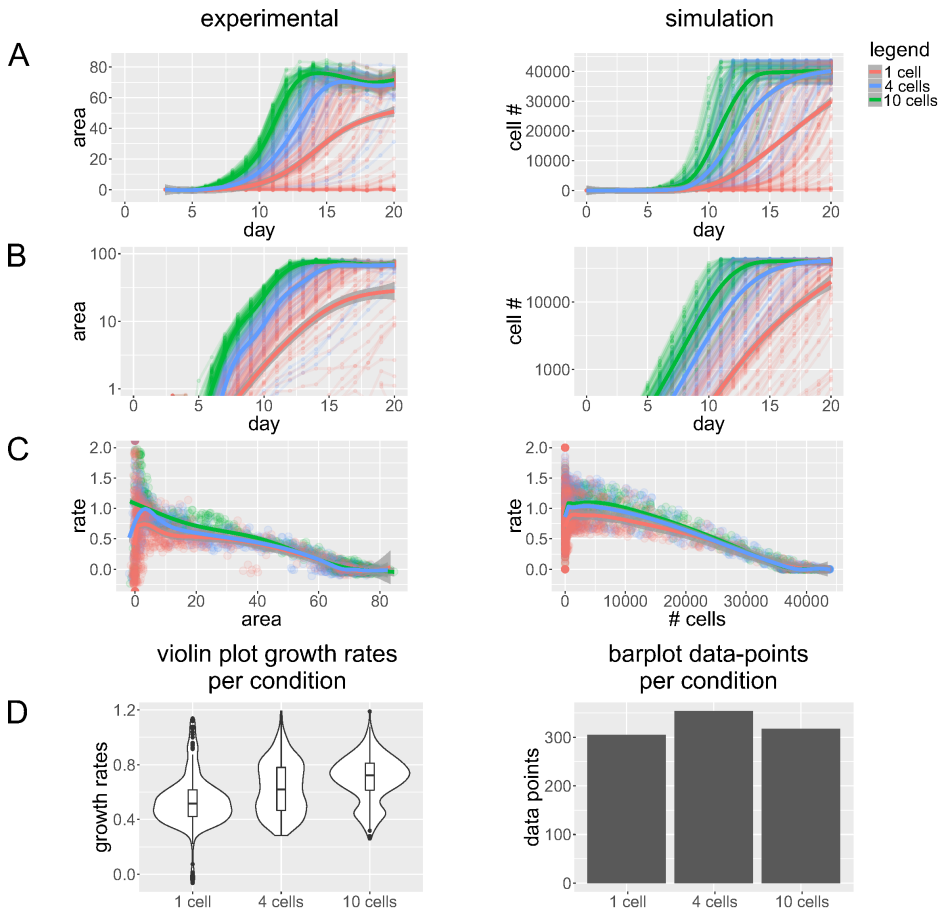


Figure 2.1: The comparison of cell growth dynamics between experimental data and simulation: (A) The area of cell occupancy (cell numbers) vs. time was presented in linear scale. Each color represented a different initial cells number. (B) The same data presented in (A) are presented in $\log_{10}$ scale. The cell growth differences among different groups were clearly visible. (C) The violin plot of cell growth rate distribution for each different seeding group. (D) The bar-plot showing the number of datapoint per initial seeding group

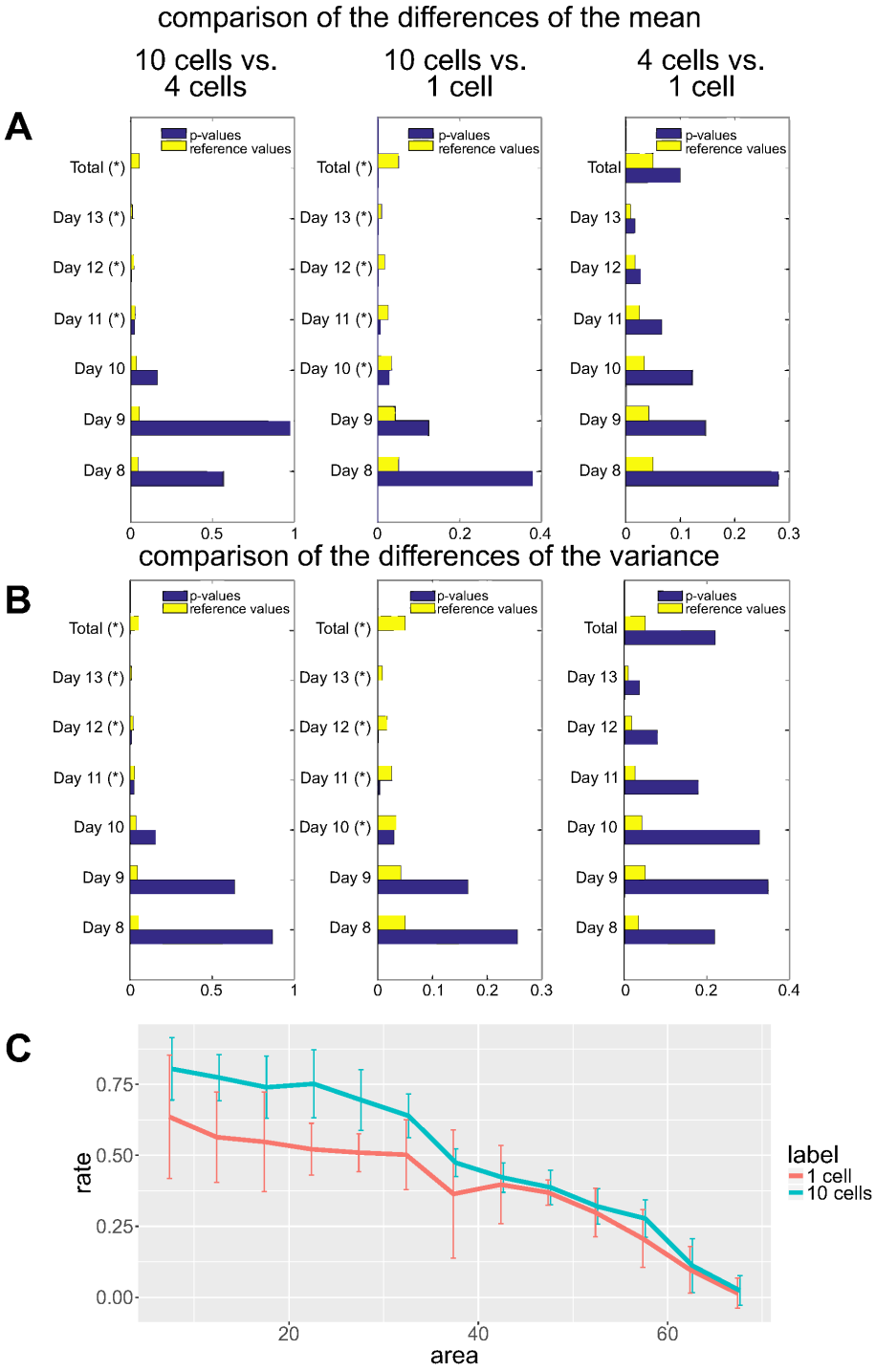


Figure 2.2: The statistical test of the means and the variance of daily cell growth for groups with different seeding cell numbers. (A) p-value for the difference of the mean areas by day. The significant differences are identified by (*). (B) p-value for the difference of the variances by day. The significant differences are identified by (*). (C) Per area growth rates of 10 seeding cells group and 1 seeding cell group with error bars.

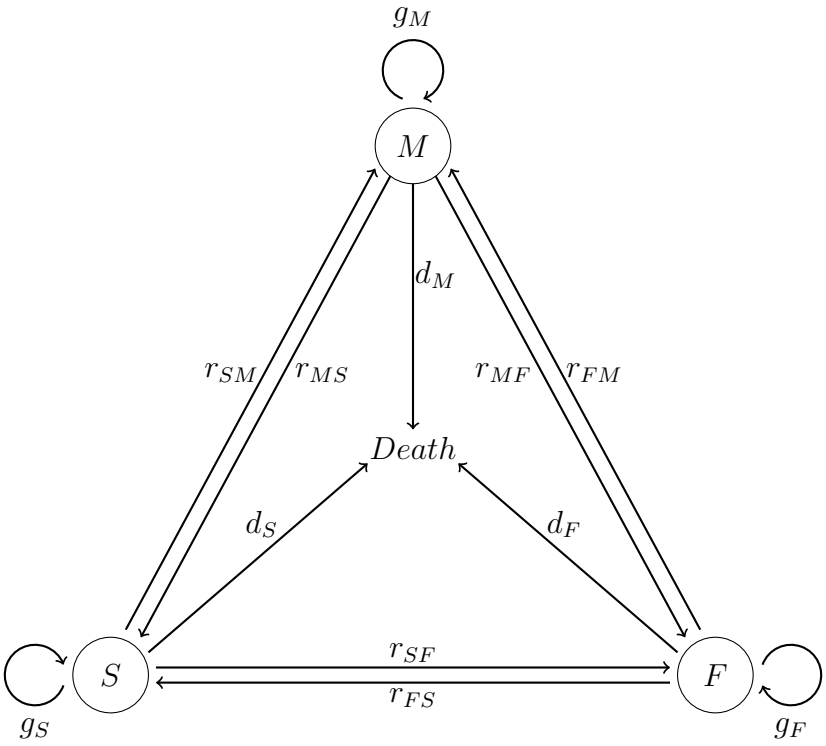


Figure 2.3: Illustration of the branching processes model. F, M, S represent fast, moderate, slow phenotypes. g_F, g_M, g_S are division rates. d_F, d_M, d_S are death rates. $r_{FM}, r_{MF}, r_{FS}, r_{SF}, r_{MS}, r_{SM}$ are transition rates between two phenotypes.

Table 2.6: Brown-Forsythe test for same variance. For each day between Day 8 and Day 13, we used Brown-Forsythe test to check whether the variance of areas of wells from different initial cells are proportional to initial cell numbers. The test results show that the differences are significant, which implies that cells are not independent.

4 and 10 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.8692	0.6362	0.1543	0.0220	0.0073	0.0000
Reference	0.0500	0.0417	0.0333	0.0250	0.0167	0.0083
Significant				Yes	Yes	Yes
1 and 10 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.2558	0.1647	0.0294	0.0027	0.0008	0.0000
Reference	0.0500	0.0417	0.0333	0.0250	0.0167	0.0083
Significant			Yes	Yes	Yes	Yes
1 and 4 seeding cells groups						
Time	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13
P-value	0.2188	0.3478	0.3275	0.1791	0.0799	0.0360
Reference	0.0333	0.0500	0.0417	0.0250	0.0167	0.0083
Significant						

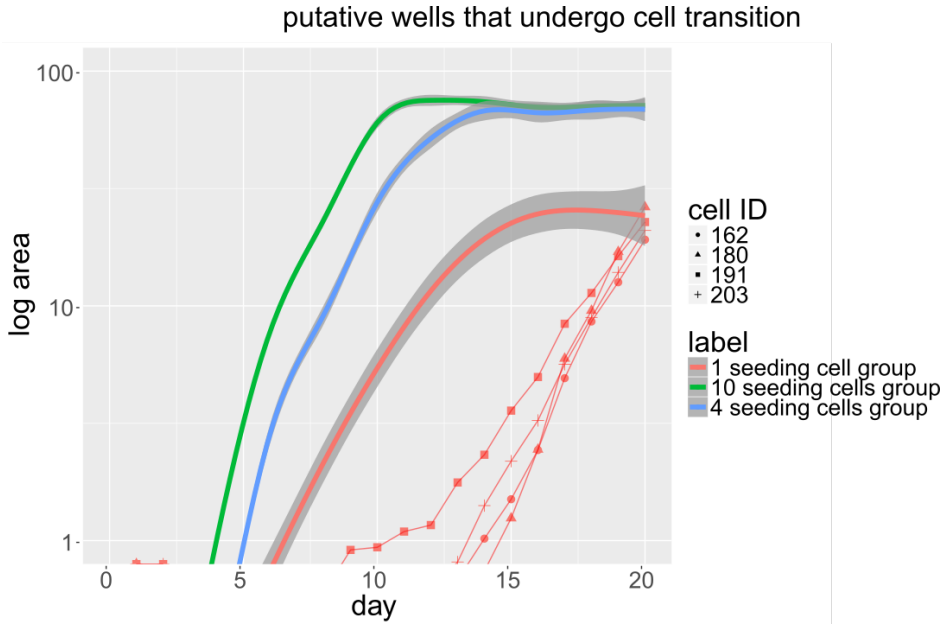


Figure 2.4: Putative wells underwent cell growth state transition. In this picture we highlighted the wells which transitions most likely happened. The wells all belong to the 1-seeding- cell group, which took considerably longer to take off compared to the population average.

Table 2.7: Performances of our model with different parameters. Here we adjusted the parameters of our model in a wide range, and observed whether the model could still reproduce some important features in the experiment. This parameter scanning shows that our model is robust under perturbations on parameters.

Fast cell proportion	Moderate cell proportion	Slow cell proportion	Death rate	Transition rate from slow cell to fast cell	10 seeding cells group all saturate	Exist late growth wells in 1 seeding cell group	Exist no growth wells in 1 seeding cell group	Growth rate difference at same population between 1 and 10 seeding cells groups
0.3	0.5	0.2	0.001	0.001	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0	0.001	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0.02	0.001	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0.05	0.001	No	Yes	Yes	Yes
0.3	0.5	0.2	0.001	0	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0.001	0.01	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0.01	0.05	Yes	Yes	Yes	Yes
0.3	0.5	0.2	0.02	0.1	Yes	Yes	Yes	Yes
0.4	0.5	0.1	0.001	0.001	Yes	Yes	Yes	Yes
0.2	0.5	0.3	0.001	0.001	Yes	Yes	Yes	Yes
0.2	0.4	0.4	0.001	0.001	No	Yes	Yes	Yes
0.3	0.4	0.3	0.001	0.001	Yes	Yes	Yes	Yes
0.4	0.4	0.2	0.001	0.001	Yes	Yes	Yes	Yes
0.5	0.3	0.2	0.001	0.001	Yes	Yes	Yes	Yes
0.4	0.4	0.2	0.01	0.05	Yes	Yes	Yes	Yes
0.3	0.2	0.5	0.02	0.1	Yes	Yes	Yes	Yes
0.2	0.1	0.7	0.02	0.1	Yes	Yes	Yes	No
0.2	0.1	0.7	0.001	0.001	No	Yes	Yes	No
0.5	0.5	0	0	0.001	Yes	Yes	No	Yes
0.9	0.1	0	0	0.001	Yes	No	No	No

2.5 Appendix

2.5.1 Parameters estimation of birth-death process

Consider a linear birth-death process which describes a cell population dynamics. All cells are independent. Each cell lives an exponential time with parameter $(b+d)$, and then with probability $b/(b+d)$ to divide, probability $d/(b+d)$ to die. Here b is the birth rate, and d is the death rate. If we consider the total number of cells, then it is a continuous-time Markov process, takes non-negative integers, and the transition rates are $q(n, n+1) = nb$, $q(n, n-1) = nd$. 0 is the only absorbing state. In the following we assume that $b > d$. Otherwise the cell population will die out with probability 1.

Now consider an experiment. We cultivate m groups of cells, each starts with n_0 cells. Then count the cell numbers at time $t_1, t_2, \cdots, t_k$. Assume they fit the birth-death process. We want to estimate the birth rate and death rate.

An obvious way is to calculate the maximal likelihood estimator (MLE). In this method, we consider all possible rates, and find the one which is most likely to produce such data, namely maximize the likelihood function. However, when data is not very small (for example, 30 groups with 5 observations), the likelihood function is very complicated, such that we cannot analytically find the MLE. The only way is to use brutal force to search. Also, we need very high accuracy. Thus the calculating needs very long time.

The transition probability of starting with n_0 cells, and having $n > 0$ cells at time t is

$$\sum_{j=0}^{\min(n_0,n)} \binom{n_0}{j} \binom{n_0+n-j-1}{n_0-1} \alpha^{n_0-j} \beta^{n-j} (1-\alpha-\beta)^j$$

where $\alpha = d[e^{(b-d)t}-1]/[de^{(b-d)t}-d]$, $\beta = b[e^{(b-d)t}-1]/[de^{(b-d)t}-d]$. When $n = 0$, the (extinction) probability is α^{n_0} [10].

Here we give a quick method. We know that the expectation of cell number at time t_i is $\mathbb{E}[N(t_i)] = n_0 \exp[(b-d)t_i]$. Then the mean cell number at each observation should be on an exponential curve. For each observation, we can use the mean cell number to calculate $b-d$. Then

the average of $b-d$ over all observations is a good estimation of $b-d$.

Starting from n_0 cells, the variance at time t_i is [10]

$$\text{Var}(t_i) = \frac{n_0(b+d)}{b-d} e^{(b-d)t_i}[e^{(b-d)t_i} - 1]$$

Once we know the estimation of $b-d$, the variance is a linear function of $b+d$. Using the same method above, we can get an estimation of $b+d$. Then we get the estimation of both birth and death rates.

2.5.2 Mean and variance estimation for single trajectory

Assume all cells are independent, but now we do not assume the population satisfies a linear birth-death process. Also, this time we only have one group of cells, with several observations with equal intervals. For example, we observe at $t = 0, 1, 2, 3, 4$, and see $1, 2, 4, 6, 9$ cells. Now the question is, how can we describe the dynamics of this process?

For one cell, denote the number of its living descendants after one time unit as X. Then the above example can be rewrite as

$X_1 = 2$, $X_2 + X_3 = 4$, $X_4 + X_5 + X_6 + X_7 = 6$, $X_8 + X_9 + X_{10} + X_{11} + X_{12} + X_{13} = 9$, where all X_i are independent.

We can use these to estimate the mean and variance of X. In fact, we will try to find the "Minimal Variance Unbiased Estimator" (MVUE). Unbiased means the estimator does not have systematic bias. Minimal Variance means the estimator is most accurate. In the following we will assume X is a normal distribution. In general case, the following method may not give MVUE, but it is still very useful.

It is natural to think the sample mean $(2+4+6+9)/(1+2+4+6) = 21/13$ is a good estimator. In fact, this is the MVUE. Assume we know all the values of X_i, not just their sum, then the MVUE of mean is $\sum X_i/n$. Since we have less information, we cannot get a better estimator. Thus the sample mean is the MVUE.

Now consider the estimator of variance. Assume $X \sim \mathcal{N}(\mu, \sigma^2)$. The observation is $S_1 = \sum_{i=1}^{k_1} X_{1i}, \cdots S_n = \sum_{i=1}^{k_n} X_{ni}$.

Theorem 2.1. *The MVUE of variance is*

$$\frac{1}{n-1}\sum_{i=1}^{n} k_i(S_i/k_i - \bar{\mu})^2$$

where $\bar{\mu}$ is the MVUE of mean $(\sum_{i=1}^{n} S_i)/(\sum_{i=1}^{n} k_i)$.

Proof. We have $S_i \sim \mathcal{N}(k_i\mu, k_i\sigma^2)$. The joint density of S_i is

$$f(S_1, \cdots, S_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2k_i\pi\sigma^2}} \exp[-\frac{1}{2k_i\sigma^2}(S_i - k_i\mu)^2]$$

$$= h(\sigma^2) \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{S_i^2}{k_i} + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} S_i + g(\mu, \sigma^2)]$$

This is an exponential family with natural parameter $\theta = (\mu/\sigma^2, 1/2\sigma^2)$ (notice that its two components are linearly independent), and sufficient statistic $T = (T_1, T_2)$, where $T_1 = \sum_{i=1}^{n} S_i$, $T_2 = \sum_{i=1}^{n}(S_i^2/k_i)$. We know that $\mathrm{E}T_1^2 = (\sum_{i=1}^{n} k_i)^2\mu^2 + (\sum_{i=1}^{n} k_i)\sigma^2$, $\mathrm{E}T_2 = (\sum_{i=1}^{n} k_i)\mu^2 + n\sigma^2$. Then

$$\mathrm{E}\{[(\sum_{i=1}^{n} k_i)T_2 - T_1^2]/[(n-1)(\sum_{i=1}^{n} k_i)]\} = \sigma^2$$

From Lehmann-Scheffé theorem, $[(\sum_{i=1}^{n} k_i)T_2 - T_1^2]/[(n-1)(\sum_{i=1}^{n} k_i)]$, as a measurable function of T_1 and T_2, is the unique MVUE of σ^2.

Finally, we can verify that $[(\sum_{i=1}^{n} k_i)T_2 - T_1^2]/[(n-1)(\sum_{i=1}^{n} k_i)] = \frac{1}{n-1}\sum_{i=1}^{n} k_i(S_i/k_i - \bar{\mu})^2$. □

Remark 2.1. *When $k_1 = \cdots = k_n = 1$, this estimator is just the sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{\mu})^2$.*

Chapter 3

PHENOTYPIC EQUILIBRIUM AS PROBABILISTIC CONVERGENCE IN MULTI-PHENOTYPE CELL POPULATION DYNAMICS

3.1 Introduction

With the same genetic background, cell population may have different cellular phenotypes. This has been one of the major topics in the research of cell population dynamics [3, 83]. Very recently much attention has been paid to the stochastic conversions between different phenotypes [29, 57]. For example, we know that cancer stem cells can give rise to cancer non-stem cells, but cancer non-stem cells can also transform back to cancer stem cells [169, 161]. Generally, we can use a branching process (stochastic model) [66, 79, 158, 159, 160] or an ODE system (deterministic model) [99] to describe the dynamics of such cell population with multiple phenotypes. However, in many experimental settings, it is difficult or even impossible to count the total cell population [20, 159, 160]. Thus in the last fifty years, people began to consider the proportions of cell individuals with distinct phenotypes instead of the absolute numbers of cells of various phenotypes [66].

We know that through multistep accumulation gene mutations, healthy cells gradually transform to malignant cancer cells, which is the current view of cancer progression [58]. Among those mutations, most of them are neutral ("passenger mutations") and have no effect on cell proliferation. Only a small portion of mutations will bring growth advantage ("driver mutations") [17]. Since the emerging of mutations can be regarded as purely stochastic, we can model such procedure with multitype branching processes (cf. Bozic et al.'s model [17, 80]). There have been some results about the emerging time for certain number of driver mutations, and the relation between the number of emerged driver mutations and emerged passenger mutations [17, 93].

In the experiments on breast cancer cell lines, Gupta et al. [57] found that the proportion of

each phenotype will tend to a certain constant regardless of the initial population states ("phenotypic equilibrium"). They built a Markovian model, assuming that the evolution of the phenotypic proportions satisfies an n-state Markov chain, and used the ergodicity of the Markov chain to explain this phenomenon [57]. However, we find that the Markovian model is just the ODE system model under a special condition. We determine this condition and its biological meaning. Furthermore, we try to remove this condition and explain the experimental phenomenon in [57] under more general context.

In the deterministic model (ODE system), we only consider the average behavior of cell population dynamics (which requires a large initial population). However, using the stochastic model (branching processes), we can study the trajectory behavior. We prove that the proportions will converge not only on average, but also almost surely. This implies that even with a small initial population, we can still observe "phenotypic equilibrium".

In the theory of multitype branching processes, people have observed similar proportion convergence phenomenon and proved such phenomenon in several limit theorems under different conditions [78, 8, 67]. Those are possible ways to explain "phenotypic equilibrium", but those required conditions may not be satisfied in experiments. Thus we improve those limit theorems by dropping redundant conditions. We will see that the conditions we need are all biologically reasonable. Therefore, we give a stochastic explanation of "phenotypic equilibrium". This result may also be of interests to probabilists.

Generally we only consider Markovian branching processes, but sometimes the biological process is not memoryless, thus we need to consider non-Markovian branching processes. We show that under some conditions, the non-Markovian branching processes can be transformed into Markovian branching processes. Using this trick, we demonstrate similar results for non-Markovian branching processes.

In Section 3.2, we will define some notations, and give the mathematical description of our models, which is based on [79] and [67]. In Section 3.3, we will describe under which condition the deterministic model becomes the Markovian model in [57]. In Section 3.4, we will prove that under some mild conditions, the "phenotypic equilibrium" phenomenon will always happen

in the Markovian branching process model. Specifically, we will improve a limit theorem about proportion convergence in multitype branching processes. We will also apply our results to Bozic et al.'s model. In Section 3.5, as an application of our conclusions, we will investigate under what conditions one of the phenotypes will die out or dominate. In Section 3.6, we will show that the above conclusions are still valid in more general cases.

3.2 Notations and model description

Notations Boldface letter, like $\mathbf{A}$, represents matrix. $\mathbf{A}'$ means the matrix transpose of $\mathbf{A}$. $\mathbf{I}$ is the identity matrix. Letter with arrow above, like $\vec{u}$, represents row vectors. $\vec{1}$ and $\vec{0}$ represent all ones and all zeros vectors. Consider the population of cells with n phenotypes: $Y_1, Y_2, \cdots, Y_n$. In the stochastic model, $\vec{X}(t) = (X_1(t), X_2(t), \cdots, X_n(t))$ is the population at time t, where $X_i(t)$ is the population of phenotype Y_i. $P_i(t) = X_i(t)/\sum_{i=1}^n X_i(t)$ is the proportion of phenotype Y_i, as long as the denominator is not zero. $\vec{P}(t) = (P_1(t), P_2(t), \cdots, P_n(t))$. In the deterministic model, $\vec{x}(t) = (x_1(t), x_2(t), \cdots, x_n(t))$ is the expected population at time t. We only consider the case where each $x_i(t)$ is nonnegative, and at least one of them is positive (to guarantee $|\vec{x}| > 0$). $|\vec{x}| = \sum_{i=1}^n x_i(t)$ is the total population. $\vec{p} = \vec{x}/|\vec{x}|$ is the proportions of different subpopulations among the total population.

Stochastic Model Assume that the population of cells have n phenotypes: $Y_1, Y_2, \cdots, Y_n$. Assume that all the cells evolve independently. (During the exponential growth period, this assumption is almost true [171].) We can present the generalized cell divisions, death and phenotypic conversions as the following reaction form:

$$Y_i \xrightarrow{\alpha_i} d_{i1}Y_1 + d_{i2}Y_2 + \cdots + d_{in}Y_n.$$

It means that for an Y_i cell, it will live an exponential time (we will consider non-exponential lifetime in Section 3.6) with expectation $1/\alpha_i$ and turn into d_{i1} Y_1 cells, d_{i2} Y_2 cells, $\cdots$, d_{in} Y_n cells, where $d_{i1}, d_{i2}, \cdots, d_{in}$ are random variables taking nonnegative integer values. $d_{i1}, d_{i2}, \cdots, d_{in}$ are

not necessarily independent, but they are assumed to be independent with the exponential reaction time of any cell.

For example, an asymmetric division $Y_1 \to Y_1+Y_2$ means $(d_{11}, d_{12}, d_{13}, \cdots, d_{1n}) = (1, 1, 0, \cdots, 0)$. A conversion $Y_1 \to Y_2$ means $(d_{11}, d_{12}, d_{13}, \cdots, d_{1n}) = (0, 1, 0, \cdots, 0)$. So the probability distribution on the possible reactions gives the joint probability distribution of $d_{i1}, d_{i2}, \cdots, d_{in}$.

In fact this is a multitype continuous-time Markovian branching process $\vec{X}(t)$ with state space $(\mathbb{Z}^*)^n$, each component of which represents the population of a phenotype, as defined in [8] and [67]. For example, if one Y_1 cell splits symmetrically, the process will move from the state $\vec{X} = (s_1, s_2, \cdots, s_n)$ to the state $\vec{X} = (s_1 + 1, s_2, \cdots, s_n)$. We require that $\mathbb{E}d_{ij}^2 < \infty$, $\forall i, j$. (In experiments, d_{ij} is bounded, thus $\mathbb{E}d_{ij}^2 < \infty$ is always true.) Then this process will always have finite value in finite time (non-explosion) with probability one [8, Section V.7.1, (3)–(4)].

Deterministic Model Now we consider the mathematical expectation of the populations, $\vec{x} = \mathbb{E}(\vec{X})$, which is nonnegative. Based on [8, Section V.7.2, (5)–(9)], we have the deterministic model, namely the following ODE system:

$$\mathrm{d}\vec{x}/\mathrm{d}t = \vec{x}\mathbf{A} \tag{3.1}$$

where $\mathbf{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \cdots & a_{n,n} \end{bmatrix}$, $a_{i,i} = \alpha_i(\mathbb{E}d_{ii} - 1) \geq -\alpha_i$, $a_{i,j} = \alpha_i \mathbb{E}d_{ij} \geq 0$ $(i \neq j)$. Define $\vec{b} = \vec{1}\mathbf{A}'$. From (3.1), we have $\mathrm{d}|\vec{x}|/\mathrm{d}t = \vec{x}\vec{b}'$.

3.3 The relationship between the Markovian model and the deterministic model

In the Markovian model [57], it is assumed that the population proportions $\vec{p}$ satisfies the Kolmogorov forward equations of an n-state Markov chain:

$$\mathrm{d}\vec{p}/\mathrm{d}t = \vec{p}\mathbf{Q} \tag{3.2}$$

where $\mathbf{Q}$ is the transition rate matrix, satisfying $\vec{1}\mathbf{Q}' = \vec{0}$. In this section, we will discuss whether such assumption can be satisfied in the deterministic model.

From (3.1), we have

$$\frac{\mathrm{d}\vec{p}}{\mathrm{d}t} = \frac{\mathrm{d}(\vec{x}/|\vec{x}|)}{\mathrm{d}t} = \frac{|\vec{x}|}{|\vec{x}|^2}\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} - \frac{\vec{x}}{|\vec{x}|^2}\frac{\mathrm{d}|\vec{x}|}{\mathrm{d}t} = \frac{\vec{x}\mathbf{A}}{|\vec{x}|} - \frac{(\vec{x}\vec{b}')\vec{x}}{|\vec{x}|^2} = \vec{p}[\mathbf{A} - (\vec{p}\vec{b}')\mathbf{I}] \tag{3.3}$$

If $\vec{b} = k\vec{1}$ for some constant k, then (3.3) becomes $\mathrm{d}\vec{p}/\mathrm{d}t = \vec{p}(\mathbf{A} - k\mathbf{I})$, and $\vec{1}'(\mathbf{A} - k\mathbf{I})' = \vec{b}' - k\vec{1}' = \vec{0}'$. Thus (3.3) has the same form of (3.2). If $\vec{b} \neq k\vec{1}$, there are non-zero quadratic terms of $p_i(t)$ in (3.3), implying that (3.3) does not have the same form of (3.2).

Notice that $\vec{b} = k\vec{1}$ means

$$\sum_{i=1}^{n} a_{1,i} = \sum_{i=1}^{n} a_{2,i} = \cdots = \sum_{i=1}^{n} a_{n-1,i} = \sum_{i=1}^{n} a_{n,i} (= k). \tag{3.4}$$

Thus we have

Theorem 3.1. *Equation (3.4) is the sufficient and necessary condition for that the proportions of different phenotypes in the deterministic model (3.1) satisfy the Kolmogorov forward equations of an n-state Markov chain.*

Now we know that the Markovian model is a special case of the deterministic model. In biology, (3.4) means that the growth rates (average number of descendants produced per unit time) of different phenotypes are the same. This condition might be well satisfied for breast cancer cells, which explains why the data fitting in [57] is satisfactory.

3.4 Asymptotic behavior in general cases

In general cases, (3.4) is not satisfied since different phenotypes may differ in cell cycling time [109, 39], then the Markovian model is invalid. Thus we need other methods to study the asymptotic behavior of the population dynamics. In this section, we will prove that under some mild conditions, the proportions of different phenotypes will tend to some constants regardless of initial population states.

From Perron-Frobenius theorem [138, 76], we know that $\mathbf{A}$ has a real eigenvalue λ_1 (called Perron eigenvalue), such that for any eigenvalue $\mu \neq \lambda_1$, Re $\mu < \lambda_1$. λ_1 has a left eigenvector $\vec{u}$= $(u_1, u_2, \cdots, u_n)$ (called Perron eigenvector), satisfying $u_i \geq 0, \forall i$ and $\sum_{i=1}^{n} u_i = 1$. When λ_1 is simple, such $\vec{u}$ is unique. We know that the set of all n-order real square matrices with repeated eigenvalue has measure 0 (as a subset of $\mathbb{R}^{n^2}$) [168]. Thus it is reasonable to assume that λ_1 is simple.

3.4.1 Deterministic model

We have proved the following theorem in Appendix B of [168].

Theorem 3.2. *Assume that λ_1 is simple. Starting from any initial value except for the point in some zero-measure set, we have $(x_1(t), x_2(t), \cdots, x_n(t))/\exp(\lambda_1 t) \to c\vec{u}$ as $t \to \infty$, where $c > 0$ is a constant. In this case, the solution of (3.4) will tend to $\vec{u}$ as $t \to \infty$. Thus (3.4) has one and only one stable fixed point $\vec{u}$ and no stable limit cycle.*

This gives a satisfactory deterministic explanation of the phenotypic equilibrium phenomenon reported in [57].

Remark 3.1. *If λ_1 is not simple, then the convergence result may not hold. Consider $\mathbf{A}$ with $a_{i,j} = 0, \forall i \neq j$ and $a_{i,i} = 1, \forall i$. Here $\lambda_1 = 1$ is not simple, $x_i(t) = x_i(0)\exp(\lambda_1 t)$, and $(x_1(t), x_2(t), \cdots, x_n(t))/\exp(\lambda_1 t) = (x_1(0), x_2(0), \cdots, x_n(0))$ will never change. Convergence to a common point will not occur.*

3.4.2 Stochastic model

Since 1960s, probabilists proved that for a continuous-time multitype branching process, $\vec{X}(t)/\mathrm{e}^{\lambda_1 t} \to W\vec{u}$ under different conditions, where W is a nonnegative random variable. In [7], [8] and [139], it is required that $\lambda_1 > 0$ and $\mathbf{A}$ is irreducible (this implies λ_1 is simple). In [8] it is proved that $W = 0$ or $W > 0$ according to whether the population will become extinct. In [78], it is required that the branching process is discrete in time. In [159, 160] it is required that the initial population

tends to infinity. Janson [67] requires that $\lambda_1 > 0$, λ_1 is simple, and assumes a special condition about communicating classes structure (see Remark 3.2). Based on [67] and [8], we will prove the convergence theorem without Janson's last assumption (Theorem 3.3). We can see the benefit of this improvement in Section 3.5.

Preliminaries

In this section, we assume that λ_1 is simple and positive. $\lambda_1 > 0$ means that the total cell population is increasing.

Sometimes, the transformation from one phenotype to another phenotype is not reversible. For example, a mature human red blood cell (which loses its nucleus) cannot transform back to a zygote. Thus we need to classify phenotypes according to communicating behaviors. In mathematical language, we need to study communicating classes of $\mathbf{A}$ when $\mathbf{A}$ is reducible.

If a cell of phenotype Y_i can produce (directly or indirectly) a cell of phenotype Y_j and vice versa, then we say phenotype Y_i communicates with phenotype Y_j ($Y_i \leftrightarrow Y_j$). Since "$\leftrightarrow$" is an equivalent relation, we can divide the n phenotypes into several disjoint sets (called communicating classes) according to $\mathbf{A}$ [104]. Then we can order the classes and rearrange the phenotypes suitably to make $\mathbf{A}$ block-triangular. (Each diagonal block corresponds to a communicating class.) Thus the eigenvalues of $\mathbf{A}$ consist of all eigenvalues of diagonal blocks. Every eigenvalue corresponds to a diagonal block, and then corresponds to a communicating class. (See [67] and [78] for details.)

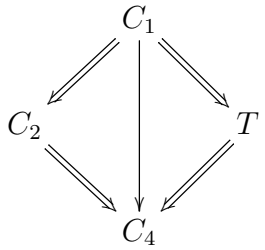
Denote the communicating class corresponding to the Perron eigenvalue λ_1 by T.

For example, consider matrix $\mathbf{A} = \begin{bmatrix} \mathbf{D_1} & \mathbf{W} & \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_2} & \mathbf{0} & \mathbf{W} \\ \mathbf{0} & \mathbf{0} & \mathbf{D_3} & \mathbf{W} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D_4} \end{bmatrix}$, where each $\mathbf{W}$ represents a different nonnegative matrix (not $\mathbf{0}$). Assume that $\mathbf{D_3}$ has the Perron eigenvalue λ_1, then $\mathbf{D_3}$ corresponds to the communicating class T. Denote the other three communicating classes by C_1, C_2, C_4.

For two communicating classes C_i and C_j, we write $C_i \Rightarrow C_j$ if there exist phenotype $X_{k_i} \in C_i$ and $X_{k_j} \in C_j$ such that $a_{k_j,k_i} > 0$. For two communicating classes C and D, we write $C \to D$ if

there exist communicating classes $C = C_1, C_2, \cdots, C_m = D$ such that $C_i \Rightarrow C_{i+1}, \forall 1 \leq i < m$. Stipulate that $C_i \Rightarrow C_i$ and $C_i \to C_i$.

Then we can illustrate the communicating classes in the example above as



For a communicating class C, define $\hat{C} = \{Y_i | Y_i \in C_j, C_j \to C\}$. In other words, $\hat{C}$ is the set of all phenotypes that can produce (directly or indirectly) phenotypes in C. In the example above $\hat{T} = C_1 \cup T$.

For a communicating class C, define $\bar{C} = \{Y_i | Y_i \in C_j, C \to C_j\}$. In other words, $\bar{C}$ is the set of all phenotypes that can be produced (directly or indirectly) by phenotypes in C. In the example above $\bar{T} = T \cup C_4$.

For the Markovian branching process $\vec{X}(\cdot)$, we say that a cell V with phenotype in $\hat{T}$ becomes "*essentially extinct*" if at some time no cell of any phenotypes in $\hat{T}$ is V or its descendants. In other words, V and its descendants become extinct inside $\hat{T}$. We say that a trajectory of the branching process $\vec{X}(\cdot)$ becomes " *essentially extinct*" if at some time no cell of any phenotypes in $\hat{T}$ remains. This means that we can never get a cell with phenotypes in T any more. If so, we cannot have the desired convergence property (Theorem 3.3), since the proportions of phenotypes in T should be positive (Lemma 3.2). Let the branching process $\vec{X}(\cdot)$ start at any initial population $\vec{X}(0)$ as long as it has some cells with phenotypes in $\hat{T}$.

Results and proofs

We now state the main result of this paper and then give the proof of it.

Theorem 3.3. *Assume that λ_1 is simple and positive. Conditioned on essential non-extinction, we have almost surely $(P_1(t), P_2(t), \cdots, P_n(t)) \to \vec{u} = (u_1, u_2, \cdots, u_n)$ as $t \to \infty$.*

Lemma 3.1. *Assume that λ_1 is simple. If for some $i \neq j$, $a_{i,j} > 0$ in (3.1), then $u_j > 0 \Rightarrow u_i > 0$.*

Proof. Without loss of generality, let $i = 1, j = 2$. Assume $u_1 = 0$, $u_2 > 0$. Let $(p_1, p_2, \cdots, p_n) = \vec{u} = (u_1, u_2, \cdots, u_n)$ in the first equation of (3.3). Then it becomes $\mathrm{d}p_1/\mathrm{d}t = \sum_{k=2}^{n} a_{1,k}u_k > 0$. However $\vec{u}$ is a fixed point of (3.3) according to Theorem 3.2, thus we should have $\mathrm{d}p_1/\mathrm{d}t = 0$, which is a contradiction. □

Lemma 3.2. *Assume that λ_1 is simple. Then $u_i > 0 \iff Y_i \in \bar{T}$.*

Proof. Apply the Perron-Frobenius theorem to $\mathbf{A}_{\bar{T}}$, the restriction of $\mathbf{A}$ on $\bar{T}$, and let $\vec{w}$ be its Perron eigenvector. $\mathbf{w}_T$, the restriction of $\vec{w}$ on T cannot be $\vec{0}$, otherwise λ_1 is an eigenvalue of $\mathbf{A}_{\bar{T}\setminus T}$, a contradiction. From Lemma 3.1 we know that $\vec{w}$ is positive. Set $u_i = w_i$ if $Y_i \in \bar{T}$, and $u_j = 0$ if $Y_j \notin \bar{T}$, then $\vec{u}$ is the Perron eigenvector of $\mathbf{A}$. Thus $u_i > 0 \iff Y_i \in \bar{T}$.

□

Lemma 3.3 (Lemma 9.8 in [67])**.** *Assume that λ_1 is simple and positive. Then we have almost surely $\mathrm{e}^{-\lambda_1 t}\vec{X}(t) \to W\vec{u}$ as $t \to \infty$, where W is a nonnegative random variable, and $\mathbb{P}(W > 0) > 0$.*

Lemma 3.4 (Lemma 9.7 (ii) and (iii) in [67], originated from Theorem V.7.2 in [8])**.** *Assume that λ_1 is simple and positive, and $\bar{T}$ contains all phenotypes, then $W = 0$ if and only if the branching process becomes essentially extinct almost surely.*

Remark 3.2. *Janson's paper [67, Section 2] has six fundamental assumptions (A1)-(A6). Assumptions (A1)-(A5) have been satisfied in this paper (regarding (A5) as "the process is not essentially extinct at time 0"). Assumption (A6) "$\bar{T}$ contains all phenotypes" is only used in Lemma 3.4. In fact, we will prove (in Lemma 3.7) that Lemma 3.4 is still correct without Assumption (A6). Thus we can drop Assumption (A6) in the main result. Assumption (A6) implies all phenotypes should have the same exponential growth rate λ_1, and no phenotype will die out or dominate (see Section 3.5), which is not necessarily satisfied in experiments. For example, in Bozic et al.'s paper [17], they consider tumor cells which gradually cumulate mutations which accelerate cell growth.*

Since tumor cells with more accelerating mutations cannot switch back to tumor cells with less such mutations, they must have different growth rates. Also, the proportion of tumor cells with less accelerating mutations will gradually decrease to 0*, which contradicts with assumption (A6).*

The following lemma is a modification of the second Borel-Cantelli lemma. We base our proof on Theorem 2.3.6 in [31].

Lemma 3.5. *Consider events* $B_1, B_2, \cdots, B_n, \cdots$*. If for any positive integers* $m < n$*, we have* $\mathbb{P}(\cap_{i=m+1}^{n} B_i^c) \leq (1-\epsilon)^{n-m}$*, where* $0 < \epsilon \leq 1$*, then* $\mathbb{P}(\limsup_{n\to\infty} B_n) = 1$*. In other words, almost surely* $\{B_n : n \geq 1\}$ *will happen infinitely often.*

Proof. Let $0 < M < N < \infty$. $\mathbb{P}(\cap_{i=M+1}^{N} B_i^c) \leq (1-\epsilon)^{N-M} \to 0$ as $N \to \infty$. So $\mathbb{P}(\cup_{i=M+1}^{\infty} B_i) = 1$ for all M, and since $\cup_{i=M+1}^{\infty} B_i \downarrow \limsup_{n\to\infty} B_n$ it follows that $\mathbb{P}(\limsup_{n\to\infty} B_n) = 1$. □

Lemma 3.6. *For almost every essentially non-extinct trajectory (according to Lemma 3.3, the set of such trajectories has positive probability), we can find an essentially non-extinct cell with phenotype in* T *within finite time. If we can find such cell at time* t*, then we can find such cell at any time* $\tau > t$*.*

Proof. If at some time t all cells with phenotypes in $\hat{T} \setminus T$ die out, then at least one of the remaining cells with phenotypes in T is not essentially extinct.

Otherwise, at each time $t = k$ ($k \in \mathbb{Z}^+$), there exists one cell E_k with phenotype in $\hat{T} \setminus T$. (For different k, E_k may be the same cell.) Let B_k ($k \in \mathbb{Z}^+$) be the event that during the time interval $[k, k+1)$, the cell E_k produces (directly or indirectly) at least one cell with phenotype in T.

If B_k happens, choose one such cell with phenotype in T and put it in a special set S. Consider any two cells F and G in S, and assume F is produced in the time interval $[i, i+1)$, G is produced in the time interval $[j, j+1)$, and $i < j$, where $i, j \in \mathbb{Z}^+$. Then E_j is the ancestor of G. Since E_j has phenotype in $\hat{T} \setminus T$, and F has phenotype in T, F cannot be the ancestor of E_j. Since E_j is still alive at time $t = j$, when F has been produced, E_j cannot be the ancestor of F. Thus F cannot be the ancestor of G. Since G is produced after F, G cannot be the ancestor of F. In sum, one cell in S cannot be the ancestor of another cell in S. Thus all cells in S are independent.

Consider two phenotypes Y_i and Y_j, and assume a cell with phenotype Y_i can produce a cell with phenotype X_j directly, namely $\mathbb{P}(d_{ij} > 0) > 0$. Because of Markovian property, within a time span of $1/n$, the probability for a cell with phenotype Y_i to produce a cell with phenotype Y_j directly is $\eta_{ij} = [1 - \exp(-\alpha_i/n)]\mathbb{P}(d_{ij} > 0) > 0$. Let $\eta = \min_{i,j}\{\eta_{i,j} : \mathbb{P}(d_{ij} > 0) > 0\}$. For a cell with phenotype in $\hat{T} \setminus T$, it can produce a cell with phenotype in T within n steps. Thus the probability of B_k is no less than η^n, regardless of what happens before time $t = k$.

Now we can use Lemma 3.5 with $\epsilon = \eta^n$, and there will be an infinite number of cells in S, except for a zero-measure set of trajectories. According to Lemma 3.3, the probability for one cell in S to become essentially extinct is less than 1, thus the probability for all cells in S to become essentially extinct is 0, and at least one cell in S is not essentially extinct, except for a zero-measure set of trajectories. □

Lemma 3.7. *Assume that λ_1 is simple and positive, then $W = 0$ if and only if the branching process becomes essentially extinct almost surely.*

Proof. $\Leftarrow$: For a trajectory $\vec{X}(\cdot)$ outside the zero-measure exclusion set of Lemma 3.3, assume that at some time $\tau \geq 0$ (dependent on the trajectory), $X_i(\tau) = 0$ for all $Y_i \in \hat{T}$. For any $Y_j \in T$, $0 = \lim_{t\to\infty} \mathrm{e}^{-\lambda_1 t} X_j(t) = W u_j$. From Lemma 3.2, $u_j > 0$. Thus $W = 0$ almost surely.

$\Rightarrow$: Assume that $\mathbb{P}(W = 0 \ \& \text{ the trajectory is not essentially extinct}) = P_0 > 0$. According to Lemma 3.6, we can find time $t_0 > 0$ large enough such that $\mathbb{P}(W = 0 \ \& \text{ the trajectory is not essentially extinct } \& \text{ there exists an essentially non-extinct cell with phenotype in } T \text{ at time } t_0) \geq P_0/2 > 0$. On this set, only consider this essentially non-extinct cell and its descendants from time $t \geq t_0$, then the population is restricted on $\bar{T}$ and we can use Lemma 3.4. Now we have $W > 0$ except for a zero-measure set of trajectories, which is a contradiction. □

From Lemma 3.3 and Lemma 3.7,we can obtain Theorem 3.3.

Remark 3.3. *The assumption of $\lambda_1 > 0$ is not too strong. If $\lambda_1 < 0$, then from Theorem 3.2, the expected populations decay to $\vec{0}$. Therefore this process will become extinct almost surely. For*

$\lambda_1 = 0$, *consider an example that each cell always have exactly one child, and the child can be any phenotype with equal probability. Then the total population is fixed, and the proportions will always fluctuate, so there is no convergence [67].*

For Gupta el al's experiment, the initial cell population is very large in cancer cell lines, thus the probability of essential extinction is quite small. Therefore, the proportions will almost always tend to the same constants. This gives a satisfactory stochastic explanation of the phenotypic equilibrium phenomenon reported in [57].

The deterministic model only reflects the average behavior of many trajectories (or equivalently a large initial population). When the cell number is relatively small, the stochasticity is not negligible, and it is not very reasonable to assume the cell number changes continuously. So the stochastic model is more effective than deterministic model. That is why we also prove the same result for stochastic model.

Remark 3.4. *In Bozic et al.'s model, we divide cells in groups by their number of driver mutations. The growth rate for cells with k driver mutations is $1-(1-s)^k/2$ ($s = 0.004$ in [17]). Cells can acquire driver mutations, but cannot lose them. Also we do not consider cell death, so there is no essential extinction. If we only consider cells with no more than n driver mutations, then the population of cells with exactly n mutations will grow with exponential growth rate $1-(1-s)^n/2$, which is larger than that of cells with less driver mutations. So cells with n driver mutations will dominate exponentially fast. Generally, the population of cells with n driver mutations will grow exponentially with rate $1-(1-s)^n/2$, and as long as the next driver mutation emerges, its proportion will decay to 0 exponentially fast. Also, we should notice that such results are valid for almost every trajectories. Here we do not consider the difference in passenger mutations, since they have no effect on cell growth, and considering them will make the Perron eigenvalue not simple. The model in [17, 80] is discrete-time, but we can see from Section 3.6 that we can still apply our results.*

3.5 When will one proportion tend to 0 *or* 1*?*

In population dynamics, we are also concerned about when one phenotype dies out or dominates. In terms of the notations in this paper, we need to consider when $P_i(t) \to 0$ or $P_i(t) \to 1$ as $t \to \infty$.

In this section, we will still assume that the Perron eigenvalue λ_1 of $\mathbf{A}$ is simple and positive. Then from Theorem 3.3, we have $(P_1(t), P_2(t), \cdots, P_n(t)) \to \vec{u}= (u_1, u_2, \cdots, u_n)$ almost surely in the stochastic model. Thus we can get the following corollaries from Lemma 3.2.

Corollary 3.4. $P_i(t) \to 0 \iff Y_i \notin \bar{T}$.

Corollary 3.5. $P_i(t) \to 1 \iff \bar{T} = T = \{Y_i\}$.

Remark 3.5. *From Corollary 3.4 we can see that the sufficient and necessary condition under which no phenotype dies out, namely* $\forall i, P_i(t) \nrightarrow 0$*, is that* $\bar{T}$ *contains all phenotypes. This is just Janson's last assumption.*

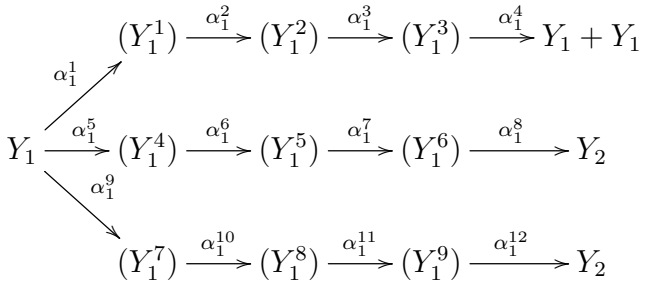
Remark 3.6. *If we find that* $P_i(t) \to 0$*,* $P_j(t) \nrightarrow 0$ *in an experiment, then we know that the phenotype* Y_j *will never transform to* Y_i *in any way. If we find that* $P_i(t) \to 1$*, then we know that the phenotype* Y_i *will never transform to any other phenotypes.*

3.6 Model generalization: non-exponential lifetime

In the previous sections, we assumed that the lifetime of a cell is exponentially distributed and independent of the type and number of its descendants. However, in real biological system, the lifetime distribution should be more like lognormal, gamma, Weibull, or exponentially modified Gaussian distribution [61, 52]. Furthermore, the time needed for division and conversion have different distributions [52]. In this way the process is a multitype Bellman-Harris branching process (also called age-dependent branching process) [95], no longer Markovian.

We can use the "device of stages" method to approximate a non-exponential random variable with several exponential random variables [22]. This indicates that through adding supplementary

sub-phenotypes, we can simulate a non-Markovian branching process with a Markovian branching process. See the example below:



Here $(Y_1^1), \cdots, (Y_1^9)$ are supplementary sub-phenotypes. We artificially assume such supplementary sub-phenotypes exist just by technical reasons. They do not have biological meanings. When we count Y_1^i as Y_1, the process has the same distribution with the original one. If we have convergence with this new process, then we also have convergence for the original process. An Y_1 cell has probability $\alpha_1^1/(\alpha_1^1+\alpha_1^5+\alpha_1^9)$ to divide into Y_1+Y_1, and probability $(\alpha_1^5+\alpha_1^9)/(\alpha_1^1+\alpha_1^5+\alpha_1^9)$ to convert into Y_2. Here we set α_1^1, α_1^5, and α_1^9 to be large enough while keeping their proportions, so that the time needed for the first step is ignorable (exponential random variable with expectation $1/(\alpha_1^1+\alpha_1^5+\alpha_1^9)$).

Now the time distribution for division $Y_1 \to Y_1 + Y_1$ is approximately $Ex(\alpha_1^2) * Ex(\alpha_1^3) * Ex(\alpha_1^4)$, where $Ex(\alpha)$ is the density function of exponential random variable with parameter α, and $*$ means convolution. Similarly, the time distribution for conversion $Y_1 \to Y_2$ is approximately $\frac{\alpha_1^5}{\alpha_1^5+\alpha_1^9} Ex(\alpha_1^6) * Ex(\alpha_1^7) * Ex(\alpha_1^8) + \frac{\alpha_1^9}{\alpha_1^5+\alpha_1^9} Ex(\alpha_1^{10}) * Ex(\alpha_1^{11}) * Ex(\alpha_1^{12})$.

According to [22], any non-negative random variable can be approximated to any accuracy by such combination of convolutions of exponential random variables. Thus we can simulate such non-Markovian branching processes to any precision with Markovian branching processes. Here the lifetime of a cell can be non-exponential, and the lifetime of a cell can depend on the type and number of its descendants.

Now we can apply Theorem 3.3 to those sub-phenotypes. The proportion of each sub-phenotype converges to a constant. Thus the proportion of each phenotype (including all its sub-phenotypes) converges to a constant. This proves the "phenotypic equilibrium" phenomenon in a more realistic

stochastic model. In addition, the conclusions in Section 3.5 are still valid.

Remark 3.7. *The most unrealistic aspect of exponential lifetime is that the density function reaches maximum at* 0*, but one cell cannot divide right after its birth. For the sum of several independent exponential variables, the density function at* 0 *is* 0*. By the law of large numbers, the density function of the sum of* n *independent exponential variables with parameter* $n\lambda$ *will have a sharp peak near* λ *when* n *is large.*

Remark 3.8. *The proportion convergence theorem for non-Markovian (age-dependent) branching processes can be proved directly, but under stronger conditions [95].*

3.7 Conclusion and discussion

We have presented a unified stochastic model for the population dynamics with cellular phenotypic conversions. We have given the sufficient and necessary condition under which the dynamical behavior of our model can be described by an n-state Markov chain. In general case, we have proved that the proportions of different phenotypes will tend to constants regardless of their initial values, and we have investigated the sufficient and necessary conditions under which one phenotype will die out or dominate. We also extend our model to non-Markovian case while keeping the above conclusions valid. In this way we explain experimental phenomenon in [57].

As remarked in Section 3.4.2, we improve a limit theorem in branching processes, which may be of theoretical interests.

Our results can also apply to cancer progression models with gene mutations, such as [17, 80, 93].

Since the phenotypic conversions have been reported in various cellular systems, such as *E.coli* [107] and cancer cells [38, 161], we hope that our model here could be applied as a general framework in the study of multi-phenotypic populations of cells.

With the improvement of experiment methods, we will accumulate more and more data for single cell. In single cell level, the stochasticity is not negligible anymore. Thus we should build detailed stochastic models (branching process would be a good framework) for cell population

dynamics. Such models will provide new insights and predictions. In the meanwhile, stochastic process related theories should attract more attention.

There are some possible improvements about this research. First, we assume that the branching process is time homogeneous, namely the birth and death rates keep the same for all time. However, as time goes on, the cell density increases, and the birth and death rates should change [171]. Thus a possible improvement is to have time-dependent or density-dependent d_{ij}. Second, we only prove the convergence for $t \to \infty$, but in experiments we only have finite observation time. Thus it is meaningful to estimate the convergence rate. Third, we only consider finite many phenotypes. We find that there is essential difficulty to build similar theory for infinite-type branching processes. However, there have been some works considering infinite many mutation states with multitype branching processes, such as [93].

3.8 Appendix

3.8.1 Fast algorithms for finding transposons in gene sequences

Introduction

A transposon is a DNA sequence that can change its relative position within the genome of a single cell. The mechanism of transposition can be either "copy and paste" or "cut and paste". However, it is not so easy to determine which genes are transposons. If there is a template sequence which can be viewed as the ancestor of other gene sequences, we can compare the other gene sequences with this template, but we cannot determine which genes change their positions since there are different ways to change the positions of genes to turn the template into the other gene sequences. In addition, when we have more than two gene sequences, we can hardly know which gene sequence is more primitive so that it can be viewed as the template. Thus, we should describe the definition of "change its relative position" more precisely.

The criterion of transposons

First, let's look at an example to have an intuitive sense of transposons. Consider two gene sequences, where number represents gene's name:

$$\{1,2,3,4\} \text{ and } \{1,4,2,3\}.$$

We will intuitively think gene 4 changes its position. However, changing genes $2, 3$'s positions can also explain this. Why do we think gene 4 moves, but not genes 2 and 3? The answer is that in the former situation, the number of genes changing their positions is smaller. Nevertheless, "the number of genes changing their positions" is not so easily determined. A proper way is to think of the fixed genes, the complement of transposons. The minimum of transposons corresponds to the maximum of fixed genes, so we reach **the criterion of transposons: find the longest common subsequence of these gene sequences, and the remainder are transposons.** In the example above, the longest common subsequence is $\{1, 2, 3\}$, so gene 4 is a transposon. Till now, we have turn the problem of finding transposons into finding the longest common subsequence of

several number sequences. In the following sections, we will consider different situations and give corresponding algorithms.

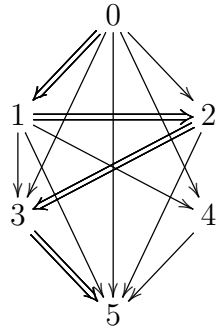
Sequences without replicated genes

Consider gene sequences without replicated genes. For several permutations of $1, 2, 3 \cdots n$ (allowing missing some numbers), we need to find the longest common subsequence of these number sequences. Constructing a directed graph as following. Vertices are numbers from 1 to n. There exists an directed edge from vertex i to j if and only if in all sequences i appears prior to j. If there exist edges $i \to j, j \to k, \cdots, l \to m$, then the path $i \to j \to k \to \cdots \to l \to m$ corresponds to the common subsequence $\{i, j, k \cdots l, m\}$. Thus the longest path in this graph corresponds to the longest common subsequence.

This graph is transitive [49]: if there exist edges $i \to j, j \to k$, then there exists edge $i \to k$, since if i is prior to j and j is prior to$k$, then i is prior to k.

Furthermore, this directed graph has no loop. If there exists a loop $i \to j \to k \to \cdots \to l \to i$, then from transitivity there exists edge $i \to i$, which means i is prior to i itself, a contradiction. No loop means that the longest path does exist, and its length is finite.

If we add 0 to the head of each sequence and $n + 1$ to the tail, then the longest common subsequence must begin at 0 and end at $n + 1$. Thus, we only need to find the longest path between two vertices in a directed graph without loop.



Graph 1

Graph 1 corresponds to the example $\{(0), 1, 2, 3, 4, (5)\}$ and $\{(0), 1, 4, 2, 3, (5)\}$, where the longest path from 0 to 5 is $0 \to 1 \to 2 \to 3 \to 5$, so the longest common sequence is $\{(0), 1, 2, 3, (5)\}$, and the transposon is 4.

We can use an iteration algorithm to find the longest path. For a vertex i, define $F(i)$ to be the vertex next to i in the longest path from i to $n + 1$, and $G(i)$ to be the length of this path. Since there exists edge $i \to n + 1$, F can be defined for all vertices except $n + 1$. Define $G(n + 1) = 0$.

The longest path is $0 \to F(0) \to F^2(0) \to F^3(0) \to \cdots \to F^{k-1}(0) \to F^k(0) = n+1$, where F^j is the jth iteration of F. The iteration equations are:

$$F(i) = j, \text{ where } j \in \{k| \text{ there exists edge } i \to k\} \text{ and maximizes } G(k).$$

$$G(i) = G(F(i)) + 1.$$

Since this graph has no loop, the iteration will terminate in finite steps and produce correct answer. This iteration can be carried out though programming languages such as C++.

Assume there are m sequences with length n.

Space to restore these sequences is $O(mn)$. The graph is restored in an $(n+2) \times (n+2)$ matrix. Functions F and G needs $O(n)$. Since n is much larger than m in general cases, the total space complexity is $O(n^2)$.

Time to construct the graph is $O(mn^2)$ since we need to check each pair $\{i, j\}$ in every sequence. In the iteration process, the calculation of each $F(i)$ and $G(i)$ needs $O(n)$, since for every i there are at most $n+1$ edges begin at i. Thus the time complexity of iteration is $O(n^2)$. The total time complexity is $O(mn^2)$.

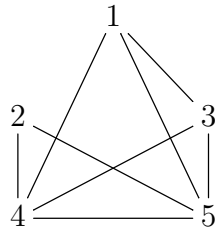
This algorithm is implemented and applied in [75].

Sequences with replicated genes

In general situations there may be replicated genes in a gene sequence. Notice that the definition of transposon is a DNA sequence that has the ability to change its position, not a certain gene copy that changes its position. Thus we should regard all copies of one gene as a whole. We should not find the longest common subsequence of these sequences, like Section 3, but find the largest number of genes that all their copies remain the same relative positions.

This time we consider an undirected graph. The vertices are different genes(not all copies). There exists edge between vertices i and j (i and j are adjacent) if and only if all copies of i and j remain the same relative positions in all sequences. A group of genes with all their copies fixed corresponds 1-1 to a group of vertices that each two vertices are adjacent, which is called "complete subgraph" in graph theory. Thus we need to find the largest complete subgraph of this

undirected graph. This is the famous "Maximum Clique Problem", which is NP-complete. We certainly cannot give a universal algorithm with acceptable complexity. However, transposons are much fewer compared with total genes in practice, so we can give an algorithm suitable for this situation.



Graph 2

Graph 2 corresponds to sequences $\{1,2,3,2,3,4,5\}$ and $\{2,1,3,3,2,4,5\}$, where the largest complete subgraph is $\{1,3,4,5\}$, so the longest common sequence is $\{1,3,3,4,5\}$, and the transposon is 2.

Assume there are n genes in these sequences. In graph theory, the degree of a vertex i is the number of vertices that are adjacent to i [49]. Notice that if a complete subgraph consists of k vertices, then each of these vertices must has degree not less than k. Suppose the largest degree of these vertices is k. Then we can check whether there are at least k vertices with degrees not less than k. If so, check whether there exists a complete subgraph of k vertices. If there is no complete subgraph of k vertices, replace k by $k-1$ and repeat the former process until finding a complete subgraph. In practice, there are only a few transposons and they generally change their relative positions violently. In this situation, there will be a few vertices with significant small degrees, and all the other vertices with large degrees form the largest complete subgraph. Thus the time complexity can be sometimes limited to $O(mn^2)$.

More strict criterion for transposons

Maybe it is not strict enough to regard a gene which only changes its position in few sequences as a transposon. We can loosen the definition of fixed genes that they should be the longest common subsequence of $m-k$ but not all m gene sequences. Certainly, k should be small enough, such as $k<5$. Then we can consider every $m-k$ out of m sequences and find the longest common subsequence, then pick the longest among these "longest". In practice, no matter whether there are replicated genes, the time complexity is $O(m^k n^2)$: the construction of graph is $O(mn^2)$, and

finding longest common subsequence is $\binom{m}{k}O(n^2) = O(m^k n^2)$.

3.8.2 Representations of branching processes

The multi-type continuous-time branching processes discussed in this chapter can be regarded as special cases of other processes. Such new representations could bring different views of branching processes.

Branching processes as chemical reactions

In such branching processes, one cell could become into many cells, therefore the cell number is not conserved. We seek to present branching processes as chemical reactions, then we need some invisible supplementary cell types.

Consider cells in n types: $Y_1, Y_2, \cdots, Y_n$.

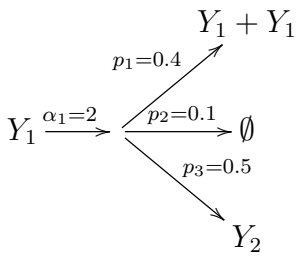
Consider a continuous-time n-type branching process $\vec{X}(t) = \{X_1(t), X_2(t), \cdots, X_n(t)\}$. Here $X_i(t)$ is the number of cells with type Y_i.

Cells proliferate as the following formula:

$$Y_i \xrightarrow{\alpha_i} d_{i1}Y_1 + d_{i2}Y_2 + \cdots + d_{in}Y_n.$$

It means that for an Y_i cell, it will live an exponential time with expectation $1/\alpha_i$ and turn into d_{i1} Y_1 cells, d_{i2} Y_2 cells, $\cdots$, d_{in} Y_n cells, where $d_{i1}, d_{i2}, \cdots, d_{in}$ are random variables taking nonnegative integer values. $d_{i1}, d_{i2}, \cdots, d_{in}$ are not necessarily independent, but they are assumed to be independent of the exponential reaction time.

For example, consider 2 types:



It means that for a Y_1 cell, it will wait for an exponential time with expectation $1/2$, and turn into 2 Y_1 cells with probability 0.4, turn into nothing (die) with probability 0.1, and turn into 1 Y_2 cell with probability 0.5.

We can use another equivalent way to describe a branching process. For the above example, a Y_1 cell has three independent exponential alarm clocks. The first one will ring after an exponential time with expectation $1.25 = (1/2)/0.4$, the second with expectation $5 = (1/2)/0.1$, and the third with expectation $1 = (1/2)/0.5$. If the ith alarm clock rings first, then the Y_1 cell will perform the ith reaction. For example, the third alarm clock rings first, then the Y_1 will turn into one Y_2 cell.

Now we can add some invisible supplementary cell types $A_1, A_2, \cdots, A_n$, and $P_1, P_2, \cdots, P_n$, such that the branching process can be described by a group of chemical reactions.

For

$$Y_1 \to d_{11}Y_1 + d_{12}Y_2 + \cdots + d_{1n}Y_n,$$

we revise it as

$$a_1A_1 + b_1P_1 + d_{12}P_2 + \cdots + d_{1n}P_n + Y_1 \to c_1A_1 + d_{11}Y_1 + d_{12}Y_2 + \cdots + d_{1n}Y_n.$$

If $d_{11} = 0$, then $a_1 = 1$, $b_1 = 0$, $c_1 = 2$. Otherwise, $a_1 = 0$, $b_1 = d_{11} - 1$, $c_1 = 0$.

For the example above, the first reaction is

$$P_1 + Y_1 \overset{0.8}{\to} 2Y_1.$$

The second reaction is

$$A_1 + Y_1 \xrightarrow{0.2} 2A_1.$$

The third reaction is

$$A_1 + P_2 + Y_1 \xrightarrow{0.2} 2A_1 + Y_2.$$

Now we get a group of conservative chemical reactions. We require that the inverse reactions are extremely slow, such that we can neglect them. Also, the concentrations of $A_1, A_2, \cdots, A_n$ and $P_1, P_2, \cdots, P_n$ are kept invariant. Then the reaction rate of a single Y_i cell is kept the same.

Branching processes as Delbrück-Gillespie processes

Delbrück-Gillespie process is a general family of processes with discrete state space. We will show that branching process is its special case.

Delbrück-Gillespie process: Assume we have n types of particles with names $X_1, X_2, \cdots, X_n$. At time t, the number vector of particles is $(x_1(t), x_2(t), \cdots, x_n(t))$. There are m different reactions, each of which has the form

$$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{in}X_n \xrightarrow{\alpha_i} b_{i1}X_1 + b_{i2}X_2 + \cdots + b_{in}X_n.$$

Here the reaction rate α_i is a function of current population $(x_1(t), x_2(t), \cdots, x_n(t))$ and the system volume V with any form. a_{ij}, b_{ij} are arbitrary natural numbers. The process evolves as following: At time $t = 0$, generate m independent exponential variables $\tau_1, \cdots, \tau_m$, where τ_i has parameter $\alpha_i(x_1(0), x_2(0), \cdots, x_n(0))$. Assume $t_1 = \tau_j$ is the smallest among all τ_i. Then the process jumps from $(x_1(0), x_2(0), \cdots, x_n(0))$ to $(x_1(0) - a_{j1} + b_{j1}, x_2(0) - a_{j2} + b_{j2}, \cdots, x_n(0) - a_{jn} + b_{jn})$ at time t_1. Then we repeat this procedure by generating new τ_i with parameters determined with population at time t_1.

Continuous-time multi-type Markov branching process: Assume we have n types of particles with names $X_1, X_2, \cdots, X_n$. At time t, the number vector of particles is $(x_1(t), x_2(t), \cdots, x_n(t))$.

There are m different reactions, each of which has the form

$$a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{in}X_n \xrightarrow{\alpha_i} b_{i1}X_1 + b_{i2}X_2 + \cdots + b_{in}X_n.$$

This process evolves as Delbrück-Gillespie process, but there are two restrictions: (1) for one j, $a_{ij} = 1$, for $k \neq j$, $a_{ik} = 0$. This means that the reactant is a single particle. (2) $\alpha_i = c_i x_j(t)$, where c_i is a positive constant. Under these two restrictions, all particles are independent.

Therefore, continuous-time multi-type Markov branching process is a special case of Delbrück-Gillespie process.

Chapter 4

ON THE BOUNDARY BETWEEN QUALITATIVE AND QUANTITATIVE METHODS FOR CAUSAL INFERENCE

4.1 Introduction

Inferring causal relationships is among the most important goals in artificial intelligence as humans tend to think causally rather than probabilistically. A formal approach to represent causal relationships uses causal directed acyclic graphs [114], where random variables are represented as nodes and causal relationships are represented as arrows. Besides qualitatively describing causal relationships via causal directed acyclic graphs, it is often desirable to obtain quantitative measures of the strength of arrows therein since they provide more detailed information. There have been many measures proposed in the literature to quantify the causal relationships between nodes in a causal directed acyclic graph, such as the average treatment effect (ATE) [102, 134], conditional mutual information (CMI) [27], causal strength (CS) [68] and part mutual information (PMI) [166].

An interesting observation is that none of these measures is both identifiable and reasonable when the causal system under consideration is degenerate. As a simple example, consider the confounder triangle $Z \to X \to Y$ with an edge $Z \to Y$, where $Z = X$ almost surely. For this example, the average treatment effect of X on Y is not identifiable, the conditional mutual information $\text{CMI}(X, Y \mid Z)$ always equals zero even when X influences Y strongly, while the causal strength and part mutual information for the arrow $X \to Y$ are not well-defined. These problems should not be viewed as drawbacks of these existing measures, since the causal effect of X on Y is not distinguishable from the causal effect of Z on Y. Instead, they reflect the intrinsic difficulty of defining causal measures for this particular degenerate system.

In this paper, we generalize the observation above by providing a formal characterization of a degenerate causal system, and argue that qualitative causal methods should be used in place

of quantitative causal methods when the causal system is degenerate. Our characterization of a degenerate causal system is based on the notion of Markov boundary. The Markov boundary of a variable W in a variable set $\mathcal{S}$ is a minimal subset of $\mathcal{S}$, conditional on which all the remaining variables in $\mathcal{S}$, excluding W, are rendered statistically independent of W [145]. Under certain criteria for a "reasonable" quantitative measure of causal effect, we show that if Y has multiple Markov boundaries in $\mathcal{S}$, its possible parents, then in general there is no measures for the causal effect of X on Y that is both reasonable and identifiable. In this case, analysts may instead use qualitative causal methods and report all the causal explanations (i.e. Markov boundaries) of the response variable.

We then propose practical approaches to determine the uniqueness of Markov boundary from data. Many methods for Markov boundary discoveries exist in the literature, such as IAMB [149], KIAMB [116], Semi-Interleaved HITON-PC [1] and TIE* [145]. However, they often require strong assumptions or output all the Markov boundaries, which is not necessary for our purpose. In this paper, *we instead derive a necessary and sufficient condition for the uniqueness of Markov boundary, based on which we develop algorithms that are computationally more tractable. Moreover, the validity of our algorithms rely on fewer assumptions.*

4.2 Background

4.2.1 Set-up

Consider a causal directed acyclic graph Γ with vertices $\mathcal{V}$. We say X is a parent of Y if the path $X \to Y$ is present in Γ, and Y is a descendant of X is a path $X \to \cdots \to Y$ is present in Γ. A variable is a descendant of itself, but not a parent of itself. For a variable W, we use $\textsc{Des}(W)$ to denote the set consisting of all descendants of W, and $\textsc{Pa}(W)$ to denote the set consisting of all parents of W. We assume the probability distribution $\mathbb{P}$ over variables $\mathcal{V}$ is Markov with respect to Γ in the sense that for every $W \in \mathcal{V}$, W is independent of $\mathcal{V} \setminus \textsc{Des}(W)$ conditional on $\textsc{Pa}(W)$ [144].

We assume that we observe independent replications of $\mathcal{V}$. Set $\mathcal{S}$ to be all the possible parents

of Y. This means that we discard all variables that are known not to be the parents of Y. In particular, this can be achieved by excluding all variables that happen after Y from our analysis. Therefore $\mathcal{S} \setminus \{X\}$ is sufficient to control for confounding between X and Y. For example, if we know the full causal directed acyclic graph, then $\mathcal{S} = \text{PA}(Y)$. If we have no knowledge of the causal directed acyclic graph, then $\mathcal{S} = \mathcal{V} \setminus \{Y\}$. Suppose X is a possible parent of Y and we are interested in the causal effect of X on Y. If the full causal directed acyclic graph Γ is known, set $\mathcal{L} = \text{PA}(Y) \setminus \{X\}$. We denote the sample space of $X, Y, \mathcal{L}$ by $\mathbb{X}, \mathbb{Y}, \mathbb{L}$, respectively.

4.2.2 Causal effect measures

We first review several causal effect measures commonly used in practice. For simplicity, we only give the definitions in the discrete case. They can be easily extended to accommodate continuous variables by replacing summation with integration.

Average treatment effect. Under our setting, if X is binary, then the average treatment effect ATE can be defined as [131]

$$\text{ATE}(X \to Y) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \sum_{l \in \mathbb{L}} -(-1)^x y f(y \mid x, l) f(l),$$

where f denotes the probability mass function. The average treatment effect describes the shift in the mean of Y caused by switching X from 0 to 1, but fails to capture the influence of X on higher order moments of the distribution of Y.

Conditional mutual information. The conditional mutual information CMI between X and Y conditional on $\mathcal{L}$ is defined as [27]

$$\text{CMI}(X, Y \mid \mathcal{L}) = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \sum_{l \in \mathbb{L}} f(x, y, l) \log \frac{f(x, y \mid l)}{f(x \mid l) f(y \mid l)}.$$

It can alternatively be expressed as

$$\text{CMI}(X, Y \mid \mathcal{L}) = \text{H}(X, \mathcal{L}) + \text{H}(Y, \mathcal{L}) - \text{H}(\mathcal{L}) - \text{H}(X, Y, \mathcal{L}), \tag{4.1}$$

where $\mathrm{H}(X) = -\sum_{x\in\mathbb{X}} f(x)\log f(x)$ is the (Shannon) entropy. It is easy to see from (4.1) that CMI is a continuous function of the joint distribution of $(X, Y, \mathcal{L})$. Furthermore, unlike ATE, CMI captures all conditional independences between variables in the sense that $\mathrm{CMI}(X, Y \mid \mathcal{L}) = 0$ if and only if $X \perp\!\!\!\perp Y \mid \mathcal{L}$.

Specifically, when the condition set $\mathcal{L} = \emptyset$, CMI degenerates to mutual information (MI): $\mathrm{MI}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y)$. We have $\mathrm{CMI}(X, Y \mid \mathcal{L}) = \mathrm{MI}(\{X\} \cup \mathcal{L}, Y) - \mathrm{MI}(Y, \mathcal{L})$.

Causal strength. The causal strength CS of X on Y is defined as [68]

$$\mathrm{CS}(X \to Y) = \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}}\sum_{l\in\mathbb{L}} f(x, y, l)\log\frac{f(y \mid x, l)}{\sum\limits_{x'\in\mathbb{X}} f(y \mid x', l)f(x')}.$$

Part mutual information. The part mutual information PMI between X and Y conditional on $\mathcal{L}$ is defined as [166]

$$\mathrm{PMI}(X, Y \mid \mathcal{L}) = \sum_{x\in\mathbb{X}}\sum_{y\in\mathbb{Y}}\sum_{l\in\mathbb{L}} f(x, y, l)\log\frac{f(x, y \mid l)}{f^*(x \mid l)f^*(y \mid l)},$$

where $f^*(x \mid l) = \sum_{y\in\mathbb{Y}} f(x \mid y, l)f(y)$, $f^*(y \mid l) = \sum_{x\in\mathbb{X}} f(y \mid x, l)f(x)$. One may also show that

$$\begin{aligned}\mathrm{PMI}(X, Y \mid \mathcal{L}) =& \sum_{x,y,l} f(x, y, l)\log\frac{f(y \mid x, l)}{\sum_{x'} f(y \mid x', l)f(x')} \\ &+ \sum_{x,y,l} f(x, y, l)\log\frac{f(x \mid y, l)}{\sum_{y'} f(x \mid y', l)f(y')} \\ &- \sum_{x,y,l} f(x, y, l)\log\frac{f(x, y \mid l)}{f(x \mid l)f(y \mid l)}.\end{aligned} \tag{4.2}$$

The first term on the right hand side of (4.2) equals $\mathrm{CS}(X \to Y)$, and the third term equals $-\mathrm{CMI}(X, Y \mid \mathcal{L})$.

The definitions of ATE and CS need the full causal directed acyclic graph. In the definitions of CMI and PMI, the full causal directed acyclic graph is not necessary, and we can substitute the conditional set $\mathcal{L}$ by other set, but then the interpretation will be irreplaceable information, not

direct causal effect.

Other measures. Other measures of causal effect between random variables include Pearson correlation coefficient and partial correlation. For time series data, effect measures commonly used in practice include Granger causality [54], conditional Granger causality [47], partial Granger causality [56], transfer entropy [135] and conditional transfer entropy [73, 40].

4.2.3 Markov blanket and Markov boundary

Definition 4.1. *A Markov blanket $\mathcal{M}$ of a variable W within an observed variable set $\mathcal{T}$ ($W \notin \mathcal{T}$) is a subset of $\mathcal{T}$ such that*

$$W \perp\!\!\!\perp (\mathcal{T} \setminus \mathcal{M}) \mid \mathcal{M}. \tag{4.3}$$

Using the notion of mutual information, condition (4.3) can be written as $\text{CMI}(W, \mathcal{T} \mid \mathcal{M}) = 0$, or equivalently, $\text{MI}(W, \mathcal{T}) = \text{MI}(W, \mathcal{M})$. In this sense, the Markov blanket $\mathcal{M}$ contains all the information of $\mathcal{T}$ on W.

The following proposition is a direct application of the weak union property of probability distributions [111].

Proposition 4.1. *Any superset of a Markov blanket is still a Markov blanket.*

Definition 4.2. *A Markov boundary is a minimal Markov blanket. In other words, it is a Markov blanket such that no proper subsets of which is also a Markov blanket.*

Even though Markov boundaries are minimal, they need not be unique. As illustrated in our confounder triangle example, both $\{X\}$ and $\{Z\}$ are Markov boundaries of Y. As another example, consider the causal directed acyclic graph in Figure 4.1. Nodes X, Y, Z take value in $\{0, 1, 2\}$, while W takes value in $\{0, 1\}$. One may verify that there exist two Markov boundaries of Y: $\{X, W\}$ and $\{Z, W\}$. See also [145].

Unlike the confounder triangle example, the two Markov boundaries of Y in Figure 4.1 do not coincide almost surely. However, they are still variation dependent since $\text{pr}(X = 2) > 0$, $\text{pr}(Z = 0) > 0$, but $\text{pr}(X = 2, Z = 0) = 0$. This variation dependence is in fact an essential property of the multiplicity of Markov boundaries.

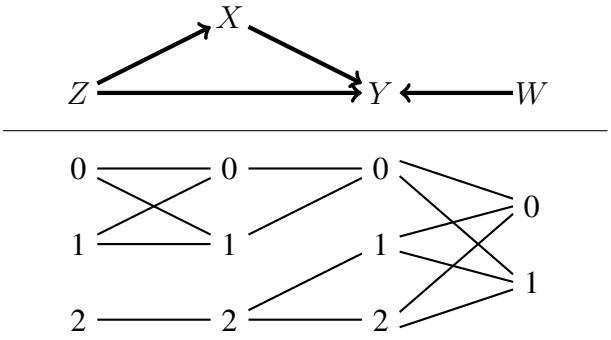


Figure 4.1: A causal directed acyclic graph for which variable Y has multiple Markov boundaries. Combinations of values connected with lines have positive joint probabilities. For example, pr$(Z = 1, X = 0, Y = 0, W = 0) > 0$, while pr$(Z = 1, X = 2, Y = 1, W = 1) = 0$.

Lemma 4.1. *For a variable set $\mathcal{T}$ with $Y \notin \mathcal{T}$, let Θ denote all Markov boundaries of Y in $\mathcal{T}$. If X is contained in at least one, but not all of the Markov boundaries of Y so that $X \in \bigcup_{\mathcal{M}\in\Theta} \mathcal{M} \setminus \bigcap_{\mathcal{M}\in\Theta} \mathcal{M}$, and $\mathcal{K} = \mathcal{T} \setminus \{X\}$, then X and $\mathcal{K}$ are variation dependent in that there exist $x \in \mathbb{X}, k \in \mathbb{K}$ such that $f(x) > 0$, $f(k) > 0$, but $f(x, k) = 0$.*

Proof. Consider two Markov boundaries $\mathcal{M}_1$, $\mathcal{M}_2$ within $\{X\} \cup \mathcal{K}$. Let $\mathcal{M}_1 = \{X\} \cup \mathcal{Z}^1$, $X \notin \mathcal{M}_2$, $\mathcal{M}_2 \setminus \mathcal{M}_1 = \mathcal{Z}^2$, $\mathcal{K} \cup \{X\} \setminus (\mathcal{M}_1 \cup \mathcal{M}_2) = \mathcal{Z}^3$, where $\mathcal{Z}^1 = \{Z_1, \ldots, Z_n\}$, $\mathcal{Z}^2 = \{Z'_1, \ldots, Z'_m\}$, $\mathcal{Z}^3 = \{Z''_1, \ldots, Z''_l\}$. Therefore $\mathcal{K} = \mathcal{Z}^1 \cup \mathcal{Z}^2 \cup \mathcal{Z}^3$.

Fix $z_0^1 \in \mathbb{Z}^1$ such that $f(z_0^1) > 0$. Assume that for $x_i \in \mathbb{X}$, $f(x_i, z_0^1) > 0$ is true for $i \in \{1, \ldots, p\}$. Assume that for $z_j^2 \in \mathbb{Z}^2$, $f(z_0^1, z_j^2) > 0$ is true for $j \in \{1, \ldots, q\}$. Consider any $y \in \mathbb{Y}$.

To obtain contradiction, we assume that $f(x_i, z_0^1, z_j^2) > 0$ for all $i \in \{1, \ldots, p\}$ and all $j \in \{1, \ldots, q\}$.

Since $X \perp\!\!\!\perp Y \mid (\mathcal{Z}^1, \mathcal{Z}^2)$ (Proposition 4.1) for all $i, r \in \{1, \ldots, p\}$ and all $j \in \{1, \ldots, q\}$,

$$f(y \mid x_i, z_0^1, z_j^2) = f(y \mid x_r, z_0^1, z_j^2).$$

Since $\mathcal{Z}^2 \perp\!\!\!\perp Y \mid (X, \mathcal{Z}^1)$ for all $r \in \{1, \ldots, p\}$ and all $j, s \in \{1, \ldots, q\}$,

$$f(y \mid x_r, z_0^1, z_j^2) = f(y \mid x_r, z_0^1, z_s^2).$$

All the conditions have positive probabilities, so the conditional probabilities are well-defined.

Then we have

$$f(y \mid x_i, z_0^1, z_j^2) = f(y \mid x_r, z_0^1, z_s^2),$$

for all $i, r \in \{1, \ldots, p\}$ and all $j, s \in \{1, \ldots, q\}$.

Since this is true for any possible values of X and $\mathcal{Z}^2$ when $\mathcal{Z}^1 = z_0^1$, we know that

$$f(y \mid x_i, z_0^1, z_j^2) = f(y \mid z_0^1).$$

Therefore, for all $z_1^1 \in \mathcal{Z}^1$ with $f(z_1^1) > 0$, all $y \in \mathbb{Y}$ and all i, j,

$$f(x_i, z_j^2, y \mid z_1^1) = f(x_i, z_j^2 \mid z_1^1) f(y \mid z_1^1)$$

is valid.

This implies that $(X, \mathcal{Z}^2) \perp\!\!\!\perp Y \mid \mathcal{Z}^1$, therefore $X \perp\!\!\!\perp Y \mid \mathcal{Z}^1$, $\text{MI}(Y, \mathcal{Z}^1) = \text{MI}(Y, (X, \mathcal{Z}^1))$. Since $\mathcal{M}_1 = \{X\} \cup \mathcal{Z}^1$, $\text{MI}(Y, (X, \mathcal{Z}^1)) = \text{MI}(Y, \mathcal{K})$. Thus $\text{MI}(Y, \mathcal{Z}^1) = \text{MI}(Y, \{X\} \cup \mathcal{K})$, implying that $\mathcal{Z}^1$ is a Markov blanket, which is a contradiction. So there exists $x \in \mathbb{X}, z_0^1 \in \mathbb{Z}^1, z_1^2 \in \mathbb{Z}^2$ such that $f(x, z_0^1) > 0$ (implies $f(x) > 0$), $f(z_0^1, z_1^2) > 0$, but $f(x, z_0^1, z_1^2) = 0$. Choose $z_1^3 \in \mathbb{Z}^3$ such that $f(z_0^1, z_1^2, z_1^3)$, and let $k = (z_0^1, z_1^2, z_1^3)$, then $f(x) > 0$, $f(k) > 0$, but $f(x, k) = 0$.

□

Remark 4.1. *Let $\mathcal{M}' \in \{\mathcal{M} \in \Theta : X \notin \mathcal{M}\}$ be a Markov boundary that does not contain X. We point out that X and $\mathcal{M}'$ may be variation independent. For example, suppose X, Z, W are independent Bernoulli variables with mean 0.5, $Y = X + Z \mod 2$ and $U = W + Y \mod 2$, where $a \mod b$ denotes the remainder of a divided by b. In this case Y has two Markov boundaries: $\{X, Z\}$ and $\{W, U\}$, and X is variation independent of $\{W, U\}$.*

We now review some conditions under which the Markov boundary is unique.

Definition 4.3 (Faithfulness)**.** *Let Λ denote the collection of conditional independence relationships shared by all probability distributions that are Markov with respect to Γ. A probability dis-*

tribution is faithful to Γ *if and only if its conditional independence relationships are characterized by* Λ *[116].*

Definition 4.4 (Intersection)**.** *A probability distribution on* $\mathcal{V}$ *satisfies the* intersection *property if and only if for any four subsets of variables* $\mathcal{P}$*,* $\mathcal{Q}$*,* $\mathcal{Z}$*,* $\mathcal{W}$ *such that* $\mathcal{P} \perp\!\!\!\perp \mathcal{Z} \mid (\mathcal{Q}, \mathcal{W})$*,* $\mathcal{P} \perp\!\!\!\perp \mathcal{W} \mid (\mathcal{Q}, \mathcal{Z})$*, we have* $\mathcal{P} \perp\!\!\!\perp (\mathcal{Z}, \mathcal{W}) \mid \mathcal{Q}$*.*

Definition 4.5 (Strict positiveness)**.** *A probability distribution on* $\mathcal{V}$ *is called* strictly positive *if and only if for any two disjoint subsets of variables* $\mathcal{X}$ *and* $\mathcal{Z}$ *such that pr*$(\mathcal{X} = x) > 0$*, pr*$(\mathcal{Z} = z) > 0$*, we have pr*$(\mathcal{X} = x, \mathcal{Z} = z) > 0$*.*

Proposition 4.2. *If a probability distribution on* $\mathcal{V}$ *(i) is faithful to* Γ*, or (ii) has intersection property, or (iii) is strictly positive, then any variable* $Z \in \mathcal{V}$ *has a unique Markov boundary in* $\mathcal{V} \setminus \{Z\}$*.*

We refer readers to [112] and [111] for a proof of Proposition 4.2. None of the three conditions in Proposition 4.2 is necessary. For example, suppose X, Y, Z, $W \in \{0, 1\}$, pr$(X = Z = Y = W = 0) = 0.5$, pr$(X = Z = 1, Y = W = 0) = 0.25$, pr$(X = Z = Y = W = 1) = 0.25$ and $\mathcal{S} = \{X, Y, Z, W\}$. Since $Y \perp\!\!\!\perp X \mid Z$, $Y \perp\!\!\!\perp Z \mid X$, $Y \not\perp\!\!\!\perp X, Z$, the joint distribution of (X, Z, Y, W) does not have intersection property and is hence not faithful. Since pr$(X = 0) > 0$, pr$(Z = 1) > 0$, pr$(X = 0, Z = 1) = 0$, this distribution is not strictly positive. However, each variable in this example has a unique Markov boundary within the other three variables.

There are many applications for which the Markov boundary of the response variable may not be unique. For instance, in breast cancer studies, it may be the case that several gene sets have nearly the same effect for survival prediction [34], such that each of the gene sets is a Markov boundary of the survival indicator. We refer interested readers to [145] for more real-world examples on multiplicity of Markov boundaries.

4.3 Quantifying causal effects with multiple Markov boundaries

4.3.1 Problems of known causal effect measures

In this section, we show that the commonly used causal effect measures may not be reasonable or identifiable when the response variable Y has multiple Markov boundaries in $\text{PA}(Y) = \mathcal{L} \cup \{X\}$.

On the one hand, for $X \in \bigcup_{\mathcal{M}\in\Theta} \mathcal{M} \setminus \bigcap_{\mathcal{M}\in\Theta} \mathcal{M}$, we have $\text{CMI}(X, Y \mid \mathcal{L}) = 0$ even when X influences Y strongly. The reason is Proposition 4.1 and the fact that $\mathcal{L}$ contains a Markov boundary. This is not desirable, as pointed out by [68, 166]. On the other hand, Lemma 4.1 shows that X and $\mathcal{L}$ are variation dependent. It follows that $f(y \mid x, l)$ may not always be defined even when $f(x) > 0$ and $f(l) > 0$. It follows that $\text{ATE}(X \to Y)$, $\text{CS}(X \to Y)$ and $\text{PMI}(X, Y \mid \mathcal{L})$ are not defined.

A natural solution to the latter problem is to assign a value in these degenerate scenarios. However, Theorem 4.1 shows that the resulting quantities are not continuous functions of the joint distribution of $(X, Y, \mathcal{L})$. In the following, for any given probability distribution $\mathbb{P}'$, we can calculate the corresponding causal quantities, denoted as $\text{CS}[\mathbb{P}'](X \to Y)$, $\text{PMI}[\mathbb{P}'](X, Y \mid \mathcal{L})$ and so on.

Theorem 4.1. *If $X \in \bigcup_{\mathcal{M}\in\Theta} \mathcal{M} \setminus \bigcap_{\mathcal{M}\in\Theta} \mathcal{M}$, then there exist two sequences of distributions on $(X, Y, \mathcal{L})$, denoted as $\{\mathbb{P}_1, \mathbb{P}_2, \ldots\}$ and $\{\mathbb{P}'_1, \mathbb{P}'_2, \ldots\}$, both of which converge to $\mathbb{P}$ under the total variation distance, but $\lim_{i\to\infty} \text{CS}[\mathbb{P}_i](X \to Y) \neq \lim_{i\to\infty} \text{CS}[\mathbb{P}'_i](X \to Y)$. The same results apply to $\text{ATE}(X \to Y)$ and $\text{PMI}(X, Y \mid \mathcal{L})$.*

Theorem 4.1 may be proved using Lemma 4.1 and the following Lemma 4.2.

Lemma 4.2. *Assume there exist $x \in \mathbb{X}, l \in \mathbb{L}$ such that $f(x) > 0$, $f(l) > 0$, but $f(x, l) = 0$. Then there exists two real numbers $g_1 < g_2$, such that for any g with $g_1 < g < g_2$, any $\delta > 0$, there exists a probability distribution $\mathbb{P}'$ with $\mathrm{d}(\mathbb{P}, \mathbb{P}') < \delta$, such that $\text{CS}[\mathbb{P}'](X \to Y) = g$. The same results apply to $\text{ATE}(X \to Y)$ and $\text{PMI}(X, Y \mid \mathcal{L})$.*

Proof. In the following we will assume there is only one pair of (x, l) such that $f(x) > 0$, $f(l) > 0$, $f(x, l) = 0$. If there are multiple pairs, we can treat them one by one.

We construct a family of probability distributions $\mathbb{P}_i^\eta$ with mass functions f_i^η based on $\mathbb{P}$. For $(x', l') \neq (x, l)$, $f_i^\eta(x', y, l') = (1 - \eta) f(x', y, l')$. $f_i^\eta(x, l) = \eta > 0$, $f_i^\eta(y_j \mid x, l) = \alpha_i^j$, where $\alpha_i^j \geq 0$, $\sum_j \alpha_i^j = 1$. Then for each i, $\text{CS}[\mathbb{P}_i^\eta](X \to Y)$ can be defined, and when $\eta \to 0$, f_i^η converges to f. The total variation distance between f and f_i^η is η.

When $\eta \to 0$,

$$\text{CS}[\mathbb{P}_i^\eta](X \to Y) = \sum_{x' \in \mathbb{X}} \sum_{y' \in \mathbb{Y}} \sum_{l' \neq l} f_i^\eta(x', y', l') \log \frac{f_i^\eta(y' \mid l', x')}{\sum_{x'' \in \mathbb{X}} f_i^\eta(y' \mid l', x'') f_i^\eta(x'')}$$

$$+ \sum_{x' \in \mathbb{X}} \sum_{y' \in \mathbb{Y}} f_i^\eta(x', y', l) \log \frac{f_i^\eta(y' \mid l, x')}{\sum_{x'' \in \mathbb{X}} f_i^\eta(y' \mid l, x'') f_i^\eta(x'')}$$

$$\to \sum_{x' \in \mathbb{X}} \sum_{y' \in \mathbb{Y}} \sum_{l' \neq l} f(x', y', l') \log \frac{f(y' \mid l', x')}{\sum_{x'' \in \mathbb{X}} f(y' \mid l', x'') f(x'')}$$

$$+ \sum_{x' \neq x} \sum_{y' \in \mathbb{Y}} f(x', y', l) \log f(y' \mid x', l) - \sum_j f(y_j, l) \log\{f(x)\alpha_i^j + \sum_{x' \neq x} f(x') f(y_j \mid x', l)\}.$$

For different i, when we let $\eta \to 0$, the only different terms are $-\sum_j f(y_j, l) \log\{f(x)\alpha_i^j + \sum_{x' \neq x} f(x') f(y_j \mid x', l)\}$. We will show that the above term is not a constant with $\{\alpha_i^j\}$. Therefore we can find two groups of $\{\alpha_i^j\}$ for $i = 1, 2$ such that $g_1 = \lim_{\eta \to 0} \text{CS}[\mathbb{P}_1^\eta](X \to Y) < \lim_{\eta \to 0} \text{CS}[\mathbb{P}_2^\eta](X \to Y) = g_2$.

If there is only one y_1 such that $f(y_1, l) > 0$, then

$$-\sum_j f(y_j, l) \log\{f(x)\alpha_i^j + \sum_{x' \neq x} f(x') f(y_j \mid x', l)\}$$

$$= -f(y_1, l) \log\{f(x)\alpha_i^1 + \sum_{x' \neq x} f(x') f(y_1 \mid x', l)\}.$$

It is not a constant when we change α_i^1.

If there are at least two values y_1, y_2 of Y, such that $f(y_1, l) > 0$, $f(y_2, l) > 0$, then we can change α_i^1 while keeping $\alpha_i^1 + \alpha_i^2 = d$, and leave other α_i^j fixed.

Set $f(y_1, l) = a_1$, $f(y_2, l) = a_2$, $f(x) = c$, $\sum_{x' \neq x} f(x') f(y_1 \mid x', l) = b_1$, $\sum_{x' \neq x} f(x') f(y_2 \mid x', l) = b_2$. All these terms are positive. Then in $-\sum_j f(y_j, l) \log\{f(x)\alpha_i^j + \sum_{x' \neq x} f(x') f(y_j \mid x', l)\}$, terms containing α_i^1 and α_i^2 are

$$-a_1 \log(c\alpha_i^1 + b_1) - a_2 \log\{c(d - \alpha_i^1) + b_2\}.$$

Its derivative with respect to α_i^1 is

$$-\frac{a_1 c}{c\alpha_i^1 + b_1} + \frac{a_2 c}{c(d - \alpha_i^1) + b_2}.$$

If the derivative always equal 0 in an interval, then we should have

$$\frac{a_1}{a_2} \equiv \frac{c\alpha_i^1 + b_1}{c(d - \alpha_i^1) + b_2},$$

which is incorrect.

Now we have two groups of $\{\alpha_i^j\}$ for $i = 1, 2$ such that $g_1 = \lim_{\eta \to 0} \text{CS}[\mathbb{P}_1^\eta](X \to Y) < \lim_{\eta \to 0} \text{CS}[\mathbb{P}_2^\eta](X \to Y) = g_2$. Then for any $g \in (g_1, g_2)$, any $\delta > 0$, we can find $\eta < \delta$ small enough such that $\text{CS}[\mathbb{P}_1^\eta](X \to Y) < g$, $\text{CS}[\mathbb{P}_2^\eta](X \to Y) > g$. Then we change $\{\alpha_1^j\}$ continuously to $\{\alpha_2^j\}$. During this process CS is always defined, and there exists $\{\alpha_3\}$ such that $\text{CS}[\mathbb{P}_3^\eta](X \to Y) = g$.

This shows that $\text{CS}(X \to Y)$ is essentially ill-defined.

Since $\text{CS}(X \to Y)$ and $\text{PMI}(X, Y \mid \mathcal{L})$ have the same non-zero terms containing $f(\cdot \mid x, l)$, the same argument shows that $\text{PMI}(X, Y \mid \mathcal{L})$ is not well-defined.

The proof of $\text{ATE}(X \to Y)$ is similar, but much more straightforward. We can change the values of $f(y_i \mid x, l)$ for different i, while keeping their sum. Since $\text{ATE}(X \to Y)$ is linear with each $f(y \mid x, l)$, and the corresponding coefficients are different, the conclusion is obvious. □

Remark 4.2. *Lemma 4.2 is similar in flavor to the Picard's great theorem: If an analytic function f has an essential singularity at a point w, then on any punctured neighborhood of w, $f(z)$ takes on all possible complex values, with at most a single exception. In this sense,* CS *and* PMI *are "essentially singular" at the probability distribution that implies multiple Markov boundaries for Y.*

To get intuition into Theorem 4.1, consider the following example: $X, Z \in \{1, \ldots, 2n\}, Y \in \{0, 1, 2, 3\}$. The causal directed acyclic graph is $X \to Y \leftarrow Z$ with an edge $Z \to X$. For $i = 1, \ldots, n$, $\text{pr}(X = i, Y = 0, Z = i) = \zeta[1/n - 2n\epsilon - (2n - 1)\delta]/3$, $\text{pr}(X = i, Y = 2, Z = i) = 2(1 - \zeta)[1/n - 2n\epsilon - (2n - 1)\delta]/3$. For $i = n + 1, \ldots, 2n$, $\text{pr}(X = i, Y = 0, Z = i) =$

$2\zeta[1/n - 2n\epsilon - (2n-1)\delta]/3$, pr$(X = i, Y = 2, Z = i) = (1-\zeta)[1/n - 2n\epsilon - (2n-1)\delta]/3$. For $i = 1, \ldots, 2n$, pr$(X = i, Y = 1, Z = i) = \zeta\epsilon/2$, pr$(X = i, Y = 3, Z = i) = (1-\zeta)\epsilon/2$. For $i \neq j$, pr$(X = i, Y = 0, Z = j) = \zeta\delta/2$, pr$(X = i, Y = 1, Z = j) = \zeta\epsilon/2$, pr$(X = i, Y = 2, Z = j) = (1-\zeta)\delta/2$, pr$(X = i, Y = 3, Z = j) = (1-\zeta)\epsilon/2$.

Letting $\epsilon \to 0$, $\delta \to 0$, then CS$(X \to Y)$ and PMI$(X, Y \mid Z)/2$ both converges to

$$\log(3n) + \frac{2}{3}\log 2 - \frac{2-\zeta}{3}\log(2-\zeta) - \frac{1+\zeta}{3}\log(1+\zeta)$$

$$-\frac{1}{3}\zeta\log\{1 + \frac{\delta}{\delta+\epsilon}(3n+\zeta-2)\} - \frac{2}{3}(1-\zeta)\log\{2 + \frac{\delta}{\delta+\epsilon}(3n+\zeta-2)\}$$

$$-\frac{2}{3}\zeta\log\{2 + \frac{\delta}{\delta+\epsilon}(3n-\zeta-1)\} - \frac{1}{3}(1-\zeta)\log\{1 + \frac{\delta}{\delta+\epsilon}(3n-\zeta-1)\}.$$

By changing the value of δ/ϵ, PMI$(X, Y \mid Z)/2$ and CS$(X \to Y)$ can take any value in an interval. When n is large enough, the interval is approximately $(0, \log n)$.

When $\delta = \epsilon = 0$, $0 < \zeta < 1$, Y has two Markov boundaries: $\{X\}$ and $\{Z\}$. When δ, ϵ, ζ tend to 0, H$(Y) \to 0$, but CS and PMI can still take large values, which is problematic.

We can see that near a distribution with multiple Markov boundaries, CS and PMI will change drastically. Therefore calculating CS or PMI of a distribution which is close to another distribution with multiple Markov boundaries might not be numerically feasible.

Theorem 4.1 is also valid for PMI with another condition set $\mathcal{J}$, especially when we do not have the full causal directed acyclic graph. Here the condition is that Y has multiple Markov boundaries within $\mathcal{J} \cup \{X\}$.

4.3.2 Criteria for reasonable causal effect measures

We propose the following criteria for reasonable definitions of the strength of arrows in a causal graph.

C0. The strength of $X \to Y$ is identifiable from the joint distribution of Y and its possible parents $\mathcal{S}$.

C1. If there is unique Markov boundary $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \notin \mathcal{M}$, then the strength of

$X \to Y$ is 0.

C2. If there is unique Markov boundary $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \in \mathcal{M}$, then the strength of $X \to Y$ is at least $\text{CMI}(X, Y \mid \mathcal{M} \setminus \{X\})$.

C3. The strength of $X \to Y$ is a continuous function of the joint distribution of Y and $\mathcal{S}$, under total variation distance.

C0 is similar to the "Locality" postulate P2 in [68]. If Z is not a parent of Y, then it impacts Y through some parents of Y. In other words, the knowledge of all possible parents should suffice to determine the strength of $X \to Y$. C1 is reasonable since Markov boundary contains all the information on Y from $\mathcal{S}$. It is natural to say that X has no causal effect on Y if $X \notin \mathcal{M}$. On the other hand, if X is contained in the unique Markov boundary $\mathcal{M}$, then it contains non-trivial information of Y since $\text{MI}(Y, \mathcal{S} \setminus \{X\}) < \text{MI}(Y, \mathcal{S})$. In this case, it is reasonable to assign a positive value to $X \to Y$. In C2, the concrete form of the lower bound is not important. We only require that it is positive, and relies only on the joint distribution of $\mathcal{M}$ and Y. Finally, if C3 fails, then a small observational error might lead to significant error in the estimated strength of arrow $X \to Y$.

4.3.3 An impossibility result

In this section, we show that with multiple Markov boundaries, reasonable causal effect measures are impossible, since the above criteria are incompatible.

Theorem 4.2. *Assume there is no extra knowledge of the causal directed acyclic graph. Assume X is contained in some Markov boundary, but not all the Markov boundaries. Then we cannot define the strength of $X \to Y$ satisfying criteria C0, C1, C2 and C3.*

To provide intuition, we first show Theorem 4.2 for a simple distribution. Consider X, Y, Z which only take $0, 1, 2$. $\text{pr}(X = 0, Y = 0, Z = 0) = 1/2 - \epsilon$, $\text{pr}(X = 1, Y = 2, Z = 1) = 1/4$, $\text{pr}(X = 1, Y = 0, Z = 1) = 1/4$, $\text{pr}(X = 2, Y = 2, Z = 0) = \epsilon$, where ϵ is a small positive number. Here $\{X\}$ is the unique Markov boundary of Y, so from criterion C2, the strength of $X \to Y$ should be at least $\text{CMI}(X, Y \mid \emptyset) = \text{MI}(X, Y) \approx 0.216$. Now consider a perturbation of the above example: $\text{pr}(X = 0, Y = 0, Z = 0) = 1/2 - \epsilon$, $\text{pr}(X = 1, Y = 2, Z = 1) = 1/4$,

$\text{pr}(X = 1, Y = 0, Z = 1) = 1/4$, $\text{pr}(X = 0, Y = 2, Z = 2) = \epsilon$. Here we just switch X and Z in the joint distribution. Now $\{Z\}$ is the unique Markov boundary of Y, so from criterion C1, the strength of $X \to Y$ should be 0. We can see that with an arbitrarily small perturbation, the strength of $X \to Y$ has to change drastically. Thus we cannot give a satisfactory value when $\epsilon = 0$, while keeping the strength continuous. Therefore criterion C3 is violated. Near $\epsilon = 0$, such quantity cannot be uniformly continuous, so no matter how we increase the accuracy of measuring joint distribution, we might still have significant errors.

The key idea above is to show that under some arbitrarily small perturbation, there is only one Markov boundary $\mathcal{M}$ which contains X. Let the magnitude of perturbation tend to 0, then the strength of $X \to Y$ in the original distribution should be at least $\text{CMI}(X, Y \mid \mathcal{M} \setminus \{X\})$. However, we can also add arbitrarily small perturbation such that there is only one Markov boundary, and it does not contain X. When we let the magnitude of perturbation tend to 0, we can see that the strength of $X \to Y$ in the original distribution should be 0. Such dilemma finishes the proof. First we need the description of perturbations and some lemmas.

Definition 4.6. *Given a variables Z_i. Let ξ_i be an independent Bernoulli variable with mean ϵ. If $\xi_i = 0$, then $Z_i^\epsilon = Z_i$. If $\xi_i = 1$, then $Z_i^\epsilon = U_i$, where U_i is a random variable that is independent of Z_i and ξ_i. When we add ϵ-noise on multiple variables $Z_i, Z_j, \ldots$, those variables $\xi_i, \xi_j, \ldots, U_i, U_j, \ldots$ are independent, and they are independent with $X, Y, Z_1, \ldots, Z_k$.*

We can see that substituting X by X^ϵ does not affect the joint distribution of $Y, Z_1, \ldots, Z_k$. If we add ϵ-noise on m variables, then the total variation distance of joint distributions before and after substituting is at most $m\epsilon$. Therefore, when we let $\epsilon \to 0$, the perturbed joint distribution will converge to the original distribution.

Lemma 4.3 (Strict Data Processing Inequality)**.** *$\mathcal{S}_1$ is a group of variables without X, Y. If we add ϵ-noise on X to get X^ϵ, then $\text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1) \leq \text{CMI}(X, Y \mid \mathcal{S}_1)$, and the equality holds if and only if $\text{CMI}(X, Y \mid \mathcal{S}_1) = 0$.*

This is a special case of "data processing inequality" in information theory [21]. The intuition is that transmitting data through a noisy channel cannot increase information, namely "garbage in,

garbage out". However, the original data processing inequality only indicates that the sufficient and necessary condition under which the equality holds is $\text{CMI}(X, Y \mid X^\epsilon, \mathcal{S}_1) = 0$, which is rather abstract and difficult to verify, since it depends on the form of ϵ noise. Here we propose a full proof, since we need to show that adding non-zero ϵ-noise will always decrease (instead of just "not increase") the information of Y, unless there is no information to lose. In other words, we propose the concrete form of the sufficient and necessary condition under which the equality holds. That is why the proof is much lengthier than that of the original data processing inequality.

The proofs for discrete and continuous X are different, therefore we state them separately. Whether Y is discrete or continuous does not matter, therefore we assume Y is discrete/continuous when X is discrete/continuous. We impose some restrictions to simplify the proofs. If Z_i is discrete, then U_i is an arbitrary discrete random variable which takes all and only the values of Z_i with positive probabilities. If Z_i is continuous, then U_i is continuous, and its density function is always positive.

Proof. When X is discrete:

$\text{CMI}(X, Y \mid \mathcal{S}_1) = \sum_{s_1} \text{pr}(\mathcal{S}_1 = s_1)\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1)$. For a fixed s_1, assume X can take values $1, \ldots, r'$, U_X can take $1, \ldots, r', \ldots, r$. Y can take values $1, \ldots, t$ with positive probabilities. Denote $\text{pr}(X = i, Y = j \mid \mathcal{S}_1 = s_1)$ by p_{ij}. Define $p_{-j} = \sum_i p_{ij}$, $p_{i-} = \sum_j p_{ij}$. With ϵ-noise, $p^\epsilon_{-j} = p_{-j}$, $p^\epsilon_{ij} = (1-\epsilon)p_{ij} + \epsilon q_i p_{-j}$, $p^\epsilon_{i-} = (1-\epsilon)p_{i-} + \epsilon q_i$. Then we have

$$\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) = \sum_{j=1}^{t}\sum_{i=1}^{r'} p_{ij} \log \frac{p_{ij}}{p_{i-}p_{-j}},$$

$$\text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1 = s_1) = \sum_{j=1}^{t}\sum_{i=1}^{r} \{(1-\epsilon)p_{ij} + \epsilon q_i p_{-j}\} \log \frac{(1-\epsilon)p_{ij} + \epsilon q_i p_{-j}}{\{(1-\epsilon)p_{i-} + \epsilon q_i\}p_{-j}}.$$

$$\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) - \text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1 = s_1) =$$

$$\sum_{j=1}^{t}\sum_{i=1}^{r}\Big[(1-\epsilon+q_i\epsilon)p_{ij}\log\frac{p_{ij}}{p_{i-}p_{-j}}+\sum_{k\neq i}q_i\epsilon p_{kj}\log\frac{p_{kj}}{p_{k-}p_{-j}}$$

$$-\{(1-\epsilon)p_{ij}+\epsilon q_i p_{-j}\}\log\frac{(1-\epsilon)p_{ij}+\epsilon q_i p_{-j}}{p_{-j}[(1-\epsilon)p_{i-}+\epsilon q_i]}\Big].$$

For fixed i, j and $k = 1, \ldots, r$, set

$$a_{ij}^k = \frac{p_{kj}}{p_{k-}p_{-j}},$$

$$b_{ij}^k = \frac{\epsilon q_i p_{k-}}{(1-\epsilon)p_{i-}+\epsilon q_i} \quad \text{for} \quad k \neq i,$$

$$b_{ij}^i = \frac{(1-\epsilon+q_i\epsilon)p_{i-}}{(1-\epsilon)p_{i-}+\epsilon q_i},$$

$$c_{ij} = p_{-j}\{(1-\epsilon)p_{i-}+\epsilon q_i\}.$$

Here we know that $p_{-j} > 0$, $(1-\epsilon)p_{i-} + \epsilon q_i > 0$. If $p_{k-} = 0$, namely $k = r'+1, \ldots, r$, then we stipulate $a_{ij}^k = 1$. In this case, the corresponding $b_{ij}^k = 0$.

Then we have

$$\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) - \text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1 = s_1)$$

$$= \sum_{j=1}^{t}\sum_{i=1}^{r} c_{ij}\{\sum_{k=1}^{r} b_{ij}^k a_{ij}^k \log a_{ij}^k - (\sum_{k=1}^{r} a_{ij}^k b_{ij}^k)\log(\sum_{k=1}^{r} a_{ij}^k b_{ij}^k)\} \geq 0.$$

The last step is Jensen's inequality, since $a_{ij}^k \geq 0$, $b_{ij}^k \geq 0$, $\sum_{k=1}^{r} b_{ij}^k = 1$, $c_{ij} > 0$, $f(x) = x\log x$ is strictly convex down when $x \geq 0$ (stipulate $0\log 0 = 0$).

The equality holds if and only if for each j, $a_{1j} = a_{2j} = \cdots = a_{r'j}$, which means p_{ij}/p_{i-} are equal for all $i \leq r'$. Since $\sum_{i=1}^{r'} p_{i-}(p_{ij}/p_{i-}) = p_{-j}$, $\sum_{i=1}^{r'} p_{i-} = 1$, we have $p_{ij}/p_{i-} = p_{-j}$ for each i, j such that $p_{i-} > 0$ and $p_{-j} > 0$. This is equivalent with that X and Y are independent conditioned on $\mathcal{S}_1 = s_1$.

$\text{CMI}(X, Y \mid \mathcal{S}_1) = 0$ if and only if X and Y are independent conditioned on any possible value of $\mathcal{S}_1$. Therefore, $\text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1) \leq \text{CMI}(X, Y \mid \mathcal{S}_1)$, and the equality holds if and only if $\text{CMI}(X, Y \mid \mathcal{S}_1) = 0$.

When X is continuous:

$$\text{CMI}(X, Y \mid \mathcal{S}_1) = \int_{-\infty}^{\infty} \text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) h(s_1) \mathrm{d}s_1,$$

where $h(s_1)$ is the probability density function of $\mathcal{S}_1$. For a fixed s_1, denote the joint probability density function of X, Y conditioned on $\mathcal{S}_1 = s_1$ by $p(x, y)$. Define $p_1(x) = \int_{-\infty}^{\infty} p(x, y) \mathrm{d}y$, $p_2(y) = \int_{-\infty}^{\infty} p(x, y) \mathrm{d}x$. With ϵ-noise, the joint probability density function of X, Y conditioned on $\mathcal{S}_1 = s_1$ is $(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)$. Notice that $\int_{-\infty}^{\infty} q(x) \mathrm{d}x = 1$, $\int_{-\infty}^{\infty} [(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)] \mathrm{d}x = p_2(y)$. Then we have

$$\begin{aligned}
&\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) - \text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1 = s_1) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) \log \frac{p(x, y)}{p_1(x) p_2(y)} \mathrm{d}x \mathrm{d}y \\
&- \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)\} \log \frac{(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)}{\{(1-\epsilon)p_1(x) + \epsilon q(x)\} p_2(y)} \mathrm{d}x \mathrm{d}y \\
= &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Big[(1-\epsilon) p(x, y) \log \frac{p(x, y)}{p_1(x) p_2(y)} + q(x) \epsilon \Big\{ \int_{-\infty}^{\infty} p(x_0, y) \log \frac{p(x_0, y)}{p_1(x_0) p_2(y)} \mathrm{d}x_0 \Big\} \\
&- \{(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)\} \log \frac{(1-\epsilon)p(x, y) + \epsilon q(x) p_2(y)}{\{(1-\epsilon)p_1(x) + \epsilon q(x)\} p_2(y)} \Big] \mathrm{d}x \mathrm{d}y.
\end{aligned}$$

For fixed x, y, we can define a probability measure on $\mathbb{R}$, $\mu_{x,y}(x_0)$, which is a mixture of discrete and continuous type measures. For the discrete component, it has probability $(1-\epsilon)p_1(x)/\{(1-\epsilon)p_1(x) + \epsilon q(x)\}$ to take x. For the continuous component, the probability density function at x_0 is $q(x)\epsilon p_1(x_0)/\{(1-\epsilon)p_1(x) + \epsilon q(x)\}$. Define $F_{x,y}(x_0) = p(x_0, y)/\{p_1(x_0) p_2(y)\}$. If $p_1(x_0) = 0$ or $p_2(y) = 0$, stipulate $F_{x,y}(x_0) = 1$.

Now we have

$$\begin{aligned}
&\text{CMI}(X, Y \mid \mathcal{S}_1 = s_1) - \text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1 = s_1) \\
= &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \{(1-\epsilon)p_1(x) + \epsilon q(x)\} p_2(y) \Big[\int_{-\infty}^{\infty} F_{x,y}(x_0) \log F_{x,y}(x_0) \mathrm{d}\mu_{x,y}(x_0)
\end{aligned}$$

$$-\Big\{\int_{-\infty}^{\infty} F_{x,y}(x_0)\mathrm{d}\mu_{x,y}(x_0)\Big\}\log\Big\{\int_{-\infty}^{\infty} F_{x,y}(x_0)\mathrm{d}\mu_{x,y}(x_0)\Big\}\Big]\mathrm{d}x\mathrm{d}y \geq 0.$$

The last step is the probabilistic form of Jensen's inequality, since $F_{x,y}(x_0)$ is non-negative and integrable with probability measure $\mu_{x,y}(x_0)$, $\{(1-\epsilon)p_1(x)+\epsilon q(x)\}p_2(y) > 0$ if $p_2(y) > 0$, and $f(x) = x\log x$ is strictly convex down when $x \geq 0$ (stipulate $0\log 0 = 0$).

The equality holds if and only if for $p_1(x_0) > 0$ and $p_2(y) > 0$, $F_{x,y}(x_0)$ is a constant with x_0, which means $p(x_0, y)/p_1(x_0)$ is a constant. Since $\int_{-\infty}^{\infty} p_1(x_0)p(x_0, y)/p_1(x_0)\mathrm{d}x_0 = p_2(y)$, $\int_{-\infty}^{\infty} p_1(x_0) = 1$, we have $p(x_0, y)/p_1(x_0) = p_2(y)$ for each x_0, y such that $p_1(x_0) > 0$ and $p_2(y) > 0$. This is equivalent with that X and Y are independent conditioned on $\mathcal{S}_1 = s_1$.

$\textsc{cmi}(X, Y \mid \mathcal{S}_1) = 0$ if and only if X and Y are independent conditioned on any possible value of $\mathcal{S}_1$, except a zero-measure set. Therefore, $\textsc{cmi}(X^\epsilon, Y \mid \mathcal{S}_1) \leq \textsc{cmi}(X, Y \mid \mathcal{S}_1)$, and the equality holds if and only if $\textsc{cmi}(X, Y \mid \mathcal{S}_1) = 0$.

□

Expand $\textsc{cmi}(X^\epsilon, Y \mid \mathcal{S}_1) \leq \textsc{cmi}(X, Y \mid \mathcal{S}_1)$, then we have $\textsc{mi}(\{X^\epsilon\} \cup \mathcal{S}_1, Y) \leq \textsc{mi}(\{X\} \cup \mathcal{S}_1, Y)$. Especially, the total information $\textsc{mi}(\mathcal{S}, Y)$ will not increase if we add ϵ-noise on possible parents of Y.

Lemma 4.4. *Assume Y has multiple Markov boundaries within $\mathcal{S} = \{X, Z_1, \ldots, Z_k\}$. Choose one Markov boundary: $\mathcal{M}_0$. If we add ϵ noise on each variable in $\mathcal{S} \setminus \mathcal{M}_0$, then in the new distribution, $\mathcal{M}_0$ is the unique Markov boundary.*

Proof. Remember that a Markov boundary $\mathcal{M}$ is a minimal subset of $\mathcal{S}$ such that $\textsc{mi}(\mathcal{M}, Y) = \textsc{mi}(\mathcal{S}, Y)$. Denote $\mathcal{S}$ with ϵ-noise on $Z_i \notin \mathcal{M}_0$ by $\mathcal{S}^\epsilon$. Since $\textsc{mi}(\mathcal{M}_0, Y) = \textsc{mi}(\mathcal{S}, Y)$, $\textsc{mi}(\mathcal{M}_0, Y) \leq \textsc{mi}(\mathcal{S}^\epsilon, Y)$, $\textsc{mi}(\mathcal{S}^\epsilon, Y) \leq \textsc{mi}(\mathcal{S}, Y)$, we have $\textsc{mi}(\mathcal{S}^\epsilon, Y) = \textsc{mi}(\mathcal{S}, Y)$. Therefore, $\mathcal{M}_0$ is still a Markov boundary after adding ϵ-noise. Assume in the new distribution, there is another Markov boundary, then it contains a variable with ϵ-noise: Z_i^ϵ. Denote this Markov boundary by $\{Z_i^\epsilon\} \cup \mathcal{S}_1$. Therefore, $\textsc{cmi}(Z_i^\epsilon, Y \mid \mathcal{S}_1) > 0$. However, from Lemma 4.3, this implies $\textsc{cmi}(Z_i^\epsilon, Y \mid \mathcal{S}_1) < \textsc{cmi}(Z_i, Y \mid \mathcal{S}_1)$, namely $\textsc{mi}(\{Z_i^\epsilon\} \cup \mathcal{S}_1, Y) < \textsc{mi}(\{Z_i\} \cup \mathcal{S}_1, Y)$. But $\textsc{mi}(\{Z_i^\epsilon\} \cup \mathcal{S}_1, Y) = \textsc{mi}(\mathcal{S}^\epsilon, Y) = \textsc{mi}(\mathcal{S}, Y) \geq \textsc{mi}(\{Z_i\} \cup \mathcal{S}_1, Y)$, which is a contradiction. □

Proof of Theorem 4.2. Lemma 4.4 shows that if $X \in \mathcal{M}_0$, $X \notin \mathcal{M}_1$, then we can add ϵ-noise on $\mathcal{S} \setminus \mathcal{M}_0$, such that $\mathcal{M}_0$ is the only Markov boundary. Therefore, in the perturbed distribution, the strength of $X \to Y$ is at least $\text{CMI}(X, Y \mid \mathcal{M}_0 \setminus \{X\})$, required by criterion C2. Let $\epsilon \to 0$, the joint distribution converges to the original joint distribution, and the continuous criterion C3 shows that the strength of $X \to Y$ in the original distribution is also no less than $\text{CMI}(X, Y \mid \mathcal{M}_0 \setminus \{X\})$. Similarly, we can add ϵ-noise on $\mathcal{S} \setminus \mathcal{M}_1$, such that $\mathcal{M}_1$ is the only Markov boundary. Therefore, in the perturbed distribution, the strength of $X \to Y$ is 0, required by criterion C1. When we let $\epsilon \to 0$, criterion C3 requires that the strength of $X \to Y$ in the original distribution is also 0. This is the contradiction we need to finish the proof of Theorem 4.2. □

In Theorem 4.2, we assume there is no extra knowledge of the causal directed acyclic graph. If we know partial or full causal directed acyclic graph, then we know some variables do not cause the response variable Y directly. After discarding these variables, we have similar results.

Theorem 4.3. *Assume we have partial or full knowledge of the causal directed acyclic graph. Discard all variables that have been known not causing Y directly. For the remaining variables, namely all the possible parents of Y, assume X is contained in some Markov boundary, but not all Markov boundaries. Then we cannot define the strength of $X \to Y$ satisfying criteria C0, C1, C2 and C3.*

The proof is the same with Theorem 4.2. The reason of discarding variables that do not cause Y directly is to guarantee the joint distribution with ϵ-noise is consistent with the causal directed acyclic graph. Also it is required by criterion C0. Consider an example with causal directed acyclic graph $Z \to X \to Y$, where $X = Z$. Here Z does not cause Y directly. If we do not discard Z, then there are two Markov boundaries for Y: $\{X\}$ and $\{Z\}$. However, in this case the strength of $X \to Y$ is well-defined. If we try to utilize the proof of Theorem 4.2, in one step we need to perturb the joint distribution by adding ϵ-noise on X. Then we have $\text{CMI}(Z, Y \mid X^\epsilon) > 0$, but the causal directed acyclic graph requires $Z \perp\!\!\!\perp Y \mid X^\epsilon$. Therefore adding ϵ-noise on X is prohibited, and the proof is invalid.

In sum, when there are multiple Markov boundaries, quantifying the causal effect is an ill-posed problem. We can only qualitatively describe the causal relationship. That is why the uniqueness of Markov boundary is important.

4.4 Algorithms for testing the uniqueness of Markov boundary

4.4.1 An assumption-free algorithm

We have seen that when Markov boundary is not unique, causal quantities are problematic. The next question is to determine the condition under which Markov boundary is unique.

Our first algorithm is based on the following necessary and sufficient condition for the uniqueness of Markov boundary.

Definition 4.7 (Essential variable)**.** *A variable $W \in \mathcal{S}$ is called an* essential variable *with respect to Y, if $\text{CMI}(Y, \mathcal{S} \mid \mathcal{S} \setminus \{W\}) > 0$. It means that W itself contains some irreplaceable information of Y. Denote the set of all essential variables by $\mathcal{E}$.*

Lemma 4.5. *$\mathcal{E}$ is the intersection of all Markov boundaries of Y within $\mathcal{S}$.*

Proof. Assume there exists a Markov boundary $\mathcal{M}$ such that $W \in \mathcal{E}$, $W \notin \mathcal{M}$. Since $\text{CMI}(Y, \mathcal{S} \mid \mathcal{M}) = \text{CMI}(Y, \mathcal{S} \mid \mathcal{S}) = 0$, we also have $\text{CMI}(Y, \mathcal{S} \mid \mathcal{S} \setminus \{W\}) = 0$ (Proposition 4.1), which is a contradiction.

If $W \notin \mathcal{E}$, then $\text{CMI}(Y, \mathcal{S} \mid \mathcal{S} \setminus \{W\}) = 0$, $\mathcal{S} \setminus \{W\}$ is a Markov blanket, which means it contains a Markov boundary. This Markov boundary does not contain W. □

Theorem 4.4. *The Markov boundary of Y is unique if and only if $\text{CMI}(Y, \mathcal{S} \mid \mathcal{E}) = 0$. In this case, the unique Markov boundary is $\mathcal{E}$.*

Proof. If Markov boundary is unique, then $\mathcal{E}$ is just the Markov boundary, therefore $\text{CMI}(Y, \mathcal{S} \mid \mathcal{E}) = 0$.

If $\text{CMI}(Y, \mathcal{S} \mid \mathcal{E}) = 0$, then $\mathcal{E}$ is a Markov blanket, which means it should contain a Markov boundary. But $\mathcal{E}$ should be contained in every Markov boundary, therefore $\mathcal{E}$ itself is a Markov

boundary. $\mathcal{E}$ as a Markov boundary cannot be a proper subset of another Markov boundary, thus the only Markov boundary is $\mathcal{E}$. □

Theorem 4.4 is the foundation of Algorithm 1. We construct $\mathcal{E}$ by checking whether $\text{CMI}(Y, \mathcal{S} \mid \mathcal{S} \setminus \{W\}) > 0$ for each W, then check whether $Y \perp\!\!\!\perp \mathcal{S} \mid \mathcal{E}$.

Algorithm 1: An assumption-free algorithm for determining the uniqueness of Markov boundary

(1) **Input**
    Observations of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y
(2) **Set** $\mathcal{E} = \emptyset$
(3) **For** $i = 1, \ldots, k$,
    Test whether $X_i \perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$
    If $X_i \not\perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$
        $\mathcal{E} = \mathcal{E} \cup \{X_i\}$
(4) **If** $Y \perp\!\!\!\perp \mathcal{S} \mid \mathcal{E}$
    output: Y has a unique Markov boundary
  Else
    output: Y has multiple Markov boundaries

To perform the condition independence tests in Algorithm 1, one may use G-test [100] or Pearson's χ^2-test if variables in $\mathcal{S}$ are discrete. For continuous variables, we can use KCI-test [164], maximal nonlinear conditional correlation [65] or a weighted Hellinger distance based test [146].

The conditional independence relationship test in Algorithm 1 concerns all observed variables, so that Algorithm 1 may be highly data inefficient [116] in that to achieve a satisfactory accuracy, it needs a large amount of observations. We now consider algorithms with better data efficiency.

4.4.2 Algorithm for producing a single Markov boundary

Alternatives for detecting the uniqueness of Markov boundaries usually depend on an algorithm that generates one Markov boundary. Therefore we first consider such algorithms.

There are many algorithms to produce Markov boundary, such as IAMB [149], KIAMB [116], Semi-Interleaved HITON-PC [1]. There is a summary of different algorithms in [145].

Remark 4.3. *There are also other approaches for finding Markov boundaries, such as MBOR [24], BLCD [92], PCMB [116] and GLL-PC [1]. However, these algorithms require faithfulness of the observed data distribution so that they cannot be used to detect uniqueness of Markov boundary.*

IAMB and KIAMB can produce one Markov boundary if the distribution has composition property. KIAMB has a parameter $0 \leq K \leq 1$. When $K = 1$, it coincides with IAMB. When $K < 1$, it is stochastic. When $K = 0$, every Markov boundary has a positive probability to be produced in one execution. When $K < 1$, if there are multiple Markov boundaries, then at least two of them have positive probabilities to be produced in one execution [116].

Semi-Interleaved HITON-PC requires that there exists a Markov boundary whose members are all dependent and conditional dependent with the response variable, unless some variables have exact the same information [1, 145]. Under this condition, multiple Markov boundaries is also possible.

According to [116], KIAMB outperforms IAMB considerably often. The complexity of Semi-Interleaved HITON-PC grows exponentially with the size of the Markov boundary, therefore it is much slower than KIAMB [145]. Thus we utilize KIAMB in the implementation of Algorithm 3 below, where a general Markov boundary producing algorithm is needed.

The validity of all these algorithms relies on extra assumptions. Instead, we propose an assumption-free algorithm to produce one Markov boundary.

Here the Δ is a measure of association level of two random variables. In this paper we choose the negative p-value of the conditional independence test as the association level.

Proof of correctness of Algorithm 2. Obviously the output $\mathcal{M}$ is a Markov blanket. In the last step, we have checked that $X_0 \not\perp\!\!\!\perp Y \mid \mathcal{M}\backslash\{X_0\}$. For $X_i \in \mathcal{M}$, since $\Delta(X_i, Y \mid \mathcal{M}\backslash\{X_i\}) \geq \Delta(X_0, Y \mid \mathcal{M} \setminus \{X_0\})$, we also have $X_i \not\perp\!\!\!\perp Y \mid \mathcal{M} \setminus \{X_i\}$. Therefore the output of Algorithm 2 is a Markov boundary. $\square$

Algorithm 2: An assumption-free algorithm for producing one Markov boundary

(1) **Input**
    Observations of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y
(2) **Set** $\mathcal{M}_0 = \mathcal{S}$
(3) **Repeat**
    Set $X_0 = \arg\min_{X \in \mathcal{M}_0} \Delta(X, Y \mid \mathcal{M}_0 \setminus \{X_i\})$
    If $X_0 \perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$
        Set $\mathcal{M}_0 = \mathcal{M}_0 \setminus \{X_0\}$
    Until $X_0 \not\perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$
(4) **Output** $\mathcal{M}_0$ is a Markov boundary

In Algorithm 2, there are some conditional independence tests which concern all or almost all variables, which is data inefficient. This is the price we pay for requiring no condition. As a result, these tests may have lower power detecting conditional dependencies. Therefore, we need to avoid executing such tests on almost all variables if the true result is not independent. Our solution is to feed these tests the variables with the least association levels, namely with the highest possibilities to be independent, so as to reduce the error rate.

4.4.3 Data efficient algorithms

TIE* is an algorithm to produce all Markov boundaries [145], which needs to utilize another algorithm which can correctly produce one Markov boundary (for example KIAMB). To test the uniqueness of Markov boundary, we can execute TIE* until it produces two Markov boundaries, or it terminates with the unique Markov boundary. This is Algorithm 3.

Proof of correctness of Algorithm 3. Markov boundary is not unique if and only if there exists variable $X_i \in \mathcal{M}_0$ which is not essential, namely $\text{MI}(Y, \mathcal{S} \setminus \{X_i\}) = \text{MI}(Y, \mathcal{S})$. Since $\text{MI}(Y, \mathcal{S} \setminus \{X_i\}) = \text{MI}(Y, \mathcal{M}_i)$ and $\text{MI}(Y, \mathcal{M}_0) = \text{MI}(Y, \mathcal{S})$, it is equivalent to $\text{MI}(Y, \mathcal{M}_i) = \text{MI}(Y, \mathcal{M}_0)$, namely $\text{CMI}(Y, \mathcal{M}_0 \mid \mathcal{M}_i) = 0$. □

Specifically, in Algorithm 3, if we set Ω to be Algorithm 2, then we acquire an assumption free algorithm to test the uniqueness of Markov boundnary, which is denoted as Algorithm 3-AF (AF

Algorithm 3: A general algorithm for determining the uniqueness of Markov boundary

(1) **Input**
Observations of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y
Algorithm Ω which can correctly produce one Markov boundary
(2) **Set** $\mathcal{M}_0 = \{X_1, \ldots, X_m\}$ to be the result of Algorithm Ω on $\mathcal{S}$
(3) **For** $i = 1, \ldots, m$,
Set $\mathcal{M}_i$ to be the result of Algorithm Ω on $\mathcal{S} \mid \{X_i\}$
If $Y \perp\!\!\!\perp \mathcal{M}_0 \mid \mathcal{M}_i$
Output Y has multiple Markov boundaries
Terminate
(4) **Output** Y has a unique Markov boundary

means assumption free). Algorithm 3 with Ω set as KIAMB is denoted as Algorithm 3-KI.

In Algorithm 3, we first produce one Markov boundary $\mathcal{M}_0 = \{X_1, X_2, \ldots, X_m\}$. Then for each $X_i \in \mathcal{M}_0$, we produce a Markov boundary within $\mathcal{S} \setminus \{X_i\}$, $\mathcal{M}_i$, and check if $\mathcal{M}_i$, which is supposed to contain all the information in $\mathcal{S} \setminus \{X_i\}$, has the same information with $\mathcal{S}$. In fact, for $X_i \in \mathcal{M}_0$, we can directly test whether $Y \perp\!\!\!\perp X_i \mid \mathcal{S} \setminus \{X_i\}$. This is Algorithm 4.

Algorithm 4: An assumption-free algorithm for determining the uniqueness of Markov boundary

(1) **Input**
Observations of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y
(2) **Set** $\mathcal{M}_0 = \{X_1, \ldots, X_m\}$ to be the result of Algorithm 2 on $\mathcal{S}$
(3) **For** $i = 1, \ldots, m$,
If $Y \perp\!\!\!\perp X_i \mid \mathcal{S} \setminus \{X_i\}$
If $X_i \not\perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$
Output Y has multiple Markov boundaries
Terminate
(4) **Output** Y has a unique Markov boundary

Proof of correctness of Algorithm 4. For a Markov boundary $\mathcal{M}_0$, it is the unique Markov boundary if and only if it coincides with $\mathcal{E}$. Therefore, we only need to check whether there exists a variable $X_i \in \mathcal{M}_0$ which is not essential, namely $\text{CMI}(Y, X_i \mid \mathcal{S}, \mathcal{M} \setminus \{X_i\}) = 0$. □

4.5 Simulation results and analysis

4.5.1 Simulation setup

We compare the performances of different algorithms for testing the uniqueness of Markov boundary on randomly generated data, and then conduct some theoretical analyses. We implement Algorithm 1, Algorithm 3 with Ω set as KIAMB (denoted as Algorithm 3-KI), Algorithm 3 with Ω set as Algorithm 2 (denoted as Algorithm 3-AF) and Algorithm 4. We use R 3.3.2 and an R package "infotheo" by P. E. Meyer. The significance level α is set to 0.05. In KIAMB, the parameter K is set as 0.8, which is adopted in [116, 145]. We generate various random data with different mechanisms, and count the frequency of success for each algorithm. Since all algorithms are theoretically correct if the needed assumptions are satisfied, when the number of observations is large enough, these algorithms can almost always produce correct results. To evaluate these algorithms, we need to consider dataset with limited number of observations. We will also consider a case where the needed assumption of KIAMB is violated. In each case, we run each algorithm 500 times for each number of observations.

In all simulations, we consider 1 response variable Y and 10 possible parents of Y, $\mathcal{S} = \{X_1, \ldots, X_{10}\}$. All variables are Bernoulli with mean 0.5.

We consider three cases with different mechanisms. In each case, we execute all algorithms with multiple numbers of observations: 300, 500, 800, 1000, 1500, 2000, 3000, 4000, 5000, 6000, 7000, 10000, 12000, 14000, 17000, 20000, 23000.

Case A, $X_1, \ldots, X_{10}$ are independent. $Y = X_1$ with probability 0.8, $Y = X_2$ with probability 0.1, $Y = X_3$ with probability 0.1. There is unique Markov boundary: $\{X_1, X_2, X_3\}$. The composition property is satisfied.

Case B, $X_1, X_2, X_4, X_5, \ldots, X_{10}$ are independent. $X_3 = X_4$. $Y = X_1$ with probability 0.8, $Y = X_4$ with probability 0.1, $Y = X_5$ with probability 0.1. There are multiple Markov boundaries: $\{X_1, X_3/X_4, X_5\}$. The composition property is satisfied.

Case C, define $Z = X_1 + X_2 \mod 2$. $X_1, \ldots, X_8$ are independent. $Y = Z$ with probability 0.8, $Y = X_3$ with probability 0.1, $Y = X_4$ with probability 0.1. $X_9 = Z$ with probability 0.95,

$X_9 = 1 - Z$ with probability 0.05. $X_{10} = X_9$. There is unique Markov boundary $\{X_1, X_2, X_3, X_4\}$. We have $X_1 \perp\!\!\!\perp Y \mid (X_3, \ldots, X_{10})$, $X_2 \perp\!\!\!\perp Y \mid (X_3, \ldots, X_{10})$, but $(X_1, X_2) \not\perp\!\!\!\perp Y \mid (X_3, \ldots, X_{10})$, therefore the composition property, which is required by KIAMB, is violated. Theoretically, in this case, Algorithm 3-KI cannot recognize the information of X_1 and X_2, and will claim there are multiple Markov boundaries: $\{X_3, X_4, X_9/X_{10}\}$. Even if we increase the number of observations, Algorithm 3-KI still cannot provide correct results.

4.5.2 Simulation results and analysis

See Figures 4.2,4.3,4.4 for success rates and average time costs per execution of different algorithms in Cases A, B, C.

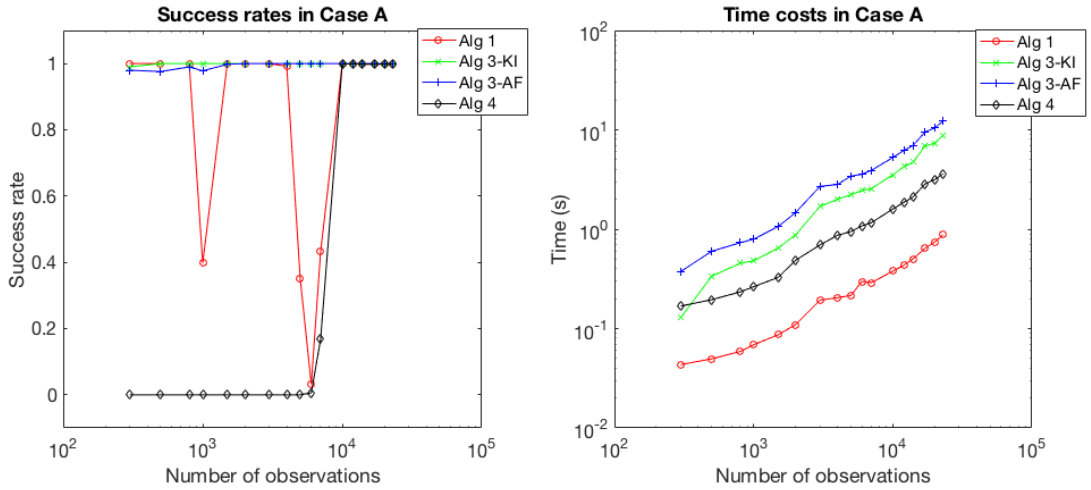


Figure 4.2: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green 'x'), 3-AF (blue '+'), 4 (black diamond) with different numbers of observations in Case A. Number of observations and time costs are in logarithm.

See Tables 4.1, 4.2, 4.3 for detailed performances of different algorithms in different cases.

In all situations, The time costs are ascending for Algorithm 1, 4, 3-KI, 3-AF. Time cost increases with the number of observations. Compared to Algorithm 1 and 4, Algorithm 3-AF and 3-KI calculates much more CMI, which significantly increases its time cost. Besides, when we

Table 4.1: Success rates and time costs (in seconds) of Algorithms 1, 3-KI, 3-AF, 4 in Case A

Number of	Success rate				Time cost (s)			
observations	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4
300	1	0.99	0.98	0	0.0434	0.129	0.3756	0.1684
500	1	1	0.976	0	0.0496	0.3368	0.6006	0.1954
800	1	1	0.99	0	0.0588	0.4542	0.7302	0.2338
1000	0.398	1	0.978	0	0.0688	0.4808	0.7944	0.2636
1500	1	1	0.998	0	0.0866	0.645	1.0636	0.327
2000	1	1	1	0	0.1094	0.8754	1.4664	0.486
3000	1	1	1	0	0.1932	1.6968	2.6992	0.6946
4000	0.992	1	1	0	0.2046	2.0004	2.8158	0.867
5000	0.35	1	1	0	0.2152	2.2218	3.4218	0.9474
6000	0.032	1	1	0.006	0.298	2.4618	3.575	1.0806
7000	0.432	1	1	0.168	0.2896	2.5534	3.9036	1.1618
10000	1	1	1	0.998	0.3778	3.5292	5.256	1.5946
12000	1	1	1	1	0.434	4.263	6.1956	1.8626
14000	1	1	1	0.998	0.503	4.7278	6.971	2.1094
17000	1	1	1	0.998	0.6408	6.84	9.467	2.8498
20000	1	1	1	0.998	0.742	7.3668	10.4418	3.1482
23000	1	1	1	0.998	0.8768	8.7754	12.2506	3.6182

Table 4.2: Success rates and time costs (in seconds) of Algorithms 1, 3-KI, 3-AF, 4 in Case B

Number of	Success rate				Time cost (s)			
observations	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4
300	0	0.258	0.232	1	0.0358	0.1662	0.3986	0.1656
500	0	0.714	0.682	1	0.045	0.2358	0.4932	0.1922
800	0	0.99	0.952	0.996	0.0566	0.3542	0.5904	0.2278
1000	0	0.998	0.978	0.998	0.0612	0.3844	0.6624	0.254
1500	0	1	0.998	1	0.0758	0.5296	0.812	0.3148
2000	0	1	1	1	0.0942	0.6768	1.0474	0.409
3000	0	1	1	1	0.1238	1.0706	1.3846	0.5328
4000	0	1	1	1	0.162	1.341	1.8696	0.7224
5000	0	1	1	1	0.1862	1.5834	2.214	0.8606
6000	0	1	1	1	0.2204	1.8316	2.547	0.9752
7000	0	1	1	1	0.237	2.0884	2.889	1.103
10000	0.258	1	1	1	0.3756	3.576	4.655	1.741
12000	0.022	1	1	1	0.4504	3.8158	4.9024	1.8228
14000	0.012	1	1	1	0.4702	3.9478	5.5814	2.0734
17000	0.662	1	1	1	0.6304	5.982	7.8514	2.907
20000	1	1	1	1	0.7302	6.6294	8.3958	3.3172
23000	1	1	1	1	0.804	6.8026	9.036	3.4956

Table 4.3: Success rates and time costs (in seconds) of Algorithms 1, 3-KI, 3-AF, 4 in Case C

Number of	Success rate				Time cost (s)			
observations	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4	Alg 1	Alg 3-KI	Alg 3-AF	Alg 4
300	1	0.006	0.9	0	0.033	0.1394	0.4388	0.1628
500	1	0.002	0.98	0	0.046	0.238	0.6228	0.1768
800	1	0.006	0.996	0	0.0514	0.2712	0.8648	0.209
1000	1	0.002	1	0	0.056	0.441	0.9822	0.2298
1500	0	0.004	1	0	0.0692	0.7014	1.2396	0.2872
2000	0	0.006	1	0	0.0876	0.9118	1.5828	0.3682
3000	0.172	0.004	1	0	0.124	1.3462	2.1356	0.4936
4000	0.998	0.002	1	0	0.1642	1.6796	2.8578	0.6868
5000	1	0.004	1	0	0.1876	2.038	3.3824	0.7672
6000	1	0.002	1	0	0.2092	2.3474	3.8966	0.8992
7000	1	0.004	1	0	0.2478	2.6932	4.4848	1.0098
10000	0.968	0.006	1	0.004	0.3514	4.1476	6.1496	1.4352
12000	0.998	0.002	1	0.794	0.3806	4.3554	6.9754	1.7346
14000	1	0.002	1	1	0.4616	5.4454	8.1958	1.9872
17000	1	0.002	1	1	0.6096	7.6286	11.4054	2.677
20000	1	0.002	1	1	0.7244	8.4308	12.7722	3.0216
23000	1	0.002	1	1	0.864	10.2032	15.0404	3.6288

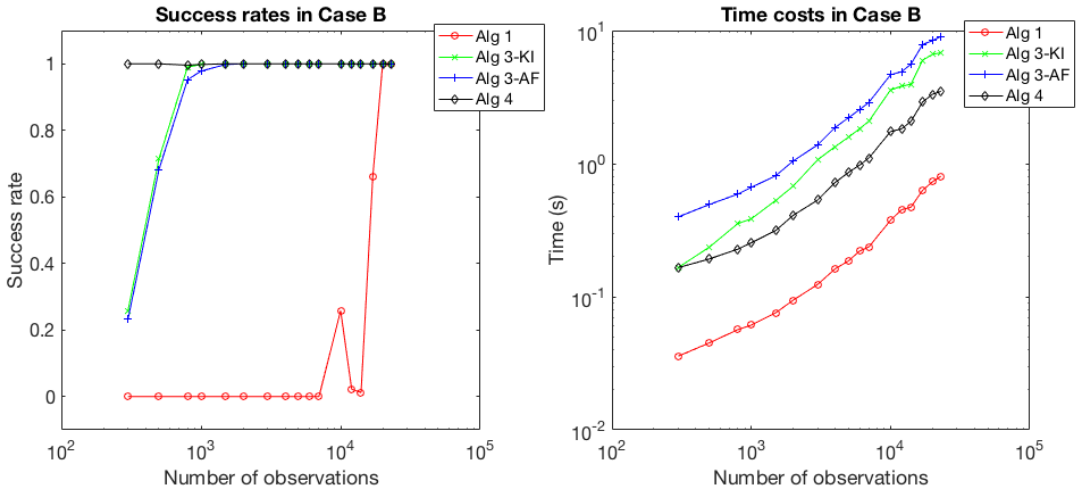


Figure 4.3: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green 'x'), 3-AF (blue '+'), 4 (black diamond) with different numbers of observations in Case B. Number of observations and time costs are in logarithm.

execute multiple tests, we might need to make false discover rate adjustment [14], but in most algorithms, the number of tests cannot be predicted, which might be problematic.

For success rate, Algorithm 3-AF is always satisfactory. Algorithm 3-KI performs slightly better than Algorithm 3-AF if the composition property is satisfied (Case A, Case B), but performs disastrously if the composition property is violated (Case C). In Case C, Algorithm 3-KI still has a very small probability to produce correct answer, since it is stochastic. Algorithm 1 tends to claim that Markov boundary is unique, while Algorithm 4 tends to claim that Markov boundary is not unique, regardless of whether the Markov boundary is unique or not.

In the last step of Algorithm 1, $Y \perp\!\!\!\perp \mathcal{S} \mid \mathcal{E}$ means Markov boundary is unique, while in the last step of Algorithm 4, $Y \perp\!\!\!\perp X_i \mid \mathcal{S} \backslash \{X_i\}$ means Markov boundary is not unique. For such tests which concern all variables, when the number of observations is not large enough, they cannot reject the conditional independence hypothesis, even if variables are dependent. Therefore, Algorithm 1 tends to claim Markov boundary is unique (false negative), Algorithm 4 tends to claim Markov boundary is not unique (false positive), whether Markov boundary is really unique or not. In Algorithm 3-AF, generally the last step test does not concern many variables. Although in the first

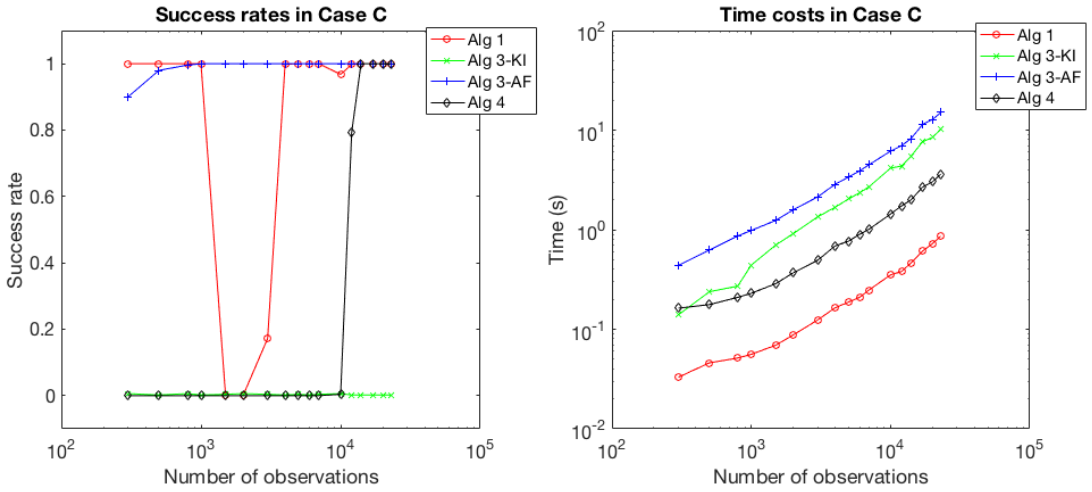


Figure 4.4: Success rates and average time costs per execution (in seconds) of Algorithms 1 (red circle), 3-KI (green 'x'), 3-AF (blue '+'), 4 (black diamond) with different numbers of observations in Case C. Number of observations and time costs are in logarithm.

few steps some tests concern almost all variables, these tests do not directly determine the output, and we have argued that our implementation can significantly lower the error rate. In Algorithm 3-KI, generally there is no test with many variables. This explains why Algorithm 3-KI is slightly better than Algorithm 3-AF under composition property, and why on average Algorithm 3-AF and 3-KI are better than Algorithm 1 and 4.

Wrong conditional independence tests are necessary but not sufficient for wrong final result (except Algorithm 3-KI in Case C). In some cases, we can see that when we increase the number of observations, the success rate of an algorithm might first decrease and then increase. The reason is that, the success rate of each test is positively related to the number of observations, but the success rate of the whole algorithm has complicated relationship with wrong tests. Initially this algorithm has several tests with wrong results, but these wrong tests happen to make the correct output. When we increase the number of observations, some of these tests become correct, but the remaining wrong tests lead to wrong output. Finally when the number of observations is large enough, all tests are correct, and the output is correct.

On average, Algorithm 3-KI performs the best when the composition property is satisfied.

However, if we set Ω to be any algorithm which requires additional assumption, such as KIAMB, we need to guarantee that theoretically algorithm Ω produces a correct Markov boundary in each execution, otherwise Algorithm 3 might produce wrong results, just like Case C. To make sure algorithm Ω can produce correct result, we can prove the distribution satisfies the condition required by algorithm Ω. However, generally such verification process is slower and less efficient than Algorithm 1, 4 or 3-AF. If we do not verify the requirement of algorithm Ω, then we need to verify the result of Ω is a Markov boundary after each execution. If the Markov boundary produced by Ω is fake, then we need to apply Algorithm 2 to produce a correct Markov boundary, which is cumbersome.

In sum, if we know that the condition required by Ω is satisfied with no or very small cost, then utilizing Ω in Algorithm 3 is better. Otherwise, Algorithm 3-AF is better. Also, we can directly utilize Ω in Algorithm 3, to obtain a less reliable result. Since multiple Markov boundaries are problematic, we prefer false positive (regard unique as non-unique) rather than false negative (regard non-unique as unique). In this sense, Algorithm 4 has the lowest false negative rate. If time cost is much more concerned than accuracy, then Algorithm 1 is preferred.

As expected, all these algorithms are faster than executing full TIE* algorithm.

4.6 Appendix

4.6.1 Review of causal quantities

Assume we have several random variables $X, Y, Z, \cdots$, and there exist causal relationships between variables. Regard variables as vertices, and regard the causal relationship "X directly causes Y" as directed edge $X \to Y$. Then we have a directed acyclic graph (DAG) [144]. The question is, with or without prior knowledge of the DAG, how to quantitatively describe the strength of causality between two variables X and Y (no direction). If we have known that there is no causal relationship $Y \to X$, then this quantity describes the strength of causality from X to Y.

We will define causal quantities on discrete random variables. Most of them can be generalized to continuous random variables.

Quantities with no prior knowledge of DAG

Assume we only know the joint distribution of several random variables, and know nothing about their causal relationships. All the quantities are symmetric with X and Y.

Correlation coefficient (ρ) Assume we only have two random variables X and Y. The (Pearson) correlation coefficient ρ is defined as

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{[\mathbb{E}(X^2) - (\mathbb{E}(X))^2][\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2]}}.$$

We have $-1 \leq \rho(X,Y) \leq 1$. Notice that if X and Y are independent, then $\rho(X,Y) = 0$, but the inverse is not true. Correlation coefficient can only detect linear relationships.

Mutual information (MI) Assume we only have two random variables X and Y. The mutual information MI is defined as [21]

$$\text{MI}(X,Y) = \sum_{y \in Y} \sum_{x \in X} \mathbb{P}(x,y) \log(\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}),$$

where $\mathbb{P}(x, y)$ is the joint distribution of X and Y, and $\mathbb{P}(x)$, $\mathbb{P}(y)$ are the marginal distributions. This is just the Kullback-Leibler (KL) divergence of $\mathbb{P}(x, y)$ from $\mathbb{P}(x)\mathbb{P}(y)$.

The (Shannon) entropy H of a random vector $\vec{X}$ is defined as

$$\mathrm{H}(\vec{X}) = \sum_{\vec{x} \in \vec{X}} -\mathbb{P}(\vec{x}) \log \mathbb{P}(\vec{x}).$$

The conditional entropy of $\vec{Y}$ conditioned on $\vec{X}$ is defined as

$$\mathrm{H}(\vec{Y}|\vec{X}) = \sum_{\vec{x} \in \vec{X}} \mathbb{P}\mathrm{H}(\vec{Y}|\vec{X} = \vec{x}) = \mathrm{H}(X, Y) - \mathrm{H}(X).$$

Using entropy, we have an expression of MI:

$$\mathrm{MI}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X).$$

$\mathrm{MI}(X, Y) \geq 0$, and it equals 0 if and only if X and Y are independent.

We also have $\mathrm{MI}(X, Y) \leq \min\{\mathrm{H}(X), \mathrm{H}(Y)\}$.

If we have more variables, then ρ and MI, which only consider two variables, are not suitable. For example, if the true causal relationship is $X \leftarrow Z \rightarrow Y$, then we might have $\rho(X, Y) \neq 0$ and $\mathrm{MI}(X, Y) > 0$, but X and Y are not causal related.

Partial correlation (PC) Assume we have several random variables $X, Y, Z_1, \cdots, Z_k$. The partial correlation (PC) between X and Y conditioned on $Z_1, \cdots, Z_k$ is defined as [143]

$$\mathrm{PC}(X, Y|Z_1, \cdots, Z_k) = \rho(R_X, R_Y),$$

where R_X (R_Y) is the linear regression residual of X (Y) with $Z_1, \cdots, Z_k$.

Still, PC can only detect linear relationships.

Conditional mutual information (CMI) Assume we have several random variables $X, Y, Z_1, \cdots, Z_k$. The Conditional Mutual information (CMI) between X and Y conditioned on $Z_1, \cdots, Z_k$ is defined as [27]

$$\text{CMI}(X, Y|Z_1, \cdots, Z_k) = \sum_{x,y,z_1,\cdots,z_k} \mathbb{P}(x, y, z_1, \cdots, z_k) \log(\frac{\mathbb{P}(x, y|z_1, \cdots, z_k)}{\mathbb{P}(x|z_1, \cdots, z_k)\mathbb{P}(y|z_1, \cdots, z_k)}).$$

We also have

$$\text{CMI}(X, Y|Z_1, \cdots, Z_k) = \text{H}(X, Z_1, \cdots, Z_k)+\text{H}(Y, Z_1, \cdots, Z_k)-\text{H}(Z_1, \cdots, Z_k)-\text{H}(X, Y, Z_1, \cdots, Z_k).$$

$\text{CMI}(X, Y|Z_1, \cdots, Z_k) \geq 0$, and it equals 0 if and only if X and Y are independent conditioned on $Z_1, \cdots, Z_k$.

Notice that $\text{CMI}(X, Y|Z) = \text{H}(X|Z) - \text{H}(X|Y, Z) \leq \text{H}(X|Z) \leq H(X)$, so $\text{CMI}(X, Y|Z) \leq \min\{\text{H}(X), \text{H}(Y)\}$.

There are several methods of estimating CMI from observation in [151, 133, 82, 23].

If the case is $X \to Z \to Y$, $X \leftarrow Z \leftarrow Y$ or $X \leftarrow Z \to Y$, then X and Y are independent conditioned on Z, namely $\text{CMI}(X, Y|Z) = 0$. Thus there is no direct relation between X and Y.

However, when the case is $X \to Z \leftarrow Y$, we can see that X and Y are independent, but they are not necessarily independent conditioned on Z.

In this case, generally $\text{CMI}(X, Y|Z) > 0$, but X and Y are not directly related.

Part mutual information (PMI) Assume we have several random variables $X, Y, Z_1, \cdots, Z_k$. The Part Mutual information (PMI) between X and Y conditioned on $Z_1, \cdots, Z_k$ is defined as [166]

$$\text{PMI}(X, Y|Z_1, \cdots, Z_k) = \sum_{x,y,z_1,\cdots,z_k} \mathbb{P}(x, y, z_1, \cdots, z_k) \log(\frac{\mathbb{P}(x, y|z_1, \cdots, z_k)}{\mathbb{P}^*(x|z_1, \cdots, z_k)\mathbb{P}^*(y|z_1, \cdots, z_k)}),$$

where

$$\mathbb{P}^*(x|z_1,\cdots,z_k)=\sum_y \mathbb{P}(x|z_1,\cdots,z_k,y)\mathbb{P}(y),$$

$$\mathbb{P}^*(y|z_1,\cdots,z_k)=\sum_x \mathbb{P}(y|z_1,\cdots,z_k,x)\mathbb{P}(x).$$

We have

$$\begin{aligned}\text{PMI}(X,Y|Z_1,\cdots,Z_k)=&\sum_{x,y,z_1,\cdots,z_k}\mathbb{P}(x,y,z_1,\cdots,z_k)\log\frac{\mathbb{P}(y|z_1,\cdots,z_k,x)}{\sum_{x'}\mathbb{P}(y|z_1,\cdots,z_k,x')\mathbb{P}(x')}\\&+\sum_{x,y,z_1,\cdots,z_k}\mathbb{P}(x,y,z_1,\cdots,z_k)\log\frac{\mathbb{P}(x|z_1,\cdots,z_k,y)}{\sum_{y'}\mathbb{P}(x|z_1,\cdots,z_k,y')\mathbb{P}(y')}\\&-\sum_{x,y,z_1,\cdots,z_k}\mathbb{P}(x,y,z_1,\cdots,z_k)\log(\frac{\mathbb{P}(x,y|z_1,\cdots,z_k)}{\mathbb{P}(x|z_1,\cdots,z_k)\mathbb{P}(y|z_1,\cdots,z_k)}).\end{aligned}$$

The third term on the right hand side is just $\text{CMI}(X,Y|Z_1,\cdots,Z_k)$, which is non-negative. So

$$\begin{aligned}\text{PMI}(X,Y|Z_1,\cdots,Z_k)\leq&\sum_{x,y,z_1,\cdots,z_k}\mathbb{P}(x,y,z_1,\cdots,z_k)\log\frac{\mathbb{P}(y|z_1,\cdots,z_k,x)}{\mathbb{P}(y|z_1,\cdots,z_k,x)p(x)}\\&+\sum_{x,y,z_1,\cdots,z_k}\mathbb{P}(x,y,z_1,\cdots,z_k)\log\frac{\mathbb{P}(x|z_1,\cdots,z_k,y)}{\mathbb{P}(x|z_1,\cdots,z_k,y)\mathbb{P}(y)}\\=&\text{H}(X)+\text{H}(Y).\end{aligned}$$

We also have $\text{PMI}(X,Y|Z_1,\cdots,Z_k)\geq\text{CMI}(X,Y|Z_1,\cdots,Z_k)\geq 0$ [166].

Quantities with partial prior knowledge of DAG

Assume we know the joint distribution of several random variables, and also have partial information about the causal DAG. Here we only consider the most common case: we have time orders of these variables (generally as time series). So if X is prior to Y in time, then there is no causal arrow $Y\to X$.

Assume we have several stationary time series $\{X(k)\}$, $\{Y(k)\}$, $\{Z(k)\},\cdots$, where $k=\cdots,-1,\,0,1,2,\cdots$. We want to detect the relationships between the present of one series, like

$X(n)$, and its causes. Here we only need to consider the past of all these series, $\{\cdots, X(n-2), X(n-1)\}$, $\{\cdots, Y(n-2), Y(n-1)\}$, $\{\cdots, Z(n-2), Z(n-1)\}$. Methods mentioned in this subsection regard the past of a time series, like $\{\cdots, Y(n-2), Y(n-1)\}$, as a single random vector, not as separated random variables.

Conditional Granger causality (CGC) We conduct linear regression of $X(n)$ with the past of X, Z, and $X(n)$ with the past of X, Y, Z:

$$X(n) = \sum_{i=1}^{\infty} a_{1i} X(n-i) + \sum_{i=1}^{\infty} c_{1i} Z(n-i) + R_{1n},$$

$$X(n) = \sum_{i=1}^{\infty} a_{2i} X(n-i) + \sum_{i=1}^{\infty} b_{2i} Y(n-i) + \sum_{i=1}^{\infty} c_{2i} Z(n-i) + R_{2n},$$

where R_{1n}, R_{2n} are the regression residuals.

The Conditional Granger causality (GC) from Y to X conditioned on Z is defined as [47]

$$\text{CGC}(Y \to X) = \log \frac{\text{Var}(R_{1n})}{\text{Var}(R_{2n})} \geq 0.$$

CGC can only detect linear relationships.

If we only consider $X(k)$ and $Y(k)$, then it is the original Granger causality (GC) [54].

Conditional transfer entropy (CTE) The conditional transfer entropy (CTE) from Y to X conditioned on Z is defined as [135, 73, 40]

$$\text{CTE}(Y \to X|Z) = \sum \mathbb{P}(X(n), X_{n-1}^{(k)}, Y_{n-1}^{(l)}, Z_{n-1}^{(m)}) \log \frac{\mathbb{P}(X(n)|X_{n-1}^{(k)}, Y_{n-1}^{(l)}, Z_{n-1}^{(m)})}{\mathbb{P}(X(n)|X_{n-1}^{(k)}, Z_{n-1}^{(m)})},$$

where $X_{n-1}^{(k)} = (X(n-k), X(n-k+1), \cdots, X(n-1))$, $Y_{n-1}^{(l)} = (Y(n-l), Y(n-l+1), \cdots, Y(n-1))$, $Z_{n-1}^{(m)} = (Z(n-m), Z(n-m+1), \cdots, Z(n-1))$. Here k, l, m are the lengths of history we consider. We can see that CTE is just the CMI of the present of X and the past of Y, conditioned

on the past of X and Z [108]:

$$\text{CTE}(Y \to X|Z) = \text{CMI}(X(n), Y_{n-1}^{(l)}|X_{n-1}^{(k)}, Z_{n-1}^{(m)}).$$

We have $\text{CTE}(Y \to X|Z) \leq \text{H}(X(n))$.

If we only consider $X(k)$ and $Y(k)$, then it is the original transfer entropy (TE) [135].

Quantities with full prior knowledge of DAG

Assume we know the joint distribution of several random variables, and also the full causal DAG. Here the joint distribution and the DAG are compatible (satisfy the Causal Markov Condition) [68]. The question is to determine the strength of a given causal arrow.

We should only utilize the information of the direct causes of the target variable.

Causal strength (CS) Assume that $X, Z_1, \cdots, Z_k$ are all the vertices (variables) which point to (directly cause) Y, namely there exists arrows from $X, Z_1, \cdots, Z_k$ to Y. Then the causal strength of arrow $X \to Y$ is defined as

$$\text{CS}(X \to Y|Z_1, \cdots, Z_k) = \sum_{x,y,z_1,\cdots,z_k} \mathbb{P}(x, y, z_1, \cdots, z_k) \log \frac{\mathbb{P}(y|z_1, \cdots, z_k, x)}{\sum_{x'} \mathbb{P}(y|z_1, \cdots, z_k, x')\mathbb{P}(x')}.$$

We have $0 \leq \text{CMI}(X, Y|Z_1, \cdots, Z_k) \leq \text{CS}(X \to Y|Z_1, \cdots, Z_k) \leq \text{H}(X)$.

We also have $PMI(X, Y|Z_1, \cdots, Z_k) \geq CS(X \to Y|Z_1, \cdots, Z_k)$ [166].

4.6.2 Continuity and performance on highly correlated variables of causal quantities

In this subsection, we will check the performance of different causal quantities on highly correlated variables. Also, we will check the (uniform) continuity since it determines whether computing such quantities is well-conditioned.

Assume all the variables like X, Y, Z can take at most k values. Then the joint distribution of X, Y, Z is a non-negative vector with length k^3 and sum 1. We use the total variation distance. For

two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, it is defined as

$$d(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2} \sum_x |\mathbb{P}_1(x) - \mathbb{P}_2(x)|.$$

Linear quantities

The case of highly correlated variables is called multicollinearity (or collinearity) in multiple regression [51].

Correlation coefficient In the definition of correlation coefficient (ρ), covariance and variance are continuous functions of the joint distribution. So the only possible case of discontinuity is that at least variance equals zero.

If both variance are zero, then ρ is not continuous. Let $X = 0$ with probability $1 - 2\epsilon$, $X = 1$ and $X = -1$ with probability ϵ. When $Y = X$, $\rho(X, Y) = 1$, and when $Y = -X$, $\rho(X, Y) = -1$, as long as $\epsilon > 0$. However, when $\epsilon \to 0$, both joint distributions converge to $X = Y = 0$ with probability 1. So when $\text{Var}(X) = \text{Var}(Y) = 0$, we cannot give a value to $\rho(X, Y)$ to make it continuous. Also, near this case, $\rho(X, Y)$ is not uniformly continuous, so an arbitrarily small perturbation might change ρ significantly.

If only one variance is zero, we can prove that $\rho(X, Y) = 0$ and it is continuous. See Appendix A for proof.

We can see that when $\text{Var}(X) > 0$, and $Y = nX$, then $\rho(X, Y) = 1, 0, -1$ when $n > 0$, $n = 0$, $n < 0$. When $\text{Var}(X) = 0$, $Y = nX$, ρ is not well-defined.

Partial correlation The joint distribution of the residuals from the linear regression is continuous with the original joint distribution. The problem appears when we calculate the correlation coefficient of the residuals.

If $X = aZ + b$, $Y = cZ + d$, then both residuals have variance 0. In this case, correlation coefficient is not continuous, so PC is also not continuous. This is the only ill-defined case.

If only one variable of X, Y is linearly related to Z, then the variance of one residual is 0, but

the other is not. In this case, the PC is 0.

(Conditional) Granger causality Still, the residuals of linear regressions are continuous with the joint distribution. The only problem is when $\text{Var}(R_{2n}) = 0$. If $\text{Var}(R_{1n}) > 0$, then under a small perturbation, $\text{Var}(R_{2n}) = 0$ will change significantly (compared to its own magnitude), and the calculated GC will change significantly, but will always be quite large. So in this case, calculating the exact value of GC is ill-conditioned, but since it is always large, sometimes it is enough.

The real problematic case is $\text{Var}(R_{1n}) = \text{Var}(R_{2n}) = 0$. Under a small perturbation, these two variances will still be very small, but their ratio might take any positive values larger than 1. So in this case, the calculated GC can take value in $[0, +\infty)$.

In sum, if only $\text{Var}(R_{2n})$ is almost 0, the exact value is not reliable, but the statement "GC is large" is reliable. If both variances are almost 0, then GC is not reliable at all.

Mutual information based quantities

Mutual information Recall that the entropy is defined as $H(X) = \sum_x -\mathbb{P}(x) \log \mathbb{P}(x)$. For function $f(p) = -p \log p$, if we define $f(0) = 0$, then it is uniformly continuous on $p \in [0, 1]$.

Mutual information of X and Y is $\text{MI}(X, Y) = \text{H}(X) + \text{H}(Y) - \text{H}(X, Y)$, so it is a uniformly continuous function of the joint distribution. Thus there is no ill-posed case, and any small perturbation on the joint distribution will only produce small change of MI. We can always reduce the output error to a given level by increasing the accuracy of input to a pre-determined level.

When $X = Y$, $\text{H}(X, Y) = \text{H}(X) = \text{H}(Y)$, so $\text{MI}(X, Y) = \text{H}(X) = \text{H}(Y)$.

Conditional mutual information We have $\text{CMI}(X, Y|Z) = \text{H}(X, Z) + \text{H}(Y, Z) - \text{H}(Z) - \text{H}(X, Y, Z)$. So it is also a uniformly continuous function of the joint distribution.

Another expression is $\text{CMI}(X, Y|Z) = \text{H}(X|Z) - \text{H}(X|Y, Z)$, namely, under the condition of Z, how much extra information of X does Y contain. So if Y (or X) and Z are highly correlated, CMI is almost 0.

(Conditional) Transfer entropy Transfer entropy and conditional transfer entropy are essentially CMI. Therefore they are uniformly continuous with the joint distribution.

If $Y_{n-1}^{(l)}$ is highly correlated with $X_{n-1}^{(k)}$ or $Z_{n-1}^{(m)}$, then CTE$(Y \to X|Z)$ is almost zero. Similarly, if $Y_{n-1}^{(l)}$ is highly correlated with $X_{n-1}^{(k)}$, then TE$(Y \to X)$ is almost zero.

Remark 4.4. *For joint normal distributions, ρ and MI are equivalent, PC and CMI are equivalent, (C)GC and (C)TE are equivalent [11]. Here equivalent means the two quantities have a deterministic relation.*

Causal strength and part mutual information

The continuity and performance on highly correlated variables of CS and PMI have been discussed in previous sections. Here we consider another example: X and Z can take $1, 2, \cdots, n$, Y can take 0 or 1. If $x = z$, then $\mathbb{P}(X = x, Y = 0, Z = z) = 1/n - n\epsilon - (n-1)\delta$, $\mathbb{P}(X = x, Y = 1, Z = z) = \epsilon$. If $x \neq z$, then $\mathbb{P}(X = x, Y = 0, Z = z) = \delta$, $\mathbb{P}(X = x, Y = 1, Z = z) = \epsilon$.

If we let $\epsilon \to 0$, $\delta \to 0$, then

$$\text{PMI}(X, Y|Z) = \text{CS}(X \to Y|Z) \to \log \frac{n}{1 + (n-1)\frac{\delta}{\delta+\epsilon}}.$$

By changing the value of δ/ϵ, PMI$(X, Y|Z)$ and CS$(X \to Y|Z)$ can take any value between 0 and $\log n$. However, the total information of Y, H$(Y) \to 0$. Any quantity that has the name "mutual information" should be the information that is shared by X and Y in some sense, so H(Y), the total information of Y, should be a natural upper bound. Even if we divide PMI by a constant, it might still exceed H(Y). So we think part mutual information is not a proper name. Also, we should ask, when Y is almost a constant, is it reasonable to assign a large value to the strength of causal relationship $X \to Y$?

Consider another example with joint distribution in Table 4.4. Let $\epsilon \to 0$, $\delta \to 0$, $\epsilon/\delta \to 0$, then PMI$(X, Y|Z) \to 2 \log 2$. However, $H(X, Y, Z) = H(X) = H(Y) = \log 2$. So we even do not have $PMI(X, Y|Z) \leq H(X, Y, Z)$.

Table 4.4: Joint distribution in subsection 4.6.2

		$Z=0$	$Z=1$
$X=0$	$Y=0$	$(1-\epsilon)(1-\delta)/2$	$\delta(1-\epsilon)/2$
	$Y=1$	$\epsilon(1-\delta)/2$	$\epsilon\delta/2$
$X=1$	$Y=0$	$\epsilon\delta/2$	$\epsilon(1-\delta)/2$
	$Y=1$	$\delta(1-\epsilon)/2$	$(1-\epsilon)(1-\delta)/2$

Summary of causal quantities

We summarize the mentioned causal quantities in Table 4.5.

For those quantities with variables as input (ρ, PC, MI, CMI, PMI, CS), X, Y, Z are random variables. For those quantities with time series as input (GC, CGC, TE, CTE), $X = \{X(1), \cdots, X(n-1)\}$, $Y = Y(n)$, $Z = \{Y(1), \cdots, Y(n-1)\}$, $W = \{Z(1), \cdots, Z(n-1)\}$.

Some quantities can only detect linear relationships (ρ, PC, GC, CGC). Although they might have nonlinear versions [56], they still can only detect relationships specified by a given model. Such quantities are not model-free. However, since they rely on pre-selected models, they need less data to calculate. Other model-free quantities can detect any relationships, but they need more data to calculate.

When some variables are the same or nearly the same, some quantities are ill-defined or ill-conditioned. Some other quantities have values 0 or nearly 0. We will specify which cases they are.

Table 4.5: Summary of Causal Quantities

	Input	DAG	Model Free	Lower Bound	Upper Bound	Ill-defined	Equals 0
Correlation Coefficient $\rho(X,Y)$	Two Variables	No	No	-1	1	$X=Y$ =const	
Partial Correlation $PC(X,Y\|Z)$	Multiple Variables	No	No	-1	1	$X=Y=Z$	$X=Z$ $Y=Z$
Granger Causality $GC(X\to Y\|Z)$	Two Time Series	Partial	No	0	$+\infty$	$Y=Z$	$X=Z$
Conditional Granger Causality CGC $(X\to Y\|Z,W)$	Multiple Time Series	Partial	No	0	$+\infty$	$Y=Z$	$X=Z$
Mutual Information $MI(X,Y)$	Two Variables	No	Yes	0	$\min$ {H(X), H(Y)}		
Conditional Mutual Information $CMI(X,Y\|Z)$	Multiple Variables	No	Yes	0	$\min$ {H(X), H(Y)}		$X=Z$ $Y=Z$
Transfer Entropy $TE(X\to Y\|Z)$	Two Time Series	Partial	Yes	0	$\min$ {H(X), H(Y)}		$X=Z$ $Y=Z$
Conditional Transfer Entropy CTE $(X\to Y\|Z,W)$	Multiple Time Series	Partial	Yes	0	$\min$ {H(X), H(Y)}		$X=Z$ $Y=Z$
Part Mutual Information $PMI(X,Y\|Z)$	Multiple Variables	No	Yes	0	H(X) +H(Y)	$X=Z$ $Y=Z$	
Causal Strength $CS(X\to Y\|Z)$	Multiple Variables	Full	Yes	0	H(X)	$X=Z$	

Chapter 5

ENTROPY PRODUCTIONS AND THEIR MATHEMATICAL REPRESENTATIONS: CLAUSIUS' VS. KELVIN'S VIEWS OF THE SECOND LAW AND IRREVERSIBILITY

5.1 Introduction

There is a growing awareness toward a slow shifting in the foundation of the Second Law of Thermodynamics, from a macroscopic postulate concerning heat as a form of random mechanical motion [110] to a derivable mathematical discovery based on the stochastic dynamics of mesoscopic systems [69, 137]. The first significant attempt in this direction was carried out by L. Boltzmann through the equation that now bears his name and the H-theorem it derives. The theory is applicable to gas dynamics; a fundamental assumption underlying the classic work is a *stosszahlansatz* [55, 28]. Rigorous mathematical breakthrough on Boltzmann's equation only became available very recently [153]. In 1950s, Bergmann and Lebowitz set up a general stochastic theory for closed as well as open mechanical systems that are consistent with Hamiltonian dynamics, and easily obtained an H-theorem like result [16]. It becomes increasingly clear in recent years that a stochastic description of the Nature is a very effective analytic tool from mathematics.

For dynamics that can be represented in terms of a Markov process, a rather coherent system of *mesoscopic theory of entropy productions* has emerged. See [42, 150, 36, 165, 123] and references cited within. More recently, when this theory is applied to general chemical reaction systems represented by stochastic kinetics of elementary reactions, a result that is consistent with and further generalizes Gibbsian macroscopic chemical thermodynamics has been obtained, as a mathematical limit by merely allowing the molecular numbers to be infinite [44, 45]. In particular it was able to show, rigorously for the first time, that for each and every elementary, reversible reaction with instantaneous forward and backward rates R^+ and R^-, the macroscopic entropy production rate is

$(R^+ - R^-)\log(R^+/R^-)$ [132, 89, 13, 85].

In a Markov description of a driven system, irreversible kinetic *cycles* have been identified as fundamental to entropy production [63, 71, 2, 126]. From the standpoint of an observer who simultaneously follows the system's internal stochastic dynamics as well as the external driving mechanism, there is a dissipation associated with a "falling weight" [86], e.g., the work being done by an external agent as a spontaneous process. There are two different perspectives that fittingly echo the two fundamental statements of the Second Law of Thermodynamics, from Kelvin and Planck and from Clausius respectively [119]:

> "*It is impossible to construct an engine which will work in a complete cycle, and produce no effect except the raising of a weight and the cooling of a heat-reservoir.*"

> "*Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time.*"

In the present paper, we shall show that in the setting of irreversible Markov processes, both with finite state space and on a continuous n-torus with local potential, counting kinetic cycles constitutes a *lift* of the Markov processes, respectively, into either an infinite-state Markov process or diffusion on $\mathbb{R}^n$. The lifted Markov process satisfies detailed balance; it has many different invariant measures. However, it has one natural potential function and thus a corresponding Gibbsian invariant measure. This "no-flux" Gibbs measure is non-normalizable; its unbounded potential function provides a rigorous notion of an "internal energy" function φ_x. This is a new feature of the present theory that is different from previous works that usually assume the existence of a unique stationary probability measure as $t \to \infty$.

We shall show that, in the limit of $t \to \infty$, the positive stationary entropy production rate in the original Markov process is precisely the change in mean internal energy, $E \equiv \mathbb{E}[\varphi]$, of the free energy dissipation $\dot{F} = \dot{E} - \dot{S}$ in the lifted system. The energetic part of the free energy grows linearly with t; the entropic part of the free energy dissipation vanishes as $t \to \infty$ in Cesàro's sense $\dot{S} \equiv S(t)/t \to 0$. Surprisingly, in rigorous mathematics, we have not been able to show the stronger assertion that $\mathrm{d}S(t)/\mathrm{d}t \to 0$ in general except some very special cases.

Our mathematical theory, therefore, rigorously establishes an equivalence between the two famous statements concerning the Second Law of Thermodynamics: A cyclic view of dissipation [63, 15] and a non-stationary view of irreversibility. In fact, there is a continuous surjective map between the trajectories before and after the lifting: In the long-time behavior, the cycle completion and entropy production in the former is precisely represented by the potential change in the latter. Alternatively, dissipation in the former is due to indistinguishability of the locally equivalent states in the latter.

The paper is structured as follows: In Sec. 5.2, we prove an embedding theorem that establishes a minimal lift of a continuous time, finite state Markov chain to a detailed balance process with a proper potential function. An equivalence between the path-dependent entropy production in the former and the potential difference in the latter is provided. Then in Sec. 5.3, we prove the theorem that, in the limit of $t \to \infty$, equating the entropy production rate $\bar{e_p}(t)$ of the finite system with the Cesàro limit of $e_p(t)$ from the lifted system. For lifted, detailed balance systems, $e_p(t)$ can be expressed as $-\mathrm{d}F/\mathrm{d}t$ where free energy $F(t) = \mathrm{D}_{\mathrm{KL}}\big(p(t), \mu\big)$ is the relative entropy of $p(t)$ with respect to the Gibbs measure $\mu = e^{-\varphi}$. We show as $t \to \infty$, $F(t) = E(t) - S(t)$ has a linearly decreasing $E(t)$ and a sublinear $S(t)$ controlled by $\log t$. Therefore in the long time limit $\bar{e_p}$ equals to $-\dot{E}$.

Sec. 5.4 is a mathematical generalization of Sec. 5.3 to diffusion processes on n-torus and their lifting to $\mathbb{R}^n$. The paper concludes with Sec. 5.5.

5.2 The lifted Markov chain and its infinite state space

In this section, we study the lifting of a continuous time, finite state Markov chain. First, we need some prerequisite in different fields.

5.2.1 Prerequisites

Graph theory

Consider an undirected connected simple finite graph (V, E), where V is the vertices set, E is the edges set. Simple means that there is at most one edge between two vertices, and there is no edge which connects a vertex to itself.

Definition 5.1. *A* cycle *in graph (V, E) is a sequence of distinct vertices $(x_0, x_1, \cdots, x_k)$, where there exists an edge connecting x_i and x_{i+1} ($i = 0, 1, \cdots, k$, regarding x_{k+1} as x_0). Here we do not distinguish between a cycle with its inverse or shift, such as $(x_k, x_{k-1}, \cdots, x_0)$, $(x_1, x_2, \cdots, x_k, x_0)$.*

Definition 5.2. *A* tree *is a connected graph without cycle.*

Definition 5.3. *A subgraph of a graph (V, E) is called a* spanning tree *if it is a tree, and contains all the vertices of (V, E). Any undirected connected simple finite graph has at least one spanning tree.*

Definition 5.4. *The* first Betti number *of graph (V, E) is $b(V, E) = |E| - |V| + 1$.*

The following Lemma 5.1 combines Theorem 1.5.1 and Theorem 1.5.3 in [26].

Lemma 5.1. *$b(V, E) \geq 0$. $b(V, E) = 0$ if and only if (V, E) is a tree. $b(V, E) = 1$ if and only if (V, E) contains exactly one cycle.*

Lemma 5.2. *For graph (V, E), we can find $b(V, E)$ cycles $c_1, c_2, \cdots, c_{b(V,E)}$, and each c_i contains an edge e_i^* which is not contained in any cycle c_j with $j \neq i$.*

Proof. Consider a spanning tree (V, E') of (V, E). By Lemma 5.1, we know that $|E| - |E'| = b(V, E)$. Consider an edge e_i^* which is contained in (V, E) but not (V, E') (we have $b(V, E)$ of them). Adding each e_i^* to (V, E') will let this spanning tree have first Betti number 1, which means it forms exactly one cycle c_i. Now we get the desired $b(V, E)$ cycles. We call such edge e_i^* as "special edge". □

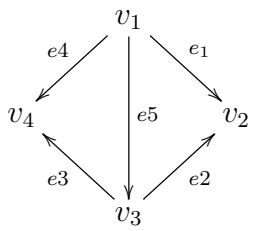
Algebraic graph theory

The definitions and notations in this part are from [120].

Now consider a directed connected simple finite graph (V, E). The only difference is that now each edge e is assigned an orientation. The inverse edge is denoted as $-e$. Define ∂ to be a $|V|\times|E|$ matrix which describes the relationship between V and E. For $v \in V$, $e \in E$, ∂_{ve} equals 1 if edge e goes into vertex v, -1 if edge e comes out of vertex v, and 0 otherwise.

Definition 5.5. *An* algebraic cycle C *is an element in the null space of* ∂*.*

Consider the following graph:



The corresponding ∂ is $\begin{bmatrix} -1 & 0 & 0 & -1 & -1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$

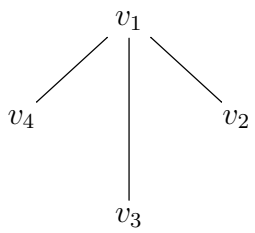
The cycle v_1, v_2, v_3, v_4 $(e_1, -e_2, e_3, -e_4)$ corresponds to an algebraic cycle, $(1, -1, 1, -1, 0)$.

It is easy to see that each cycle corresponds to an algebraic cycle. In the following we do not distinguish between a cycle and its corresponding algebraic cycle.

Lemma 5.3. *The* $b(V, E)$ *cycles in Lemma 5.2 constitute a basis of the algebraic cycle space.*

Proof. From [120], we know that the dimension of algebraic cycle space is $b(V, E) = |E|-|V|+1$. Furthermore, we can see that the $b(V, E)$ cycles in Lemma 5.2 are linearly independent, since each of them has a unique "special edge". □

In the above example, consider this spanning tree:



This means we can choose $c_1 = (v_1, v_2, v_3)\ (1, -1, 0, 0, -1)$ (corresponds to edge e_2) and $c_2 = (v_1, v_3, v_4)\ (0, 0, 1, -1, 1)$ (corresponds to edge e_3) as a basis. Then cycle $c_3 = (v_1, v_2, v_3, v_4)$ $(1, -1, 1, -1, 0)$ can be expressed as $c_3 = c_1 + c_2$.

In the following we only consider cycles (sometimes as algebraic cycles).

Potential of Markov chain

For a Markov chain, we define the potential gain of a trajectory $i_1, \cdots, i_k$ as

$$\sum_{j=1}^{k-1} \log \frac{q_{i_{j+1} i_j}}{q_{i_j i_{j+1}}}.$$

If there exists a function on state space, $f(i)$, such that the potential gain of a trajectory is the difference of this function on the two end states, then $f(i)$ is a global potential of the Markov chain.

Global potential exists if and only if the potential gain of a closed trajectory is 0. In general, a finite Markov chain does not satisfy this condition.

A global potential is proper if different states have different potentials. (In this paper, potential is always calculated symbolically.)

Algebraic topology

Definition 5.6. *Let X be a topological space. A covering space of X is a topological space C together with a continuous surjective map $p : C \to X$, such that for every $x \in X$, there exists an open neighborhood U of x, such that $p^{-1}(U)$ is a union of disjoint open sets in C, each of which is mapped homeomorphically onto U by p. A path in X can be uniquely lifted to C with a given starting point.*

Definition 5.7. *A covering space is a universal covering space if it is simply connected. General space, such as connected graph or n dimensional torus, has universal covering space. If exists, universal covering space is unique.*

For a finite graph, its covering space is still a graph. Each vertex in the covering space has an image vertex in the original graph, and they have the same neighbors. We say that the covering space is locally isomorphic to the original graph.

5.2.2 Embedding a Markov chain into an n-torus

Motivation and results

Consider a continuous time irreducible Markov chain with transition rate matrix $Q = \{q_{ij}\}$. We require that $q_{ij} > 0$ if and only if $q_{ji} > 0$. Then we can define a graph (V, E). Vertices are states of this Markov chain, and edges are possible transitions. We would like to study the reversibility of a trajectory. We define the potential gain of a trajectory $i_1, \cdots, i_k$ as

$$\sum_{j=1}^{k-1} \log \frac{q_{i_{j+1}i_j}}{q_{i_j i_{j+1}}}.$$

When a trajectory finishes a cycle, the potential gain is not zero in general, although it returns to its starting point. The aim is to find a new expression of the Markov chain, such that we can determine the potential gain of a trajectory by its starting point and ending point.

This means there exists a global potential, therefore potential gain is path-independent. The new expression should still be a Markov chain, and locally isomorphic to the original Markov chain.

A simple way is to expand all the cycles, such that there is no cycle in the new Markov chain. To be precise, this is the universal covering space of the original Markov chain. Covering space guarantees local isomorphism, and universal implies there is no cycle.

A problem of universal covering space is that different states may have the same potential, namely the potential is not proper. A natural idea is to glue states together if they have the same potential, but the result is difficult to study.

Now the goal is to find a new Markov chain, which is locally isomorphic to the original Markov chain, and has a proper global potential.

To illustrate our idea, consider a 3-state Markov chain, with one cycle $1-2-3-1$. We can embed the corresponding graph into S^1, and then lift it to $\mathbb{R}$. Now it is $\cdots$–1–2–3–1–2–3–1–$\cdots$. There exists a proper global potential. As long as we know the ends of a trajectory, we know the potential gain of this trajectory.

This problem implies relationship with the fundamental group of the original Markov chain. Notice that the potential gain of a trajectory is invariant if we exchange the order of cycles in the trajectory. Therefore we need the abelianization of the fundamental group. If the original Markov chain has first Betti number n, then the abelianization of its fundamental group is $\mathbb{Z}^n$. This is the fundamental group of n dimensional torus $\mathbb{T}^n$. Therefore we consider combining the Markov chain with $\mathbb{T}^n$.

Theorem 5.1 (torus version)**.** *For a Markov chain with $q_{ij} > 0 \iff q_{ji} > 0$, it can be embedded into n-torus $\mathbb{T}^n$, such that any closed path has zero potential gain if and only if it is homotopy trivial, namely it could continuously transform into a single point. (For $n = 1$, only the cycle can be embedded.)*

Theorem 5.2 (lifted version)**.** *With the same condition above, one can find a Markov chain with a proper global potential, and it is locally isomorphic with the original Markov chain.*

Proofs

We will prove the lifted version directly. We will use the example in Sec. 5.2.1.

Regard $\mathbb{T}^n$ as the unit hypercube $[0,1]^n$ with opposite hypersurfaces glued together. For the original Markov chain with first Betti number $b(V,E) = n$, choose a spanning tree, and embed it into $\mathbb{T}^n$. For each edge of the original Markov chain that is not in the spanning tree (we have n of such special edges), assign a pair of opposite hypersufraces to it. Draw this edge in $\mathbb{T}^n$ while crossing the corresponding hypersufrace once. Now we have embedded the original Markov chain into $\mathbb{T}^n$.

Consider the universal covering space of $\mathbb{T}^n$, $\mathbb{R}^n$. Correspondingly, the embedded Markov chain is lifted into $\mathbb{R}^n$. The lifted Markov chain is connected and locally isomorphic to the original

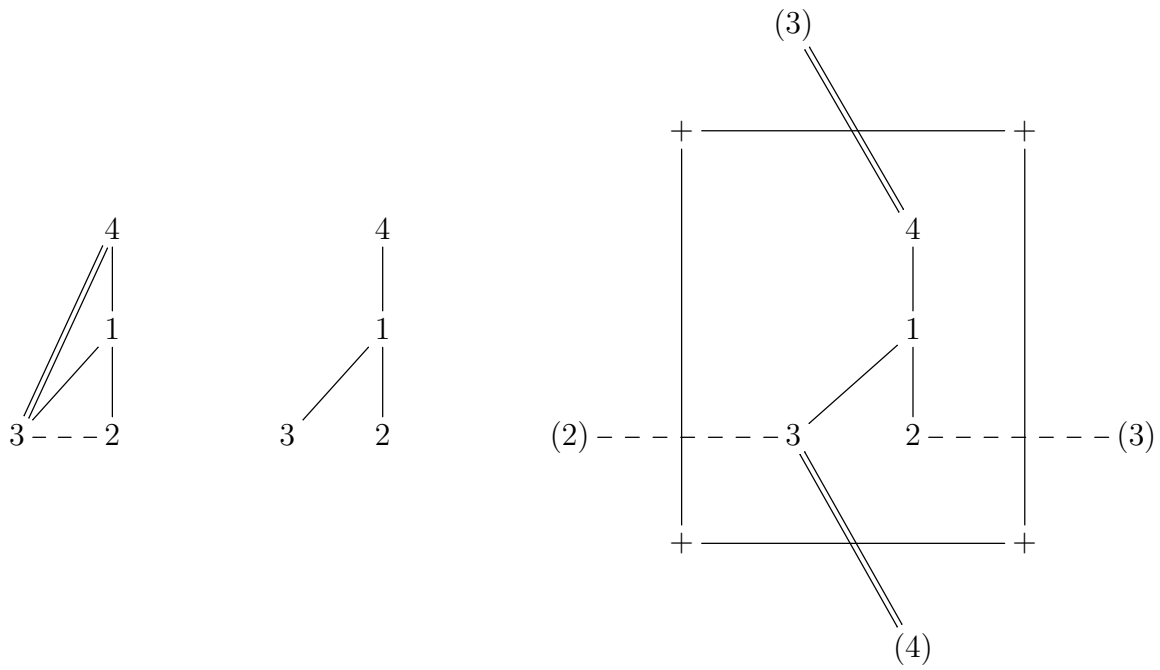


Figure 5.1: In this example, the Markov chain (left) has first Betti number $b(V, E) = 2$, therefore we can choose a spanning tree (middle), which corresponds to two special edges $2 - -3$ and $3 == 4$. Embed the spanning tree into $\mathbb{T}^2$, which is a square in $\mathbb{R}^2$ with opposite boundaries glued. Assign special edge $2 - -3$ to vertical boundaries, and $3 == 4$ to horizontal boundaries. Then connect 2 and 3 across the vertical boundaries, connect 3 and 4 across the horizontal boundaries. Now we have embedded the Markov chain into $\mathbb{T}^2$ (right).

Markov chain. A trajectory of the original Markov chain can be lifted into the new Markov chain. A trajectory of the lifted Markov chain can be folded back to the original Markov chain. We can assign an n-tuple coordinate to each unit hypercube. On the lifted Markov chain, moving along special edges is the only way to change the coordinate.

Consider a closed trajectory in the lifted Markov chain. The net number of each special edge appears in the trajectory equals the net number of corresponding hypersurface crossed, which is 0. Therefore, the net number of all special edges are 0. Fold this trajectory back to the original Markov chain, then it is an algebraic cycle. It is the linear combination of the cycles in the basis, where the basis is determined by the spanning tree. Now the coefficient of each cycle in the basis is the net number of corresponding special edge, which is 0. Therefore The folded trajectory as an

algebraic cycle is all 0. Thus the closed trajectory in the lifted Markov chain has 0 potential gain.

If a trajectory in the lifted Markov chain has 0 potential gain, then first its starting point and ending point should be of the same state. Otherwise the total number of edges containing the starting point state is odd, a contradiction. Also, if the starting point and ending point are different, then at least one component of their coordinates are different, which means the net number of the corresponding special edge is not 0. Thus the potential gain is not 0.

This finishes the proof of the lifted version. We prove that a path in the lifted Markov chain has 0 potential gain if and only if it is closed. Furthermore, a closed path in $\mathbb{T}^n$ is homotopy trivial if and only if its lifting in $\mathbb{R}^n$ is still closed (recall that the fundamental group of $\mathbb{T}^n$ is $\mathbb{Z}^n$). Therefore we also prove the torus version.

Properties

Consider a closed path in the original Markov chain. Using the notion of the "derived chain" [71], we can count all cycles completed in this path. Now decompose cycles with cycles in the basis, and count their net numbers. Then we know the net number of each cycle in the basis, which is the net number of the corresponding special edge. Also, the coordinate changes if and only if it passes the corresponding special edge.

This means that for the lifted path, the coordinate difference of its ends is just the net number of cycles completed (winding number). Two paths with the same ends have the same winding number.

When the path is not closed, the coordinate difference may not be exactly the winding number. For example, the trajectory $1-2-3-1-3$ has one cycle completed, and $1-2-3$ has no cycle completed. However, the absolute error for each cycle in the basis is at most 1. Therefore, counting cycles becomes counting the special edges, which is much simplified.

5.3 Thermodynamic quantities of Markov chains

With a Markov process and its lift established, we now consider the entropy production of the corresponding Markov processes.

5.3.1 Stationary distributions and measures of finite-state Markov chain and its lifting

We consider a continuous-time finite-state Markov chain. The chain is irreducible, and for any states i, j, the transition rates satisfy $q_{ij} > 0 \Leftrightarrow q_{ji} > 0$.

We lift this Markov chain to be an n-dimensional Markov chain with infinite states, where n is the first Betti number of the former. State i is lifted to i_α, where $\alpha \in \mathbb{Z}^n$. The lifted initial distribution is compatible with the original distribution: $\bar{p}_i(0) = \sum_\alpha p_{i_\alpha}(0)$. Then we have $\bar{p}_i(t) = \sum_\alpha p_{i_\alpha}(t)$.

The distribution of lifted Markov chain satisfies

$$\frac{\mathrm{d}p_{i_\alpha}}{\mathrm{d}t} = \sum_{j_\beta \sim i_\alpha} [p_{j_\beta} q_{ji} - p_{i_\alpha} q_{ij}].$$

We write $j_\beta \sim i_\alpha$ if j_β and i_α are adjacent.

For the finite Markov chain, starting from any initial distribution $\bar{p}_i(0)$, it will converge to the unique stationary distribution $\bar{\pi}_i$.

The lifted chain has an invariant measure

$$\pi_{i_\alpha} = \bar{\pi}_i,$$

where i_α is a copy of state i in the original chain.

Since the lifted Markov chain has a global potential φ_{i_α}, one could construct a detailed balance stationary measure $\mu_{i_\alpha} = \exp(-\varphi_{i_\alpha})$, such that $\mu_{i_\alpha} q_{ij} = \mu_{j_\beta} q_{ji}$.

To further study the stationary distributions and measures, we need to consider the relative entropy of $p_i(t)$ with respect to any stationary measure θ_{i_α},

$$\mathrm{D_{KL}}(p,\theta) = \sum_i \sum_\alpha p_{i_\alpha}(t) \log \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}}.$$

Remark of summation indices: In $\sum_{j_\beta \sim i_\alpha}$, the adjacent pair (j_β, i_α) is the same as (i_α, j_β). In $\sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha}$, (j_β, i_α) and (i_α, j_β) are counted separately. Thus

$$\sum_{j_\beta \sim i_\alpha} f_{i_\alpha j_\beta} = \frac{1}{2} \sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} f_{i_\alpha j_\beta}$$

for any $f_{i_\alpha j_\beta}$ with $f_{i_\alpha j_\beta} = f_{j_\beta i_\alpha}$. $\sum_{i,j}$ is the same as $\sum_i \sum_j$.

The following Lemma 5.4 can be found in many references [97, 154, 148].

Lemma 5.4. *$D_{KL}(p,\theta)$ is monotonically decreasing.*

Proof.

$$\begin{aligned}
& \frac{\mathrm{d}}{\mathrm{d}t} \sum_i \sum_\alpha p_{i_\alpha}(t) \log \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} \\
= & \sum_i \sum_\alpha \frac{\mathrm{d}p_{i_\alpha}(t)}{\mathrm{d}t} \log \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} + \sum_i \sum_\alpha \frac{\mathrm{d}p_{i_\alpha}(t)}{\mathrm{d}t} \\
= & -\sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} [p_{i_\alpha}(t) q_{ij} - p_{j_\beta}(t) q_{ji}] \log \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} \\
= & -\frac{1}{2} \sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} \left[p_{i_\alpha}(t) q_{ij} - p_{j_\beta}(t) q_{ji}\right] \left[\log \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} - \log \frac{p_{j_\beta}(t)}{\theta_{j_\beta}}\right]
\end{aligned}$$

$$\begin{aligned}
= & -\sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} p_{i_\alpha}(t) q_{ij} \log \frac{p_{i_\alpha}(t)\theta_{j_\beta}}{\theta_{i_\alpha} p_{j_\beta}(t)} \leq -\sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} p_{i_\alpha}(t) q_{ij} \left[1 - \frac{\theta_{i_\alpha} p_{j_\beta}(t)}{p_{i_\alpha}(t)\theta_{j_\beta}}\right] \\
= & -\sum_{i,j} \bar{p}_i(t) q_{ij} + \sum_j \sum_\beta \sum_{i_\alpha \sim j_\beta} \frac{q_{ij}\theta_{i_\alpha} p_{j_\beta}(t)}{\theta_{j_\beta}} = -\sum_{i,j} \bar{p}_i(t) q_{ij} + \sum_j \sum_\beta \frac{p_{j_\beta}(t)}{\theta_{j_\beta}} \sum_i q_{ji} \theta_{j_\beta} \\
= & -\sum_{i,j} \bar{p}_i(t) q_{ij} + \sum_{j,i} \bar{p}_j(t) q_{ji} = 0.
\end{aligned}$$

The inequality is from $\log y \geq 1 - 1/y$. The equality holds if and only if $p_{i_\alpha}(t) = c\theta_{i_\alpha}$ for a

constant c. □

Then we can prove the following result:

Proposition 5.1. *The lifted Markov chain has no stationary probability distribution.*

Proof. Assume there exists a stationary probability distribution η_{i_α}. Let $p(t)$ be the stationary distribution, and θ be π, then $\mathrm{D_{KL}}(p, \pi)$ is a constant. This is true only if the equality holds in Lemma 5.4, which means $\eta_{i_\alpha}\pi_j = \eta_{j_\beta}\pi_i$. Thus η and π only differ by a constant multiple. π is non-normalizable, so is η. □

5.3.2 Instantaneous entropy production rate, free energy, and housekeeping heat

For the finite Markov chain, one could define several thermodynamic quantities: entropy production rate, free energy, and housekeeping heat.

Definition 5.8. *The instantaneous free energy of finite Markov chain $\bar{F}(t)$ is defined as $\bar{F}(t) = D_{KL}(\bar{p}, \bar{\pi})$.*

Definition 5.9. *The instantaneous entropy production rate of finite Markov chain $\bar{e_p}(t)$ is defined as [71]*

$$\bar{e_p}(t) = \sum_{i\sim j} \left[\bar{p}_i(t)q_{ij} - \bar{p}_j(t)q_{ji}\right] \log \frac{\bar{p}_i(t)q_{ij}}{\bar{p}_j(t)q_{ji}}.$$

This definition is derived from the original idea of entropy production rate that it describes the difference between a process and its time inverse.

Definition 5.10. *The instantaneous housekeeping heat of finite Markov chain $\bar{Q}_{hk}(t)$ is defined as $\bar{e_p}(t) + \mathrm{d}\bar{F}(t)/\mathrm{d}t$.*

For the lifted Markov chain with probability distribution $p_{i_\alpha}(t)$, since there is no stationary distribution, one could choose a stationary measure instead. In general stationary measure is not unique, therefore one could have different versions of free energy and housekeeping heat.

Definition 5.11. *The instantaneous free energy with respect to stationary measure θ, $F^\theta(t)$, is defined as $F^\theta(t) = D_{KL}(f, \theta)$. Its time derivative is*

$$\mathrm{d}F^\theta(t)/\mathrm{d}t = -\sum_{i_\alpha \sim j_\beta} [p_{i_\alpha}(t)q_{ij} - p_{j_\beta}(t)q_{ji}] \log \frac{p_{i_\alpha}(t)\theta_{j_\beta}}{p_{j_\beta}(t)\theta_{i_\alpha}}.$$

Definition 5.12. *The instantaneous entropy production rate $e_p(t)$ is defined as [71]*

$$e_p(t) = \sum_{i_\alpha \sim j_\beta} \left[p_{i_\alpha}(t)q_{ij} - p_{j_\beta}(t)q_{ji}\right] \log \frac{p_{i_\alpha}(t)q_{ij}}{p_{j_\beta}(t)q_{ji}}.$$

Definition 5.13. *The instantaneous housekeeping heat with respect to stationary measure θ, $Q^\theta_{hk}(t)$, is defined as*

$$Q^\theta_{hk}(t) = e_p(t) + \mathrm{d}F^\theta(t)/\mathrm{d}t = \sum_{i_\alpha \sim j_\beta} \left[p_{i_\alpha}(t)q_{ij} - p_{j_\beta}(t)q_{ji}\right] \log \frac{\theta_{i_\alpha}q_{ij}}{\theta_{j_\beta}q_{ji}}.$$

From Lemma 5.4, $\mathrm{d}F^\theta(t)/\mathrm{d}t \leq 0$. From the definition of instantaneous entropy production rate, $e_p(t) \geq 0$. For $Q^\theta_{hk}(t)$, we have the same result.

Proposition 5.2. $Q^\theta_{hk}(t) \geq 0$.

Proof.

$$\begin{aligned} Q^\theta_{hk}(t) &= \sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} p_{i_\alpha}(t)q_{ij} \log \frac{\theta_{i_\alpha}q_{ij}}{\theta_{j_\beta}q_{ji}} \geq \sum_i \sum_\alpha \sum_{j_\beta \sim i_\alpha} p_{i_\alpha}(t)q_{ij} \left[1 - \frac{\theta_{j_\beta}q_{ji}}{\theta_{i_\alpha}q_{ij}}\right] \\ &= \sum_{i,j} \bar{p}_i(t)q_{ij} - \sum_i \sum_\alpha \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} \sum_{j_\beta \sim i_\alpha} q_{ji}\theta_{j_\beta} = \sum_{i,j} \bar{p}_i(t)q_{ij} - \sum_i \sum_\alpha \frac{p_{i_\alpha}(t)}{\theta_{i_\alpha}} \sum_j q_{ij}\theta_{i_\alpha} \\ &= \sum_{i,j} \bar{p}_i(t)q_{ij} - \sum_{i,j} \bar{p}_i(t)q_{ij} = 0. \end{aligned}$$

The inequality is from $\log y \geq 1 - 1/y$. □

Thus we have the decomposition

$$e_p(t) = Q^{\theta}_{hk}(t) + [-\mathrm{d}F^{\theta}(t)/\mathrm{d}t],$$

where each term is non-negative. This is also valid for the finite Markov chain version.

5.3.3 Time limits of thermodynamic quantities

Since $\bar{p}(t)$ converges to $\bar{\pi}$, $\bar{F}(t)$ and $\mathrm{d}\bar{F}(t)/\mathrm{d}t$ converge to 0, $\bar{e_p}(t)$ and $\bar{Q}_{hk}(t)$ converge to the stationary entropy production rate

$$\bar{e_p} = \sum_{i\sim j}[\bar{\pi}_i q_{ij} - \bar{\pi}_j q_{ji}] \log \frac{\bar{\pi}_i q_{ij}}{\bar{\pi}_j q_{ji}}.$$

For the lifted Markov chain, $p(t)$ does not converge to a stationary distribution, therefore we do not have the stationary version of these quantities. However, we can still study their behavior as $t \to \infty$.

If we set θ to be the periodic stationary measure π, then $Q^{\pi}_{hk}(t)$ converges to

$$\sum_{i\sim j}[\bar{\pi}_i q_{ij} - \bar{\pi}_j q_{ji}] \log \frac{\bar{\pi}_i q_{ij}}{\bar{\pi}_j q_{ji}},$$

which is just $\bar{e_p}$.

If we set θ to be the detailed balance stationary measure μ, then $Q^{\mu}_{hk}(t) \equiv 0$ since $\mu_{i_\alpha} q_{ij} = \mu_{j_\beta} q_{ji}$.

For the time limit of $e_p(t)$, we have the following theorem. The proof is in the next part.

Theorem 5.3. *Assume the initial distribution $p_{i_\alpha}(0)$ has finite covariance matrix. Then $e_p(t)$ converges to $\bar{e_p}$ in Cesàro's sense that* $\lim_{T\to\infty} \frac{1}{T}\int_0^T e_p(t)\mathrm{d}t = \bar{e_p}$.

In summary, $e_p = Q^{\theta}_{hk} + (-\mathrm{d}F^{\theta}/\mathrm{d}t)$, where e_p, Q^{θ}_{hk} and $-\mathrm{d}F^{\theta}/\mathrm{d}t$ are non-negative.

$e_p \to \bar{e_p}$ in Cesàro's sense, $\mathrm{d}F^{\mu}/\mathrm{d}t = -e_p \to -\bar{e_p}$ in Cesàro's sense, $Q^{\mu}_{hk} \equiv 0$, $\mathrm{d}F^{\pi}/\mathrm{d}t \to 0$ in Cesàro's sense, $Q^{\pi}_{hk} \to \bar{e_p}$ in general sense.

Therefore, the periodic stationary measure π and the detailed balance stationary measure μ reach the maximum and minimum of Q^{θ}_{hk}, $\bar{e_p}$ and 0, as $t \to \infty$.

For the two special stationary measures θ and μ, the expectation of their Radon-Nikodym derivative is

$$\sum_{i_\alpha} p_{i_\alpha}(t) \frac{\mu_{i_\alpha}}{\pi_{i_\alpha}}.$$

Since

$$\mathrm{d}F^{\pi}(t)/\mathrm{d}t - \mathrm{d}F^{\mu}(t)/\mathrm{d}t = \sum_{i_\alpha \sim j_\beta} \left[p_{i_\alpha}(t)q_{ij} - p_{j_\beta}(t)q_{ji}\right] \log \frac{\mu_{i_\alpha}\pi_{j_\beta}}{\mu_{j_\beta}\pi_{i_\alpha}} = Q^{\rho}_{hk}(t) - Q^{\mu}_{hk}(t),$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t} \sum_{i_\alpha} p_{i_\alpha}(t) \frac{\mu_{i_\alpha}}{\pi_{i_\alpha}} = \mathrm{d}F^{\pi}(t)/\mathrm{d}t - \mathrm{d}F^{\mu}(t)/\mathrm{d}t = Q^{\pi}_{hk}(t) - Q^{\mu}_{hk}(t) \to \bar{e_p}.$$

5.3.4 Proof of Theorem 5.3

The proof consists of the following Lemmas 5.5-5.8. The key idea is that we have

$$e_p(t) - \bar{e_p}(t) = \frac{\mathrm{d}\bar{F}(t)}{\mathrm{d}t} - \frac{\mathrm{d}F^{\pi}(t)}{\mathrm{d}t}. \tag{5.1}$$

We have $\mathrm{d}\bar{F}(t)/\mathrm{d}t \to 0$ and $\bar{e_p}(t) \to \bar{e_p}$. For $F^{\pi}(t)$, we prove $\mathrm{d}F^{\pi}(t)/\mathrm{d}t < 0$, and $F^{\pi}(t) > -C\log t$ for large t. Thus

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \frac{\mathrm{d}F^{\pi}(t)}{\mathrm{d}t} \mathrm{d}t = 0,$$

therefore we have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T e_p(t) \mathrm{d}t = \bar{e_p}.$$

These results are not enough, however, for proving $\mathrm{d}F^{\pi}(t)/\mathrm{d}t \to 0$. Thus we do not have $e_p(t) \to \bar{e_p}$.

For the lifted Markov chain, the entropy of distribution $p_{i_\alpha}(t)$ is

$$h[p(t)] = \sum_{i_\alpha} -p_{i_\alpha}(t) \log p_{i_\alpha}(t).$$

Lemma 5.5. *Assume the initial distribution $p(0)$ has finite entropy. Then $\lim_{T\to\infty} \frac{1}{T}\int_0^T e_p(t)\mathrm{d}t - \bar{e_p} = 0$ is equivalent with $\lim_{T\to\infty} h[p(T)]/T = 0$.*

Proof. We have

$$\frac{\mathrm{d}F^\pi(t)}{\mathrm{d}t} + e_p(t) = \frac{1}{2}\sum_{i,j}[\bar{p}_i(t)q_{ij} - \bar{p}_j(t)q_{ji}] \log \frac{\bar{\pi}_i q_{ij}}{\bar{\pi}_j q_{ji}} = \frac{\mathrm{d}\bar{F}(t)}{\mathrm{d}t} + \bar{e_p}(t).$$

Therefore

$$\begin{aligned} & \frac{1}{T}\int_0^T e_p(t)\mathrm{d}t - \frac{1}{T}\int_0^T \bar{e_p}(t)\mathrm{d}t \\ = \ & \frac{1}{T}\Big[F^\pi(0) - F^\pi(T) + \bar{F}(T) - \bar{F}(0)\Big]. \end{aligned}$$

Since $\bar{p}(t)$ converges to $\bar{\pi}$, $\bar{F}(t)$ converges to 0, therefore $\bar{F}(T)$ is bounded.

$F^\pi(0) - \bar{F}(0) = h[\bar{p}(0)] - h[p(0)]$, which is finite.

$F^\pi(T) + h[p(T)] = -\sum_i \bar{p}_i(T) \log \bar{\pi}_i$, which is bounded. Thus $\lim_{T\to\infty} F^\pi(T)/T = -\lim_{T\to\infty} h[p(T)]/T$.

We also have $\lim_{T\to\infty} \frac{1}{T}\int_0^T \bar{e_p}(t)\mathrm{d}t = \bar{e_p}$.

Therefore $\lim_{T\to\infty} \frac{1}{T}\int_0^T e_p(t)\mathrm{d}t - \bar{e_p} = -\lim_{T\to\infty} h[p(T)]/T$.

□

We have a famous result that the maximal entropy under fixed variance is achieved by normal distributions [136]:

Lemma 5.6. *For a continuous probability density function p on $\mathbb{R}^n$ with fixed covariance matrix Σ, its entropy $h[p]$ satisfies*

$$h[p] \le \frac{1}{2}\Big[n + \log\left(2^n \pi^n \det \Sigma\right)\Big].$$

The equality holds if and only if p is an n-dimensional normal distribution with covariance matrix Σ.

The last step is using variance to bound entropy. But variance does not naturally exist for all Markov chains. Also, the distribution is discrete. We can utilize the embedding of Markov chain into n-torus, then expand it. Now the lifted Markov chain is embedded into $\mathbb{R}^n$. Since we only embed finite states into a torus, we can put hypercubes centered at each state, such that these hypercubes do not intersect. Denote the length of these hypercubes by l. Then we construct a Markov process with finite density function.

The initial density is 0 outside all hypercubes, and equals $p_{i_\alpha}(0)/l^n$ in the hypercube centered at state i_α. For a point inside the hypercube centered at state i_α, it has transition rate q_{ij} to jump to an adjacent hypercube centered at state j_β. The destination is uniformly distributed in this hypercube. Therefore, the density of this new Markov process at time t is 0 outside all hypercubes, and equals $p_{i_\alpha}(t)/l^n$ in the hypercube centered at state i_α. At any time, the entropies of the Markov chain distribution and this Markov process distribution only differ by a constant $n \log l$. Now the process is in $\mathbb{R}^n$, and the distribution is continuous.

Lemma 5.7. *Consider the Markov process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$ defined above with initial distribution $p(\boldsymbol{x}, 0)$. Assume the initial distribution has finite covariance matrix. Then there exist constants C, T_0 such that for any $T > T_0$, $i, j = 1, \cdots, n$, $|Cov[\boldsymbol{X}(T)]_{ij}| \leq CT^2$.*

Proof. For any jump, the step length at each coordinate direction has an upper bound L. Then starting from any initial distribution, for $i = 1, \cdots, n$, $\Delta t > 0$ and any $0 \leq \Delta t' \leq \Delta t$,

$$\mathrm{Var}\Big[\boldsymbol{X}_i(t+\Delta t') - \boldsymbol{X}_i(t)\Big] \leq \Big[\boldsymbol{X}_i(t+\Delta t') - \boldsymbol{X}_i(t)\Big]^2 \leq \mathbb{E}N^2(t, t+\Delta t') \leq \mathbb{E}N^2(t, t+\Delta t),$$

where $N(t, t+\Delta t)$ is the number of jumps in time interval $[t, t+\Delta t)$.

Now set $q = \max q_{ij}$, and define a new process with transition rates $q'_{ij} = q$. Denote the number of jumps in time interval $[t, t+\Delta t)$ by $N_0(t, t+\Delta t)$, then we have $N(t, t+\Delta t) \leq N_0(t, t+\Delta t)$.

$N_0(t, t+\Delta t)$ is a Poisson variable with parameter $q\Delta t$, therefore $\mathbb{E}N_0^2(t, t+\Delta t) = q\Delta t + (q\Delta t)^2 < \infty$.

Then we can choose a constant G and a small enough Δt such that for any $0 \le \Delta t' \le \Delta t$, $i = 1, \cdots, n$,

$$\text{Var}\Big[\boldsymbol{X}_i(t+\Delta t') - \boldsymbol{X}_i(t)\Big] \le G\Delta t,$$

regardless of the value of $\boldsymbol{X}(t)$.

Denote $D = \max_i \text{Var}[\boldsymbol{X}_i(0)]$.

For a fixed $T > 0$, set $m = \lceil T/\Delta t \rceil$. Then

$$\text{Cov}[\boldsymbol{X}(T)] = \text{Cov}\{\boldsymbol{X}(0) + [\boldsymbol{X}(\Delta t) - \boldsymbol{X}(0)] + \cdots + [\boldsymbol{X}(T) - \boldsymbol{X}((m-1)\Delta t)]\}.$$

For two random variables Y, Z, we have $|\text{Cov}(Y, Z)| \le \sqrt{\text{Var}[Y]\text{Var}[Z]}$.

Applying this inequality to $\text{Cov}[\boldsymbol{X}(T)]$, we have

$$|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le D + 2m\sqrt{2DG\Delta t} + 2m^2 G\Delta t$$

When T is large enough, $|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le 3(T/\Delta t)^2 G\Delta t$. □

When $|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le CT^2$, $|\det \text{Cov}[\boldsymbol{X}(T)]| \le n! C^n T^{2n}$. Combining the above two lemmas, we have

Lemma 5.8. *For the lifted Markov process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, assume the initial distribution $p(\boldsymbol{x}, 0)$ has finite covariance matrix. Then its entropy at time T, $h(T)$, is controlled by $C'' \le h(T) \le C' \log T$ for T large enough, where C' and C'' are constants. Therefore $\lim_{T\to\infty} h(T)/T = 0$.*

This finishes the proof of Theorem 5.3.

5.3.5 Entropy production as energy dissipation

Consider the free energy with detailed balance stationary measure $\mu = \exp(-\varphi)$

$$F^{\mu}(t) = \sum_{i_\alpha} p_{i_\alpha}(t) \log \frac{p_{i_\alpha}(t)}{\mu_{i_\alpha}} = \sum_{i_\alpha} p_{i_\alpha}(t) \log p_{i_\alpha}(t) + \sum_{i_\alpha} p_{i_\alpha}(t) \varphi_{i_\alpha}.$$

Here $S(t) = -\sum_{i_\alpha} p_{i_\alpha}(t) \log p_{i_\alpha}(t)$ is the entropy, and $E(t) = \sum_{i_\alpha} p_{i_\alpha}(t) \varphi_{i_\alpha}$ is the mean potential energy. Thus $F^{\mu}(t) = E(t) - S(t)$.

For $E(t)$, we have the following result:

Proposition 5.3. *The time derivative of $E(t)$ converges to the negative stationary entropy production rate,*

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} \to -\bar{e_p}.$$

Proof.

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t} + \bar{e_p}(t) = \sum_{i_\alpha \sim j_\beta} [p_{i_\alpha}(t) q_{ij} - p_{j_\beta}(t) q_{ji}] \log \frac{\mu_{i_\alpha} \bar{p}_i(t) q_{ij}}{\mu_{j_\beta} \bar{p}_j(t) q_{ji}} \to \frac{1}{2} \sum_{i,j} (\bar{\pi}_i q_{ij} - \bar{\pi}_j q_{ji}) \log \frac{\bar{\pi}_i}{\bar{\pi}_j}.$$

The last term is the time derivative of $\sum_i \bar{p}_i(t) \log \bar{p}_i(t)$ when $\bar{p}_i(t) = \bar{\pi}_i$, which is 0. □

In the decomposition of free energy $F^{\mu}(t) = E(t) - S(t)$, The first term is asymptotically linear with t, and the second term is sub-linear with t (controlled by $C \log t$).

The entropy production of the finite Markov chain, which cannot be described by system status quantities directly, is reflected by the free energy/potential energy dissipation of the lifted Markov chain.

5.4 Lifting and thermodynamic quantities of multidimensional diffusion processes

5.4.1 Diffusion processes on Euclidean space and torus

Consider a time-homogeneous diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$:

$$\mathrm{d}\boldsymbol{X}(t) = \boldsymbol{\Gamma}(\boldsymbol{X})\mathrm{d}\boldsymbol{B}(t) + \boldsymbol{b}(\boldsymbol{X})\mathrm{d}t,$$

where $\boldsymbol{B}(t)$ is an n-dimensional standard Brownian motion. The drift parameter $\boldsymbol{b}(\boldsymbol{x})$ is $\mathbb{R}^n \to \mathbb{R}^n$, C^∞, with period 1 for each component. The diffusion parameter $\boldsymbol{\Gamma}(\boldsymbol{x})$ is $\mathbb{R}^n \to \mathbb{R}^{n\times n}$, non-degenerate for each $\boldsymbol{x}$, C^∞, and 1-periodic for each component. We shall also denote $\boldsymbol{D}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{\Gamma}(\boldsymbol{x})\boldsymbol{\Gamma}^T(\boldsymbol{x})$. It is positive definite for each $\boldsymbol{x}$. All vectors are $n \times 1$.

The transition probability density function $f\big(\boldsymbol{x}, t|\boldsymbol{x}_0\big)$ of the diffusion process $\boldsymbol{X}(t)$ is the fundamental solution to the linear, Kolmogorov forward equation:

$$\frac{\partial f(\boldsymbol{x},t)}{\partial t} = -\nabla\cdot\Big[\boldsymbol{b}(\boldsymbol{x})f(\boldsymbol{x},t)\Big] + \nabla\cdot\nabla\cdot\left[\boldsymbol{D}(\boldsymbol{x})f(\boldsymbol{x},t)\right]. \tag{5.2}$$

For an $n\times n$ matrix $\boldsymbol{M}$ with i-th row $\boldsymbol{M}_i$, $\nabla\cdot\boldsymbol{M}$ is defined as $n\times 1$ vector $(\nabla\cdot\boldsymbol{M}_1, \cdots, \nabla\cdot\boldsymbol{M}_n)^T$.

In parallel, consider a time-homogeneous diffusion process $\bar{\boldsymbol{X}}(t)$ on $\mathbb{T}^n$, where $\mathbb{T}^n$ is defined as $[0,1)^n$:

$$\mathrm{d}\bar{\boldsymbol{X}}(t) = \bar{\boldsymbol{\Gamma}}(\bar{\boldsymbol{X}})\mathrm{d}\bar{\boldsymbol{B}}(t) + \bar{\boldsymbol{b}}(\bar{\boldsymbol{X}})\mathrm{d}t.$$

Here $\bar{\boldsymbol{B}}(t)$ is an n-dimensional standard Brownian motion on $\mathbb{T}^n$. $\bar{\boldsymbol{\Gamma}}(\cdot)$ and $\bar{\boldsymbol{b}}(\cdot)$ are the restrictions of $\boldsymbol{\Gamma}(\cdot)$ and $\boldsymbol{b}(\cdot)$ on $\mathbb{T}^n$. Similarly, the transition probability density function for $\bar{\boldsymbol{X}}(t)$, $\bar{f}(\bar{\boldsymbol{x}}, t|\bar{\boldsymbol{x}}_0)$ satisfies the Kolmogorov forward equation:

$$\frac{\partial \bar{f}(\bar{\boldsymbol{x}},t)}{\partial t} = -\nabla\cdot\Big[\bar{\boldsymbol{b}}(\bar{\boldsymbol{x}})\bar{f}(\bar{\boldsymbol{x}},t)\Big] + \nabla\cdot\nabla\cdot\left[\bar{\boldsymbol{D}}(\bar{\boldsymbol{x}})\bar{f}(\bar{\boldsymbol{x}},t)\right], \tag{5.3}$$

in which $\bar{\boldsymbol{D}}(\cdot)$ and $\bar{\boldsymbol{b}}(\cdot)$ are the restrictions of $\boldsymbol{D}(\cdot)$ and $\boldsymbol{b}(\cdot)$ on $\mathbb{T}^n$.

For the above diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$ with periodic diffusion and drift, we can "fold" the trajectories to $\mathbb{T}^n$ by

$$\bar{\boldsymbol{X}}(t) = \boldsymbol{X}(t) \mod 1,$$

where $\mod 1$ is for every component. Since $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\Gamma}(\cdot)$ are 1-periodic, the folded process on $\mathbb{T}^n$ is exactly the above diffusion process $\bar{\boldsymbol{X}}(t)$ on $\mathbb{T}^n$. Therefore, the corresponding density

function on $\mathbb{T}^n$ is

$$\bar{f}(\bar{\boldsymbol{x}},t)=\sum_{i_1=-\infty}^{+\infty}\cdots\sum_{i_n=-\infty}^{+\infty}f\big(\bar{\boldsymbol{x}}+i_1\boldsymbol{e}_1+\cdots+i_n\boldsymbol{e}_n,t\big), \tag{5.4}$$

where $\boldsymbol{e}_k$ is an elementary n-vector, with 1 as its k-th component and 0 for other components. Eq. 5.4 can be directly verified based on the linearity of Kolmogorov forward equation.

For the diffusion process $\bar{\boldsymbol{X}}(t)$ on $\mathbb{T}^n$, we can also lift it to a diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, and the above relationship between $f(\boldsymbol{x},t)$ and $\bar{f}(\bar{\boldsymbol{x}},t)$ is still valid.

5.4.2 Stationary distributions and measures

The diffusion process $\bar{\boldsymbol{X}}(t)$ on $\mathbb{T}^n$ has a stationary distribution $\bar{\rho}(\bar{\boldsymbol{x}})$. Its 1-periodic continuation to $\mathbb{R}^n$,

$$\rho(\boldsymbol{x})=\bar{\rho}(\boldsymbol{x}\bmod 1),$$

is a stationary measure of the diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$.

To further study the stationary distributions and measures, we need to consider the relative entropy of $f(\boldsymbol{x},t)$ with respect to any stationary measure $\nu(\boldsymbol{x})$

$$\mathrm{D}_{\mathrm{KL}}[f(t),\nu]=\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\log\frac{f(\boldsymbol{x},t)}{\nu(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}.$$

Lemma 5.9. *$D_{KL}[f(t),\nu]$ is monotonically decreasing with t.*

Proof.

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}[f(t),\nu]&=\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^n}f\log\frac{f}{\nu}\mathrm{d}\boldsymbol{x}=\int_{\mathbb{R}^n}\frac{\partial f}{\partial t}\log\frac{f}{\nu}\mathrm{d}\boldsymbol{x}+\int_{\mathbb{R}^n}\frac{\partial f}{\partial t}\mathrm{d}\boldsymbol{x}\\&=\int_{\mathbb{R}^n}[-\nabla\cdot(\boldsymbol{b}f)+\nabla\cdot\nabla\cdot(\boldsymbol{D}f)]\log\frac{f}{\nu}\mathrm{d}\boldsymbol{x}\\&=\int_{\partial\mathbb{R}^n}\nabla\cdot\left[-\boldsymbol{b}f\log\frac{f}{\nu}+\nabla\cdot(\boldsymbol{D}f)\log\frac{f}{\nu}\right]\mathrm{d}S-\int_{\mathbb{R}^n}[-\boldsymbol{b}f+\nabla\cdot(\boldsymbol{D}f)]\cdot\nabla\left(\log\frac{f}{\nu}\right)\mathrm{d}\boldsymbol{x}\\&=-\int_{\mathbb{R}^n}[-\boldsymbol{b}f+f\nabla\cdot\boldsymbol{D}+\boldsymbol{D}\nabla f]\cdot\left(\frac{\nu}{f}\nabla\frac{f}{\nu}\right)\mathrm{d}\boldsymbol{x}\end{aligned}$$

$$
\begin{aligned}
&= -\int_{\mathbb{R}^n} \left[-\boldsymbol{b}\nu + \nu\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\frac{\nu}{f}\nabla f \right] \cdot \nabla\left(\frac{f}{\nu}\right) \mathrm{d}\boldsymbol{x} \\
&= -\int_{\mathbb{R}^n} \left[-\boldsymbol{b}\nu + \nabla\cdot(\boldsymbol{D}\nu) - \boldsymbol{D}\nabla\nu + \boldsymbol{D}\frac{\nu}{f}\nabla f \right] \cdot \nabla\left(\frac{f}{\nu}\right) \mathrm{d}\boldsymbol{x} \\
&= -\int_{\partial\mathbb{R}^n} \frac{f}{\nu}[\nabla\cdot(\boldsymbol{D}\nu) - \boldsymbol{b}\nu]\mathrm{d}S + \int_{\mathbb{R}^n} \frac{f}{\nu}\nabla\cdot[-\boldsymbol{b}\nu + \nabla\cdot(\boldsymbol{D}\nu)]\,\mathrm{d}\boldsymbol{x} \\
&\qquad -\int_{\mathbb{R}^n} \left[\boldsymbol{D}\left(\frac{\nu}{f}\nabla f - \nabla\nu\right) \right] \cdot \nabla\left(\frac{f}{\nu}\right) \mathrm{d}\boldsymbol{x} \\
&= -\int_{\partial\mathbb{R}^n} \nabla\cdot\left[\boldsymbol{D}\left(\frac{\nu}{f}\nabla f - \nabla\nu\right)\frac{f}{\nu} \right] \mathrm{d}S + \int_{\mathbb{R}^n} \frac{f}{\nu}\nabla\cdot\left[\boldsymbol{D}\left(\frac{\nu}{f}\nabla f - \nabla\nu\right) \right] \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} -\frac{f}{\nu}\nabla\cdot\left(\boldsymbol{D}f\nabla\frac{\nu}{f} \right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\partial\mathbb{R}^n} -\nabla\cdot\left(\frac{f}{\nu}\boldsymbol{D}f\nabla\frac{\nu}{f} \right) \mathrm{d}S + \int_{\mathbb{R}^n} \left(\boldsymbol{D}f\nabla\frac{\nu}{f} \right)\cdot\nabla\frac{f}{\nu}\mathrm{d}\boldsymbol{x} \\
&= -\int_{\mathbb{R}^n} (f\nabla\nu - \nu\nabla f)^T \frac{\boldsymbol{D}}{f\nu^2} (f\nabla\nu - \nu\nabla f)\,\mathrm{d}\boldsymbol{x},
\end{aligned}
$$

which is non-positive. It is 0 if and only if $f\nabla\nu = \nu\nabla f$, namely $\nabla\log f = \nabla\log\nu$, thus $f = c\nu$ for a constant c. □

The above lemma is also valid for diffusion on torus, thus $\mathrm{D}_{\mathrm{KL}}[\bar{f}(t), \bar{\rho}]$ is monotonically decreasing for any $\bar{f}$. If the diffusion on torus has another stationary distribution $\bar{\theta}$, then $\mathrm{D}_{\mathrm{KL}}[\bar{\theta}, \bar{\rho}]$ is a constant. However it should decrease unless $\bar{\rho} = c\bar{\theta}$, which means $\bar{\theta} = \bar{\rho}$. Therefore, the diffusion on torus has unique stationary distribution, and any initial distribution will converge to it.

The lifted diffusion on $\mathbb{R}^n$ has no stationary distribution. An intuition is that $\mathbb{R}^n$ is not compact, and the density function $f(\boldsymbol{x}, t)$ will converge to 0 at each $\boldsymbol{x}$ as $t \to \infty$.

Proposition 5.4. *The lifted diffusion process has no stationary probability distribution.*

Proof. Assume there exists a stationary probability distribution p. Let $f(t) = p$, then $\mathrm{D}_{\mathrm{KL}}(p, \rho)$ is a constant. This is true only if the equality holds in Lemma 5.9, which means p and ρ only differ by a constant multiple. However ρ is non-normalizable, so is f. □

Notice that

$$\sum_{i_1=-\infty}^{+\infty} \cdots \sum_{i_n=-\infty}^{+\infty} f(\bar{\boldsymbol{x}} + i_1\boldsymbol{e}_1 + \cdots + i_n\boldsymbol{e}_n, t) = \bar{f}(\bar{\boldsymbol{x}}, t)$$

will converge to $\bar{\rho}(\bar{\boldsymbol{x}})$.

Different from lifted Markov chain, the detailed balance measure does not always exist for lifted diffusion.

If the detailed balance measure exists, then the probability flux satisfies

$$\boldsymbol{J} = \nabla \cdot (\boldsymbol{D}f) - \boldsymbol{b}f = \boldsymbol{0}.$$

This is equivalent with

$$-\boldsymbol{D}^{-1}(\nabla \cdot \boldsymbol{D} - \boldsymbol{b}) = \nabla \log f.$$

In general, this is impossible, since $\boldsymbol{D}^{-1}(\nabla \cdot \boldsymbol{D} - \boldsymbol{b})$ may not be curl-free (conservative).

The idea of lifting is that, the process has asymmetric cycles. For Markov chain, since cycle number is finite, we can expand all of them, such that in the lifted Markov chain, there is no asymmetric cycle. For diffusion process on $\mathbb{T}^n$, there are n non-trivial basic cycles, which might be asymmetric. We expand these cycles and lift the process into $\mathbb{R}^n$. However, there are still infinite many local cycles in the lifted process, which are homotopic to a single point. In such local cycles, the curl is not always 0, therefore these cycles might be asymmetric, and we cannot expand all of them [128].

Assume that $-\boldsymbol{D}^{-1}(\nabla \cdot \boldsymbol{D} - \boldsymbol{b}) = \nabla g(\boldsymbol{x})$ is curl-free in $\mathbb{R}^n$. Then there exists a detailed balance stationary measure, $\mu(\boldsymbol{x}) = ce^{g(\boldsymbol{x})}$, where c is any positive number. Then $\boldsymbol{J} = \nabla \cdot (\boldsymbol{D}\mu) - \boldsymbol{b}\mu = \boldsymbol{0}$.

5.4.3 Instantaneous entropy production rate, free energy, and housekeeping heat

For the diffusion process $\bar{\boldsymbol{X}}(t)$ on $\mathbb{T}^n$, one could define several thermodynamics quantities: entropy production rate, free energy and housekeeping heat.

Definition 5.14. *The instantaneous free energy on torus $\bar{F}(t)$ is defined as $\bar{F}(t) = D_{KL}(\bar{f}, \bar{\rho})$.*

Definition 5.15. *The instantaneous entropy production rate on torus $\bar{e_p}(t)$ is defined as [71]*

$$\bar{e_p}(t) = \int_{\mathbb{T}^n} \frac{1}{\bar{f}}[-\bar{\boldsymbol{b}}\bar{f} + \bar{f}\nabla\cdot\bar{\boldsymbol{D}} + \bar{\boldsymbol{D}}\nabla\bar{f}]^T\bar{\boldsymbol{D}}^{-1}[-\bar{\boldsymbol{b}}\bar{f} + \bar{f}\nabla\cdot\bar{\boldsymbol{D}} + \bar{\boldsymbol{D}}\nabla\bar{f}]\mathrm{d}\bar{\boldsymbol{x}}.$$

This definition is derived from the original idea of entropy production rate that it describes the difference between a process and its time inverse.

Definition 5.16. *The instantaneous housekeeping heat on torus $\bar{Q}_{hk}(t)$ is defined as $\bar{e_p}(t)+\mathrm{d}\bar{F}(t)/\mathrm{d}t$.*

For the lifted diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$ with probability density function is $f(\boldsymbol{x},t)$, since there is no stationary distribution, one could choose a stationary measure instead. In general stationary measure is not unique, therefore one could have different versions of free energy and housekeeping heat.

Definition 5.17. *The instantaneous free energy with respect to stationary measure ν, $F^\nu(t)$, is defined as $F^\nu(t) = D_{KL}(f,\nu)$. Its time derivative is*

$$\mathrm{d}F^\nu(t)/\mathrm{d}t = -\int_{\mathbb{R}^n}\big[-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f\big]\cdot\left[\frac{\nabla f}{f} - \frac{\nabla\nu}{\nu}\right]\mathrm{d}\boldsymbol{x}.$$

Definition 5.18. *The instantaneous entropy production rate $e_p(t)$ is defined as [71]*

$$e_p(t) = \int_{\mathbb{R}^n}(-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f)^T f^{-1}\boldsymbol{D}^{-1}(-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f)\mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^n}\big[-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f\big]\cdot\left[\frac{\nabla f}{f} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]\mathrm{d}\boldsymbol{x}.$$

Definition 5.19. *The instantaneous housekeeping heat with respect to stationary measure ν, $Q^\nu_{hk}(t)$, is defined as*

$$Q^\nu_{hk}(t) = e_p(t) + \mathrm{d}F^\nu(t)/\mathrm{d}t = \int_{\mathbb{R}^n}\big[-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f\big]\cdot\left[\frac{\nabla\nu}{\nu} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]\mathrm{d}\boldsymbol{x}.$$

From Lemma 5.9, $\mathrm{d}F^\nu(t)/\mathrm{d}t \le 0$. Since $\boldsymbol{D}$ is positive definite, $e_p(t) \ge 0$. For $Q^\nu_{hk}(t)$, we have the same result.

Proposition 5.5. $Q^{\nu}_{hk}(t) \geq 0$.

Proof.

$$Q^{\nu}_{hk}(t) = \int_{\mathbb{R}^n} \big[-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f\big]\cdot\left[\frac{\nabla\nu}{\nu} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]\mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^n} f\left[\frac{\nabla\nu}{\nu} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]^T\boldsymbol{D}\left[-\boldsymbol{D}^{-1}\boldsymbol{b} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} + \frac{\nabla f}{f}\right]\mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^n} f\left[\frac{\nabla\nu}{\nu} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]^T\boldsymbol{D}\left[-\boldsymbol{D}^{-1}\boldsymbol{b} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} + \frac{\nabla\nu}{\nu}\right]\mathrm{d}\boldsymbol{x}$$

$$+ \int_{\mathbb{R}^n} f\left[\frac{\nabla\nu}{\nu} + \boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D} - \boldsymbol{D}^{-1}\boldsymbol{b}\right]^T\boldsymbol{D}\left[-\frac{\nabla\nu}{\nu} + \frac{\nabla f}{f}\right]\mathrm{d}\boldsymbol{x}.$$

Since $\boldsymbol{D}$ is positive definite, the first term is non-negative. The second term equals

$$\int_{\mathbb{R}^n}[\boldsymbol{D}\nabla\nu + \nu\nabla\cdot\boldsymbol{D} - \boldsymbol{b}\nu]^T\frac{f}{\nu}\left[-\frac{\nabla\nu}{\nu} + \frac{\nabla f}{f}\right]\mathrm{d}\boldsymbol{x}$$

$$= \int_{\mathbb{R}^n}[\boldsymbol{D}\nabla\nu + \nu\nabla\cdot\boldsymbol{D} - \boldsymbol{b}\nu]^T\nabla\left(\frac{f}{\nu}\right)\mathrm{d}\boldsymbol{x}$$

$$= -\int_{\mathbb{R}^n}\frac{f}{\nu}\nabla\cdot[\boldsymbol{D}\nabla\nu + \nu\nabla\cdot\boldsymbol{D} - \boldsymbol{b}\nu]\mathrm{d}\boldsymbol{x} = 0,$$

since ν is a stationary measure, $\nabla\cdot\nabla\cdot(\boldsymbol{D}\nu) - \nabla\cdot(\boldsymbol{b}\nu) = 0$. □

Thus we have the decomposition

$$e_p(t) = Q^{\nu}_{hk}(t) + [-\mathrm{d}F^{\nu}(t)/\mathrm{d}t],$$

where each term is non-negative. This is also valid for the torus version.

5.4.4 Time limits of thermodynamic quantities

Since $\bar{f}(t)$ converges to $\bar{\rho}$, $\bar{F}(t)$ and $\mathrm{d}\bar{F}(t)/\mathrm{d}t$ converge to 0, $\bar{e_p}(t)$ and $\bar{Q}_{hk}(t)$ converge to the stationary entropy production rate

$$\bar{e_p} = \int_{\mathbb{T}^n} \frac{1}{\bar{\rho}}[-\bar{\boldsymbol{b}}\bar{\rho} + \bar{\rho}\nabla\cdot\bar{\boldsymbol{D}} + \bar{\boldsymbol{D}}\nabla\bar{\rho}]^T\bar{\boldsymbol{D}}^{-1}[-\bar{\boldsymbol{b}}\bar{\rho} + \bar{\rho}\nabla\cdot\bar{\boldsymbol{D}} + \bar{\boldsymbol{D}}\nabla\bar{\rho}]\mathrm{d}\bar{\boldsymbol{x}}.$$

For the lifted diffusion process, $f(t)$ does not converge to a stationary distribution, therefore we do not have the stationary version of these quantities. However, we can still study their behavior as $t \to \infty$.

If we set ν to be the periodic stationary measure ρ, then $Q^{\rho}_{hk}(t)$ converges to

$$\int_{\mathbb{T}^n} [-\bar{\boldsymbol{b}}\bar{\rho} + \bar{\rho}\nabla\cdot\bar{\boldsymbol{D}} + \bar{\boldsymbol{D}}\nabla\bar{\rho}]\cdot\left[\frac{\nabla\bar{\rho}}{\bar{\rho}} + \bar{\boldsymbol{D}}^{-1}\nabla\cdot\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\bar{\boldsymbol{b}}\right]\mathrm{d}\bar{\boldsymbol{x}},$$

which is just $\bar{e_p}$.

If we set ν to be the detailed balance stationary measure μ (if exists), then $Q^{\mu}_{hk}(t) \equiv 0$ since $\nabla\cdot(\boldsymbol{D}\mu) - \boldsymbol{b}\mu = \boldsymbol{0}$.

For the time limit of $e_p(t)$, we have the following theorem. The proof is in the next part.

Theorem 5.4. *For any initial distribution $f(\boldsymbol{x}, 0)$ that has a finite covariance matrix, the entropy production rate of diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, $e_p(t)$, also converges to $\bar{e_p}$ in Cesàro's sense that*

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T e_p(t)\mathrm{d}t = \bar{e_p}.$$

In summary, $e_p = Q^{\nu}_{hk} + (-\mathrm{d}F^{\nu}/\mathrm{d}t)$, where e_p, Q^{ν}_{hk} and $-\mathrm{d}F^{\nu}/\mathrm{d}t$ are non-negative.

$e_p \to \bar{e_p}$ in Cesàro's sense, $\mathrm{d}F^{\mu}/\mathrm{d}t = -e_p \to -\bar{e_p}$ in Cesàro's sense, $Q^{\mu}_{hk} \equiv 0$, $\mathrm{d}F^{\rho}/\mathrm{d}t \to 0$ in Cesàro's sense, $Q^{\rho}_{hk} \to \bar{e_p}$ in general sense.

Therefore, the periodic stationary measure ρ and the detailed balance stationary measure μ reach the maximum and minimum of Q^{ν}_{hk}, $\bar{e_p}$ and 0, as $t \to \infty$.

When $-\boldsymbol{D}^{-1}(\nabla\cdot\boldsymbol{D} - \boldsymbol{b})$ is not curl-free, μ does not exist, and the minimum of Q^{ν}_{hk} is larger

than 0.

For the two special stationary measures ρ and μ, the expectation of their Radon-Nikodym derivative is

$$\int_{\mathbb{R}^n} f(t)\frac{\mu}{\rho}\mathrm{d}\boldsymbol{x}.$$

Since

$$\mathrm{d}F^{\rho}(t)/\mathrm{d}t - \mathrm{d}F^{\mu}(t)/\mathrm{d}t = \int_{\mathbb{R}^n} \left[-\boldsymbol{b}f + f\nabla\cdot\boldsymbol{D} + \boldsymbol{D}\nabla f\right]\cdot\left[\frac{\nabla\rho}{\rho} - \frac{\nabla\mu}{\mu}\right]\mathrm{d}\boldsymbol{x} = Q^{\rho}_{hk}(t) - Q^{\mu}_{hk}(t),$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^n} f(t)\frac{\mu}{\rho}\mathrm{d}\boldsymbol{x} = \mathrm{d}F^{\rho}(t)/\mathrm{d}t - \mathrm{d}F^{\mu}(t)/\mathrm{d}t = Q^{\rho}_{hk}(t) - Q^{\mu}_{hk}(t) \to \bar{e_p}.$$

5.4.5 Proof of Theorem 5.4

The proof consists of the following lemmas. The key idea is that we have

$$e_p(t) - \bar{e_p}(t) = \frac{\mathrm{d}\bar{F}(t)}{\mathrm{d}t} - \frac{\mathrm{d}F^{\rho}(t)}{\mathrm{d}t}.$$

Also $\bar{e_p}(t) \to \bar{e_p}$, $\mathrm{d}\bar{F}(t)/\mathrm{d}t \to 0$.

For $F^{\rho}(t)$, we prove $\mathrm{d}F^{\rho}(t)/\mathrm{d}t \leq 0$, and $F^{\rho}(t) > -C\log t$ for large t. Thus

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \frac{\mathrm{d}F^{\rho}(t)}{\mathrm{d}t}\mathrm{d}t = 0,$$

therefore we have

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T e_p(t)\mathrm{d}t = \bar{e_p}.$$

But these are not enough for $\mathrm{d}F^{\rho}(t)/\mathrm{d}t \to 0$, therefore we do not have $e_p(t) \to \bar{e_p}$.

For a distribution $q(\boldsymbol{x})$, its entropy is defined as

$$h[q] = \int_{\mathbb{R}^n} -q(\boldsymbol{x})\log q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Lemma 5.10. *Assume the initial distribution $f(\boldsymbol{x},0)$ has finite covariance matrix for any $\boldsymbol{x}$. Then $\lim_{T\to\infty}\frac{1}{T}\int_0^T e_p(t)\mathrm{d}t-\bar{e_p}=0$ is equivalent with $\lim_{T\to\infty}F^\rho(T)/T=0$.*

Proof. We have

$$e_p(t)-\bar{e_p}(t)=\int_{\mathbb{R}^n}\frac{1}{f}(\nabla f)^T\boldsymbol{D}\nabla f\mathrm{d}\boldsymbol{x}-\int_{\mathbb{T}^n}\frac{1}{\bar{f}}(\nabla\bar{f})^T\bar{\boldsymbol{D}}\nabla\bar{f}\mathrm{d}\bar{\boldsymbol{x}}=-\frac{\mathrm{d}F^\rho(t)}{\mathrm{d}t}+\frac{\mathrm{d}\bar{F}(t)}{\mathrm{d}t}.$$

Thus

$$\frac{1}{T}\int_0^T e_p(t)\mathrm{d}t-\frac{1}{T}\int_0^T\bar{e_p}(t)\mathrm{d}t=\frac{1}{T}[F^\rho(0)-F^\rho(t)+\bar{F}(T)-\bar{F}(0)].$$

Since $\bar{f}(\bar{\boldsymbol{x}},t)$ converges to $\bar{\rho}(\bar{\boldsymbol{x}})$, $\bar{F}(t)$ converges to 0, therefore $\bar{T}$ is bounded.

$F^\rho(0)-\bar{F}(0)=h[\bar{f}(0)]-h[f(0)]$. Since $f(\boldsymbol{x},0)$ has finite covariance matrix, Lemma 5.6 shows that $h[f(0)]$ is finite, so as $h[\bar{f}(0)]$.

We also have $\lim_{T\to\infty}\frac{1}{T}\int_0^T\bar{e_p}(t)\mathrm{d}t=\bar{e_p}$.

Therefore $\lim_{T\to\infty}\frac{1}{T}\int_0^T e_p(t)\mathrm{d}t-\bar{e_p}=\lim_{T\to\infty}F^\rho(T)/T$.

□

The next step is to control $F^\rho(T)$ for large T. Since $F^\rho(T)=-h[f(T)]-\int_{\mathbb{R}^n}f(\boldsymbol{x},T)\log\rho\mathrm{d}\boldsymbol{x}$, and $\int_{\mathbb{R}^n}f(\boldsymbol{x},T)\log\rho\mathrm{d}\boldsymbol{x}$ converges to $\int_{\mathbb{T}^n}\rho(\bar{\boldsymbol{x}})\log\rho(\bar{\boldsymbol{x}})\mathrm{d}\bar{\boldsymbol{x}}$, which is finite, we only need to show $h[f(T)]/T\to 0$.

Since $f(\boldsymbol{x},t)\le\bar{f}(\bar{\boldsymbol{x}},t)$, $\bar{f}(\bar{\boldsymbol{x}},t)$ converges to $\bar{\rho}(\bar{\boldsymbol{x}})$, $f(\boldsymbol{x},t)$ has a uniform upper bound for large t. Therefore $h[p(t)]$ has a finite lower bound for large t. We only need to control $h[f(t)]$ from above. From Lemma 5.6, we need to control the covariance matrix of the diffusion process.

Lemma 5.11. *Consider the diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$ with initial distribution $f(\boldsymbol{x},0)$. Assume the initial distribution $f(\boldsymbol{x},t)$ has finite covariance matrix. Then there exist constants C,T_0 such that for any $T>T_0$, $i,j=1,\cdots,n$, $|Cov[\boldsymbol{X}(T)]_{ij}|\le CT^2$.*

Proof. For the diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, we have the infinitesimal mean

$$\mathbb{E}[\boldsymbol{X}(t+\Delta t)-\boldsymbol{X}(t)\mid\boldsymbol{X}(t)=\boldsymbol{x}_0]=\boldsymbol{b}(\boldsymbol{x}_0)\Delta t+O(\Delta t^2).$$

We also have the infinitesimal variance

$$\mathbb{E}\{[\boldsymbol{X}(t+\Delta t)-\boldsymbol{X}(t)][\boldsymbol{X}(t+\Delta t)-\boldsymbol{X}(t)]^T \mid \boldsymbol{X}(t)=\boldsymbol{x}_0\} = \boldsymbol{\Gamma}(\boldsymbol{x}_0)\boldsymbol{\Gamma}(\boldsymbol{x}_0)^T\Delta t + O(\Delta t^2).$$

Therefore the covariance matrix satisfies

$$\text{Cov}[\boldsymbol{X}(t+\Delta t) \mid \boldsymbol{X}(t)=\boldsymbol{x}_0] = \boldsymbol{\Gamma}(\boldsymbol{x}_0)\boldsymbol{\Gamma}(\boldsymbol{x}_0)^T\Delta t + O(\Delta t^2).$$

Set $G = \max_{\boldsymbol{x},i}[\Gamma(\boldsymbol{x})\Gamma(\boldsymbol{x})^T]_{ii} < \infty$. Then we can choose a small enough Δt such that for any $0 \le \Delta t' \le \Delta t$, $i = 1, \cdots, n$

$$\text{Var}[\boldsymbol{X}_i(t+\Delta t') - \boldsymbol{X}_i(t)] \le 2G\Delta t,$$

regardless of the value of $\boldsymbol{X}(t)$.

Denote $D = \max_i \text{Var}[\boldsymbol{X}_i(0)]$.

For a fixed $T > 0$, set $m = \lceil T/\Delta t \rceil$. Then

$$\text{Cov}[\boldsymbol{X}(T)] = \text{Cov}\{\boldsymbol{X}(0) + [\boldsymbol{X}(\Delta t) - \boldsymbol{X}(0)] + \cdots + [\boldsymbol{X}(T) - \boldsymbol{X}((m-1)\Delta t)]\}.$$

For two random variables Y, Z, we have $|\text{Cov}(Y,Z)| \le \sqrt{\text{Var}[Y]\text{Var}[Z]}$.

Applying this inequality to $\text{Cov}[\boldsymbol{X}(T)]$, we have

$$|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le D + 2m\sqrt{2DG\Delta t} + 2m^2G\Delta t$$

When T is large enough, $|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le 3(T/\Delta t)^2 G\Delta t$. □

When $|\text{Cov}[\boldsymbol{X}(T)]_{ij}| \le CT^2$, $|\det \text{Cov}[\boldsymbol{X}(T)]| \le n!C^nT^{2n}$. Now we have

Lemma 5.12. *For the diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, assume the initial distribution $f(\boldsymbol{x}, 0)$ has finite covariance matrix. Then its entropy at time T, $h(T)$, is controlled by $C'' \le h(T) \le C' \log T$ for T large enough, where C' and C'' are constants. Therefore $\lim_{T\to\infty} h(T)/T = 0$.*

This finishes the proof of Theorem 5.4.

In general we do not have $e_p(t) \to \bar{e_p}$. However, we can prove $e_p(t) \to \bar{e_p}$ for a special case.

Proposition 5.6. *For the diffusion process $\boldsymbol{X}(t)$ on $\mathbb{R}^n$, if $\boldsymbol{b}(\boldsymbol{x})$ and $\boldsymbol{\Gamma}(\boldsymbol{x})$ are constants, initial distribution $f(\boldsymbol{x}, 0)$ has finite covariance matrix, then $\lim_{t\to\infty} e_p(t) = \bar{e_p}$.*

Proof. We need to prove $\mathrm{d}F^\rho(t)/\mathrm{d}t \to 0$. Since $\rho(\boldsymbol{x})$ is a constant function in this case, we only need $r(t) = \mathrm{d}h[f(t)]/\mathrm{d}t \to 0$. In the following we will prove $\mathrm{d}r(t)/\mathrm{d}t \leq 0$. Then $r(t)$ is positive and decreasing, therefore it has a limit c. Thus

$$c = \lim_{T\to\infty} \frac{1}{T} \int_0^T r(t)\mathrm{d}t = 0.$$

Set $f_0(\boldsymbol{x}, t) = f(\boldsymbol{x} - \boldsymbol{b}t, t)$, then it is the density function of the no-drift process $\mathrm{d}\boldsymbol{X}_0 = \boldsymbol{\Gamma}\mathrm{d}\boldsymbol{B}(t)$. Entropy is invariant under translation, $h[f(t)] = h[f_0(t)]$, therefore we can set $\boldsymbol{b} = 0$. Since $\boldsymbol{\Gamma}$ is non-degenerate, we have singular value decomposition $\boldsymbol{\Gamma} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}, \boldsymbol{V}$ are orthonormal, and $\boldsymbol{\Sigma}$ is diagonal and positive. Now $\mathrm{d}\boldsymbol{U}^T\boldsymbol{X}_0 = \boldsymbol{\Sigma}\mathrm{d}\boldsymbol{V}^T\boldsymbol{B}$. $\boldsymbol{V}^T\boldsymbol{B}$ is still an n-dimensional standard Brownian motion. Since entropy is invariant under rotation, $\boldsymbol{Y} = \boldsymbol{U}^T\boldsymbol{X}_0$ has the same entropy with $\boldsymbol{X}_0$. We can set $\boldsymbol{\Gamma}$ to be diagonal and positive. Finally, set $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1}\boldsymbol{X}$, then $\mathrm{d}\boldsymbol{Z} = \mathrm{d}\boldsymbol{B}$. Now the entropy of $\boldsymbol{Z}$ and $\boldsymbol{X}$ only differ by a constant $\det(\boldsymbol{\Sigma})$, which does not affect the derivative. Therefore, we only need to prove $\mathrm{d}^2h[f(t)]/\mathrm{d}t^2 \geq 0$ for process $\mathrm{d}\boldsymbol{X} = \mathrm{d}\boldsymbol{B}$, where the Kolmogorov forward equation is heat equation $f_t = \Delta f/2$.

The following proof is from [37].

We shall abbreviate $\mathrm{d}f/\mathrm{d}x_i$ by f_i in the following to prevent double subscripts. We also assume that the initial condition is sufficiently nice to guarantee integration by parts being valid.

$$\frac{\mathrm{d}h[f(t)]}{\mathrm{d}t} = \int_{\mathbb{R}^n} \frac{\nabla f \cdot \nabla f}{2f} \mathrm{d}\boldsymbol{x}.$$

$$\begin{aligned}\frac{\mathrm{d}^2 h[f(t)]}{\mathrm{d}t^2} &= \int_{\mathbb{R}^n}\left(\frac{\nabla f\cdot\nabla f_t}{f}-\frac{\nabla f\cdot\nabla f}{2f^2}f_t\right)\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^n}\left(\frac{1}{2}\sum_{i,j}\frac{f_i f_{ijj}}{f}-\frac{1}{4}\sum_{i,j}\frac{f_i f_i f_{jj}}{f^2}\right)\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Integrate by parts yields,

$$\begin{aligned}\frac{\mathrm{d}^2 h[f(t)]}{\mathrm{d}t^2} &= \int_{\mathbb{R}^n}\left(-\frac{1}{2}\sum_{i,j}f_{ij}\frac{ff_{ij}-f_i f_j}{f^2}+\frac{1}{4}\sum_{i,j}f_j\frac{2f^2 f_i f_{ij}-2f_i^2 f f_j}{f^4}\right)\mathrm{d}\boldsymbol{x}\\ &= -\frac{1}{2}\int_{\mathbb{R}^n}\left(\frac{f_{ij}^2}{f}-2\frac{f_i f_j f_{ij}}{f^2}+\frac{f_i^2 f_j^2}{f^3}\right)\mathrm{d}\boldsymbol{x}\\ &= -\frac{1}{2}\int_{\mathbb{R}^n}\left(\frac{f_{ij}}{\sqrt{f}}-\frac{f_i f_j}{f\sqrt{f}}\right)^2\mathrm{d}\boldsymbol{x}\ \le\ 0.\end{aligned}$$

□

5.4.6 Entropy production as energy dissipation

Consider the free energy with detailed balance stationary measure $\mu=\exp(-\varphi)$ (if exists)

$$F^{\mu}(t)=\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\log\frac{f(\boldsymbol{x},t)}{\mu(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}=\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\log f(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}-\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\log\mu(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Here $S(t)=-\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\log f(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}$ is the entropy, and $E(t)=\int_{\mathbb{R}^n}f(\boldsymbol{x},t)\varphi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is the mean potential energy. Thus $F^{\mu}(t)=E(t)-S(t)$.

For $E(t)$, we have the following result:

Proposition 5.7. *The time derivative of $E(t)$ converges to the negative stationary entropy production rate,*

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t}\to-\bar{e_p}.$$

Proof.

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t}+\bar{e_p}(t)=\int_{\mathbb{R}^n}[\nabla\cdot(\boldsymbol{D}f)-\boldsymbol{b}f]\cdot\left(\frac{\nabla\mu}{\mu}\right)\mathrm{d}\boldsymbol{x}$$

$$+\int_{\mathbb{R}^n}[\nabla\cdot(\boldsymbol{D}f)-\boldsymbol{b}f]\cdot\left(\frac{\nabla\bar{f}}{\bar{f}}+\boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D}-\boldsymbol{D}^{-1}\boldsymbol{b}\right)\mathrm{d}\boldsymbol{x}$$

$$=\int_{\mathbb{T}^n}\left[\nabla\cdot(\bar{\boldsymbol{D}}\bar{f})-\bar{\boldsymbol{b}}\bar{f}\right]\cdot\left(\frac{\nabla\bar{f}}{\bar{f}}\right)\mathrm{d}\bar{\boldsymbol{x}}=-\int_{\mathbb{T}^n}\nabla\cdot\left[\nabla\cdot(\bar{\boldsymbol{D}}\bar{f})-\bar{\boldsymbol{b}}\bar{f}\right]\log\bar{f}\mathrm{d}\bar{\boldsymbol{x}}$$

$$=-\int_{\mathbb{T}^n}\nabla\cdot\left[\nabla\cdot(\bar{\boldsymbol{D}}\bar{\rho})-\bar{\boldsymbol{b}}\bar{\rho}\right]\log\bar{\rho}\mathrm{d}\bar{\boldsymbol{x}}=0.$$

In the integral $\int_{\mathbb{R}^n}\cdots\mathrm{d}\boldsymbol{x}$, $\bar{f}(\boldsymbol{x})$ is the 1-periodic extension of $\bar{f}(\bar{\boldsymbol{x}})$. We also apply the facts that $\nabla\cdot\left[\nabla\cdot(\bar{\boldsymbol{D}}\bar{\rho})-\bar{\boldsymbol{b}}\bar{\rho}\right]=0$ and $\frac{\nabla\mu}{\mu}+\boldsymbol{D}^{-1}\nabla\cdot\boldsymbol{D}-\boldsymbol{D}^{-1}\boldsymbol{b}=\boldsymbol{0}$.

□

In the decomposition of free energy $F^{\mu}(t)=E(t)-S(t)$, The first term is asymptotically linear with t, and the second term is sub-linear with t (controlled by $C\log t$).

The entropy production of the diffusion process on $\mathbb{T}^n$, which cannot be described by system status quantities directly, is reflected by the free energy/potential energy dissipation of the lifted diffusion process on $\mathbb{R}^n$.

5.5 Discussion and summary of conclusions

5.5.1 Gibbs potential and Kirkwood's potential of mean force

Relative entropy w.r.t. the Lebesgue measure, e.g., Gibbs-Shannon's entropy, is not the most appropriate information characteristics for system with a nontrivial invariant measure. This insight has already in the work of classical thermodynamics, where entropy as the "thermodynamic potential" of an isolated system with a priori equal probability is replaced by the free energy as the proper thermodynamic potential for an isothermal system. In Gibbs' theory of chemical thermodynamics, chemical potential of a chemical special i has actually *three* parts: $\mu_i=\mu_i^o+k_BT\log x_i=h^o-Ts^o+k_BT\log x_i$.

5.5.2 The nature of nonequilibrium dissipation

The nature of "thermodynamic dissipation" has long been debated. A notion that is generally agreed upon was put forward by Onsager [105], who clearly identified a dissipation with a transport process, with both nonzero *thermodynamic force* and *thermodynamic flux*. In fact, they are necessarily vanishing simultaneously in a thermodynamic equilibrium with detailed balance. This gives rise to the reciprocal relation in the linear regime near equilibrium. Macroscopic transport processes, however, can be classified into two gross types: Those induced by a nonequilibrium initial condition and those driven by an active forcing. This distinction, we believe, is precisely behind Clausius' and Kelvin's statements of the Second Law of thermodynamics, as well as behind the irreversibility formulated by Boltzmann and I. Prigogine, respectively. Still, in the latter case, the precise physical step(s) at which dissipation occurs has generated a wide range of arguments: It is attributed to the external driving force, to the transport processes themselves inside a system, and to the "boundary" where a system in contact with its nonequilibrium environment [6]. Our mathematical theory clearly indicates that all these perspectives are not incorrect, but a more precise, complete notion really is to identify nonequilibrium *cycles* which have been independently discovered by T. L. Hill [63] and by Laudauer and Bennett [15], in biochemistry and computation respectively.

A cyclic macroscopic transport process driven by a sustained nonequilibrium environment is of course an idealization of the reality: A battery has to be re-charged, the chemical solution that sustains a chemostat has to be replenished. In fact, almost all engineering systems do not use a continuously charged energy source, but rather rely on the principle of quasi-stationarity [43]: A narrow range of decreasing external driving force is acceptable. Therefore, by clearly recognizing the source of a driving force, the cycles inside a finite driven system can and should be identified with a spontaneous "downhill" of an external process, as clearly shown in the present paper. In fact, a perfect stationary driving force, which is represented by the house-keeping heat in the present work, corresponds to an unbounded potential energy function on a non-compact space.

Mathematically, this lifting of a driven system with discrete state space or continuous n-torus

state spaces has been rigorously established in the present work. Generalization of this result to $\mathbb{R}^n$ without local potential is technically challenging, but that should not prevent our understanding of the nature of the physics of nonequilibrium dissipation. In fact, we propose a modernized, combined Clausius-Kelvin statement:

> "*A mesoscopic engine that works in completing irreversible internal cycles statistically has necessarily an external effect that passes heat from a warmer to a colder body.*"

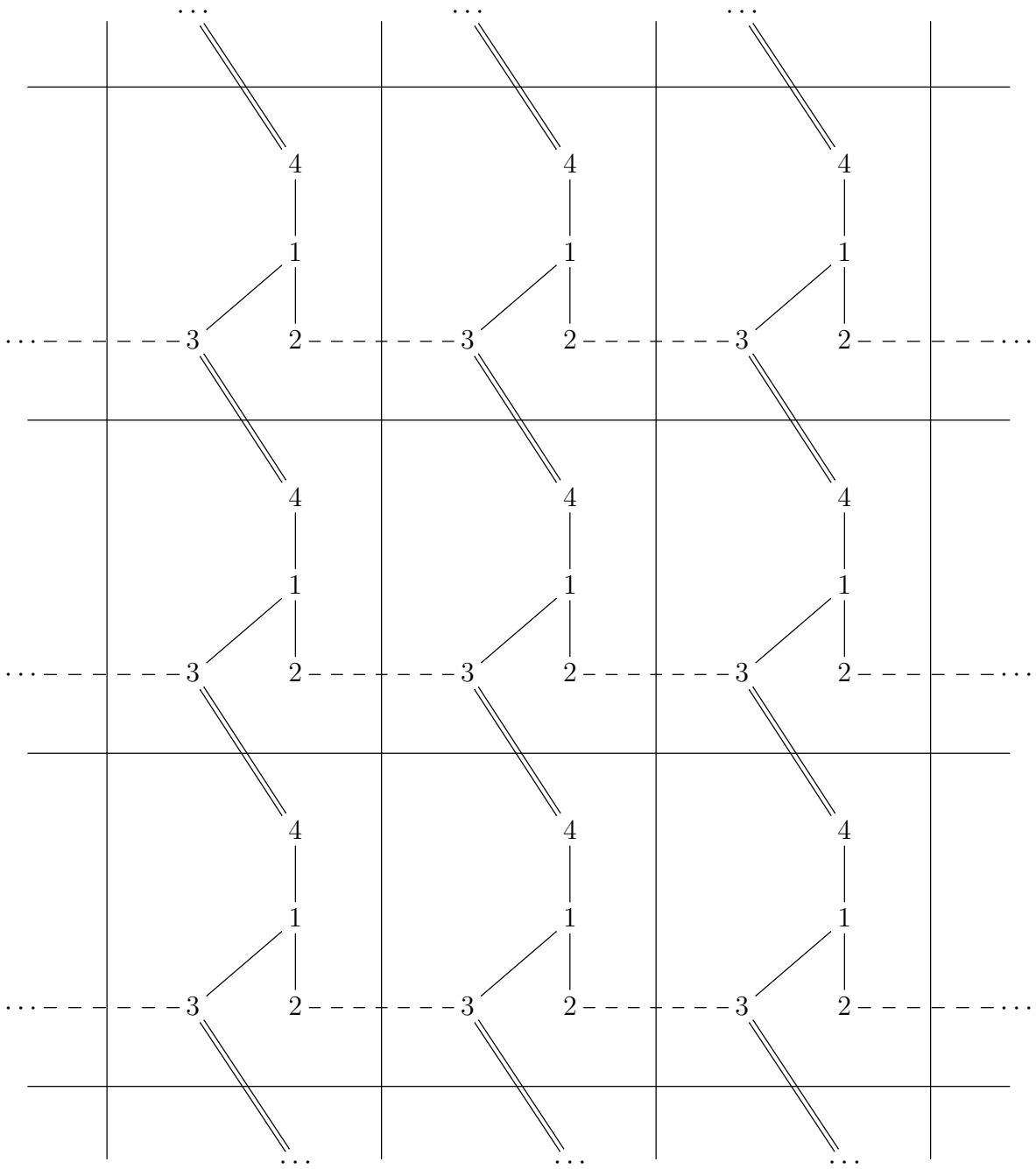


Figure 5.2: Continued with Figure 5.1. $\mathbb{T}^2$ is lifted to its universal covering space $\mathbb{R}^2$. Accordingly the embedded Markov chain is lifted to a covering space (not universal), which is a larger Markov chain.

Chapter 6

FUTURE WORKS

6.1 General problems

In recent years, with the dramatic advances in financial mathematics, diffusion processes have become one of widely studied and applied types of stochastic processes with continuous state space. On the other hand, significant applications of continuous-time, discrete-state Markov processes have also developed in single-molecule biophysics [127], and continuous-time integer-valued Markov processes in the stochastic kinetics of intracellular biochemical reactions [121, 122]. One expects this latter approach, a generalized multi-dimensional birth-death process on $\mathbb{N}^n$, to also have a significant role to play in the mathematical theory of Single-Cell Biology. It provides a natural generalization, in terms of T. G. Kurtz's law of large numbers [4], of the deterministic differential equations based population dynamics. Stochastic dynamics with discrete state space also calls for more combinatorics techniques and discrete mathematics.

As the biology of transcriptomics and proteomics enter the single-cell level [117], a single cell is no longer the negligible one individual in a total population, but a wealth of statistical information. The interplay between individual cell phenotype switchings and the population of interacting cells with differential growth gives rise to an *ecological population dynamics* of cells in a tissue, normal and/or malignant. In the limit of large tissues, laws of large numbers in Kurtz's type yield the traditional deterministic dynamic models for nonlinear population dynamics. One expects a much more realistic and mathematically sophisticated population dynamics theory to emerge.

Cancer biology also suggests a novel type of stochastic dynamics: processes that continuously "grow" thus without a long-time stationary probability. One needs a stochastic theory beyond the perspective of *nonequilibrium steady state* [46]. The next, natural idealization is to consider

processes that have a stationary "drift" in the long-time limit. The distinction between stationary phenotypic distribution of an "individual cell" and stationary proportions of independent cells in a growing population with differential rates of proliferation constitutes the "pure Darwinian selection", as already noted succinctly by Dr. Sui Huang [170]. Several mathematical issues in connection to this general problem arose in this dissertation: The law of large numbers in Chapter 3, and the observation of entropy and mean potential energy growing as $\log t$ and t respectively in the lifted system in Chapter 5. In a long-time limit, drift dominates stationary growth. There might indeed be a connection between Darwinian natural selection's fitness, the theory of large deviations, and entropy production of stationary drifting processes, as hypothesized by Dr. Hong Qian [124, 125].

6.2 Specific problems

6.2.1 Data–based stochastic cell growth

For the stochastic cell growth project at the Institute for Systems Biology, more experiments are in progress or in preparation. Currently the data is only on counting the number of cells. The future data will include clear linage between individual cells and their descendants. Intracellular gene expression patterns within single cells will be measured. In addition, barcodes will be injected into gene sequences, such that the genealogy can be determined by genes. With the gene expression data, we might be able to verify our current hypothesis concerning the heterogeneity among cells, and further estimate transition rates.

In current experimental data, the correlation coefficient of populations at different times is relatively small, compared with that of the model we use. This might be a possible direction of proposing more valid models.

In [62], it is reported that during the metastasis of cancer cells, the pioneer cells which are killed by immune systems can affect the function of immune systems, such that the later arrived cancer cells will be less vulnerable. Inspired by this, we can include the number of dead cells in our models. This might produce more complicated behaviors.

6.2.2 Mechanism–based stochastic cell growth

For the theory of branching process, there is another approach that we do not mention much, the generating function method. Through this method, more analysis tools could be introduced, and it is possible to prove further theorems, especially in the non-Markovian case.

Stochastic processes have much broader applications in cell biology than what we discuss here. One plan is to study other stochastic processes with biology background, such as coalescent processes. Such research results might be applied to population biology directly.

6.2.3 Data statistics and causality

Chapter 4 only considers the negative aspect of causal inference, namely when the data is not coherent with quantitative causal methods. However, the core concept in this chapter, Markov boundary, is essential in applications of causal inference, even in machine learning. Therefore, it is a good start for the positive aspect of causal inference, namely designing causal methods and determining the condition under which it is valid.

6.2.4 Entropy production of stochastic processes

In Chapter 5, we prove that the entropy production rate of the lifted process converges to that of the original process. However, the convergence is only in Cesàro's sense. We believe the convergence is also valid in a stronger version. Based on more than a year of struggle and exploration, however, I suspect that a proof might need much more delicate techniques.

We have shown that there exist deep relationships between algebraic topology and the entropy theory of stochastic processes. One could search further applications of algebraic topology, especially homology theory, in non-symmetric Markov processes [74].

Diffusion processes in Chapter 5 are limited to $\mathbb{T}^n$ and $\mathbb{R}^n$. However, there have been studies that extend entropy theory to diffusion processes on general closed manifolds [128]. The local flux can be lifted through connection to a circle bundle. In such way, the process is totally lifted, such that for a lifted trajectory, the non-triviality of its entropy production is equivalent with the

non-triviality of its topology. We guess that through such total lifting, our results in Chapter 5 are still valid.

BIBLIOGRAPHY

[1] C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *J. Mach. Learn. Res.*, 11(Jan):171–234, 2010.

[2] B. Altaner. Foundations of stochastic thermodynamics. *arXiv preprint arXiv:1410.3983*, 2014.

[3] S. J. Altschuler and L. F. Wu. Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559–563, 2010.

[4] D. F. Anderson and T. G. Kurtz. *Stochastic Analysis of Biochemical Systems*. Springer, 2015.

[5] S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Stat.*, 25(2):505–541, 1997.

[6] B. Andresen, P. Salamon, and R. S. Berry. Thermodynamics in finite time. *Phys. Today*, 37(9):62–70, 1984.

[7] K. B. Athreya. Some results on multitype continuous time Markov branching processes. *Ann. Math. Stat.*, 39(2):347–357, 1968.

[8] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer-Verlag, Berlin, 1972.

[9] L. A. Baccala, K. Sameshima, G. Ballester, A. C. Do Valle, and C. Timo-Iaria. Studying the interaction between brain structures via directed coherence and Granger causality. *Appl. Signal Process.*, 5(1):40, 1998.

[10] N. T. J. Bailey. *The Elements of Stochastic Processes with Applications to the Natural Sciences*, volume 25. John Wiley & Sons, 1990.

[11] L. Barnett, A. B. Barrett, and A. K. Seth. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.*, 103(23):238701, 2009.

[12] R. Bartoszynski, B. W. Brown, C. M. McBride, and J. R. Thompson. Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. *Ann. Stat.*, 9(5):1050–1060, 1981.

[13] D. A. Beard and H. Qian. Relationship between thermodynamic driving force and one-way fluxes in reversible processes. *PLOS ONE*, 2(1):e144, 2007.

[14] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188, 2001.

[15] C. H. Bennett. Notes on Landauer's principle, reversible computation, and Maxwell's demon. *Stud. Hist. Philos. Sci. B*, 34(3):501–510, 2003.

[16] P. G. Bergmann and J. L. Lebowitz. New approach to nonequilibrium processes. *Phys. Rev.*, 99(2):578, 1955.

[17] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, R. Karchin, K. W. Kinzler, B. Vogelstein, and M. A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci.*, 107(43):18545–18550, 2010.

[18] M. B. Brown and A. B. Forsythe. Robust tests for the equality of variances. *J. Am. Stat. Assoc.*, 69(346):364–367, 1974.

[19] W. H. Clark. Tumour progression and the nature of cancer. *Br. J. Cancer*, 64(4):631, 1991.

[20] E. Clayton, D. P. Doupé, A. M. Klein, D. J. Winton, B. D. Simons, and P. H. Jones. A single type of progenitor cell maintains normal epidermis. *Nature*, 446(7132):185–189, 2007.

[21] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[22] D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes*. Methuen & Co. Ltd., London, 1965.

[23] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE T. Inform. Theory*, 45(4):1315–1321, 1999.

[24] S. Rodrigues de Morais and A. Aussem. A novel Markov boundary based feature subset selection algorithm. *Neurocomput.*, 73(4):578–584, 2010.

[25] A. Dewanji, E. G. Luebeck, and S. H. Moolgavkar. A generalized Luria–Delbrück model. *Math. Biosci.*, 197(2):140–152, 2005.

[26] R. Diestel. *Graph Theory*. Springer-Verlag, New York, 2nd edition, 2000.

[27] R. L. Dobrushin. General formulation of Shannon's main theorem in information theory. *Amer. Math. Soc. Trans.*, 33:323–438, 1963.

[28] J. R. Dorfman. *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*. Cambridge University Press, 1999.

[29] R. V. dos Santos and L. M. da Silva. The noise and the KISS in the cancer stem cells niche. *J. Theor. Biol*, 335:79–87, 2013.

[30] O. J. Dunn. Multiple comparisons among means. *J. Am. Stat. Assoc.*, 56(293):52–64, 1961.

[31] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, Cambridge, 4th edition, 2010.

[32] M. Egeblad, E. S. Nakasone, and Z. Werb. Tumors as organs: Complex tissues that interface with the entire organism. *Dev. Cell*, 18(6):884–901, 2010.

[33] M. Eigen, J. McCaskill, and P. Schuster. Molecular quasi-species. *J. Phys. Chem.*, 92(24):6881, 1988.

[34] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics*, 21(2):171–178, 2004.

[35] A. Einstein. *Investigations on the Theory of the Brownian Movement*. Courier Corporation, 1956.

[36] M. Esposito and C. Van den Broeck. Three detailed fluctuation theorems. *Phys. Rev. Lett.*, 104(9):090601, 2010.

[37] L. C. Evans. *Entropy and Partial Differential Equations*. Createspace Independent Pub, 2014.

[38] I. J. Fidler and M. L. Kripke. Metastasis results from preexisting variant cells within a malignant tumor. *Science*, 197(4306):893–895, 1977.

[39] C. M. Fillmore and C. Kuperwasser. Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. *Breast Cancer Res.*, 10(2):R25, 2008.

[40] S. Frenzel and B. Pompe. Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99(20):204101, 2007.

[41] R. A. Gatenby, J. J. Cunningham, and J. S. Brown. Evolutionary triage governs fitness in driver and passenger mutations and suggests targeting never mutations. *Nat. Commun.*, 5:5499, 2014.

[42] H. Ge and H. Qian. Physical origins of entropy production, free energy dissipation, and their mathematical representations. *Phys. Rev. E*, 81(5):051133, 2010.

[43] H. Ge and H. Qian. Heat dissipation and nonequilibrium thermodynamics of quasi-steady states and open driven steady state. *Phys. Rev. E*, 87:062125, 2011.

[44] H. Ge and H. Qian. Mesoscopic kinetic basis of macroscopic chemical thermodynamics: A mathematical theory. *Phys. Rev. E*, 94(5):052150, 2016.

[45] H. Ge and H. Qian. Mathematical formalism of nonequilibrium thermodynamics for nonlinear chemical reaction systems with general rate law. *J. Stat. Phys.*, 166(1):190–209, 2017.

[46] H. Ge, M. Qian, and H. Qian. Stochastic theory of nonequilibrium steady states. Part II: Applications in chemical biophysics. *Phys. Rep.*, 510(3):87–118, 2012.

[47] J. F. Geweke. Measures of conditional linear dependence and feedback between time series. *J. Am. Stat. Assoc.*, 79(388):907–915, 1984.

[48] P. Glansdorff, I. Prigogine, and R. N. Hill. Thermodynamic theory of structure, stability and fluctuations. *Am. J. Phys.*, 41(1):147–148, 1973.

[49] C. Godsil and G. F. Royle. *Algebraic Graph Theory*. Springer Science & Business Media, 2013.

[50] L. R. Goldberg, A. N. Kercheval, and K. Lee. T-statistics for weighted means in credit risk modeling. *J. Risk Finance*, 6(4):349–365, 2005.

[51] A. S. Goldberger. *A Course in Econometrics*. Harvard University Press, Cambridge, 1991.

[52] A. Golubev. Exponentially modified gaussian (EMG) relevance to distributions related to cell proliferation and differentiation. *J. Theor. Biol.*, 262:257–266, 2010.

[53] B. Gourévitch, R. Le Bouquin-Jeannès, and G. Faucon. Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biol. Cybern.*, 95(4):349–369, 2006.

[54] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[55] V. Griffing and K. F. Herzfeld. *Fundamental Physics of Gases*. Princeton University Press, 2015.

[56] S. Guo, A. K. Seth, K. M. Kendrick, C. Zhou, and J. Feng. Partial Granger causality–eliminating exogenous inputs and latent variables. *J. Neurosci. Meth.*, 172(1):79–93, 2008.

[57] P. B. Gupta, C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, and E. S. Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, 2011.

[58] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 2000.

[59] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, 2011.

[60] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2016.

[61] E. D. Hawkins, M. L. Turner, M. R. Dowling, C. van Gend, and P. D. Hodgkin. A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proc. Natl. Acad. Sci.*, 104:5032–5037, 2007.

[62] M. B. Headley, A. Bins, A. Nip, E. W. Roberts, M. R. Looney, A. Gerard, and M. F. Krummel. Visualization of immediate immune responses to pioneer metastatic cells in the lung. *Nature*, 531(7595):513–517, 2016.

[63] T. L. Hill. *Free Energy Transduction in Biology: The Steady-State Kinetic and Thermodynamic Formalism*. Academic Press, 1977.

[64] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117(4):500–544, 1952.

[65] T.-M. Huang. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Stat.*, 38(4):2047–2091, 2010.

[66] P. Jagers. The proportions of individuals of different kinds in two-type populations. A branching process problem arising in biology. *J. Appl. Probab*, 6(2):249–260, 1969.

[67] S. Janson. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stoch. Process. Appl.*, 110(2):177–245, 2004.

[68] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Ann. Stat.*, 41(5):2324–2358, 2013.

[69] C. Jarzynski. Equalities and inequalities: Irreversibility and the second law of thermodynamics at the nanoscale. *Annu. Rev. Condens. Matter Phys.*, 2(1):329–351, 2011.

[70] C. Jia, D.-Q. Jiang, and M.-P. Qian. Cycle symmetries and circulation fluctuations for discrete-time and continuous-time Markov chains. *Ann. Appl. Probab.*, 26(4):2454–2493, 2016.

[71] D.-Q. Jiang, M. Qian, and M.-P. Qian. *Mathematical Theory of Nonequilibrium Steady States: on the Frontier of Probability and Dynamical Systems*. Springer-Verlag, Berlin, 2004.

[72] D.-Q. Jiang, Y. Wang, and D. Zhou. Phenotypic equilibrium as probabilistic convergence in multi-phenotype cell population dynamics. *PLOS ONE*, 12(2):e0170916, 2017.

[73] A. Kaiser and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166(1):43–62, 2002.

[74] S. L. Kalpazidou. *Cycle representations of Markov processes*. Springer Science & Business Media, 2007.

[75] Y. Kang, C. Gu, L. Yuan, Y. Wang, Y. Zhu, X. Li, Q. Luo, J. Xiao, D.-Q. Jiang, M.-P. Qian, A. A. Khan, F. Chen, Z. Zhang, and J. Yu. Flexibility and symmetry of prokaryotic genome rearrangement reveal lineage-associated core-gene-defined genome organizational frameworks. *MBio*, 5(6):e01867–14, 2014.

[76] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes*. Academic Press, New York, 2nd edition, 1975.

[77] R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.*, 90(431):928–934, 1995.

[78] H. Kesten and B. P. Stigum. Limit theorems for decomposable multi-dimensional Galton-Watson processes. *J. Math. Anal. Appl.*, 17(2):309–338, 1967.

[79] M. Kimmel and D. E. Axelrod. *Branching Processes in Biology*. Springer, New York, 2002.

[80] M. Kimmel and S. Corey. Stochastic hypothesis of transition from inborn neutropenia to AML: Interactions of cell population dynamics and population genetics. *Front. Oncol.*, 3:89, 2013.

[81] A. L. Koch. Mutation and growth rates from Luria-Delbrück fluctuation tests. *Mutat. Res. -Fund. Mol. M.*, 95(2-3):129–143, 1982.

[82] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69(6):066138, 2004.

[83] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.

[84] D. E. Lea and C. A. Coulson. The distribution of the numbers of mutants in bacterial populations. *J. Genet.*, 49(3):264–285, 1949.

[85] V. V. Levchenko, R. Fleming, H. Qian, and D. A. Beard. An annotated english translation of 'Kinetics of stationary reactions' [MI Temkin, Dolk. Akad. Nauk SSSR. 152, 156 (1963)]. *arXiv preprint arXiv:1001.2861*, 2010.

[86] E. H. Lieb and J. Yngvason. A fresh look at entropy and the second law of thermodynamics. *Phys. Today*, 53(4):32–38, 200.

[87] A. J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci.*, 6(7):410–415, 1920.

[88] G. Luebeck and S. H. Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *Proc. Natl. Acad. Sci.*, 99(23):15095–15100, 2002.

[89] J.-L. Luo, C. Van Den Broeck, and G. Nicolis. Stability criteria and fluctuations around nonequilibrium states. *Z. Phys. B*, 56(2):165–170, 1984.

[90] S. E. Luria and M. Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[91] W. J. Mackillop. The growth kinetics of human tumours. *Clin. Phys. Physiol. M.*, 11(4A):121, 1990.

[92] S. Mani and G.F. Cooper. Causal discovery using a bayesian local causal discovery algorithm. *Medinfo*, 11(Pt 1):731–735, 2004.

[93] T. O. McDonald and M. Kimmel. A multitype infinite-allele branching process with applications to cancer evolution. *J. Appl. Probab.*, 52(3):864–876, 2015.

[94] C. E. Meacham and S. J. Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[95] C. J. Mode. *Multitype Branching Processes: Theory and Applications*. American Elsevier Pub. Co., New York, 1971.

[96] S. H. Moolgavkar and G. Luebeck. Two-event model for carcinogenesis: Biological, mathematical, and statistical considerations. *Risk Anal.*, 10(2):323–341, 1990.

[97] T. Morimoto. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.*, 18(3):328–331, 1963.

[98] D. Mumford. The dawning of the age of stochasticity. In V. I. Arnold, M. Atiyah, P. Lax, and B. Mazur, editors, *Mathematics: Frontiers and Perspectives*, pages 197–218. American Mathematical Society, Providence, 2000.

[99] J. D. Murray. *Mathematical Biology I: An Introduction*. Springer-Verlag, Berlin, Heidelberg, 2001.

[100] R.E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

[101] P. K. Newton, J. Mason, K. Bethel, L. A. Bazhenova, J. Nieva, and P. Kuhn. A stochastic Markov chain model to describe lung cancer growth and metastasis. *PLOS ONE*, 7(4):e34637, 2012.

[102] J. Neyman. On the application of probability theory to agricultural experiments: Essay on principles. *Stat. Sci.*, 5:463–472, 1990.

[103] G. Nicolis and I. Prigogine. *Self-Organization in Nonequilibrium Systems*. Wiley, New York, 1977.

[104] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

[105] L. Onsager. Reciprocal relations in irreversible processes. I. *Phys. Rev.*, 37(4):405, 1931.

[106] A.-M. M. Oswald, B. Doiron, and L. Maler. Interval coding. I. Burst interspike intervals as indicators of stimulus intensity. *J. Neurophysio.*, 97(4):2731–2743, 2007.

[107] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. van Oudenaarden. Multistability in the lactose utilization network of *Escherichia coli*. *Nature*, 427(6976):737–740, 2004.

[108] M. Paluš and M. Vejmelka. Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. *Phys. Rev. E*, 75(5):056211, 2007.

[109] L. Patrawala, T. Calhoun, R. Schneider-Broussard, J. Zhou, K. Claypool, and D. G. Tang. Side population is enriched in tumorigenic, stem-like cancer cells, whereas ABCG2$^+$ and ABCG2$^-$ cancer cells are similarly tumorigenic. *Cancer Res.*, 65(14):6207–6219, 2005.

[110] W. Pauli and C. P. Enz. *Thermodynamics and the Kinetic Theory of Gases*. Courier Corporation, 2000.

[111] J. Pearl. *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann, San Mateo, 1988.

[112] J. Pearl and A. Paz. *Graphoids: A Graph-based Logic for Reasoning about Relevance Relations*. University of California (Los Angeles). Computer Science Department, 1985.

[113] J. Pearl and T. Verma. A theory of inferred causation. *Stud. Log. Found. Math.*, 134:789–811, 1995.

[114] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[115] H. Pearson. Cell biology: The cellular hullabaloo. *Nature*, 453(7192):150–153, 2008.

[116] J.M. Peña, R. Nilsson, J. Björkegren, and J. Tegnér. Towards scalable and data efficient learning of Markov boundaries. *Int. J. Approx. Reason.*, 45(2):211–232, 2007.

[117] J. M. Perkel. Single-cell biology: The power of one. *Science*, 350(6261):696–698, 2015.

[118] A. O. Pisco and S. Huang. Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer*, 112(11):1725–1732, 2015.

[119] M. Planck. *Treatise on Thermodynamics*. Dover, 3rd edition, 1910.

[120] M. Polettini. Cycle/cocycle oblique projections on oriented graphs. *Lett. in Math. Phys.*, 105(1):89–107, 2015.

[121] H. Qian. Cellular biology in terms of stochastic nonlinear biochemical dynamics: Emergent properties, isogenetic variations and chemical system inheritability. *J. Stat. Phys.*, 141(6):990–1013, 2010.

[122] H. Qian. Nonlinear stochastic dynamics of mesoscopic homogeneous biochemical reaction systems–an analytical theory. *Nonlinearity*, 24(6):R19–R49, 2011.

[123] H. Qian. A decomposition of irreversible diffusion processes without detailed balance. *J. Math. Phys.*, 54(5):053302, 2013.

[124] H. Qian. Fitness and entropy production in a cell population dynamics with epigenetic phenotype switching. *Quant. Biol.*, 2(1):47–53, 2014.

[125] H. Qian and H. Ge. Book review for "Random Perturbations of Dynamical Systems". *SIAM Rev.*, 55(3):569–574, 2013.

[126] H. Qian, S. Kjelstrup, A. B. Kolomeisky, and D. Bedeaux. Entropy production in mesoscopic stochastic thermodynamics: Nonequilibrium kinetic cycles driven by chemical potentials, temperatures, and mechanical forces. *J. Phys. Condens. Matter*, 28(15):153004, 2016.

[127] H. Qian and S. C. Kou. Statistics and related topics in single-molecule biophysics. *Annu. Rev. Stat. Appl.*, 1:465–492, 2014.

[128] M. Qian and Z.-D. Wang. The entropy production of diffusion processes on manifolds and its circulation decompositions. *Commun. Math. Phys.*, 206(2):429–445, 1999.

[129] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[130] T. Richardson and P. Spirtes. Ancestral graph markov models. *Ann. Stat.*, 30(4):962–1030, 2002.

[131] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period–application to control of the healthy worker survivor effect. *Math. Model.*, 7(9-12):1393–1512, 1986.

[132] J. Ross and R. S. Berry. *Thermodynamics and Fluctuations Far from Equilibrium*. Springer, 2008.

[133] M. S. Roulston. Significance testing of information theoretic functionals. *Physica D*, 110(1):62–66, 1997.

[134] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66(5):688, 1974.

[135] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85(2):461, 2000.

[136] P. Schuster. *Stochasticity in Processes*. Springer, 2016.

[137] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75(12):126001, 2012.

[138] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, New York, 2nd edition, 1981.

[139] R. T. Smythe. Central limit theorems for urn models. *Stoch. Process. Appl.*, 65(1):115–137, 1996.

[140] C. Sonnenschein and A. M. Soto. Somatic mutation theory of carcinogenesis: Why it should be dropped and replaced. *Mol. Carcinog.*, 29(4):205–211, 2000.

[141] J. F. Speer, V. E. Petrosky, M. W. Retsky, and R. H. Wardwell. A stochastic numerical model of breast cancer growth that simulates clinical data. *Cancer Res.*, 44(9):4124–4130, 1984.

[142] S. Spina, V. Giorno, P. Román-Román, and F. Torres-Ruiz. A stochastic model of cancer growth subject to an intermittent treatment with combined effects: Reduction in tumor size and rise in growth rate. *Bull. Math. Biol.*, 76(11):2711–2736, 2014.

[143] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.*, 9(1):62–72, 1991.

[144] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[145] A. Statnikov, N.I. Lytkin, J. Lemeire, and C.F. Aliferis. Algorithms for discovery of multiple Markov boundaries. *J. Mach. Learn. Res.*, 14(Feb):499–566, 2013.

[146] L. Su and H. White. A nonparametric Hellinger metric test for conditional independence. *Econ. Theory*, 24(04):829–864, 2008.

[147] D. P. Tabassum and K. Polyak. Tumorigenesis: It takes a village. *Nat. Rev. Cancer*, 15(8):473–483, 2015.

[148] L. F. Thompson and H. Qian. Potential of entropic force in Markov systems with nonequilibrium steady state, generalized Gibbs function and criticality. *Entropy*, 18(8):309, 2016.

[149] I. Tsamardinos and C.F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[150] C. Van den Broeck and M. Esposito. Three faces of the second law. II. Fokker-Planck formulation. *Phys. Rev. E*, 82(1):011144, 2010.

[151] M. Vejmelka and M. Paluš. Inferring the directionality of coupling with conditional mutual information. *Phys. Rev. E*, 77(2):026214, 2008.

[152] T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence*, pages 323–330. Morgan Kaufmann Publishers Inc., 1992.

[153] C. Villani. H-theorem and beyond: Boltzmann's entropy in today's mathematics. In G. Gallavotti, W. L. Reiter, and J. Yngvason, editors, *Boltzmann's Legacy*, pages 129–143. European Mathematical Society, Zürich, 2008.

[154] J. Voigt. Stochastic operators, information, and entropy. *Commun. Math. Phys.*, 81(1):31–38, 1981.

[155] H. W. Watson and F. Galton. On the probability of the extinction of families. *J. Royal Anthropol. Inst.*, 4:138–144, 1875.

[156] B. L. Welch. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[157] N. Wiener. The theory of prediction. In E. F. Beckenbach, editor, *Modern Mathematics for Engineers*, pages 125–139. McGraw-Hill New York, 1956.

[158] A. Y. Yakovlev, M. Mayer-Proschel, and M. Noble. A stochastic model of brain cell differentiation in tissue culture. *J. Math. Biol.*, 37(1):49–60, 1998.

[159] A. Y. Yakovlev and N. M. Yanev. Relative frequencies in multitype branching processes. *Ann. Appl. Probab.*, 19(1):1–14, 2009.

[160] A. Y. Yakovlev and N. M. Yanev. Limiting distributions for multitype branching processes. *Stoch. Anal. Appl.*, 28(6):1040–1060, 2010.

[161] G. Yang, Y. Quan, W. Wang, Q. Fu, J. Wu, T. Mei, J. Li, Y. Tang, C. Luo, Q. Ouyang, S. Chen, L. Wu, T.K. Hei, and Y. Wang. Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human SW620 and MCF-7 cancer cell populations. *Br. J. Cancer*, 106(9):1512–1519, 2012.

[162] H. Ye, R. J. Beamish, S. M Glaser, S. C. H. Grant, C.-H. Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proc. Natl. Acad. Sci.*, 112(13):E1569–E1576, 2015.

[163] E. D. Yorke, Z. Fuks, L. Norton, W. Whitmore, and C. C. Ling. Modeling the development of metastases from primary and locally recurrent tumors: Comparison with a clinical data base for prostatic cancer. *Cancer Res.*, 53(13):2987–2993, 1993.

[164] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.

[165] X.-J. Zhang, H. Qian, and M. Qian. Stochastic theory of nonequilibrium steady states and its applications. Part I. *Phys. Rep.*, 510(1):1–86, 2012.

[166] J. Zhao, Y. Zhou, X. Zhang, and L. Chen. Part mutual information for quantifying direct associations in networks. *Proc. Natl. Acad. Sci.*, 113(18):5130–5135, 2016.

[167] Q. Zheng. Progress of a half century in the study of the Luria–Delbrück distribution. *Math. Biosci.*, 162(1):1–32, 1999.

[168] D. Zhou, Y. Wang, and B. Wu. A multi-phenotypic cancer model with cell plasticity. *J. Theor. Biol.*, 357:35–45, 2014.

[169] D. Zhou, D. Wu, Z. Li, M. P. Qian, and M. Q. Zhang. Population dynamics of cancer cells with cell state conversions. *Quant. Biol.*, 1(3):201–208, 2013.

[170] J. X. Zhou, A. O. Pisco, H. Qian, and S. Huang. Nonequilibrium population dynamics of phenotype conversion of cancer cells. *PLOS ONE*, 9(12):e110714, 2014.

[171] M. H. Zwietering, I. Jongenburger, F. M. Rombouts, and K. van't Riet. Modeling of the bacterial growth curve. *Appl. Environ. Microbiol.*, 56(6):1875–1881, 1990.