Development and Analysis of Markov Chains on Manifolds

Elnaz Karimian Sichani

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Ph.D. degree in Mathematics and Statistics *

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

*The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Random sampling from a probability distribution is a fundamental computational task that is widely applied in various disciplines. It has applications in statistics, machine learning, probability, and other areas that involve stochastic modeling. MCMC is a way to (approximately) sample. Given a target distribution, the MCMC method typically consists of two phases. First, construct a Markov chain whose stationary distribution is either the target distribution or a close approximation of it. Next, simulate the chain for a sufficient number of steps to ensure that it has mixed and produced an approximate sample from the target distribution. However, for an MCMC algorithm to be efficient, the Markov chain must quickly reach its steady state. This is a major concern in computational science. Therefore, understanding the time it takes for the Markov chain to reach its equilibrium distribution, known as the "mixing time," is essential. This thesis explores the most commonly employed techniques for bounding the mixing time of Markov chains, such as the spectral and profile methods, geometric bounds, comparison methods, coupling, and decomposition techniques. We present some new results on convergence bounds of continuous-state Markov chains and prove analogous relationships between different notions of distance from stationarity in discrete to continuous state spaces. We then investigate the sharpness of the spectral profile bound on the mixing time of continuous-state Markov chains and find that it is precise up to a factor of $\log \log$ of the initial density. This finding has applications in the comparison of Markov chains.

Finally, in this thesis, we discuss the application of simulation models and model fitting

in synthetic data generation (SDG). We propose a generative model for producing synthetic longitudinal data that integrates efficient simulation techniques, including MCMC, sequential trees, and copula, to sample a latent factor matrix within the framework of the state-of-the-art generalized canonical polyadic (GCP) tensor decomposition.

Dedications

To my beloved family:

My dear father, my wonderful and caring mother, my amazing and supportive husband, and my two adorable and wonderful children.

You mean the world to me, but I don't tell you enough.

Acknowledgements

This thesis was completed under the supervision of Professor Aaron Smith and Professor Mahmoud Zarepour. I would like to express my profound gratitude, sincere respect, and deep appreciation to my supervisors for their invaluable supervision.

Dear Professor Smith, I would like to express my heartfelt gratitude for your unwavering patience, understanding, immense support, and encouragement throughout my PhD journey. Your comprehensive and insightful feedback has been invaluable, shaping my understanding of the subject matter. I am deeply grateful to you.

Dear Professor Zarepour, I would like to extend my sincere thanks for your skillful and insightful guidance, as well as your encouragement and support. I am deeply grateful to you.

Furthermore, I would like to extend my heartfelt gratitude and sincere appreciation to Professor David Asher Levin, Professor Raluca Balan, Professor Rafal Kulik, and Professor Yiqiang Zhao for their invaluable contributions as members of my committee and for providing thoughtful and insightful feedback.

I would like to express my gratitude to the University of Ottawa and the Department of Mathematics and Statistics for providing me with various financial opportunities and a great academic environment during my research process.

Finally, I want to convey my heartfelt gratitude and appreciation to my family for their unfaltering support and unwavering belief in me during the highs and lows of my PhD journey. To my beloved parents, especially to my wonderful mother, for her endless love,

countless sacrifices, and continuous blessings. To my beloved husband, I deeply appreciate your unconditional support, patience, sacrifice, profound understanding, and encouragement. To my precious children, who persistently looked into the completion of my PhD. You have been a constant source of inspiration, motivation, and joy to me. To my dear mother-in-law and beloved aunty for their continuous blessings and encouragement.

I am deeply grateful to all those who offered their support and encouragement throughout my PhD journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this introductory chapter, we will provide an overview of Markov chains and their applications in Markov chain Monte Carlo (MCMC) methods. In addition, we will give a preview of our contributions and provide an outline of the chapters in this thesis.

1.1 Motivation

It is commonly understood that when we shuffle a deck of cards, our intention is to achieve a random arrangement of them. We are using a (presumably) stochastic process in which the states represent different orderings of the cards. The objective is to reach a final ordering that is random and independent of the initial ordering. Metropolis et al. [101] first developed a comparable computational technique, known as Markov chain Monte Carlo, for sampling from a set in statistical physics. Later, statisticians like Hastings [62] improved the algorithms before being introduced to the broader statistical community by Gelfand and Smith [52] in 1990 [52]. In order to sample from the distribution π using MCMC algorithms, we run a Markov chain with the stationary distribution π until it reaches near-stationarity. At this point, we can draw samples from the chain.

A Markov chain is a stochastic process $\{X_k\}_{k=0}^{\infty}$ known for its property of being "mem-

oryless" (also referred to as the Markov property). This means that the future trajectory of the chain is independent of its past, given its current state. This process occurs on some agreed-on set called the state space, denoted by $\mathcal{X}$ in this thesis. A common method for representing all of this is by using a probability kernel. This kernel is a mapping function $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ (where $\mathcal{B}(\mathcal{X})$ is a sigma-algebra on $\mathcal{X}$) that satisfies

1. $P(x, \cdot)$ represents a probability measure on $\mathcal{X}$ for every x in $\mathcal{X}$.
2. $P(\cdot, A)$ is a function that is $\mathcal{B}(\mathcal{X})$-measurable function, for all $A \in \mathcal{B}(\mathcal{X})$.

Consider a distribution π with density. A common computational task is estimating the expectation of a function $f : \mathcal{X} \to \mathbb{R}$. This means approximating $\pi(f) := \mathbb{E}_\pi[f(X)] = \int_{\mathcal{X}} f(x)d\pi(x)$. In many instances, this could be quite a challenging endeavor. Numerical integration is also impossible in high dimensions due to the well-known curse of dimensionality.

The classical Monte Carlo approximation to $\pi(f)$ relies on a sampling technique that generates i.i.d. random variables $Z_i \sim \pi$ for $i = 1, \ldots, N$. Given such samples, the random variable $\hat{\pi}(f) := \frac{1}{N} \sum_{i=1}^{N} f(Z_i)$ is an unbiased estimate of $\pi(f)$, with a variance proportional to $1/N$. By the strong law of large numbers, with probability 1:

$$\hat{\pi}(f) \to \pi(f) \quad \text{as } N \to \infty.$$

Furthermore, if $\pi(f^2) < \infty$, then there is a central limit theorem (CLT) for $\hat{\pi}(f)$; that is, $\sqrt{N}(\hat{\pi}(f) - \pi(f)) \xrightarrow{d} N(0, \sigma^2)$ with the usual sample variance of $\hat{\pi}(f)$ is a consistent estimate of σ^2. However, it is not possible to directly sample from many distribution functions. When the form of π is complicated and the dimension is large, it is hard to obtain i.i.d. samples from π, such as rejection sampling [55], which works well in low dimensions and fails when the number of dimensions goes up.

The Markov chain Monte Carlo technique builds an easily-simulated Markov chain transition probabilities $P(x, dy)$ for $x, y \in \mathcal{X}$, such that

$$\int_{x\in\mathcal{X}} \pi(dx)P(x,dy) = \pi(dy).$$

Under certain conditions outlined in [132, Section 3.2], if we run the Markov chain for a long time, then for large n the distribution of X_n will be approximately stationary distribution π. Then we can restart sampling $Z_i = X_{n+(i-1)}, i = 1, 2, \ldots$, and then estimate $\hat{\pi}(f)$.

When using the MCMC method, two important questions come up: (i) how to construct such a Markov chain, and (ii) how long it takes for the Markov chain to converge sufficiently close to the stationary distribution? (Or, what is an appropriate burn-in?). These problems have been the focus of much research over the years, as shown by studies like [151, 145]. It's worth noting that constructing this sort of Markov chain is quite straightforward [132, Section 2]. Thus, the rate of convergence is a significant concern when it comes to MCMC algorithms, and the "mixing time" of the underlying Markov chain plays a crucial role in determining the running time of the algorithms.

When it comes to estimating the mixing times of Markov chains, there is actually no right approach. Different strategies have been proposed, each with its own set of advantages and disadvantages. Some of the most frequently used methods include the spectral gap method, geometric and profile methods, coupling techniques, comparison approach, and decomposition methods. The comparison approach has been widely used to estimate the mixing time of a desired chain in cases where studying the Markov chain directly is difficult, but there is a simpler, related known chain whose spectral gap is much easier to bound. Decomposition methods break the states of the chain into different parts and come up with a mixing time estimation for the entire chain. Other approaches include evolving sets and drift-and-minorization. The expository paper by Rosenthal [135] provides a survey of techniques for studying the convergence rate of Markov chains. For a good sketch of fundamental methods, we also refer the reader to [74, 77], a book by Levin et al. [91], a monograph by Aldous and Fill [7], and a study on geometric walks by Vempala [154].

Many of the most popular approaches to bounding the mixing time of a Markov chain

are closely related to the spectrum of the underlying transition matrix [91], both because they are easy to use and stable under natural changes to the underlying Markov chain [88, 91]. This raises the obvious question [83]: is it possible to find bounds that are both (almost) sharp and recognizably "spectral" or "geometric" in nature? Many studies, such as [64, 88, 66, 65, 3], have investigated similar problems, with the details depending on which notion of mixing must be approximated and what information the "spectral" or "geometric" bounds are allowed to use.

In this thesis, we also explore the application of model fitting and simulation for synthetic data generation (SDG), as obtaining and sharing real data due to privacy concerns might be challenging and has led to a growing interest in synthetic data generation. Synthetic data is not actual data, and researchers can access it more rapidly with far fewer privacy constraints.

There are privacy concerns surrounding traditional de-identification methods[148]. Although there is no single definition of privacy, the basic concept is simple: it represents the level of "protection" against unexpected access to potentially sensitive information about individuals [125].

SDG involves constructing a model based on real data to capture its underlying structure and distribution. The model is then used to generate artificial data that mimics the patterns and characteristics of the original data. Effective models produce synthetic data with statistical properties that closely resemble the real data [48]. Synthetic data, when properly generated, does not have a one-to-one mapping to the original data, hence having the ability to preserve privacy by reducing the risk of identity disclosure [129, 48, 72, 126, 138]. However, overfitting the model to the original data may generate a synthetic record that matches a real observation. Therefore, it is also crucial to prevent the model from overfitting to reduce the chance of identity disclosure.

Existing generative models have limitations in terms of computational efficiency and adaptability to different types of data. Some models do well on small data sets, while others

require a large amount of training data. In addition, the synthesis of longitudinal health data can be quite challenging because patients may have lengthy sequences of events and come from diverse populations. Hence, there is a need to construct a generative model that can efficiently synthesize longitudinal health data, considering different types of variables and structural complexities. To address these limitations, we proposed a generative model that addresses the challenge of synthesizing massive longitudinal health data by instead synthesizing a non-longitudinal and significantly smaller data set [84].

1.2 Structure of the Thesis

The outline of this thesis is as follows:

In Chapter 2, we discuss the convergence rate of Markov chains and methods for bounding the mixing time. We present some new results on the convergence rate of continuous-state Markov chains. Further, we develop the relationship between the notions of distance from stationarity for the continuous-state Markov chains. Some of the results presented in this chapter will be employed in Chapter 3.

In Chapter 3, we study the precision of certain bounds on the mixing time. We investigate the sharpness of the spectral profile bound presented by Goel et al. [56] and Chen et al. [26] on the L^2 mixing time of Markov chains on continuous state spaces. We show that this bound is sharp up to a factor of log log of the starting density. This result extends the findings of Kozma [88], which showed the analogous result for the original spectral profile bound of Goel et al. [56] for Markov chains on finite state spaces. We discuss the application of our finding to the comparison of Markov chains. This chapter contributed to our work [83].

In Chapter 4, we discuss the application of simulation models in synthetic data generation. We provide a generative model for producing synthetic longitudinal data that integrates suitable simulation techniques such as Hamiltonian Monte Carlo (HMC), sequential trees, and copula to sample a latent factor matrix within the framework of the state-of-the-art generalized canonical polyadic (GCP) tensor factorization. We validate the model for generating synthetic longitudinal health data by using the publicly available MIMIC-III data set, which contributed to our work [84].

Finally, Chapter 5 presents a summary of the thesis and outlines plans for future research problems.

Chapter 2

Mixing Time

2.1 Introduction

Sampling from high-dimensional distributions can be quite challenging. A commonly used approach is Markov chain Monte Carlo (MCMC), which involves constructing a Markov chain with the desired distribution as its stationary distribution. MCMC algorithms, such as the Gibbs sampler and the Hamiltonian Monte Carlo (HMC), are state-of-the-art sampling methods from high-dimensional distributions [35, 18]. These methods have applications in various scientific fields, including physics [43], computer vision [164], machine learning [8], and Bayesian statistics [21].

How can we assess the efficiency of Markov chain Monte Carlo algorithms? The time it takes for the Markov chains to reach the stationary state, known as the "mixing time," is essential and has been a subject of great interest in the study of Markov chain Monte Carlo algorithms [35, 13]. To ensure the efficiency of MCMC algorithms, it is crucial for the constructed Markov chain to have rapid mixing [154]. This means that the equilibrium distribution is achieved quickly, starting from any initial state. Analyzing the mixing time for commonly used MCMC algorithms has been a persistent mathematical challenge, as obtaining good samples requires an enormous amount of time for the mixing process. As

a result, it is highly desirable to study techniques for analyzing the rate of convergence of Markov chains. Additionally, in order to discuss the concept of mixing time, we first need a way to measure how far we are from stationarity. The mixing time may be measured using different notions of distances from stationary, such as total variation distance, L^2, or L^∞ distances [54].

Scientific and statistical applications of MCMC techniques tend to be more common in continuous state spaces, while mathematical mixing analysis has mainly focused on discrete state spaces. However, certain efforts, such as [80, 7, 162, 136], focused on developing flexible and universal tools.

In this chapter, we study the mixing time of Markov chains and explore the most widely used techniques that have been proposed to analyze the convergence rate of chains with discrete state spaces. We discuss and extend some methods for estimating mixing time that were used in developing discrete theory for continuous-state chains. We point out that some of the results in this chapter are new and apply to Chapter 3. We present some results on convergence bounds of mixing time of continuous-state Markov chains and extend the relationship between different notions of distance from stationarity in discrete to continuous state spaces.

2.1.1 Related Works

So far, researchers have established various approaches for bounding the mixing time of Markov chains on discrete state spaces. The study of the mixing time of discrete-state chains motivated the study of the convergence rate of Markov chains on continuous state spaces.

These methods include geometric techniques such as the Poincaré inequality. The Poincaré inequality relates the variance of a function on the state space to its Dirichlet form and provides an upper bound on the spectral gap of the Markov chain. In 1991, Diaconis and Stroock [40] provided geometric bounds for the eigenvalues of reversible Markov

chains by developing a geometric quantity for Poincaré inequality on discrete state space, called the canonical paths method. Another geometric approach is the conductance, also known as Cheeger inequality, which links the spectral gap to the bottleneck ratio. Jerrum and Sinclair [75] developed the approach of bounding mixing time using worst-case conductance for discrete-state chains. Lovász and Simonovits [95] then extended this concept to continuous space settings. Besides, Lawler and Sokal [90] provided a general version of Cheeger's inequality that applies to discrete-time Markov chains and continuous-time Markovian jump processes in general state space, both reversible and non-reversible. Yuen [161] provided an upper bound for the Cheeger constant of reversible Markov chains (discrete or continuous time) on $\mathbb{R}^n$. In 2002, Yuen [163] generalized a technique from discrete-time to compute the Cheeger bound on the convergence rates of some homogeneous continuous-time Markov processes. The survey by Vempala [154] and its references provide a detailed review of conductance-based approaches for continuous-state Markov chains.

The spectral method is another technique that provides a way to estimate the mixing time of a Markov chain by analyzing its eigenvalues. For instance, Qin et al. [124], proposed a method to accurately estimate the spectral gap of a trace-class Markov operator in general state space. The method relies on the fact that the second largest eigenvalue, and consequently the spectral gap of these operators, can be bounded by simple functions of the power sums of the eigenvalues. Further, Davis and Hobert [33], in their recent study, utilized the method introduced in [124] to construct an asymptotically valid confidence interval for an upper bound of the spectral gap of the trace-class chains.

There are more sophisticated approaches for bounding mixing time of discrete-state chains that have recently emerged, including those based on the conductance profile by Lovász and Kanna [94], the spectral and conductance profile by Goel et al. [56], and [105] by Montenegro and Tetali. Morris and Peres [106] established a sharp bound on the mixing time of Markov chains on discrete state space based on the conductance profile that applies to non-reversible chains and then extended their findings to the continuous case as well. Goel et al. [56] also provided conductance bounds for non-reversible Markov chains as part

of their study. Chen et al. [26] extended the spectral and conductance profile methods presented by Goel et al. [56] for reversible Markov chains from discrete state to continuous-state chains, particularly demonstrating solid non-asymptotic mixing time's bounds for Hamiltonian Monte Carlo (HMC).

Another well-known method of bounding mixing time is the coupling technique, which was first introduced by Doeblin [41] to analyze the convergence rate of Markov chains. This method has broad application in other areas of probability as well. One may find comprehensive details on the coupling and its application in the lectures of Lindvall [93].

Another strategy is the comparison method, which was introduced by Diaconis and Saloff-Coste [37, 38] to compare the spectral gaps of reversible and finite state Markov chains by relating them to another chain with known properties on the same state space. This technique is particularly useful for understanding the behavior of complex chains by comparing them to simpler and known chains. The comparison method with a special focus on non-reversible chains was proposed by Dyer et al. [45]. The comparison theory for continuous-state Markov chains on $\mathbb{R}^n$ was presented by Yuen [161].

The decomposition method is another approach to analyzing the mixing time, where the idea is to break down a Markov chain into pieces so that the mixing time of restricted Markov chains for each piece is easier to study. This concept has been around for over 15 years, but just recently it has received more attention. The most frequently used decomposition bounds were presented by [97, 99] as overlapping and disjoint decompositions, respectively.

2.2 Notations and Basic Definitions

Markov chains are basic mathematical models that describe the evolution of random phenomena over time. A Markov chain X is a sequence of random variables $\{X_0, X_1, \ldots\}$ that take values in a state space $\mathcal{X}$ and satisfy the Markov property. This property is represented as:

$$\mathbb{P}\{X_k \in A \mid X_0, X_1, \ldots, X_{k-1}\} = \mathbb{P}\{X_k \in A \mid X_{k-1}\},$$

for any measurable set $A \subset \mathcal{X}$, where $\mathbb{P}\{\cdot \mid B\}$ denotes the conditional probability given B.

Throughout this chapter, we consider discrete-time Markov chains. We present the basic facts and some notations for Markov chains in the following subsections.

2.2.1 Discrete State Space

Here are some definitions and notations for situations in which the state space $\mathcal{X}$ is discrete. Consider a Markov chain with the stationary distribution π on a discrete state space $\mathcal{X}$, and let P be its transition probability matrix (or Markov kernel), thus:

- $P(x, y) \geq 0, \quad \forall x, y \in \mathcal{X},$
- $\sum_{y \in \mathcal{X}} P(x, y) = 1, \quad \forall x \in \mathcal{X},$
- $\sum_{x \in \mathcal{X}} \pi(x) P(x, y) = \pi(y), \quad \forall y \in \mathcal{X}.$

A Markov chain with transition probability matrix P is called irreducible if there exists an integer $k > 0$ such that $P^k(x, y) > 0$ for all $x, y \in \mathcal{X}$. A Markov chain with transition matrix P and stationary distribution π satisfying the below detailed balance equation is called reversible.

$$\pi(x)P(x,y) = \pi(y)P(y,x), \quad \forall x, y \in \mathcal{X}.$$

The period of state x is the greatest common divisor of $\left\{k \in \mathbb{N} : P^k(x,x) > 0\right\}$ [91, 112], and an aperiodic Markov chain is defined to be a chain with all states having period 1. Let $0 < \zeta < 1$. A ζ-lazy chain with a transition probability matrix P is defined as the chain remaining in the current state with probability at least ζ in each move, i.e. $P(x,x) \geq \zeta$, $\forall x \in \mathcal{X}$. Thus, a lazy Markov chain is aperiodic.

An irreducible and aperiodic Markov chain is called ergodic, and there is a unique stationary distribution for ergodic Markov chains.

Assuming P is a transition matrix on $\mathcal{X}$, we consider P as an operator on functions $f : \mathcal{X} \to \mathbb{R}$ by defining:

$$(Pf)(x) = \sum_{y \in \mathcal{X}} P(x,y) f(y).$$

The following is an important definition and notion in the context of Markov chains.

Definition 2.2.1. *Let $f, g : \mathcal{X} \to \mathbb{R}$. The Dirichlet form associated with a reversible transition matrix P and a stationary distribution π is defined as follows:*

$$\mathcal{E}_P(f,g) = \frac{1}{2} \sum_{x,y \in \mathcal{X}} [f(x) - f(y)][g(x) - g(y)]\pi(x)P(x,y) = \langle (I - P)f, g \rangle_\pi, \tag{2.2.1}$$

where $\langle \cdot, \cdot \rangle_\pi$ is an inner product with respect to π. In particular, for a reversible Markov chain:

$$\mathcal{E}_P(f) = \mathcal{E}_P(f,f) = \frac{1}{2} \sum_{x,y \in \mathcal{X}} [f(x) - f(y)]^2 \pi(x)P(x,y).$$

In this thesis, the Dirichlet form for a Markov chain may also be denoted as $\mathcal{E}(f,g)$ without a subscript. The subscript just indicates the Markov chain associated with the Dirichlet form.

Consider a Markov chain on a state space $\mathcal{X}$, the geometric properties that influence mixing time are referred to as bottlenecks. The bottleneck ratio of a Markov chain (conductance in the computer science literature) is defined to be:

$$\Phi^{\star} = \min_{Z:\pi(Z)\leq\frac{1}{2}} \frac{Q\left(Z, Z^{c}\right)}{\pi(Z)}, \tag{2.2.2}$$

where $\pi(Z) = \sum_{x\in Z}\pi(x)$ and Q (the edge measure) is:

$$Q(x,y) = \pi(x)P(x,y), \quad Q(B,C) = \sum_{x\in B, y\in C} Q(x,y). \tag{2.2.3}$$

The conductance, a well-known geometric tool for bounding the mixing time of a Markov chain, is the likelihood of moving out of a set after one step.

Finally, in this subsection, we present the notions of distances from stationary in discrete state spaces, which are important in measuring the rate of convergence to stationary.

Given a distribution π on $\mathcal{X}$, the L^p norm of a function $g : \mathcal{X} \to \mathbb{R}$ is defined as:

$$\|g\|_p := \begin{cases} \left[\sum_{x\in\mathcal{X}} |g(x)|^p \pi(x)\right]^{1/p} & 1 \leq p < \infty, \\ \max_{x\in\mathcal{X}} |g(x)| & p = \infty. \end{cases}$$

Definition 2.2.2. *The L^p mixing time $\tau_p(\epsilon)$ for a Markov chain with transition kernel P and stationary distribution π is given by:*

$$\tau_p(\epsilon) = \min\left\{k \in \mathbb{N} : \max_{x\in\mathcal{X}} d_p\left(P^k(x,\cdot),\pi\right) \leq \epsilon\right\},$$

where

$$d_p(P^k(x,\cdot),\pi) = \left\|\frac{P^k(x,\cdot)}{\pi(\cdot)} - 1\right\|_p, \quad 1 \leq p \leq \infty.$$

The total variation distance between any two distributions μ and ν on $\mathcal{X}$ is defined by:

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{B \subseteq \mathcal{X}} |\mu(B) - \nu(B)|,$$

which is related to the L^1 distance as follows: $\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} d_1(\mu, \nu)$.

The total variation distance measures the maximum error caused while approximating μ by ν to forecast the probability of an event. Note that the maximum value that the total variation distance between two measures can have is 1.

Definition 2.2.3. *For a Markov chain with transition kernel P and stationary distribution π, the mixing time at level $\epsilon \in (0,1)$ is given by:*

$$\tau_{\mathrm{mix}}(\epsilon) = \min\{k \in \mathbb{N} : \max_{x \in \mathcal{X}} \left\| P^k(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \epsilon\},$$

and $\tau_{\mathrm{mix}} := \tau_{\mathrm{mix}}(1/4)$ is called the mixing time of a Markov chain.

In the following subsection, we present some notations and definitions for situations where the state space $\mathcal{X}$ is continuous.

2.2.2 Continuous State Space

We consider a continuous-state Markov chain defined on a measurable state space $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with a transition probability kernel $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ and with stationary probability measure π. The probability of entering any measurable set $A \in \mathcal{B}(\mathcal{X})$ from any state $x \in \mathcal{X}$ is denoted by $P(x, A)$. Hence we have:

- $\int_{y \in \mathcal{X}} P(x, dy) = 1, \quad \forall x \in \mathcal{X},$
- $\int_{\mathcal{X}} P^k(x, dy) = 1, \quad \forall k \in \mathbb{N}$ and $x \in \mathcal{X},$

where P^k refers to the k-step transition kernel.

We denote by $\mathcal{B}(\mathcal{X})$ a σ-algebra on $\mathcal{X}$. In the case where $\mathcal{X}$ is a metric space, this will be the usual Borel σ-algebra.

For any set $A \in \mathcal{B}(\mathcal{X})$ and any state $x \in \mathcal{X}$, P^k is defined recursively as $P^{k+1}(x, A) = \int_{z \in \mathcal{X}} P^k(x, dz)P(z, A)$. Let δ_x denotes the Dirac-delta distribution at x. For simplicity, we may write $\delta_x P^k(A)$ as an abbreviation for $P^k(x, A)$.

A probability measure π on $\mathcal{X}$ is called stationary for the transition kernel P if

$$\int_{x \in \mathcal{X}} P(x, A)\pi(dx) = \pi(A), \quad \forall A \in \mathcal{B}(\mathcal{X}).$$

We suppose that the invariant probability measure π has density with respect to a reference measure.

P is reversible with respect to π if:

$$\int_{x \in A} \int_{y \in B} \pi(dx)P(x, dy) = \int_{x \in A} \int_{y \in B} \pi(dy)P(y, dx), \quad \forall A, B \in \mathcal{B}(\mathcal{X}).$$

It is clear that if P is reversible with respect to π, then π is a stationary distribution of P [163]. Moreover, reversible P is self-adjoint with respect to the inner product induced by π.

A Markov chain is called irreducible if for each $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$ such that $\pi(A) > 0$, there is an integer $k > 0$ so that $P^k(x, A) > 0$.

To extend some methods of bounding the convergence rate of Markov chains on discrete state space to the continuous case, we may need to make a better assumption, for which we provide the following definition introduced by Vempala [154]. We say that a Markov chain with state space $\mathcal{X}$ and stationary distribution π has a β-warm start if its initial distribution μ is somewhat spread out, as defined below:

Definition 2.2.4 (Warm start)**.** *A distribution μ is a β-warm start for a Markov chain with stationary distribution π if:*

$$\sup_{A \in \mathcal{B}(\mathcal{X})} \frac{\mu(A)}{\pi(A)} \leq \beta,$$

where $\mathcal{B}(\mathcal{X})$ denotes the Borel σ-algebra of the state space $\mathcal{X}$.

Note that the initial distribution μ is often referred to as β-warm start or a warm start with constant β. As the warmness parameter β decreases, the initial distribution approaches the stationary distribution, making the sampling process easier.

To understand the concept of mixing time, we first need a way to measure how far we are from stationarity. Given probability measures on a state space $\mathcal{X}$; μ, the distribution of the chain after a given number of steps, and π the stationary distribution of the chain, there are different notions of distance between μ and π that can be used. The most straightforward notion is the total variation distance, whose generic form was introduced in the previous section.

The following definitions are also required to measure the mixing time of a Markov chain:

Definition 2.2.5 (L^p-distance)**.** *The L^p-distance between any two distributions ν and ν' is defined as follows:*

$$d_p(\nu, \nu') = \left(\int_{\mathcal{X}} \left| \frac{d\nu}{d\nu'}(x) - 1 \right|^p \nu'(dx) \right)^{\frac{1}{p}},$$

for $1 \leq p < \infty$, and

$$d_\infty(\nu, \nu') = \operatorname{ess\,sup}_x \left| \frac{d\nu}{d\nu'}(x) - 1 \right|,$$

where $\frac{d\nu}{d\nu'}$ is the Radon-Nikodym derivative.

Let μP indicate the distribution of the next state of a Markov chain with transition kernel P given a distribution μ on the current state of the chain. For every $A \in \mathcal{B}(\mathcal{X})$,

we have $\mu P^k(A) = \int_{\mathcal{X}} P^k(x, A)\mu(dx)$. We denote its density with respect to the stationary measure π by $h_{\mu,k}(x) := \frac{d(\mu P^k)}{d\pi}(x)$ and assume it exists.

Definition 2.2.6 (L^p mixing time)**.** *The L^p mixing time of a Markov chain on continuous state space with transition kernel P and the stationary distribution π from an initial distribution μ is defined as:*

$$\tau_p(\epsilon; \mu, P) = \min\left\{k \in \mathbb{N} : d_p(\mu P^k, \pi) \leq \epsilon\right\},$$

where $\epsilon > 0$ is an error tolerance.

The L^p mixing time from an initial state x is defined as follows:

$$\tau_p(\epsilon; P) = \sup_x \tau_p(\epsilon; \delta_x, P) = \min\left\{k \in \mathbb{N} \mid \sup_{x\in\mathcal{X}} d_p(\delta_x P^k, \pi) \leq \epsilon\right\}. \tag{2.2.4}$$

Assumption 2.2.7 ([26], Smooth chain)**.** *A Markov chain satisfies the smooth chain assumption if its transition probability kernel $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0, 1]$ can be represented in the form:*

$$P(x, dy) = \tilde{P}(x, dy) + \alpha_x \delta_x(dy), \quad \forall x, y \in \mathcal{X}, \tag{2.2.5}$$

where δ_x is the Dirac-delta function at x and α_x is the probability that the chain stays at the current state x after one step. $\tilde{P}$ is a transition kernel whose distribution has density with respect to π.

We then present the definition of ζ-lazy chain:

Definition 2.2.8 (ζ-lazy chain)**.** *A Markov chain with transition kernel P is called ζ-lazy for a Markov chain with transition kernel $\tilde{P}$ if it can be written in the form of:*

$$P(x, A) = (1 - \alpha_x)\tilde{P}(x, A) + \alpha_x \delta_x(A), \quad \forall x \in \mathcal{X} \text{ and } A \in \mathcal{B}(\mathcal{X}), \tag{2.2.6}$$

for some $\zeta \le \alpha_x \le \sup_x \alpha_x = \alpha < 1$, *where* $\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$.

We say a Markov chain is "exactly ζ-lazy" if, during each step, the chain stays in the current state with exactly ζ probability, i.e. $\alpha_x = \zeta$, $\forall x \in \mathcal{X}$ in Definition 2.2.8. In practice, a lazy chain may not be used due to the slow convergence rate caused by the lazy steps. However, it is a useful assumption for theoretical analysis of the mixing rate up to constant factors.

Many notations of discrete-state Markov chains are easily generalized to the continuous setting, with integration replacing summations. Therefore, the definition of the Dirichlet form for continuous-state Markov chains is as follows:

Definition 2.2.9 (Dirichlet form)**.** *For a π-reversible transition kernel P, define the Dirichlet form on $L^2(\pi)$ by*

$$\mathcal{E}(f,f) = \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} (f(x)-f(y))^2\,\pi(dx)P(x,dy) = \langle (I-P)f, f\rangle_\pi,$$

where $\langle\cdot,\cdot\rangle_\pi$ is the inner product on $L^2(\pi)$.

The Dirichlet form for the Markov chain P can also be represented with a subscript notation as $\mathcal{E}_P(f,f)$.

Given a Markov chain with transition probability $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$, its stationary flow $\phi : \mathcal{B}(\mathcal{X}) \to [0,1]$ is defined as:

$$\phi(S) = \int_{x\in S} P\left(x, S^c\right)\pi(dx), \qquad \text{for any } S \in \mathcal{B}(\mathcal{X}). \tag{2.2.7}$$

We can also define the stationary flow (the edge measure) in the continuous case as following:

$$Q(S, S^c) = \int_{x\in S}\int_{y\in S^c} \pi(dx)P(x,dy).$$

Some of the notations presented in this subsection will be recalled and used in Chapter 3. Throughout this chapter, unless otherwise mentioned, we consider Markov chains on general state spaces. Furthermore, it is assumed that all chains under discussion are irreducible.

2.3 Relationship Between Notions of Distance from Stationarity

There are different ways to measure the distance between two probability measures μ and ν on a state space $\mathcal{X}$. These measures can be useful when comparing the distribution of a chain after a certain number of steps with its stationary distribution. The relationship between various notions of distance between a Markov chain distribution and its stationary distribution in discrete state space has been extensively studied, such as the study of Gibbs and Su [54]. While some of these results can be easily extended to the continuous case, others require further investigation and calculation.

The most intuitive and straightforward notion is total variation distance, which is defined in Definition 2.2.3. Here are some fundamental properties of the total variation distance. The following is reasonably widely known such as Proposition 3 of [132].

Lemma 2.3.1. *Assume that μ and ν are probability measures on $\mathcal{X}$, having density with respect to a σ-finite reference measure. We then have:*

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2} d_1(\mu, \nu).$$

Proof. From Definition 2.2.3, we have:

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{B \in \mathcal{B}(\mathcal{X})} |\mu(B) - \nu(B)|.$$

Let $B \in \mathcal{B}(\mathcal{X})$ be any measurable set and $A = \{x \in \mathcal{X} : \frac{d\mu(x)}{d\nu(x)} \geq 1\}$. Then, by imitating the proof of [12, Lemma 1.1]:

$$\mu(B) - \nu(B) = [\mu(B \cap A) - \nu(B \cap A)] + [\mu(B \cap A^c) - \nu(B \cap A^c)].$$

By definition of A, we get $A^c = \{x \in \mathcal{X} : \frac{d\mu(x)}{d\nu(x)} < 1\}$, which implies

$$\mu(B \cap A^c) - \nu(B \cap A^c) \leq 0.$$

Hence

$$\begin{aligned}\mu(B) - \nu(B) &\leq (\mu - \nu)(B \cap A) \leq (\mu - \nu)(A) \\ &= \mu(A) - \nu(A),\end{aligned}$$

since $\mu - \nu$ is non-negative on A. This proves that

$$\sup_{B \in \mathcal{B}(\mathcal{X})} |\mu(B) - \nu(B)| = \mu(A) - \nu(A).$$

We observe that $\mu(A) + \mu\left(A^c\right) = 1 = \nu(A) + \nu\left(A^c\right)$, thus

$$\mu(A) - \nu(A) = \nu\left(A^c\right) - \mu\left(A^c\right).$$

We then deduce

$$\begin{aligned}\|\mu - \nu\|_{\mathrm{TV}} = \sup_{B \in \mathcal{B}(\mathcal{X})} |\mu(B) - \nu(B)| &= \frac{1}{2}\left([\mu(A) - \nu(A)] + \left[\nu\left(A^c\right) - \mu\left(A^c\right)\right]\right) \\ &= \frac{1}{2}\int_{x \in A} |d\mu(x) - d\nu(x)| + \frac{1}{2}\int_{x \in A^c} |d\mu(x) - d\nu(x)| \\ &= \frac{1}{2}\int_{x \in \mathcal{X}} |d\mu(x) - d\nu(x)| \\ &= \frac{1}{2} d_1(\mu, \nu),\end{aligned}$$

and this completes the proof. ∎

Therefore, $\frac{1}{2}d_1(\mu, \nu)$ is the total variation distance between two distributions μ and ν; we may also use $d_{TV}(\mu, \nu)$ or $\|\mu - \nu\|_{TV}$ to denote this distance.

The following lemma is an immediate result of a remark in [130], which is helpful when using the L^2 distance.

Lemma 2.3.2. *Consider a Markov chain with transition probability kernel P and the stationary distribution π on state space $\mathcal{X}$. Let $h_{\delta_x,k}(y) := \frac{d(\delta_x P^k)}{d\pi}(y)$ and suppose it exists. Then*

$$d_2^2(\delta_x P^k, \pi) = \|\delta_x P^k\|_{L^2(\pi)}^2 - 1, \quad \forall x \in \mathcal{X} \text{ and } k \in \mathbb{N}.$$

Proof.

It is obvious that $\mathbb{E}_\pi [h_{\delta_x,k}] = 1$, thus

$$\begin{aligned}
d_2^2(\delta_x P^k, \pi) &= \int_{\mathcal{X}} \left| \frac{d(\delta_x P^k)}{d\pi}(y) - 1 \right|^2 \pi(dy) \\
&= \int_{\mathcal{X}} |h_{\delta_x,k}(y) - 1|^2 \, \pi(dy) \\
&= \mathbb{E}_\pi [h_{\delta_x,k} - 1]^2 \\
&= \mathrm{Var}_\pi [h_{\delta_x,k}] \\
&= \mathbb{E}_\pi [h_{\delta_x,k}]^2 - \mathbb{E}_\pi^2 [h_{\delta_x,k}] \\
&= \int_{\mathcal{X}} h_{\delta_x,k}^2(y) \pi(dy) - 1 \\
&= \int_{\mathcal{X}} \left(\frac{d(\delta_x P^k)}{d\pi}(y) \right)^2 \pi(dy) - 1 \\
&= \|\delta_x P^k\|_{L^2(\pi)}^2 - 1.
\end{aligned}$$

∎

The L^2 distance can be bounded by the total variation distance if the two distributions have a bounded ratio, as stated in the following lemma derived from the study of Wu et al. [158].

Lemma 2.3.3. *Consider a reversible, irreducible, and half-lazy continuous-state Markov chain P with the stationary distribution π. Let μ be a β-warm start. Then, the L^2 distance of μP^k with respect to π is bounded by the total variation distance via*

$$d_2(\mu P^k, \pi) \leq \sqrt{2\beta \, \mathrm{d_{TV}}(\mu P^k, \pi)}. \tag{2.3.1}$$

The proof is omitted because a similar proof will be used for the next lemma. By using Lemma 2.3.3, we present a lemma that can be utilized in the application of the results in Chapter 3.

Lemma 2.3.4. *Consider a reversible, irreducible, and half-lazy continuous-state Markov chain P with the stationary distribution π. Denote its non-lazy version as $\tilde{P}$ in (2.2.6). Assume that $\delta_x\tilde{P}$ is a β-warm start for every $x \in \mathcal{X}$. Then, the L^2 distance of $\delta_x\tilde{P}P^k$ with respect to π is bounded by the total variation distance via*

$$\sup_x d_2(\delta_x\tilde{P}P^k, \pi) \leq \sqrt{2\beta \sup_x \mathrm{d_{TV}}(\delta_x\tilde{P}P^k, \pi)}. \tag{2.3.2}$$

Proof.

This proof is derived from a part of the proof of Theorem 2 in the work of Wu et al. [158]. $\delta_x\tilde{P}$ is β-warm start for every $x \in \mathcal{X}$. It is straightforward to check that $\delta_x\tilde{P}P^k$ is also β-warm ($\sup_{A\in\mathcal{B}(\mathcal{X})} \frac{\delta_x\tilde{P}P^k}{\pi}(A) \leq \beta$ for every $x \in \mathcal{X}$). Therefore, we obtain

$$\begin{aligned} d_2^2(\delta_x\tilde{P}P^k, \pi) &= \int_{\mathcal{X}} \left(\frac{d(\delta_x\tilde{P}P^k)}{d\pi}(y) - 1 \right)^2 \pi(dy) \\ &\leq (\beta - 1) \int_{\mathcal{X}} \left| \frac{d(\delta_x\tilde{P}P^k)}{d\pi}(y) - 1 \right| \pi(dy) \\ &\leq \beta \int_{\mathcal{X}} \left| \frac{d(\delta_x\tilde{P}P^k)}{d\pi}(y) - 1 \right| \pi(dy) \\ &= 2\beta\, d_{\mathrm{TV}}(\delta_x\tilde{P}P^k, \pi). \end{aligned}$$

∎

Thus, one should note that

$$\sup_x \tau_2(\sqrt{2\beta\epsilon}; \delta_x\tilde{P}, P) \leq \frac{1}{2} \sup_x \tau_1(\epsilon; \delta_x\tilde{P}, P). \tag{2.3.3}$$

We define the ratio of the density of the Markov chain P at the k-th iteration with respect to the density of the stationary distribution π:

$$h_{\delta_x,k}(y) := \frac{d\delta_x P^k}{d\pi}(y).$$

We then present the relationship between L^2 and L^∞ distances for reversible chains as follows:

Proposition 2.3.5. *For a reversible Markov chain with transition probability kernel P and the stationary distribution π:*

$$\sup_{x,y} h_{\delta_x,2k}(y) - 1 \leq \sup_{x\in\mathcal{X}} d_2(\delta_x P^k, \pi) \sup_{y\in\mathcal{X}} d_2(\delta_y P^k, \pi).$$

Proof. This proof is inspired by inequality (2.2) in [56]. Note that $h_{\delta_x,k}(y) = h_{\delta_y,k}(x)$ when P is reversible with respect to π. Using reversibility, we then obtain

$$\begin{aligned}
|h_{\delta_x,2k}(y) - 1| &= \left| \frac{\delta_x P^{2k}(dy) - \pi(dy)}{\pi(dy)} \right| \\
&= \left| \frac{\int_{z\in\mathcal{X}} \left(\delta_x P^k(dz) - \pi(dz)\right)\left(\delta_z P^k(dy) - \pi(dy)\right)}{\pi(dy)} \right| \\
&= \left| \int_{z\in\mathcal{X}} \left[h_{\delta_x,k}(z) - 1\right]\left[h_{\delta_y,k}(z) - 1\right] \pi(dz) \right| \\
&\overset{(i)}{\leq} \left\|h_{\delta_x,k} - 1\right\|_2 \left\|h_{\delta_y,k} - 1\right\|_2 \\
&= d_2(\delta_x P^k, \pi) d_2(\delta_y P^k, \pi),
\end{aligned}$$

where inequality (i) follows from Cauchy-Schwarz. The proof is completed by taking the supremum over x and y. ∎

Proposition 2.3.6. *For a reversible Markov chain with transition probability kernel P and the stationary distribution π:*

$$\sup_{x\in\mathcal{X}} d_2^2(\delta_x P^k, \pi) = \sup_{x\in\mathcal{X}} d_\infty(\delta_x P^{2k}, \pi).$$

Proof. This proof is inspired by the proof of Proposition 4.15 in [91].

Note that $h_{\delta_x,k}(y) = h_{\delta_y,k}(x)$ when P is reversible with respect to π. By Definition 2.2.6, we have: $d_2^2(\delta_x P^k, \pi) = \int_{\mathcal{X}} |h_{\delta_x,k}(y) - 1|^2 \, \pi(dy)$.

By the properties of the inner product, we also have

$$\int_{\mathcal{X}} |h_{\delta_x,k}(y) - 1|^2 \, \pi(dy) = \langle h_{\delta_x,k}(\cdot) - 1, h_{\delta_x,k}(\cdot) - 1 \rangle_\pi \,,$$

besides

$$\langle h_{\delta_x,k}(\cdot), 1 \rangle_\pi = \int_{y \in \mathcal{X}} h_{\delta_x,k}(y) \pi(dy) = 1. \tag{2.3.4}$$

Using reversibility, we then obtain

$$\begin{aligned} h_{\delta_x,2k}(y) := \frac{d\delta_x P^{2k}}{d\pi}(y) &= \frac{P^{2k}(x, dy)}{\pi(dy)} \\ &= \int_{z \in \mathcal{X}} \frac{P^k(x, dz)}{\pi(dz)} \frac{P^k(z, dy)}{\pi(dy)} \pi(dz) \\ &= \left\langle h_{\delta_x,k}(\cdot), h_{\delta_y,k}(\cdot) \right\rangle_\pi . \end{aligned}$$

Using (2.3.4) and properties of the inner product, we have

$$\begin{aligned} \left\langle h_{\delta_x,k}(\cdot) - 1, h_{\delta_y,k}(\cdot) - 1 \right\rangle_\pi &= \left\langle h_{\delta_x,k}(\cdot), h_{\delta_y,k}(\cdot) \right\rangle_\pi - \left\langle 1, h_{\delta_y,k}(\cdot) \right\rangle_\pi - \langle h_{\delta_x,k}(\cdot), 1 \rangle_\pi + 1 \\ &= h_{\delta_x,2k}(y) - 1, \end{aligned} \tag{2.3.5}$$

let $x = y$ and take the supremum over x yields

$$\sup_{x \in \mathcal{X}} d_2^2(\delta_x P^k, \pi) = \sup_{x \in \mathcal{X}} d_\infty(\delta_x P^{2k}, \pi). \tag{2.3.6}$$

Hence, one should note that

$$\tau_2(\sqrt{\epsilon}; P) = \frac{1}{2} \tau_\infty(2\epsilon; P).$$

∎

2.4 Techniques for Bounding Convergence Rates

In this section, we present a summary of techniques for bounding the mixing time of Markov chains. Some of the results in this section are applied in the next chapter. We also provide new results and extend some bounds and techniques from discrete-state Markov chains to continuous spaces.

2.4.1 Spectral Methods

A function f on $\mathcal{X}$ is an eigenfunction with corresponding eigenvalue λ for a transition probability kernel P, if $Pf = \lambda f$. If P is not a reversible chain, the eigenfunctions and eigenvalues may not be real. We start by gathering some basic information about the eigenvalues of transition kernels. We then discuss how eigenvalues and mixing time are related in both discrete and continuous state spaces.

Assume P is the transition matrix of a reversible Markov chain on discrete state space $\mathcal{X}$, then [91]:

1. If λ is an eigenvalue, then $|\lambda| \leq 1$.
2. If P is irreducible, then the eigenspace corresponding to the eigenvalue $\lambda = 1$ is a one-dimensional space and can be generated by the column vector $(1, 1, \ldots, 1)^T$.
3. If P is irreducible and aperiodic then -1 is not an eigenvalue.

Let $|\mathcal{X}| = n$, for a reversible Markov chain P with respect to π, there is a set of eigenfunctions $f_1, \ldots, f_n$ that are orthonormal for $\langle \cdot, \cdot \rangle_\pi$, where f_1 is the constant vector $(1, \ldots, 1)^T$. Further, all eigenvalues are real and can be arranged in descending order as follows:

$$1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq -1.$$

The following lemma shows how the eigenvalues of P relate to the distance from the stationary distribution at time k in discrete state space. This lemma has been stated in various references, such as [39, Lemma 2.1], [91, Lemma 12.2], [146, Theorem 3.1], and [12, Theorem 2.1]. The proof presented here is derived from [81].

Lemma 2.4.1. *Consider a Markov chain with transition matrix P that is reversible with respect to π on the discrete state space $\mathcal{X}$. Further, assume that $\{f_j\}_{j=1}^n$ and $\{\lambda_j\}_{j=1}^n$ are orthonormal basis of real-valued eigenfunctions and their corresponding eigenvalues, then*

$$\frac{P^k(x,y)}{\pi(y)} = \sum_{j=1}^{n} f_j(x) f_j(y) \lambda_j^k, \quad \forall k \in \mathbb{N}, \quad \forall x, y \in \mathcal{X}.$$

Proof. For any $y \in \mathcal{X}$, let δ_y be the delta function as follows:

$$\delta_y(x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x. \end{cases}$$

We can represent δ_y via basis decomposition as:

$$\delta_y = \sum_{j=1}^{n} \langle \delta_y, f_j \rangle_\pi f_j = \sum_{j=1}^{n} f_j(y) \pi(y) f_j. \tag{2.4.1}$$

Since $P^k(x,y) = \left(P^k \delta_y\right)(x)$, applying Equation (2.4.1)

$$\begin{aligned} P^k(x,y) = \left(P^k \left(\sum_{j=1}^{n} f_j(y)\pi(y) f_j\right)\right)(x) &= \left(\sum_{j=1}^{n} f_j(y)\pi(y) P^k f_j\right)(x) \\ &\overset{(i)}{=} \left(\sum_{j=1}^{n} f_j(y)\pi(y)\lambda_j^k f_j\right)(x) \\ &= \sum_{j=1}^{n} f_j(y)\pi(y)\lambda_j^k f_j(x). \end{aligned}$$

Step (i) uses the eigenvalue decomposition $P^k f_j = \lambda_j^k f_j$, $\forall j \geq 1$. ∎

The eigenvalue with the second largest absolute value plays an important role in finding a bound on the mixing time.

Definition 2.4.2. *Let P be a reversible Markov chain on finite state space. Define $\lambda_* = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } P, \neq 1\}$. The absolute spectral gap is denoted as $\gamma^* = 1 - \lambda^*$. Then the relaxation time τ_{rel} is defined by*

$$\tau_{\text{rel}} = \frac{1}{\gamma^*}.$$

The difference $\gamma = 1 - \lambda_2$ is called the spectral gap of a reversible Markov chain P, where λ_2 is the second supremum eigenvalue of P.

Assume that the Markov chain P is irreducible and reversible with respect to π on discrete state space. The relaxation time τ_{rel} and the mixing time τ_{mix} are then related in the following way [91, Theorem 12.4]:

$$\tau_{\text{mix}} \leq \tau_{\text{rel}} \log\left(\frac{4}{\pi_{\min}}\right),$$

where $\pi_{\min} = \min_x \pi(x)$, and under irreducibility assumption $\pi_{\min} > 0$.

In addition, for discrete-state ergodic and reversible chains [91, Theorem 12.5]:

$$(\tau_{\text{rel}} - 1)\log(2) \leq \tau_{\text{mix}}.$$

In the following, we address the study of the preceding discussion and results in continuous state space as presented by Diaconis et al. [36].

Assumption 2.4.3. *Assume that P is a self-adjoint operator on $L^2(\pi)$ and that $L^2(\pi)$ has an orthonormal basis of real eigenfunctions $\{f_i\}_{i\geq 1}$ with real eigenvalues $\{\lambda_i\}_{i\geq 1}$ satisfy $f_1 \equiv 1$, $\lambda_1 = 1, \lambda_i \geq 0$, $\lambda_i \downarrow 0$ so that*

$$\int_{\mathcal{X}} \frac{d(\delta_x P)}{d\pi}(y) f_i(y)\, \pi(dy) = \lambda_i f_i(x), \quad \forall x \in \mathcal{X}.$$

Assume further that P is a Hilbert-Schmidt operator on $L^2(\pi)$ (i.e., $\sum |\lambda_i|^2 < \infty$).

Lemma 2.4.4. *Consider a Markov chain with transition probability kernel P and the stationary distribution π. Suppose that $P(x,\cdot)$ has density with respect to π, and π has density with respect to a σ-finite reference measure. Assume further that P satisfies Assumption 2.4.3, we then have*

$$d_2^2(\delta_x P^k, \pi) = \sum_{i>1} \lambda_i^{2k} f_i^2(x).$$

Proof.

Define $h_{\delta_x,\ell}(y) := \frac{d(\delta_x P^\ell)}{d\pi}(y)$. By Assumption 2.4.3 and [36, Page 156], we have

$$h_{\delta_x,\ell}(y) = \sum_{i\geq 1} \lambda_i^\ell f_i(x) f_i(y), \quad \forall \ell \in \mathbb{N}, \quad \forall x, y \in \mathcal{X}.$$

Taking $\ell = 2k, k \in \mathbb{N}$ and $x = y$, we obtain

$$\begin{aligned} h_{\delta_x,2k}(x) &= \sum_{i\geq 1} \lambda_i^{2k} f_i(x) f_i(x) \\ &= \sum_{i\geq 1} \lambda_i^{2k} f_i^2(x) \\ &= \sum_{i>1} \lambda_i^{2k} f_i^2(x) + \lambda_1^{2k} f_1^2(x) \\ &= \sum_{i>1} \lambda_i^{2k} f_i^2(x) + 1, \end{aligned}$$

where the last equation follows from $f_1 \equiv 1$ and $\lambda_1 = 1$. By Assumption 2.4.3, P is a self-adjoint operator on $L^2(\pi)$, and so it is reversible with respect to π. Hence

$$d_2^2(\delta_x P^k, \pi) = h_{\delta_x,2k}(x) - 1 = \sum_{i>1} \lambda_i^{2k} f_i^2(x).$$

∎

Qin et al. [124] proposed a method for estimating the spectral gap of Markov chains with non-negative and trace-class operators. The proposed method is based on the fact that the second largest eigenvalue of the Markov operator can be bounded above and below by simple functions of the power sums of the eigenvalues.

A non-negative and trace-class operator P has at most countably many eigenvalues, which are contained in $[0, 1]$. Consider the strictly positive eigenvalues of P in decreasing order, denoted as $\lambda_1, \lambda_2, \ldots$ taking into account multiplicity. Then $\lambda_1 = 1$, and λ_2 is what we previously referred to as the chain's "second largest eigenvalue." The following is an immediate corollary of Theorem 1 in [124].

Corollary 2.4.5. *Suppose P is a Hilbert-Schmidt and reversible Markov chain with respect to stationary distribution π on state space $\mathcal{X}$. Let π have a probability density function with respect to a σ-finite measure μ (i.e, π is absolutely continuous with respect to μ), then*

$$\sum_{i \geq 1} \lambda_i^{2k} = \int_{\mathcal{X}} p^{2k}(x, x)\mu(dx), \quad \forall k \in \mathbb{N},$$

where $p(x, \cdot)$ is the Markov transition density of P.

Proof. P is reversible with respect to π, it follows that P^2 is a data augmentation (DA) operator [124, Page 1797]. P is also Hilbert-Schmidt, and we get through a simple spectral decomposition (refer to [63, Corollary 2.1]) that P^2 is trace-class, and its eigenvalues are exactly the squares of the eigenvalues of P. By applying [124, Theorem 2], we then have

$$\int_{\mathcal{X}} p^2(x, x)\mu(dx) < \infty,$$

and hence

$$\sum_{i \geq 1} \lambda_i^{2k} = \int_{\mathcal{X}} p^{2k}(x, x)\mu(dx), \quad \forall k \in \mathbb{N}.$$

∎

2.4.2 Geometric Methods

The spectral gap of a reversible Markov chain P with respect to the stationary distribution π, based on min-max principal, satisfies:

$$\gamma = \inf_{\substack{f \in L^2(\pi): \\ \mathrm{Var}_\pi(f) \neq 0 \\ \mathbb{E}_\pi(f) = 0}} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)}, \tag{2.4.2}$$

where $\mathrm{Var}_\pi(f) = \langle f, f \rangle_\pi$, and $\mathbb{E}_\pi(f) = \langle f, \mathbf{1} \rangle_\pi$.

Computing the spectral gap explicitly is challenging. However, we can estimate the spectral gap using several methods, such as canonical path, which was introduced in [76, 37] and can lead to Poincaré inequality. The Poincaré inequality, as defined in Definition 2.4.6, has been proven to be a significantly effective tool for bounding mixing time.

Definition 2.4.6. *An irreducible Markov chain with transition probability kernel P and the stationary distribution π, satisfies Poincaré inequality with constant α if:*

$$\mathrm{Var}_\pi(f) \leq \alpha \mathcal{E}_P(f,f), \quad \forall f : \mathcal{X} \to \mathbb{R}.$$

According to the variational characterization of the spectral gap and Definition 2.4.6; $\gamma \geq 1/\alpha$.

For any distinct pair of $x, y \in \mathcal{X}$, we select η_{xy} as a path from x to y (i.e. a collection of states $x_0 = x, x_1, \ldots x_k = y$ such that $Q(x_i, x_{i+1}) > 0$ for all $i = 0, \ldots k-1$) that may contain repeated vertices, but a given edge appears once in a given path. The existence of such paths is guaranteed by irreducibility. We denote $Q(e) = Q(a,b)$ when e is the edge $\{a,b\}$ (i.e. $Q(a,b) > 0$). Diaconis and Stroock [40, Proposition 1] proposed a specific α for Poincaré inequality on discrete state space $\mathcal{X}$ using the canonical paths method as follows:

$$\alpha = \max_{e} \left\{ \frac{1}{Q(e)} \sum_{\substack{x,y \\ \eta_{xy} \ni e}} |\eta_{xy}| \, \pi(x)\pi(y) \right\},$$

where $|\eta_{xy}|$ is the number of edges in the path η_{xy}. The maximum is over all $e = \{a, b\} \in \mathcal{X} \times \mathcal{X}$ such that $Q(a, b) > 0$. It should be noted that the edge e appears no more than once in a given path. We refer to [40] for more details.

On the other hand, the method of canonical paths introduces the idea that if there are many paths between pairs of vertices passing through the same edge, the bound on the mixing of the Markov chain gets worse. Cheeger's inequality provides a more formal and rigorous mathematical approach to this idea introduced by canonical paths. The idea of linking the spectral gap and conductance as the Cheeger inequality was presented as the following theorem in [75, 90]. However, twenty years ago, Cheeger [24] established the direct analogue of the inequality on a compact Riemannian manifold, which served as the inspiration for all subsequent studies.

Theorem 2.4.7 ([141, 90]). *Let P be a reversible transition matrix and γ be its spectral gap, then*

$$\frac{\Phi^{\star 2}}{2} \leq \gamma \leq 2\Phi^{\star},$$

where $\Phi^{\star}$ represents the bottleneck ratio defined in (2.2.2).

Lawler and Sokal [90] proved a general version of Cheeger's inequality for reversible and nonreversible Markov chains in general state space. Define the Cheeger's constant as:

$$k \equiv \inf_{\substack{B \in \mathcal{B}(\mathcal{X}) \\ 0 < \pi(B) < 1}} k(B),$$

with

$$k(B) \equiv \frac{\int_B P\left(x, B^{\mathrm{c}}\right) \pi(\mathrm{d}x)}{\pi(B)\pi\left(B^{\mathrm{c}}\right)}.$$

The rate of probability flow from a set B to its complement B^c is given by $k(B)$ and then normalizing it by the stationary probabilities of B and B^c.

Theorem 2.4.8. *For reversible Markov chain P with respect to the stationary measure π on a bounded subset of $\mathbb{R}^n$, the Cheeger's inequality is:*

$$k^2/8 \leq 1 - \lambda_2 \leq k,$$

where λ_2 is the second largest eigenvalue of P.

Yuen [161, 163] proposed Cheeger's k-constant for reversible Markov chains on $\mathbb{R}^n$ associated with a set of paths. It is supposed that the transition kernel satisfy the smooth chain Assumption 2.2.7, i.e. the transition probability kernel is of the form of $P(x,\ \mathrm{d}y) = \alpha(x)\delta_x(\ \mathrm{d}y) + p(x,y)\mathrm{d}y$. Furthermore, it is assumed that the invariant distribution π has density $q(y)$ with respect to the Lebesgue measure. It is also required the existence of a collection of paths, $\eta = \{\eta_{xy}\}$, satisfying certain regularity conditions as presented in Assumption 2.4.9.

Let $S = \{x \in \mathbb{R}^n : q(x) > 0\}$. For any $x, y \in S$, choose a path $\eta_{xy} : \{0, \ldots, b_{xy}\} \to S$ such that $\eta_{xy}(0) = x, \eta_{xy}(1), \ldots, \eta_{xy}\left(b_{xy}\right) = y$ and $p(\eta_{xy}(i-1), \eta_{xy}(i)) > 0$ for all $i = 1, \ldots, b_{xy}$. Denote the set of such paths as $\eta = \{\eta_{xy}\}$. Define $Q(u,v) = p(u,v)q(u)$, and $\|\eta_{xy}\|_{\varepsilon} = \sum_{(u,v)\in\eta_{xy}} Q(u,v)^{-2\varepsilon}$ for $\eta_{xy} \in \eta$ and $\varepsilon \in \mathbb{R}$. The pair (u,v) is an ith edge of a path η_{xy} iff $u = \eta_{xy}(i-1)$, $v = \eta_{xy}(i)$ for some i, and $Q(u,v) > 0$. Also, suppose E_i to be the set of all ith edges and $E = \bigcup E_i$.

Assumption 2.4.9 (First Regularity Condition)**.** *Let $V = \{(x,y,i) : x, y \in S,\ 1 \leqslant i \leqslant b_{xy}\}$. Define $T : V \to S^2 \times \mathbb{N}^2$ by $T(x,y,i) = (\eta_{xy}(i-1), \eta_{xy}(i), b_{xy}, i)$. We then say that η, a collection of paths for an irreducible Markov chain P, satisfies the first regularity condition if the following conditions are met:*

- *T is a one-to-one map onto $T(V)$.*
- *Fix $b, i \in \mathbb{N}$ such that $(u, v, b, i) \in T(V)$ for some $(u, v) \in S \times S$ and let $W_{bi} = \{(u, v) : (u, v, b, i) \in T(V)\} \subset E$. Define a one-to-one map $L_{bi} : W_{bi} \to S \times S$ by $L_{bi}(u, v) = (x, y)$, where $T(x, y, i) = (u, v, b, i)$ can be extended to a bijection of open sets and also has continuous partial derivatives almost everywhere with respect to the Lebesgue measure on $\mathbb{R}^n \times \mathbb{R}^n$ for any values of b and i.*

If η satisfies the first regularity condition, then the paths are "differentiable" in a way that even a small change in an edge will result in a path that is continuous, and the Jacobian for this change is well defined. The Jacobian, $J_{bi}(u, v)$, of the change of variable $(x, y) = L_{bi}(u, v)$ is given by

$$J_{bi}(u,v) = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} & \frac{\partial x_1}{\partial v_1} & \cdots & \frac{\partial x_1}{\partial v_n} \\ \vdots & \vdots & \vdots & & \vdots & \\ \frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n} & \frac{\partial x_n}{\partial v_1} & \cdots & \frac{\partial x_n}{\partial v_n} \\ \frac{\partial y_1}{\partial u_1} & \cdots & \frac{\partial y_1}{\partial u_n} & \frac{\partial y_1}{\partial v_1} & \cdots & \frac{\partial y_1}{\partial v_n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial u_1} & \cdots & \frac{\partial y_n}{\partial u_n} & \frac{\partial y_n}{\partial v_1} & \cdots & \frac{\partial y_n}{\partial v_n} \end{vmatrix}$$

where $u = (u_1, \ldots, u_n)$ and so on. We may denote $J_{bi}(u, v)$ by $J_{xy}(u, v)$ as given $(u, v) \in E$, and there is a one-to-one relationship between (x, y) and (b, i). This condition can be easily met for paths that are deemed satisfactory. The geometric constants used by Yuen [161] are denoted as:

$$A_\varepsilon = \operatorname{essup}_{(u,v)\in E} \left\{ Q(u,v)^{-(1-2\varepsilon)} \sum_{\eta_{xy} \ni (u,v)} \|\eta_{xy}\|_\varepsilon \, q(x) q(y) \, |J_{xy}(u,v)| \right\},$$

the sum is taken over all (x, y) such that $\eta_{xy} \ni (u, v)$ and the essential supremum with respect to the Lebesgue measure μ on $\mathbb{R}^n \times \mathbb{R}^n$, is defined as $\operatorname{esssup}\{f(x) : x \in X\} = \inf\{a :$

$\mu\{x : f(x) > a\} = 0\}$. Consequently, Yuen [163] provided a lower bound for the spectral gap of an irreducible Markov chain P on $\mathbb{R}^n$.

Theorem 2.4.10. *Consider the set of paths $\eta = \{\eta_{xy}\}$ that satisfy Assumption 2.4.9 for an irreducible Markov chain P on $\mathbb{R}^n$. For any $\varepsilon \in \mathbb{R}$, we then have*

$$\gamma \geq 1/A_\varepsilon,$$

where γ is the spectral gap of the Markov chain P.

In addition, we discuss the results of Wu et al.[158] regarding a lower bound on the mixing time in χ^2-divergence of reversible Markov chains on general state spaces using spectral gap and Dirichlet form. The findings indicate that getting the lower bound may be reduced to handling the spectral gap. Previous studies have provided similar findings, as demonstrated by Goel et al. [56, Lemma 3.1], Coulhon et al. [31, Proposition 4.2], Coulhon and Grigor'yan [32, Proposition 4.3], and Coulhon [30]. However, because the majority of these conclusions do not apply directly to Markov chains in infinite-dimensional spaces, they cannot be applied to the general state space. Wu et al. [158, Theorem 6] developed a new lower bound on mixing time by combining principles from previous findings, as follows:

Theorem 2.4.11. *Consider P to be the transition kernel of a reversible Markov chain with a stationary distribution π. Given an initial distribution $\mu_0 \ll \pi$ satisfying $d_2(\mu_0, \pi) < \infty$, then*

$$d_2^2\left(\mu_0 P^k, \pi\right) \geq d_2^2\left(\mu_0, \pi\right)\left(1 - \frac{\mathcal{E}_{P^2}\left(h_0, h_0\right)}{d_2^2\left(\mu_0, \pi\right)}\right)^k,$$

where $h_0 = \frac{d\mu_0}{d\pi}$ and $\mathcal{E}_{P^2}$ denotes the Dirichlet form of the two-step transition kernel P.

The proof is inspired by Coulhon and Grigor'yan [32] on-diagonal lower bounds for heat kernels and Markov chains.

2.4.3 Profile Methods

Profile methods aim to use the fact that Markov chains mix at various rates on different sizes and scales, and particularly mix faster on small sets. In discrete state space, Goel et al. [56] presented both the continuous and discrete time versions of the spectral profile upper bound on the mixing time.

Corollary 2.4.13 was presented by Goel et al. [56], for bounding the L^∞ mixing time of Markov chains using the spectral profile by introducing the following notations:

Definition 2.4.12 ([56]). *Define the spectral profile* $\Gamma : [\pi_{\min}, \infty) \to \mathbb{R}$ *by*

$$\Gamma(v) = \inf_{\pi_{\min} \le \pi(A) \le v} \gamma(A). \tag{2.4.3}$$

For a non-empty subset $A \subset \mathcal{X}$*, the spectral gap for set* A *is defined as:*

$$\gamma(A) = \inf_{f \in c_0^+(A)} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\pi(f)}, \tag{2.4.4}$$

where $c_0^+(A) = \{f \in L^2(\pi) : \mathrm{supp}(f) \subset A, f \ge 0, f \not\equiv \text{constant}\}$*, and* $\pi_{\min} = \min_x \pi(x)$*.*

Corollary 2.4.13 ([56], Corollary 2.1). *Let* P *be a* ζ*-lazy, irreducible and reversible transition kernel of a discrete-state Markov chain on* $\mathcal{X}$*. Then the* L^∞ *mixing time,* $\tau_\infty(\epsilon)$*, is bounded as:*

$$\tau_\infty(\epsilon) \le \int_{4\pi_{\min}}^{4/\epsilon} \frac{2dv}{\zeta v \Gamma(v)},$$

where Γ *indicates the spectral profile* (2.4.3) *of the chain.*

Kozma [88] investigated the sensitivity of the above spectral profile upper bound on the L^∞ mixing time of a reversible continuous-time random walk. In Chapter 3, we extend the result of Kozma [88] from the discrete-space spectral profile bound of [56] to the continuous-space setting.

Recently, there has been an extension of the profile bounds of the mixing time for Markov chains from discrete-state to continuous-state. Chen et al. [26] adapted the spectral profile

technique introduced by Goel et al. [56] to the continuous state setting in the following way: Define

Definition 2.4.14 ([26]). *For non-empty subsets $A, \Omega \subset \mathcal{X}$, the Ω-restricted spectral gap for the set A is given by:*

$$\gamma_\Omega(A) = \inf_{f \in c_0^+(A \cap \Omega)} \frac{\mathcal{E}(f,f)}{\mathrm{Var}_\pi(f)}, \tag{2.4.5}$$

where $c_0^+(A \cap \Omega) = \{f \in L^2(\pi) : \mathrm{supp}(f) \subset A \cap \Omega, f \geq 0, f \not\equiv \text{ constant }\}$.

Consequently, the Ω-restricted spectral profile is defined as:

$$\Gamma_\Omega(v) = \inf_{A\,:\,\pi(A\cap\Omega)\in[0,v]} \gamma_\Omega(A), \quad \forall v \in [0,\infty). \tag{2.4.6}$$

Note that the spectral profile is restricted to the set Ω and when considering Ω as $\mathcal{X}$, Definition 2.4.14 aligns with the standard definition of the restricted spectral gap and spectral profile, as presented by Goel et al. [56], for finite-state Markov chains to the continuous-state chains.

Similar to Corollary 2.4.13, Chen et al. [26] presented a spectral profile upper bound for the L^2 mixing time of continuous-state Markov chains. This was obtained by using the restricted spectral profile and the warm start.

Lemma 2.4.15 ([26], Lemma 11). *Let P be an exactly ζ-lazy, reversible and irreducible Markov chain satisfies the smooth chain Assumption 2.2.7 on a continuous state space $\mathcal{X}$, with a warm start μ_0 with constant β. Given an error tolerance ϵ, and a set $\Omega \in \mathcal{B}(\mathcal{X})$ such that $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{3\beta^2}$, then the L^2 mixing time, $\tau_2(\epsilon; \mu_0, P)$, is bounded as:*

$$\tau_2(\epsilon; \mu_0, P) \leq \int_{4/\beta}^{8/\epsilon^2} \frac{dv}{\zeta \cdot v\Gamma_\Omega(v)},$$

where Γ_Ω indicates the Ω-restricted spectral profile (2.4.6) of the chain.

In addition, Chen et al. [26] developed the conductance profile method proposed by Goel et al. [56] from discrete state to continuous-state chains. They made the adjustments required to fit the general setting.

We present the conductance profile definition of Chen et al. [26], which employs the idea that Markov chains usually mix faster on small sets, and the conductance measures the set that is the most difficult to escape. In other words, the lack of bottlenecks in the Markov chain's state space indicates rapid mixing. Moreover, working with the conductance profile is usually easier than working with the spectral profile, and it may also be beneficial in bounding the mixing time using the decomposition technique described in section 2.4.6.

Definition 2.4.16. *The Ω-restricted conductance profile is given by:*

$$\Phi_{\Omega}(v) = \inf_{A\,:\,\pi(A\cap\Omega)\in(0,v]} \frac{\phi(A)}{\pi(\Omega\cap A)}, \quad \forall v \in \left(0, \frac{\pi(\Omega)}{2}\right],$$

where ϕ is defined in Equation (2.2.7). Further, the truncated conductance profile $\hat{\Phi}$ is defined by:

$$\hat{\Phi}_{\Omega}(v) = \begin{cases} \Phi_{\Omega}(v), & v \in \left(0, \frac{\pi(\Omega)}{2}\right] \\ \Phi_{\Omega}\left(\frac{\pi(\Omega)}{2}\right), & v \in \left[\frac{\pi(\Omega)}{2}, \infty\right). \end{cases} \tag{2.4.7}$$

$\Phi_{\mathcal{X}}(\frac{1}{2})$ is also known as the conductance or the isoperimetric constant.

According to Lemma 9 of [26], the spectral profile and the conductance profile of a Markov chain have the following relationship:

$$\Gamma_{\Omega}(v) \geq \begin{cases} \frac{1}{2}\Phi^2_{\Omega}(v), & v \in \left[0, \frac{\pi(\Omega)}{2}\right] \\ \frac{1}{4}\Phi^2_{\Omega}\left(\frac{\pi(\Omega)}{2}\right), & v \in \left(\frac{\pi(\Omega)}{2}, \infty\right). \end{cases}$$

The worst-case conductance bound for the chain is the main focus of the standard conductance-based analysis. Chen et al. [26] used the Ω-restricted conductance profile to express the bounds, as it is common for a Markov chain to have poor conductance solely in

regions that have a small probability under the target distribution. The following lemma is the first statement for continuous state space chains introduced by Chen et al. [26].

Lemma 2.4.17 ([26], Lemma 3). *Let P be an exactly ζ-lazy, reversible and irreducible Markov chain satisfies the smooth chain Assumption 2.2.7 on a continuous state space $\mathcal{X}$, with a warm start μ_0 with constant β. Given an error tolerance ϵ, and a set $\Omega \in \mathcal{B}(\mathcal{X})$ such that $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{3\beta^2}$, then the L^2 mixing time, $\tau_2(\epsilon; \mu_0, P)$, is bounded as:*

$$\tau_2(\epsilon; \mu_0, P) \leq \int_{4/\beta}^{8/\epsilon^2} \frac{8dv}{\zeta \cdot v\hat{\Phi}_\Omega^2(v)},$$

where $\hat{\Phi}_\Omega$ denotes the truncated Ω-restricted conductance profile (2.4.7) of the chain.

The proof of the preceding lemma is derived from an appropriate extension of the concepts used by Goel et al. [56] for discrete state chains. Lemma 2.4.17 established a connection between the mixing time and the conductance profile, which can be seen as point-wise conductance.

The following is presented in [26] as an immediate result of Lemma 2.4.17.

Corollary 2.4.18. *Let P be an exactly ζ-lazy, reversible and irreducible Markov chain that satisfies the smooth chain Assumption 2.2.7 on a continuous state space $\mathcal{X}$, with a warm start μ_0 with constant β. If the Ω-restricted conductance profile is bounded as*

$$\Phi_\Omega(v) \geq \sqrt{B \log\left(\frac{1}{v}\right)}, \qquad \text{for } v \in \left[\frac{4}{\beta}, \frac{1}{2}\right],$$

for some $\beta > 0$ and Ω such that $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{3\beta^2}$, then with a β-warm start, we have

$$\tau_2(\epsilon; \mu_0, P) \leq \frac{64}{\zeta B} \log\left(\frac{\log \beta}{2\epsilon}\right).$$

2.4.4 Coupling Techniques

The coupling, which is a helpful and strong way of analyzing the convergence rate of Markov chains, was first introduced by Doeblin [41], and it has been widely applied in other fields of probability. Thorisson [150] and Lindvall [93] have provided further and extensive details on coupling and its application.

A pair of random variables (X, Y) on the same probability space is called a coupling of two probability measures μ and ν if the marginal distribution of X is μ and the marginal distribution of Y is ν. Thus, we define a sequence $(X_t, Y_t)_{t\geq 0}$ as a coupling of two Markov chains with transition kernel P so that $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are both Markov chains with transition kernel P. Couplings are useful because comparing distributions reduces to comparing random variables. This is a great technique to determine upper bounds on the total variation distance. The following, Theorem 5.4 by Levin et al. [91] describes the primary tool and the relationship between the coupling of two Markov chains and their total variation distance. The discussions in the book [91] focus on discrete state spaces, but it is worth noting that the proof of this theorem does not rely on this assumption [143].

Theorem 2.4.19 ([91]). *Suppose (X_t, Y_t) to be a coupling of Markov chains with transition kernel P on a state space $\mathcal{X}$ so that if $X_s = Y_s$, then $X_t = Y_t$ for all $t \geq s$. Let τ be the coupling time, which is the first time that $X_t = Y_t$ for all $t \geq \tau$. Assume that $X_0 = x$ and $Y_0 = y$, then*

$$\left\|P^t(x,\cdot) - P^t(y,\cdot)\right\|_{\mathrm{TV}} \leq \mathbb{P}_{x,y}\left\{\tau > t\right\},$$

where $\mathbb{P}_{x,y}$ is the probability on the space in which X_t and Y_t are both defined.

The following lemma is the immediate outcome of Theorem 2.4.19.

Lemma 2.4.20 (Fundamental Coupling Lemma). *Suppose (X_t, Y_t) to be a coupling of Markov chains with transition kernel P and the stationary distribution π on a state space*

$\mathcal{X}$ so that if $X_s = Y_s$, then $X_t = Y_t, \forall t \geq s$. Assume that $X_0 = x$ and Y_0 is distributed according to the stationary distribution π, then

$$\sup_x \left\| P^t(x, \cdot) - \pi(\cdot) \right\|_{\mathrm{TV}} \leq \sup_x \mathbb{P}_x \left\{ \tau > t \right\}.$$

The purpose of using this theory is to construct a coupling where X_t and Y_t quickly collide and stick to each other. Even though the theorem is quite simple, there are some important points to address. One requirement is that X_t and Y_t stick together at τ. Adjusting a coupling to possess this quality is not always simple, refer to [137] for various poor examples. On finite state spaces, the inequality in Theorem 2.4.19 is actually equality for certain couplings, as shown in [121]. However, demonstrating the existence of such an ideal coupling is unproductive and frequently not very helpful.

It is important to keep in mind that while (X_t, Y_t) may be a coupling of Markov chains, it may not necessarily be a Markov chain itself. Define a coupling as Markovian if the joint process is a Markov chain and non-Markovian otherwise. Pitman [121] discovered an optimal coupling that is typically non-Markovian, which greatly contributes to the challenge of obtaining it. Defining a global coupling with the desired properties can be challenging when using the coupling-based approach.

2.4.5 Comparison Methods

The comparison technique introduced by Diaconis and Saloff-Coste [37, 38] has been a crucial tool in the theory of finite state Markov chains. This theory allows users to analyze the mixing of a Markov chain by considering the mixing properties of another Markov chain that has the same state space, as long as their stationary distributions are reasonably similar. In practice, this might be helpful since a Markov chain can be hard to study at times, but there is a simpler and related chain that bounding its spectral gap is much easier. The comparison method has been widely used to estimate the mixing time of the desired chain

in this scenario [105]. The right method of comparing two Markov chains is to compare their Dirichlet forms. The study by Diaconis and Saloff-Coste [37, 38] provides a general comparison between chains, where the ratios of the Dirichlet forms, as well as the stationary distributions, are both bounded.

Lemma 2.4.21 ([91], Lemma 13.18)**.** *Suppose P and $\tilde{P}$ to be reversible transition matrices on discrete state space $\mathcal{X}$, with stationary distributions π and $\tilde{\pi}$, respectively. If for all f, $\mathcal{E}_{\tilde{P}}(f) \leq \alpha \mathcal{E}_P(f)$, then*

$$\tilde{\gamma} \leq \left[\max_{x \in \mathcal{X}} \frac{\pi(x)}{\tilde{\pi}(x)}\right] \alpha\gamma,$$

where $\tilde{\gamma}$ and γ are spectral gaps (2.4.2) of the transition matrices $\tilde{P}$ and P, respectively.

In the following, we present a useful lemma, which is a kind of toy form of the comparison method. If two Markov chains' transition probabilities are related by simple inequalities, then their spectral gaps and spectral profiles will also have the corresponding relationship.

Lemma 2.4.22. *Consider two reversible Markov chains P and $\tilde{P}$ on the same finite state space $\mathcal{X}$, with the same stationary distribution π. Assume that there exist constants $0 < c_1, c_2 < \infty$ such that*

$$c_1 P(x, y) \leq \tilde{P}(x, y) \leq c_2 P(x, y), \quad \forall\, x \neq y,$$

then

$$c_1 \gamma \leq \tilde{\gamma} \leq c_2 \gamma,$$

and

$$c_1 \Gamma(v) \leq \tilde{\Gamma}(v) \leq c_2 \Gamma(v),$$

where $\tilde{\Gamma}$ and Γ are the spectral profiles (2.4.3), and $\tilde{\gamma}$ and γ denotes the spectral gaps (2.4.2) of the transition matrices $\tilde{P}$ and P, respectively.

Proof. $c_1 P(x,y) \le \tilde{P}(x,y) \le c_2 P(x,y) \quad \forall\, x \neq y$, using Equation (2.2.1):

$$\begin{aligned} c_1 \mathcal{E}_P(f,f) &= \frac{1}{2} \sum_{x,y \in \mathcal{X}} [f(x) - f(y)]^2 \pi(x) c_1 P(x,y) \\ &\le \frac{1}{2} \sum_{x,y \in \mathcal{X}} [f(x) - f(y)]^2 \pi(x) \tilde{P}(x,y) \\ &= \mathcal{E}_{\tilde{P}}(f,f) \\ &\le \frac{1}{2} \sum_{x,y \in \mathcal{X}} [f(x) - f(y)]^2 \pi(x) c_2 P(x,y) \\ &= c_2 \mathcal{E}_P(f,f). \end{aligned}$$

Divide each side by $\mathrm{Var}_\pi(f)$,

$$c_1 \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)} \le \frac{\mathcal{E}_{\tilde{P}}(f,f)}{\mathrm{Var}_\pi(f)} \le c_2 \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)}. \tag{2.4.8}$$

By taking infimum on f from all sides of inequalities (2.4.8) and using the minimax characterization of the spectral gap (2.4.2), we get:

$$\begin{aligned} c_1 \gamma &= c_1 \inf_{\substack{f \in L^2(\pi): \\ \mathrm{Var}_\pi(f) \neq 0 \\ \mathbb{E}_\pi(f) = 0}} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)} \\ &\le \inf_{\substack{f \in L^2(\pi): \\ \mathrm{Var}_\pi(f) \neq 0 \\ \mathbb{E}_\pi(f) = 0}} \frac{\mathcal{E}_{\tilde{P}}(f,f)}{\mathrm{Var}_\pi(f)} = \tilde{\gamma} \\ &\le c_2 \inf_{\substack{f \in L^2(\pi): \\ \mathrm{Var}_\pi(f) \neq 0 \\ \mathbb{E}_\pi(f) = 0}} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)} \\ &= c_2 \gamma, \end{aligned}$$

where $\mathrm{Var}_\pi(f) = \langle f, f\rangle_\pi$, and $\mathbb{E}_\pi(f) = \langle f, \mathbf{1}\rangle_\pi$.

For a non-empty subset $S \subset \mathcal{X}$, and any function f belongs to the class of $c_0^+(S)$;

$$c_0^+(S) = \{f \in L^2(\pi) : \mathrm{supp}(f) \subset S, f \geq 0, f \not\equiv \text{ constant }\},$$

using (2.4.8) and taking infimum on $c_0^+(S)$ from all sides:

$$c_1 \inf_{f\in c_0^+(S)} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)} \leq \inf_{f\in c_0^+(S)} \frac{\mathcal{E}_{\tilde{P}}(f,f)}{\mathrm{Var}_\pi(f)} \leq c_2 \inf_{f\in c_0^+(S)} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)}, \tag{2.4.9}$$

hence

$$c_1\gamma(S) \leq \tilde{\gamma}(S) \leq c_2\gamma(S). \tag{2.4.10}$$

Consequently, taking infimum over S from all sides of inequality (2.4.10), we obtain $\forall v \in [\pi_{\min}, \infty)$:

$$c_1\Gamma(v) \leq \tilde{\Gamma}(v) \leq c_2\Gamma(v),$$

where $\Gamma(v) = \inf_{S:\,\pi(S)\in[\pi_{\min},v]} \gamma(S)$ and $\tilde{\Gamma}(v) = \inf_{S:\,\pi(S)\in[\pi_{\min},v]} \tilde{\gamma}(S)$.

∎

Lemma 2.4.21 can be extended to compare chains defined on two different state spaces, as shown in the following lemma. The study by Diaconis and Saloff-Coste [39, page 718] presents such lemmas that may be proven directly.

Lemma 2.4.23. *Let P and $\tilde{P}$ be reversible transition matrices, with stationary distributions π and $\tilde{\pi}$, defined on the finite sets $\mathcal{X}$ and $\tilde{\mathcal{X}}$, respectively. Suppose that there exists a linear map:*

$$\begin{aligned} L^2(\mathcal{X},\pi) &\to L^2(\tilde{\mathcal{X}},\tilde{\pi}),\\ f &\mapsto \tilde{f}. \end{aligned}$$

Additionally, let constants $A, \tilde{A}, a > 0$ *and for all* $f \in L^2(\mathcal{X},\pi)$*, the following conditions hold:*

$$\mathcal{E}_{\tilde{P}}(\tilde{f},\tilde{f}) \le A\mathcal{E}_P(f,f), \qquad \textit{and} \qquad \mathrm{aVar}_\pi(f) \le \mathrm{Var}_{\tilde{\pi}}(\tilde{f}) + \tilde{A}\mathcal{E}_P(f,f),$$

then

$$\gamma \ge \frac{a\,\tilde{\gamma}}{A + \tilde{A}\,\tilde{\gamma}}.$$

Proof. By utilizing the minimax characterization of the spectral gap (2.4.2) and considering the two given conditions:

$$\begin{aligned} \frac{A}{a}\gamma &= \frac{A}{a} \inf_{\substack{f\in L^2(\pi):\\ \mathrm{Var}_\pi(f)\neq 0\\ \mathbb{E}_\pi(f)=0}} \frac{\mathcal{E}_P(f,f)}{\mathrm{Var}_\pi(f)}\\ &\ge \inf_{\substack{f\in L^2(\pi):\\ \mathrm{Var}_\pi(f)\neq 0\\ \mathbb{E}_\pi(f)=0}} \frac{A\mathcal{E}_P(f,f)}{\mathrm{Var}_{\tilde{\pi}}(\tilde{f}) + \tilde{A}\mathcal{E}_P(f,f)}\\ &\ge \inf_{\substack{f\in L^2(\pi):\\ \mathrm{Var}_\pi(f)\neq 0\\ \mathbb{E}_\pi(f)=0}} \frac{1}{\frac{\mathrm{Var}_{\tilde{\pi}}(\tilde{f})}{\mathcal{E}_{\tilde{P}}(\tilde{f},\tilde{f})} + \frac{\tilde{A}}{A}}\\ &= \inf_{\substack{\tilde{f}\in L^2(\tilde{\pi}):\\ \mathrm{Var}_{\tilde{\pi}}(\tilde{f})\neq 0\\ \mathbb{E}_{\tilde{\pi}}(\tilde{f})=0}} \frac{1}{\frac{\mathrm{Var}_{\tilde{\pi}}(\tilde{f})}{\mathcal{E}_{\tilde{P}}(\tilde{f},\tilde{f})} + \frac{\tilde{A}}{A}}\\ &= \frac{1}{\frac{1}{\tilde{\gamma}} + \frac{\tilde{A}}{A}}. \end{aligned}$$

Consequently,

$$\gamma \geq \frac{a\,\tilde{\gamma}}{A + \tilde{A}\,\tilde{\gamma}}.$$

∎

Smith [144] extended the research conducted by Diaconis and Saloff-Coste [37, 38], comparing two Markov chains with distinct state spaces, where one state space is a subset of the other. The main technique involves extending functions from a smaller domain to a larger one.

Moreover, Theorem 2.4.24 by Diaconis and Saloff-Coste [37] allows us to estimate the mixing time of a Markov chain by comparing it to another Markov chains' mixing behaviors via canonical paths. Assume P and $\tilde{P}$ are two reversible transition matrices with stationary distributions π and $\tilde{\pi}$ with edge sets of $E = \{(x, y) : P(x, y) > 0\}$ and $\tilde{E} = \{(x, y) : \tilde{P}(x, y) > 0\}$, respectively. Moreover, we suppose that there is one path of $\eta_{xy} \in \eta$ for each $(x, y) \in \tilde{E}$, where $\eta = (e_1, e_2, \ldots, e_m)$ is a collection of paths from x to y made of edges in E such that $e_1 = (x, x_1), e_2 = (x_1, x_2), \ldots, e_m = (x_{m-1}, y)$ for some vertices $x_1, \ldots, x_{m-1} \in \mathcal{X}$. η is also called E-paths from x to y. Then, the congestion ratio B is defined as follows:

$$B := \max_{e \in E} \left(\frac{1}{Q(e)} \sum_{\substack{x,y \\ \eta_{xy} \ni e}} \tilde{Q}(x, y)\, |\eta_{xy}| \right), \tag{2.4.11}$$

where $Q(x, y) = \pi(x)P(x, y)$ and $\tilde{Q}(x, y) = \tilde{\pi}(x)\tilde{P}(x, y)$ are edge measures and $|\eta_{xy}|$ is the number of edges in the path η_{xy}.

Theorem 2.4.24 (Comparison via Paths)**.** *Suppose P and $\tilde{P}$ to be reversible transition matrices on discrete state space $\mathcal{X}$, with stationary distributions π and $\tilde{\pi}$, respectively. Consider B as defined in (2.4.11), which is the congestion ratio for a choice of E-paths. Then for all functions $f : \mathcal{X} \to \mathbb{R}$,*

$$\mathcal{E}_{\tilde{P}}(f) \leq B\mathcal{E}_P(f),$$

consequently,

$$\tilde{\gamma} \leq \left[\max_{x\in\mathcal{X}} \frac{\pi(x)}{\tilde{\pi}(x)}\right] B\gamma,$$

where $\tilde{\gamma}$ and γ are spectral gaps of the reversible transition matrices $\tilde{P}$ and P, respectively.

So far, we have been discussing the theory of comparison in discrete state spaces. However, Mira [102] presented an ordering for the second largest eigenvalues of two reversible Markov chains with the same stationary probability measure on general state spaces. Further, Yuen [161] extended the result of Diaconis and Saloff-Coste [38] to a more general setting with not necessarily the same stationary probability measure. The analogous to Lemma 2.4.21 and Theorem 2.4.24 for continuous-state Markov chains have been presented in Theorems 3.1 and 3.2 of the work of Yuen [161].

Theorem 2.4.25 ([161], Theorems 3.2)**.** *Consider two reversible Markov chains with transition probability kernels $P(x, dy)$, $\tilde{P}(x, dy)$ on the same state space $\mathcal{X}$ with invariant probability measures, π, $\tilde{\pi}$, respectively. If there exists constants $a, A > 0$ such that*

$$\mathcal{E}_{\tilde{P}}(f, f) \leq A\mathcal{E}_P(f, f), \quad \forall f \in L^2(\pi) \quad \text{and} \quad \tilde{\pi} \geq a\pi,$$

then

$$\gamma \geq (a/A)\tilde{\gamma},$$

and so

$$\beta \leq 1 - \frac{a}{A}\left(1 - \tilde{\beta}\right),$$

where $\tilde{\gamma}$ and γ represent the spectral gaps, and $\tilde{\beta}$ and β indicate the largest eigenvalues of Markov chains $\tilde{P}$ and P, respectively.

We present the following lemma as a result of Theorem 2.4.25.

Lemma 2.4.26. *Consider P and $\tilde{P}$ are two reversible transition probability kernels on the same state space $\mathcal{X}$, with the stationary probability measure π. Suppose that there exists a constant $0 < c_1 < \infty$ such that:*

$$\sup_{x \in \mathcal{X}} \frac{\tilde{P}(x, dy)}{P(x, dy)} \leq c_1,$$

then

$$\gamma \geq \frac{\tilde{\gamma}}{c_1},$$

where $\tilde{\gamma}$ and γ are spectral gaps (2.4.2) of the transition kernels $\tilde{P}$ and P, respectively.

Proof. By the given condition that $\sup_{x \in \mathcal{X}} \frac{\tilde{P}(x,dy)}{P(x,dy)} \leq c_1$, we have for all $f \in L^2(\pi)$

$$\begin{aligned} \mathcal{E}_{\tilde{P}}(f, f) &= \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} (f(x) - f(y))^2 \pi(dx) \tilde{P}(x, dy) \\ &\leq c_1 \frac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} (f(x) - f(y))^2 \pi(dx) P(x, dy) \\ &= c_1 \mathcal{E}_P(f, f), \end{aligned}$$

applying Theorem 2.4.25

$$\gamma \geq \frac{\tilde{\gamma}}{c_1}.$$

∎

Yuen [161] developed geometric constants that fulfill the initial condition of Theorem 2.4.25 for reversible Markov chains on $\mathbb{R}^n$. It has been assumed that the transition kernels satisfy the smooth chain Assumption 2.2.7, which states that the transition probability kernels can be expressed in the form of $P(x, dy) = \alpha(x)\delta_x(dy) + p(x, y)dy$ and $\tilde{P}(x, dy) = \tilde{\alpha}(x)\delta_x(dy) + \tilde{p}(x, y)dy$. Furthermore, it is supposed that the invariant distributions $\pi, \tilde{\pi}$ have densities $q(y), \tilde{q}(y)$ with respect to the Lebesgue measure, respectively.

It also requires the existence of a collection of paths, $\eta = \{\eta_{xy}\}$, satisfying certain regularity conditions related to both P, π and $\tilde{P}, \tilde{\pi}$ as presented in Assumption 2.4.27.

Let $\eta = \{\eta_{xy}\}$ be a set of $(P, \tilde{P})$ paths that for each $x \neq y$ with $\tilde{q}(x)\tilde{p}(x, y) > 0$, there exists $b_{xy} \in \mathbb{N}$ and a path $\eta_{xy} : \{0, \ldots, b_{xy}\} \to \mathbb{R}^n$ such that $\eta_{xy}(0) = x, \eta_{xy}(1), \ldots, \eta_{xy}(b_{xy}) = y$ and $p(\eta_{xy}(i-1), \eta_{xy}(i)) > 0$ for all $i = 1, \ldots, b_{xy}$. Define $Q(u, v) = p(u, v)q(u)$ for any $u, v \in \mathbb{R}^n$, and $\|\eta_{xy}\|_\varepsilon = \sum_{(u,v)\in\eta_{xy}} Q(u, v)^{-2\varepsilon}$ for $\eta_{xy} \in \eta$ for $\varepsilon \in \mathbb{R}$. The pair (u, v) is an ith edge of a path η_{xy} iff $u = \eta_{xy}(i-1)$, $v = \eta_{xy}(i)$ for some i, and $Q(u, v) > 0$. Assume E_i to be the set of all i th edges and $E = \bigcup E_i$.

Assumption 2.4.27 ($(P, \tilde{P})$ Regularity Condition)**.**

Let $V = \{(x, y, i) : \tilde{q}(x)\tilde{p}(x, y) > 0, 1 \leq i \leq b_{xy}\}$. *Define* $T : V \to (\mathbb{R}^n)^2 \times \mathbb{N}^2$ *by* $T(x, y, i) = (\eta_{xy}(i-1), \eta_{xy}(i), b_{xy}, i)$. *We then say that* η, *a collection of paths for an irreducible Markov chain* P, *satisfies* $(P, \tilde{P})$ *regularity condition if the following conditions are met:*

- T *is a one-to-one map onto* $T(V)$.
- *Fix* $b, i \in \mathbb{N}$ *such that* $(u, v, b, i) \in T(V)$ *for some* $(u, v) \in \mathbb{R}^n \times \mathbb{R}^n$ *and let* $W_{bi} = \{(u, v) : (u, v, b, i) \in T(V)\} \subset E$. *Define a one-to-one map* $L_{bi} : W_{bi} \to \mathbb{R}^n \times \mathbb{R}^n$ *by* $L_{bi}(u, v) = (x, y)$, *where* $T(x, y, i) = (u, v, b, i)$ *can be extended to a bijection of open sets and also has continuous partial derivatives almost everywhere with respect to the Lebesgue measure on* $\mathbb{R}^n \times \mathbb{R}^n$ *for any values of* b *and* i.

The Jacobian, $J_{bi}(u, v)$, of the change of variable $(x, y) = L_{bi}(u, v)$ is given by

$$J_{bi}(u,v) = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} & \frac{\partial x_1}{\partial v_1} & \cdots & \frac{\partial x_1}{\partial v_n} \\ \vdots & \vdots & \vdots & & \vdots & \\ \frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n} & \frac{\partial x_n}{\partial v_1} & \cdots & \frac{\partial x_n}{\partial v_n} \\ \frac{\partial y_1}{\partial u_1} & \cdots & \frac{\partial y_1}{\partial u_n} & \frac{\partial y_1}{\partial v_1} & \cdots & \frac{\partial y_1}{\partial v_n} \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial u_1} & \cdots & \frac{\partial y_n}{\partial u_n} & \frac{\partial y_n}{\partial v_1} & \cdots & \frac{\partial y_n}{\partial v_n} \end{vmatrix}$$

where $u = (u_1, \ldots, u_n)$ and so on. We may denote $J_{bi}(u,v)$ by $J_{xy}(u,v)$ as given $(u,v) \in E$, and there is a one-to-one relationship between (x,y) and (b,i).

For paths that are nice enough, this condition can be easily met. The geometric constants used by Yuen [161] are denoted as:

$$A_\varepsilon = \operatorname{essup}_{(u,v)\in E} \left\{ Q(u,v)^{-(1-2\varepsilon)} \sum_{\eta_{xy} \ni (u,v)} \|\eta_{xy}\|_\varepsilon \, \tilde{q}(x)\tilde{p}(x,y) \, |J_{xy}(u,v)| \right\}.$$

We then present the following theorem, which is the immediate result of [161, Theorem 3.1 and Theorem 3.2] as an extension of Theorem 2.4.24 to a continuous state space.

Theorem 2.4.28. *Consider the reversible Markov chains P and $\tilde{P}$ on $\mathbb{R}^n$ with respect to the stationary distributions π and $\tilde{\pi}$, respectively. Let $\eta = \{\eta_{xy}\}$ be a set of paths that satisfy the regularity condition 2.4.27 for $(P, \tilde{P})$. For any $\varepsilon \in \mathbb{R}$ and $f \in L^2(\pi)$,*

$$\mathcal{E}_{\tilde{P}}(f) \leq A_\varepsilon \mathcal{E}_P(f),$$

consequently,

$$\tilde{\gamma} \leq \left[\sup_{x\in\mathbb{R}^n} \frac{q(x)}{\tilde{q}(x)} \right] A_\varepsilon \gamma,$$

where $\tilde{\gamma}$ and γ are spectral gaps of $\tilde{P}$ and P, respectively.

Computing and estimating Markov chain Monte Carlo's spectral gap is challenging, particularly in continuous state space. However, if a non-negative Markov operator is trace-class, which means that the transition kernel is compact with summable eigenvalues, then the spectral gap of such an operator can be bounded by simple functions with typically decent integral forms [124]. It should also be noted that any non-negative Markov operator that is compact has the set of all eigenvalues that is contained in the set $[0, 1]$. In certain circumstances, determining whether a Markov operator is trace-class might be difficult and tricky by direct computation given that establishing compactness is not always easy [123].

Recent studies, such as Chakraborty and Khare [23] and Pal et al. [113] have shown that a significant number of Markov operators belonging to practically relevant MCMC algorithms can be classified as trace-class using a specific formula; refer to section 4.3 of [123]. Inspired by comparison theory, we present the following theorem as a simple approach to determine whether a Markov chain in general state space belongs to the trace-class by comparing it to another trace-class chain that is both known and closely related. Therefore, the method described in [124] or the comparison theory in either a continuous or discrete setting can be employed to bound its spectral gap.

Theorem 2.4.29. *Consider two reversible Markov chains on general state space $\mathcal{X}$, with transition probability distributions P and $\tilde{P}$ and the same stationary distribution π. If*

$$\sup_{x \in \mathcal{X}} \frac{\tilde{P}(x, dy)}{P(x, dy)} \leq 1,$$

then, P is trace-class if $\tilde{P}$ is trace-class.

Proof. Let $\{f_j\}$ be an orthonormal basis of $L^2(\pi)$. A reversible $\tilde{P}$ is trace-class, hence (see [124, 29])

$$\sum_j \langle \tilde{P} f_j, f_j \rangle_\pi < \infty. \tag{2.4.12}$$

In other words, the condition mentioned as inequality (2.4.12) is the same as $\tilde{P}$ being compact with summable eigenvalues.

Given that $\sup_{x\in\mathcal{X}} \frac{\tilde{P}(x,dy)}{P(x,dy)} \leq 1$, then $\langle Pf_j, f_j\rangle_\pi < \langle \tilde{P}f_j, f_j\rangle_\pi$, $\forall j$. Applying (2.4.12), we then obtain:

$$\sum_j \langle Pf_j, f_j\rangle_\pi < \infty.$$

This is equivalent to P being a trace class operator. ∎

Next, we provide a lemma that compares the spectral gaps of a reversible Markov chain $\tilde{P}$ and its lazy version P on continuous state space, as defined in Definition 2.2.8. We then use this lemma in the next chapter.

Lemma 2.4.30. *Consider a reversible Markov chain $\tilde{P}$ with respect to a stationary distribution $\tilde{\pi}$ on the state space $\mathcal{X}$. Denote its half-lazy version as P that satisfies Definition 2.2.8 with respect to a stationary measure π. Then*

$$\gamma \geq (1-\alpha)\tilde{\gamma}, \tag{2.4.13}$$

where γ and $\tilde{\gamma}$ are spectral gap of P and $\tilde{P}$, respectively, and α is defined in Definition 2.2.8.

Proof. By using Definition 2.2.8, equation(2.2.6), we obtain

$$\pi(dx) \propto \frac{1}{1-\alpha_x}\tilde{\pi}(dx) \Rightarrow \pi(dx) = z\frac{1}{1-\alpha_x}\tilde{\pi}(dx),$$

thus

$$1 = \int_{\mathcal{X}} \pi(dx)$$

$$= \int_{\mathcal{X}} z\frac{1}{1-\alpha_x}\tilde{\pi}(dx) \Rightarrow \begin{cases} z \geq 1-\sup_x \alpha_x \overset{(i)}{=} 1-\alpha \\ z \leq 1-\inf_x \alpha_x \overset{(ii)}{=} 1-\frac{1}{2} = \frac{1}{2}. \end{cases}$$

Steps (i) and (ii) follow from Definition 2.2.8, where $\frac{1}{2} \leq \alpha_x \leq \alpha < 1$, $\forall x \in \mathcal{X}$, and $\sup_x \alpha_x = \alpha$.

Hence

$$\begin{aligned} \mathcal{E}_{\tilde{P}}(f) &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}}(f(x)-f(y))^2\tilde{\pi}(dx)\tilde{P}(x,dy) \\ &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}}(f(x)-f(y))^2\tilde{\pi}(dx)\frac{1}{1-\alpha_x}P(x,dy) \\ &\quad -\frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}}(f(x)-f(y))^2\frac{\alpha_x}{1-\alpha_x}\tilde{\pi}(dx)\delta_x(dy) \\ &= \frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}}(f(x)-f(y))^2\frac{1}{z}\pi(dx)\tilde{P}(x,dy) - 0 \\ &= \frac{1}{2z}\int_{\mathcal{X}}\int_{\mathcal{X}}(f(x)-f(y))^2\pi(dx)P(x,dy) \\ &= \frac{1}{z}\mathcal{E}_P(f). \end{aligned} \tag{2.4.14}$$

On the other hand

$$\begin{aligned} \pi(dx) &= z\frac{1}{1-\alpha_x}\tilde{\pi}(dx) \\ &\leq z\frac{1}{1-\sup_x \alpha_x}\tilde{\pi}(dx) \\ &= z\frac{1}{1-\alpha}\tilde{\pi}(dx) \end{aligned} \tag{2.4.15}$$

and

$$\begin{aligned} \pi(dx) &= z\frac{1}{1-\alpha_x}\tilde{\pi}(dx) \\ &\geq z\frac{1}{1-\inf_x \alpha_x}\tilde{\pi}(dx) \end{aligned}$$

$$= \quad 2z\tilde{\pi}(dx). \tag{2.4.16}$$

Combining (2.4.14), (2.4.15), and applying Theorem 2.4.25, we then have

$$1 - \tilde{\gamma} \geq 1 - \frac{1}{(1-\alpha)}\gamma, \tag{2.4.17}$$

alternatively

$$\gamma \geq (1-\alpha)\tilde{\gamma}, \tag{2.4.18}$$

where γ and $\tilde{\gamma}$ are spectral gaps of P and $\tilde{P}$, respectively. ∎

2.4.6 Decomposition Methods

Comparison and profile methods are commonly used strategies for predicting spectral gaps and bounding convergence in more robust norms. However, their flexibility may make them challenging to employ. In this section, we will discuss a method for analyzing Markov chains using a set of intermediate, typically simpler, Markov chains. In recent years, this strategy has proven to be effective in dealing with statistical difficulties. There have been numerous notable developments in this approach that are highly applicable to Markov chain Monte Carlo. These include addressing various problems, such as the papers by Mangoubi et al. [98], Zhuo and Gao [166], and investigating different versions of these bounds that can provide more accurate estimates for statistical problems, like the studies done by Pillai and Smith [119] on finite-state reversible Markov chains and Miracle et al. [103].

The decomposition method for bounding the mixing time is based on the intuition that studying the entire Markov chain can sometimes be challenging, but the chain may be broken down into pieces that are simpler to study. If the Markov chain flows efficiently from one piece to the next, and if each piece quickly converges to its stationary state, then the original chain is expected to rapidly converge to equilibrium [97]. Further, the decomposition method

lets us combine multiple approaches to estimate the mixing time of a Markov chain's different pieces, which is a hybrid strategy for analyzing the convergence rate of a chain.

It is important to consider that decomposition can demonstrate the quick mixing of certain Markov chains, where other methods like path coupling [20] may not be sufficient. However, it is worth noting that this approach has the potential to significantly increase the bounds on the mixing time. This is because we are analyzing the Markov chain indirectly by utilizing the projection chain (defined below) and imposing restrictions on smaller subsets of the state space. We may also employ other indirect analysis techniques, such as the comparison method on the component chains. It would be preferable to have a more straightforward analysis, but in case other methods are unsuccessful, decomposition can be helpful in demonstrating that the mixing time is bounded by a polynomial.

To provide a basis for this section, let P be a reversible transition probability kernel of a Markov chain on $\mathcal{X}$ with respect to a probability distribution π. We divide and break the state space $\mathcal{X}$ into pieces $A_1, \ldots, A_n$ such that $\cup_i A_i = \mathcal{X}$ (In general, these subsets will not be mutually exclusive.) For each $i = 1, \ldots, n$, define the restriction of P to A_i, the restricted chain P_{A_i}, as follows:

$$P_{A_i}(x, A) = P(x, A) + 1_{\{x \in A\}}(1 - P(x, A_i)), \quad x \in A_i,\ A \subset A_i,$$

where any move that leaves A_i is rejected. One should note that P_{A_i} is reversible on the state space A_i with respect to the measure that has a density proportional to the restriction of π to A_i. Then, we bound the Markov chain's mixing time on the entire state space, by bounding the mixing time of the Markov chains corresponding to each P_{A_i}, and the Markov chain of the projection of $\{A_1, \ldots, A_n\}$.

The maximum overlap, Θ, between the pieces of the state space $\mathcal{X}$, and the transition

probability of the projection chain, $\ddot{P}(i,j)$, on the state space of $\{1,\ldots,n\}$, are defined by:

$$\Theta = \max_{x\in\mathcal{X}} |\{i : x \in A_i\}|,$$
$$\ddot{P}(i,j) = \frac{\pi(A_i \cap A_j)}{\Theta\,\pi(A_i)}, \quad i \neq j.$$

Madras and Randall [97] presented the following decomposition theorem, assuming that the pieces of the state space would overlap and not be pairwise disjoint.

Theorem 2.4.31 ([97], State Decomposition Theorem)**.**

$$\gamma_P \geq \frac{1}{\Theta^2}\gamma_{\ddot{P}}\left(\min_{i=1,\ldots,n}\gamma_{P_{A_i}}\right),$$

where γ_P and $\gamma_{\ddot{P}}$ are the spectral gaps of P and $\ddot{P}$, respectively.

If $\ddot{P}$ and P_{A_i} on each piece A_i have fast mixing times, which means they approach stationary state quickly, then the original chain is rapidly mixing.

On the other hand, Jerrum et al. [78], Martin and Randall [99] introduced decomposition theorems considering the state space's pieces to be disjoint. The following theorem by Martin and Randall [99] is analogous to Theorem 2.4.31 but applies to disjoint pieces $A_1,\ldots,A_n$ of the state space $\mathcal{X}$.

Theorem 2.4.32 ([99], Theorem 4.2)**.** *We have the following:*

$$\gamma_P \geq \frac{1}{2}\gamma_{\hat{P}}\min_{i=1,\ldots,n}\gamma_{P_{A_i}},$$

where γ_P and $\gamma_{\hat{P}}$ are the spectral gaps of P and $\hat{P}$, respectively, and $\hat{P}$ is defined as

$$\hat{P}(i,j) = \frac{1}{\pi(A_i)}\sum_{x\in A_i y\in A_j}\pi(x)P(x,y).$$

Theorem 2.4.31 decomposes a Markov chain's state space, whereas the following theorem as Theorem 1.2 in [97] decomposes the stationary distribution of a Markov chain. This is particularly applicable to reversible Metropolis-Hastings chains, which we define as follows:

Let R be the transition kernel of a reversible Markov chain on $\mathcal{X}$ with respect to a probability density ρ. Let $\tilde{\rho}$ be another probability density whose support is contained inside ρ's support. Then the "Metropolis-Hastings chain for R with respect to $\tilde{\rho}$ " is the new Markov chain with transition kernel $R^{[\tilde{\rho}]}$ defined by:

$$R^{[\tilde{\rho}]}(x, dy) = R(x, dy) \min\left\{1, \frac{\tilde{\rho}(y)\rho(x)}{\tilde{\rho}(x)\rho(y)}\right\} \quad \text{if } y \neq x$$
$$R^{[\tilde{\rho}]}(x, \{x\}) = 1 - \int_{\mathcal{X}\setminus\{x\}} R^{[\tilde{\rho}]}(x, dy),$$

if $\tilde{\rho}(x)\rho(y)$ is 0, then we take $R^{[\tilde{\rho}]}(x, dy) = 0$. The kernel R is usually referred to as the "proposal kernel." The concept is that $R^{[\tilde{\rho}]}$ works by proposing a move and then calculating a ratio that determines the probability of the proposed move being accepted. The acceptance scheme guarantees that $\tilde{\rho}$ is the equilibrium distribution.

For Theorem 2.4.33, Madras and Randall [97] assumed that the chain of interest, P, is a Metropolis-Hastings for a proposal chain R with respect to a desired stationary density $\tilde{\rho}$, which means $P = R^{[\tilde{\rho}]}$. Further, they supposed that $\tilde{\rho}$ can be represented as a convex combination of a few number of densities $\tilde{\rho}_0, \ldots, \tilde{\rho}_n$, which means $\tilde{\rho}$ is a "mixture density". They considered running a Metropolis-Hastings chain for each $\tilde{\rho}_j$, using the same proposal kernel R as in the original P, the chains $R^{[\tilde{\rho}_j]}$ are the pieces of the original chain. If the $\tilde{\rho}_j$'s overlap in the sense described below, then the spectral gap of the original chain can be bounded in terms of the spectral gaps of the Metropolis-Hastings chains for the $\tilde{\rho}_j$'s.

Theorem 2.4.33 ([97], Density Decomposition Theorem)**.** *Let $\tilde{\rho}_0, \ldots, \tilde{\rho}_n$ be probability densities on $\mathcal{X}$, with respect to a common reference measure μ, and assume $c_0, \ldots, c_n$ to be positive values that add up to 1. The mixture density, $\tilde{\rho}_{\text{mix}}$, is then defined as follows:*

$$\tilde{\rho}_{\text{mix}} := \sum_{j=0}^{n} c_j \tilde{\rho}_j.$$

Suppose R to be a reversible Markov chain with respect to a mixture probability density $\tilde{\rho}_{\text{mix}}$ on $\mathcal{X}$. Further, let γ_j and γ_{mix} be the spectral gaps of the Metropolis-Hastings chain $R^{[\tilde{\rho}_j]}$

and $R^{[\tilde{\rho}_{\text{mix}}]}$, *respectively. We also assume that neighboring* $\tilde{\rho}_j$*'s have some "overlap" , which means:*

$$\int_{\mathcal{X}} \min\left\{\tilde{\rho}_j(x), \tilde{\rho}_{j+1}(x)\right\} \mu(dx) \geq \delta, \quad j = 0, \ldots, n-1 \quad \textit{for some } \delta > 0,$$

then

$$\gamma_{\text{mix}} \geq \frac{\delta}{2n} \min_{j=0,\ldots,n} c_j \gamma_j.$$

Chapter 3

Precision of Mixing Time

3.1 Introduction

Many widely used methods for bounding the mixing time of a Markov chain, discussed in Chapter 2, are closely related to the spectrum of the underlying transition matrix [91], both because they are easy to use and they are stable under natural changes to the underlying Markov chain [88, 91]. This raises the obvious question [83]: is it feasible to find bounds that are both sharp and recognizably "spectral" or "geometric" in nature?

Kozma [88] found that the spectral profile of Goel et al. [56] provides nearly-sharp bounds on the L^∞ mixing time. More precisely, it showed that the spectral profile bound is sharp up to a multiplicative factor of $\log(\log(1/\pi_{\min}))$, where $\pi_{\min}$ is the smallest value of the probability mass function (PMF) of the stationary distribution.

Chen et al. [26] generalized the spectral profile bound of Goel et al. [56] to the continuous state setting. They presented a bound for the L^2 mixing time of continuous-state Markov chains, which extends the techniques of Goel et al. [56] from discrete to continuous-state chains.

In this chapter, we aim to generalize the finding of Kozma [88] to the continuous state

scenario and investigate the sharpness of the spectral profile bound of Chen et al. [26] on the L^2 mixing time of continuous-state Markov chains. We find that the bound is sharp up to a factor of $\log\log$ of the starting density. Our result shows the robustness of the spectral profile bound in continuous state setting. Furthermore, when used in conjunction with other methods, this precision might be useful for bounding the mixing time of a Markov chain, for which bounding its mixing time is difficult using existing techniques. Our main result can be used as a comparison bound, indicating that it is possible to compare chains even when only non-spectral bounds exist for a known chain.

This chapter is organized as follows: in section 3.2, we establish our notations and define the δ-approximate L^2 mixing time of a lazy Markov chain from a single starting point on continuous state space, which will be employed throughout this chapter. In section 3.3, we present our findings on the precision of the spectral profile bound on the mixing time of Markov chains in continuous state spaces. We then discuss the application of our primary finding to the comparison of Markov chains. The final section 3.4 is devoted to proofs of the theorems and lemmas presented in this chapter.

3.1.1 Related Works

Many studies, such as [64, 88, 66, 65, 3], have investigated similar problems, with the details depending on which notion of mixing must be approximated and what information the "spectral" or "geometric" bounds are allowed to use. We mention some of the most closely-related work on how small changes in graph properties can affect the mixing times of associated random walks. The most similar to our study [83] is [88], which shows that the spectral profile bound on the uniform mixing time proposed by Goel et al. [56] is sharp up to a $\log\log$ factor.

Hermon [64] explored that even small perturbations to the transition probabilities can significantly impact the L^∞ mixing time of simple random walks on graphs with uniformly bounded degrees of size n. The study reveals that such perturbations can cause the mixing

time to increase by a factor of $\Theta(\log \log n)$.

Hermon and Kozma [65] investigated the robustness of the total variation mixing time in vertex-transitive graphs, particularly Cayley graphs, under small perturbations. Their findings indicate that for non-transitive graphs, the mixing time can vary significantly based on the starting point, especially after increasing certain edge weights. Moreover, Hermon and Peres [66] examined the sensitivity of the total variation mixing time and the presence of a cutoff. The study shows that the total variation mixing time is not invariant under quasi-isometry, even for Cayley graphs, and can be substantially altered by bounded perturbations of edge weights or metric changes.

3.2 Notations and Basic Definitions

Throughout this chapter, we consider discrete-time Markov chains. Let $\mathcal{X}$ be a Polish space with Borel-σ algebra $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Fix a transition probability kernel P with unique stationary probability measure π.

Note that, if $P(x, \{x\}) > 0$ for any point $x \in \mathcal{X}$ but π has no atoms, then the L^2 mixing time of a continuous-state Markov chain with non-zero holding probability is infinite. Thus, in a continuous-space setting, the result of Kozma [88] would be vacuous. In this chapter, we are primarily interested in such chains and define a closely related metric, informally showing that this non-zero holding probability is the only thing that goes wrong.

By a small abuse of standard notation, we define the following definitions:

Definition 3.2.1 (Half-lazy chain)**.** *A Markov chain with transition probability kernel P and the stationary distribution π is called half-lazy if it can be written in the form:*

$$P(x, \cdot) = (1 - \alpha_x)\tilde{P}(x, \cdot) + \alpha_x \delta_x(\cdot), \quad \forall x \in \mathcal{X}, \tag{3.2.1}$$

for some $\frac{1}{2} \leq \alpha_x \leq \sup_x \alpha_x = \alpha < 1$, and some transition kernel $\tilde{P}$ whose distributions $\tilde{P}(x, \cdot)$ all have densities $\tilde{p}(x, y)$ with respect to π, where δ_x is the Dirac-delta function at x.

We denote the hitting time of the set $\{x\}^c \subset \mathcal{X}$ for the half-lazy chain in Definition 3.2.1 as $T_{\{x\}^c}$. This is the first time that the chain moves out from its initial state x. We have

$$\mathbb{P}(T_{\{x\}^c} = n) = \alpha_x^{n-1}(1 - \alpha_x). \tag{3.2.2}$$

Definition 3.2.2 (Exactly Half-Lazy)**.** *A Markov chain with transition probability kernel P and the stationary distribution π is called exactly half-lazy if it can be written in the form:*

$$P(x, A) = \frac{1}{2}\tilde{P}(x, A) + \frac{1}{2}\delta_x(A), \quad \forall\, x \in \mathcal{X} \text{ and } A \in \mathcal{B}(\mathcal{X}), \tag{3.2.3}$$

for some transition kernel $\tilde{P}$ *whose distributions* $\tilde{P}(x,\cdot)$ *all have densities* $\tilde{p}(x,y)$ *with respect to* π, *where* $\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$.

We are primarily interested in such chains in this chapter and we define the following mixing time.

Definition 3.2.3 (The δ-approximate L^2 mixing time from a single starting point)**.** *Consider an exactly half-lazy Markov chain* P. *For* $x \in \mathcal{X}$, *define*

$$d_{2,\delta}(x,P,k) = \begin{cases} d_2(\delta_x \tilde{P} P^{\lceil \frac{1}{2}k \rceil}, \pi) + \delta & P^{\lceil \frac{1}{2}k \rceil}(x,\{x\}) \leq \delta \\ \infty & 0 \leq \delta < P^{\lceil \frac{1}{2}k \rceil}(x,\{x\}). \end{cases} \tag{3.2.4}$$

We then define the δ*-approximate* L^2 *mixing time of* P *with respect to its stationary distribution* π *from a single starting point* $x \in \mathcal{X}$, $\tau_{2,\delta}(\epsilon;P)$, *as*

$$\tau_{2,\delta}(\epsilon;P) = \inf\left\{ k \in \mathbb{N} \mid \sup_{x\in\mathcal{X}} d_{2,\delta}(x,P,k) \leq \epsilon \right\}. \tag{3.2.5}$$

In our study [83], we use Lemma 11 of Chen et al. [26], which is referred to as Lemma 2.4.15 in Chapter 2 of this thesis. In this lemma, we consider Ω as the full state space $\mathcal{X}$ for simplicity in notation. However, the results can easily be applied to the case of using the Ω-restricted spectral profile in Lemma 2.4.15, which we will discuss later in the next section.

3.3 Precision of the Spectral Profile Bound

Our main result, Theorem 3.3.5, is an extension of Kozma's result [88] from the original discrete-space spectral profile bound of Goel et al. [56] to the continuous-space setting of Chen et al. [26]. Kozma [88] was inspired by the observation that Faber-Krahn inequalities provide precise and sharp bounds in various interesting manifolds. We present the finding of Kozma [88, Theorem 1] below.

Let $\tau_\infty(1/2)$ denote the L^∞ mixing time of a reversible continuous-time random walk on finite graph $\mathcal{X}$ with error tolerance of $\frac{1}{2}$.

Theorem 3.3.1 ([88], Theorem 1)**.** *For any reversible continuous-time Markov chain with transition kernel P with stationary distribution π on finite state space $\mathcal{X}$:*

$$\int_{4\pi_{\min}}^{8} \frac{2dv}{v\Gamma(v)} < C\tau_\infty(1/2)\log\log(1/\pi_{\min}),$$

where C denotes an absolute positive constant which is large enough.

Inspired by Theorem 3.3.1, we investigated the precision of the spectral profile bound on the L^2 mixing time of continuous-state Markov chains. Of course, $\pi_{\min}$ is 0 in the continuous-space setting, so as stated, the result of [88] would be vacuous in continuous state spaces. In our work [83], we replaced $\pi_{\min}$ with a "warm-start" constant that is equal to $\pi_{\min}$ in the discrete-space setting. This is a popular replacement in extending geometric bounds from discrete to continuous spaces [154]. The results of our study [83] are discussed in the following.

We first establish an analogue to Lemma 3.1 of [56] in the continuous state space. This result is then applied to our Lemma 3.3.2, which is similar to Lemma 2.4.15 but from a single starting point. All proofs are deferred to section 3.4.

Lemma 3.3.2. *Consider a reversible, irreducible, and exactly half-lazy continuous-state Markov chain P with the stationary distribution π. Assume that $\delta_x\tilde{P}$ is a β-warm start*

for every $x \in \mathcal{X}$. *Given error tolerances* $\epsilon \in (0,1)$ *and* $\delta \in [0, \epsilon)$, *then*

$$\tau_{2,\delta}(\epsilon; P) \leq \max\left(2\left(\int_{4/\beta}^{8/(\epsilon-\delta)^2} \frac{2\,dv}{v\Gamma(v)}\right), \frac{2\log(\frac{1}{\delta})}{\log(2)} + 1\right), \tag{3.3.1}$$

where Γ *denotes the spectral profile* (2.4.5) *of* P.

Refer to section 3.4.1 for the proof of the preceding lemma. Throughout this chapter, the term "ρ" is used to refer to the right-hand side of inequality (3.3.1), and the symbol log is used to represent the natural logarithm.

For $S \in \mathcal{B}(\mathcal{X})$ with $\pi(S) > 0$, define a sub-stochastic kernel

$$\tilde{P}_S(x, B) = \tilde{P}(x, B \cap S), \quad \forall B \in \mathcal{B}(\mathcal{X}).$$

Note that the kernels $\tilde{P}$ and $\tilde{P}_S$ are reversible with respect to π and $\pi_{|S}(B) \equiv \frac{\pi(B \cap S)}{\pi(S)}$, respectively. This can be easily verified by utilizing exactly half-lazy Definition 3.2.2.

Next, we state an assumption on which our results rely.

Assumption 3.3.3. *For* $S \in \mathcal{B}(\mathcal{X})$ *with* $\pi(S) > 0$, *we assume that the sub-stochastic kernel* $\tilde{P}_S$ *on* $L^2(\pi_{|S})$ *is Hilbert-Schmidt. Assume that* $L^2(\pi_{|S})$ *has an orthonormal basis of eigenfunctions* $\{f_i\}_{i\geq 0}$ *of* $\tilde{P}_S$, *with real eigenvalues* $\left\{\tilde{\beta}_i\right\}_{i\geq 0}$ *satisfy* $f_0 \equiv 1$, $0 \leq \tilde{\beta}_i < 1$, $\tilde{\beta}_i \downarrow 0$ *so that*

$$\int_{\mathcal{X}} \frac{d(\delta_x \tilde{P}_S)}{d\pi_{|S}}(y) f_i(y)\, \pi_{|S}(dy) = \tilde{\beta}_i f_i(x), \quad \forall x \in \mathcal{X}. \tag{3.3.2}$$

In order to prove Theorem 3.3.5, we then propose the following lemma, which is similar to Lemma 3.1 in [56].

Lemma 3.3.4. *Let* $S \in \mathcal{B}(\mathcal{X})$ *with* $\pi(S) > 0$ *and* $k \in \mathbb{N}$. *Let* $\tilde{P}_S$ *satisfy Assumption 3.3.3, then*

$$\sup_{x\in\mathcal{X}} h_{\delta_x\tilde{P},\lceil k/2\rceil}(x) \geq \frac{(1-\tilde{\gamma}(S))^{2k}}{\pi(S)}, \tag{3.3.3}$$

where $\tilde{\gamma}(S)$ is the spectral gap of $\tilde{P}$ for the set S.

Following Lemmas 3.3.2 and 3.3.4, we can extend Theorem 1 in [88] to the continuous state space context as our Theorem 3.3.5 in this chapter. This represents the precision of the spectral profile bound for $\tau_{2,\frac{1}{8}}(\frac{1}{4};P)$.

Theorem 3.3.5. *Consider a reversible, irreducible, and exactly half-lazy Markov chain P with the stationary distribution π. Assume that $\delta_x\tilde{P}$ is a β-warm start for every $x\in\mathcal{X}$ and $\tilde{P}$ satisfies Assumption 3.3.3. Then there exists a universal constant C such that*

$$\tau_{2,\frac{1}{8}}(\frac{1}{4};P) \leq \rho \leq C(\log\lceil\log_2(\beta)\rceil+1)\tau_{2,\frac{1}{8}}(\frac{1}{4};P) + 108\lceil\log_2(\beta)\rceil + 7. \tag{3.3.4}$$

Refer to section 3.4.3 for the proof. The proof of Theorem 3.3.5 is quite similar to the proof of Theorem 1 in [88], with the substitution of Lemma 3.1 in [56] with our Lemma 3.3.4.

When looking at the right-hand side of inequality (3.3.4), note that the spectral profile bounds in continuous state space may increase by an optimal factor of log log of the warm start. This finding demonstrates that the use of spectral profile bounds does not result in a significant loss of information. In addition, it may be helpful to bound the mixing time of a Markov chain, which is challenging to do using existing methods.

Finally, we expand our Lemma 3.3.2, to include a broader class of Markov chains: half-lazy Markov chains 3.2.1 and smooth chains 2.2.7. This result extends Lemma 11 in Chen et al. [26], providing a spectral profile upper bound on the L^2 mixing time of continuous-state Markov chains from any initial state. This conclusion may be used to expand the precision theorem and serve as a foundation for future study.

Lemma 3.3.6. *Consider a reversible, irreducible, and half-lazy continuous-state Markov chain P with the stationary distribution π. Assume that $\delta_x\tilde{P}$ is a β-warm start for every*

$x \in \mathcal{X}$. Given error tolerances $\epsilon \in (0,1)$, $\delta \in [0,\epsilon)$, and a set $\Omega \in \mathcal{B}(\mathcal{X})$ with $\pi(\Omega) \geq 1 - \frac{\epsilon^2}{3\beta^2}$, then

$$\tau_{2,\delta}(\epsilon; P) \leq \max\left(2\left(\int_{4/\beta}^{8/(\epsilon-\delta)^2} \frac{2\,dv}{v\Gamma_\Omega(v)}\right),\ \frac{2\log(\frac{1}{\delta})}{\log(\alpha)} + 1\right),$$

where Γ_Ω denotes the Ω-restricted spectral profile (2.4.6) of P and $\alpha = \sup \alpha_x$ in Equation (3.2.1).

3.3.1 Applications

In this section, we present the application of our main result. Theorem 3.3.5 can be used to compare Markov chains in the strong L^2 metric. As an auxiliary result, we note in Corollary 3.3.7 how our main result can be used to get comparison bounds that are similar to the popular comparison bounds exposited in e.g. [91]. The main difference is that traditional comparison bounds, as mentioned in [91] and also discussed in Section 2.4.5, require bounding the spectrum of a "nice" kernel K and then using this to bound the spectrum of a "difficult" kernel K'. However, Corollary 3.3.7 shows the possibility of comparing chains even with only non-spectral bounds on K.

Corollary 3.3.7. *Consider two Markov chains, $\tilde{K}$ and $\tilde{K}'$, which are reversible, irreducible, and satisfy Assumption 3.3.3. Let K and K' be their respective exactly half-lazy versions, with the stationary distribution π, where $\delta_x\tilde{K}$ is a β-warm start for every x. Denote the spectral profiles of K' and K as $\Gamma_{K'}$ and Γ_K, respectively. If there exists $0 < C_1 < \infty$ so that*

$$\Gamma_{K'}(v) \geq \frac{1}{C_1}\Gamma_K(v),\ \forall v \in [0,\infty),$$

then there exists a universal constant C such that

$$\tau_{2,\frac{1}{8}}(\frac{1}{4}; K') \leq C_1\, C(\log\lceil\log_2(\beta)\rceil + 1)\left(\tau_2(\frac{1}{8}; \delta_x\tilde{K}, K) + \frac{1}{8}\right) + 108\,\lceil\log_2(\beta)\rceil + 7.$$

3.4 Proofs

In this section, we provide proof of the lemmas and theorems presented in this chapter.

3.4.1 Proof of Lemma 3.3.2

Conditioning on the first time a Markov chain with kernel P moves, we have for every $\delta \geq P^{\lceil \frac{1}{2}k \rceil}(x, \{x\})$

$$\begin{aligned}
\delta_x P^s &= \delta_x P P^{s-1} \\
&= \delta_x \left[2^{-1}\tilde{P} + 2^{-1} I \right] P^{s-1} \\
&= 2^{-1}\delta_x \tilde{P} P^{s-1} + 2^{-1}\delta_x P^{s-1} \\
&= 2^{-1}\delta_x \tilde{P} P^{s-1} + 2^{-1}\delta_x P P^{s-2} \\
&\vdots \\
&= \sum_{n=1}^{s} 2^{-n}\delta_x \tilde{P} P^{s-n} + 2^{-s}\delta_x \\
&= 2^{-s}\left(\frac{1}{2^{-s}} \sum_{n=1}^{s} \mathbb{P}(T_{\{x\}^c} = n)\delta_x \tilde{P} P^{s-n} \right) + 2^{-s}\delta_x .
\end{aligned}$$

Hence

$$\delta_x P^s = \sum_{n=1}^{s} 2^{-n}\delta_x \tilde{P} P^{s-n} + 2^{-s}\delta_x . \tag{3.4.1}$$

Since

$$\begin{aligned}
\tau_{2,\delta}(\epsilon; P) &= \inf\left\{ k \in \mathbb{N} \mid \sup_{x \in \mathcal{X}} d_{2,\delta}(x, P, k) \leq \epsilon \right\} \\
&= \inf\left\{ k \in \mathbb{N} \mid \sup_{x \in \mathcal{X}} d_2(\delta_x \tilde{P}\, P^{\lceil \frac{1}{2}k \rceil}, \pi) \leq \epsilon - \delta,\ 2^{-\lceil \frac{1}{2}k \rceil} \leq \delta \right\},
\end{aligned} \tag{3.4.2}$$

hence

$$\tau_{2,\delta}(\epsilon; P) \leq \max\left(2 \sup_{x\in\mathcal{X}} \tau_2(\epsilon-\delta; \delta_x\tilde{P}, P),\ \frac{2\log(\frac{1}{\delta})}{\log(2)} + 1\right). \tag{3.4.3}$$

Applying Lemma 2.4.15 to the term "$\sup_{x\in\mathcal{X}} \tau_2(\epsilon-\delta; \delta_x\tilde{P}, P)$" in inequality (3.4.3) completes the proof.

3.4.2 Proof of Lemma 3.3.4

By Assumption 3.3.3:

$$\frac{d(\delta_x\tilde{P}_S^\ell)}{d\pi_{|S}}(y) = \sum_{i=0}^{\infty} \tilde{\beta}_i^\ell f_i(x) f_i(y), \quad \forall x, y \in \mathcal{X} \quad \text{and} \quad \forall\, \ell \in \mathbb{N},$$

where $\tilde{\beta}_i$ and f_i are eigenvalues and orthonormal eigenfunctions of $\tilde{P}_S$, respectively. Let $y = x$, we then have

$$\frac{d(\delta_x\tilde{P}_S^\ell)}{d\pi_{|S}}(x) = \sum_{i=0}^{\infty} \tilde{\beta}_i^\ell f_i^2(x), \quad \forall x \in \mathcal{X}. \tag{3.4.4}$$

For $s, j \in \mathbb{N}_0$, $j \leq s$, define $b(s,j) = 2^{-s}\binom{s}{j}$. We have:

$$P^s = \sum_{j=0}^{s} b(s,j)\tilde{P}^j. \tag{3.4.5}$$

Thus, we can write $\delta_x\tilde{P}P^{\lceil\frac{1}{2}k\rceil}$ as a polynomial in $\tilde{P}$ as follows:

$$\delta_x\tilde{P}P^{\lceil\frac{1}{2}k\rceil} = \delta_x\tilde{P}\left(\sum_{j=0}^{\lceil\frac{1}{2}k\rceil} b(\lceil\frac{1}{2}k\rceil, j)\tilde{P}^j\right) = \sum_{j=0}^{\lceil\frac{1}{2}k\rceil} b(\lceil\frac{1}{2}k\rceil, j)\delta_x\tilde{P}^{j+1}. \tag{3.4.6}$$

Using Equation (3.4.4), we can derive that $\forall x \in S$ and $k \in \mathbb{N}$

$$\pi(S)\frac{d(\delta_x\tilde{P}P^{\lceil\frac{1}{2}k\rceil})}{d\pi}(x) \quad \overset{(i)}{=} \quad \frac{d(\delta_x\tilde{P}P^{\lceil\frac{1}{2}k\rceil})}{d\pi_{|S}}(x)$$

$$
\begin{aligned}
&\overset{(ii)}{=} \frac{d(\sum_{j=0}^{\lceil \frac{1}{2}k \rceil} b(\lceil \frac{1}{2}k \rceil, j)\delta_x \tilde{P}^{j+1})}{d\pi_{|S}}(x) \\
&\geq \frac{d(\sum_{j=0}^{\lceil \frac{1}{2}k \rceil} b(\lceil \frac{1}{2}k \rceil, j)\delta_x \tilde{P}_S^{j+1})}{d\pi_{|S}}(x) \\
&= \sum_{j=0}^{\lceil \frac{1}{2}k \rceil} b(\lceil \frac{1}{2}k \rceil, j) \frac{d(\delta_x \tilde{P}_S^{j+1})}{d\pi_{|S}}(x) \\
&\overset{(iii)}{=} \sum_{j=0}^{\lceil \frac{1}{2}k \rceil} b(\lceil \frac{1}{2}k \rceil, j) \left(\sum_{i=0}^{\infty} \tilde{\beta}_i^{j+1} f_i^2(x) \right) \\
&\overset{(iv)}{\geq} \sum_{j=0}^{\lceil \frac{1}{2}k \rceil} b(\lceil \frac{1}{2}k \rceil, j) \left(\tilde{\beta}_1^{j+1} f_1^2(x) \right) \\
&\geq f_1^2(x) \tilde{\beta}_1^{2k}.
\end{aligned}
$$

Step (i) uses $\pi_{|S}(B) = \frac{\pi(B \cap S)}{\pi(S)}$, $\forall B \subset \mathcal{X}$. Step (ii) uses Equation (3.4.6). Step (iii) applies Equation (3.4.4) and step (iv) follows from Assumption 3.3.3 that $0 \leq \tilde{\beta}_i < 1$ and the fact that $f_i^2(x) \geq 0$.

By taking the supremum over x

$$
\begin{aligned}
\sup_{x \in \mathcal{X}} h_{\delta_x \tilde{P}, \lceil k/2 \rceil}(x) &\geq \sup_{x \in S} h_{\delta_x \tilde{P}, \lceil k/2 \rceil}(x) \\
&= \sup_{x \in S} \frac{d(\delta_x \tilde{P} P^{\lceil \frac{1}{2}k \rceil})}{d\pi}(x) \\
&\geq \frac{\tilde{\beta}_1^{2k} \sup_{x \in S} f_1^2(x)}{\pi(S)} \\
&\overset{(i)}{\geq} \frac{\tilde{\beta}_1^{2k}}{\pi(S)} \\
&= \frac{(1 - \tilde{\gamma}(S))^{2k}}{\pi(S)},
\end{aligned}
$$

where $\tilde{\gamma}(S)$ is the spectral gap of $\tilde{P}$ for the set S. Inequality (i) follows from the fact that $\|f_1\|_{\pi_{|S}}^2 = 1$.

We are now equipped to prove Theorem 3.3.5.

3.4.3 Proof of Theorem 3.3.5

Fix $C' > 0$. Consider set A_s so that $0 < \pi(A_s) \leq 2^{-s}$ and

$$\gamma(A_s) \leq \inf\left\{\gamma(S) : \pi(S) \leq 2^{-s}\right\} + C', \tag{3.4.7}$$

where $\gamma(A_s)$ is the spectral gap of P for the set A_s. By monotonicity of Γ and change of variables:

$$\begin{aligned}\int_{4/\beta}^{512} \frac{2\,dv}{v\Gamma(v)} &= (2\log 2)\int_{-9}^{\log_2\beta-2} \frac{du}{\Gamma\left(2^{-u}\right)} \\ &= (2\log 2)\left[\int_{0}^{\log_2\beta-2} \frac{du}{\Gamma\left(2^{-u}\right)} + \frac{9}{\Gamma(1)}\right] \\ &\leq 2\log 2 \sum_{s=1}^{\lceil\log_2(\beta)\rceil} \frac{1}{\Gamma\left(2^{-s}\right)} + \frac{18\log 2}{\Gamma(1)}.\end{aligned} \tag{3.4.8}$$

Using Lemma 2.2 of [56] and the definition of the spectral profile:

$$\Gamma(1/2) \leq 2\,\Gamma(1). \tag{3.4.9}$$

Hence

$$\begin{aligned}\int_{4/\beta}^{512} \frac{2\,dv}{v\Gamma(v)} &\leq 2\log 2 \sum_{s=1}^{\lceil\log_2(\beta)\rceil} \frac{1}{\Gamma\left(2^{-s}\right)} + \frac{36\log 2}{\Gamma(1/2)} \\ &\leq 27 \sum_{s=1}^{\lceil\log_2(\beta)\rceil} \frac{1}{\Gamma\left(2^{-s}\right)}.\end{aligned} \tag{3.4.10}$$

We then have, for all $C' > 0$ sufficiently small,

$$\rho \leq 54 \sum_{s=1}^{\lceil\log_2(\beta)\rceil} \frac{1}{\Gamma\left(2^{-s}\right)} + \frac{2\,\log(8)}{\log(2)} + 1$$

$$\begin{aligned}
&\overset{(i)}{\leq} \quad 54 \sum_{s=1}^{\lceil \log_2(\beta) \rceil} \frac{1}{\Gamma\left(2^{-s}\right)} + \frac{2\ \log(8)}{\log(2)} + 1 \\
&\overset{(ii)}{\leq} \quad 54 \sum_{s=1}^{\lceil \log_2(\beta) \rceil} \frac{1}{\gamma(A_s) - C'} + 7 \\
&\overset{(iii)}{=} \quad 108 \sum_{s=1}^{\lceil \log_2(\beta) \rceil} \frac{1}{\tilde{\gamma}(A_s) - 2C'} + 7. \qquad (3.4.11)
\end{aligned}$$

Step (i) follows from equation (3.4.9). Step (ii) follows from (3.4.7). Step (iii) applies the exactly half-lazy Definition 3.2.2. Let $C' \to 0$, and applying Lemma 3.3.4 for the set A_s we have for any $k \in \mathbb{N}$

$$\begin{aligned}
\sup_{x \in \mathcal{X}} h_{\delta_x \tilde{P}, \lceil k/2 \rceil}(x) &\geq \frac{(1 - \tilde{\gamma}(A_s))^{2k}}{\pi(A_s)} \\
&= \frac{\exp(2k \log(1 - \tilde{\gamma}(A_s))}{\pi(A_s)}. \qquad (3.4.12)
\end{aligned}$$

Imitating the steps of [88], we can deduce the following:

$$\begin{aligned}
\tau_{2,\frac{1}{8}}(\frac{1}{4}; P) &\overset{(i)}{\geq} c \frac{\log\left(\pi(A_s)\right)}{\log(1 - \tilde{\gamma}(A_s))} \\
&\geq c \frac{\log\left(2^{-s}\right)}{\log(1 - \tilde{\gamma}(A_s))} \\
&\overset{(ii)}{\geq} c \frac{s \log 2}{\frac{1}{1 - \tilde{\gamma}(A_s)} - 1}, \qquad (3.4.13)
\end{aligned}$$

where c denotes an absolute positive constant. Step (i) follows from inequality (3.4.12), Equation (3.4.3), and Proposition 2.3.5 in Chapter 2 for the reversible chain $Q = \tilde{P} P^{\lceil k/2 \rceil}$. Step (ii) applies $\log(\frac{1}{x}) \leq \frac{1}{x} - 1, \forall x > 0$.

Note that $\frac{1}{1 - \tilde{\gamma}(A_s)} - 1 > 0$. Therefore, we obtain the following using inequality (3.4.13)

$$\tau_{2,\frac{1}{8}}(\frac{1}{4}; P) \times \left(\frac{1}{1 - \tilde{\gamma}(A_s)} - 1 \right) \quad \geq \quad c \log 2\, s.$$

As a result, we get

$$\frac{1}{\tilde{\gamma}(A_s)} \leq c' \frac{\tau_{2,\frac{1}{8}}(\frac{1}{4}; P)}{s} + 1, \tag{3.4.14}$$

where c' denotes an absolute positive constant. Finally, by combining (3.4.11) and (3.4.14), we have

$$\begin{aligned}
\rho &\leq 108 \sum_{s=1}^{\lceil \log_2(\beta) \rceil} \left(c' \frac{\tau_{2,\frac{1}{8}}(\frac{1}{4}; P)}{s} + 1 \right) + 7 \\
&\overset{(i)}{\leq} 108\, c' (\log \lceil \log_2(\beta) \rceil + 1) \tau_{2,\frac{1}{8}}(\frac{1}{4}; P) + 108 \lceil \log_2(\beta) \rceil + 7 \\
&= C (\log \lceil \log_2(\beta) \rceil + 1) \tau_{2,\frac{1}{8}}(\frac{1}{4}; P) + 108 \lceil \log_2(\beta) \rceil + 7,
\end{aligned}$$

where C denotes an absolute positive constant. Inequality (i) follows by Riemann approximation. This completes the proof of Theorem 3.3.5.

3.4.4 Proof of Lemma 3.3.6

Using Definition 3.2.1, we can apply a telescoping sum to obtain the following:

$$\begin{aligned}
\delta_x P^s &= \delta_x P P^{s-1} \\
&= \delta_x \left[(1-\alpha_x)\tilde{P} + \alpha_x I \right] P^{s-1} \\
&= (1-\alpha_x)\delta_x \tilde{P} P^{s-1} + \alpha_x \delta_x P^{s-1} \\
&= (1-\alpha_x)\delta_x \tilde{P} P^{s-1} + \alpha_x \delta_x P P^{s-2} \\
&\vdots \\
&\overset{(i)}{=} \sum_{n=1}^{s} (\alpha_x)^{n-1} (1-\alpha_x) \delta_x \tilde{P} P^{s-n} + \alpha_x^s \delta_x \\
&\overset{(ii)}{=} (1-\alpha_x^s) \left(\frac{1}{(1-\alpha_x^s)} \sum_{n=1}^{s} \mathbb{P}(T_{\{x\}^c} = n) \delta_x \tilde{P} P^{s-n} \right) + \alpha_x^s \delta_x.
\end{aligned} \tag{3.4.15}$$

Step (i) follows from induction. In step (ii), $T_{\{x\}^c}$ denotes the first time the chain moves out from its initial state x, and $\mathbb{P}(T_{\{x\}^c} = n) = \alpha_x^{n-1}(1-\alpha_x)$.

Using Equation (3.4.15) and Definition 3.2.3, we then have for every $x \in \mathcal{X}$, and every $\delta \geq \alpha_x^{\lceil \frac{k}{2} \rceil}$

$$d_{2,\delta}(\delta_x, P, k) = d_2(\delta_x \tilde{P} P^{\lfloor \frac{1}{2}k \rfloor}, \pi) + \delta. \tag{3.4.16}$$

Since

$$\begin{aligned} \tau_{2,\delta}(\epsilon; P) &= \inf\left\{k \in \mathbb{N} \mid \sup_{x\in\mathcal{X}} d_{2,\delta}(x, P, k) \leq \epsilon\right\} \\ &= \inf\left\{k \in \mathbb{N} \mid \sup_{x\in\mathcal{X}} d_2(\delta_x \tilde{P}\, P^{\lceil \frac{1}{2}k \rceil}, \pi) \leq \epsilon - \delta,\ \alpha_x^{\lceil \frac{k}{2} \rceil} \leq \delta\right\}, \end{aligned} \tag{3.4.17}$$

hence

$$\tau_{2,\delta}(\epsilon; P) \leq \max\left(2 \sup_{x\in\mathcal{X}} \tau_2(\epsilon - \delta; \delta_x \tilde{P}, P),\ \frac{2\log(\delta)}{\log(\alpha)} + 1\right), \tag{3.4.18}$$

where $\alpha = \sup \alpha_x$ in Equation (3.2.1). Apply Lemma 2.4.15 to the term "$\sup_{x\in\mathcal{X}} \tau_2(\epsilon - \delta; \delta_x \tilde{P}, P)$" in inequality (3.4.18), and complete the proof.

3.4.5 Proof of Corollary 3.3.7

$$\begin{aligned} \tau_{2,\frac{1}{8}}(\frac{1}{4}; K') \leq \rho' &\overset{(i)}{\leq} C_1 \rho \\ &\overset{(ii)}{\leq} C_1 C(\log\lceil \log_2(\beta)\rceil + 1)(\tau_2(\frac{1}{8}; \delta_x \tilde{K}, K) + \frac{1}{8}) + 108\,\lceil \log_2(\beta)\rceil + 7. \end{aligned}$$

Step (i) follows from $\Gamma_{K'}(v) \geq \frac{1}{C_1}\Gamma_K(v),\ \forall v \in [0, \infty)$. Step (ii) follows from the precision inequality in Theorem 3.3.5.

Chapter 4

Simulation and Its Application

4.1 Introduction

This chapter discusses our work [84] on developing a generative model that is capable of producing synthetic longitudinal health data. We aimed to build a generative model that is efficient and versatile, capable of synthesizing various types of variables and handling different data structures. Our primary goal was to ensure that the synthesized data had adequate utility. This utility is commonly described as the ability to replicate the patterns observed in the original data. Furthermore, we aimed to prevent overfitting in the model, which helps to avoid the unintentional duplication of records from the original data, reducing the risk of identity disclosure.

4.1.1 Background

Synthetic data generation (SDG) is a technique for creating fake data sets that mimic the patterns and distributions found in real data. Although the technical concepts behind synthetic data generation have been around for quite some time, their practical application has only recently started to gain popularity. One reason is that this type of data either addresses

difficult problems that were previously challenging to solve, or it solves them more efficiently and affordably. All of these problems revolve around data access; it might be difficult to obtain real data at times. Sharing actual data for secondary purposes might be challenging due to legal or ethical considerations which can result in delays in data set sharing or access approvals. But synthetic data is not actual data, and researchers can access it more rapidly with far fewer privacy constraints.

A kind of synthetic data is generated from real data set. This means that the analyst has some real data and then builds a model to capture its distributions and structure. After creating the model, the synthetic data is sampled or generated from it. If the model is a good representation of the actual data, synthetic data will have statistical properties that are similar to those of real data. The synthetic data does not have a one-to-one mapping to the actual data, and therefore has the potential of privacy preserving properties [129, 48, 72, 126, 138]. However, overfitting the model to the original data might result in a synthetic record that matches a real observation [48]. Thus, we must avoid overfitting the model in order to reduce the chance of identity disclosure.

Electronic health records, also known as EHRs, are rapidly becoming an invaluable source of detailed information about patients. This is because EHRs could help solve many problems in healthcare, like making clinical decisions faster and making sure patients are safer. Unfortunately, researchers often struggle to find quality health data for their work, and the process of properly de-identifying EHRs before sharing them with researchers requires expertise and time. A common way to make data anonymous is to take out specific information that can be used to identify people, like names or social security numbers. However, Sweeney [148] showed that getting rid of explicit identifiers is not enough to protect people's privacy because attributes that don't directly identify a person, like date of birth, race, ZIP code, or gender, can be linked to public databases to reveal personal information. Therefore, data synthesis has become an interesting way to simulate non-identifiable health data. This makes it easier for researchers to access data faster and with fewer privacy concerns.

The healthcare domain presents unique challenges for synthetic data generation due to factors like noisy data, missing measurements, evolving patient populations, heterogeneous data, lengthy sequences of events, and rare disease cases. In many circumstances, SDG is a better solution to the data access problem than other available methods in the context of health data. For example, applying privacy-enhancing technologies (PET) like differential privacy to machine learning models in healthcare settings presents challenges in preserving data utility and addressing class imbalance [27, 147]. A recent study by Suriyakumar et al. [147] found that the unique complexities of health data, such as imbalanced classes and sensitive patient information, pose significant obstacles in balancing privacy and model utility. As a result, further research is required to develop effective differential privacy approaches for healthcare applications.

Different generative models have been proposed so far. Drechsler and Reiter [42] and Reiter [128] introduced decision trees as one of the first machine-learning techniques for synthetic data generation. It has been frequently used to synthesize small and non-longitudinal health and social sciences data. Generative adversarial networks (GANs) have become increasingly popular as generative models in both research and practice [159, 104, 67]. However, their main limitations lie in their unstable training process, which could result in mode collapse [104] where the generator fails to capture the entire distribution of the training data, resulting in repetitive or limited samples. Training GANs can be quite challenging, and it is a more difficult problem than optimizing an objective function. It has been observed that excessive training of GAN-based models can negatively impact the quality of synthetic data as well [159].

Diffusion models [68, 160] were initially designed as latent variable models for continuous data sets. They can provide a framework for creating synthetic continuous data. To produce synthetic discrete data using diffusion models, the data needs first to be presented as binary bits, followed by training a continuous diffusion model [25]. Diffusion models can be quite challenging in terms of computation and training time, especially when compared to GANs. Additionally, they involve numerous adjustments and settings that need to be carefully tuned

to obtain optimal samples. Another issue is the need for large-scale training datasets. It can be even more challenging to capture multimodal distributions [25].

As pointed out, existing generative models have certain drawbacks when it comes to computational efficiency and data adaptability. While some models do well with smaller data sets, others require a substantial amount of training data. In addition, the synthesis of longitudinal health data can be quite challenging due to the fact that patients may have lengthy sequences of events and come from diverse populations. Thus, further research is required to develop generative models that can efficiently synthesize longitudinal health data while considering various variables and structural complexities. Applying a dimensionality reduction technique to build a generative model might be advantageous. These methods convert the high-dimensional description of the data into a low dimension without losing important information and phenotypes [28]. In the medical context, phenotypes are used to describe relevant variations and features [133]. Matrix factorization [61] is one of the unsupervised dimensionality reduction techniques that have been emerging. However, matrix factorization does not necessarily detect associations within a data set because the data may not be accurately represented as matrices and tensor decompositions [107] have drawn growing attention because of their interpretability and flexibility in accommodating high-dimensional data.

We provide a short background on tensors, and tensor decompositions in the following.

4.1.1.1 Tensor Decomposition

A tensor is a multidimensional array that generalizes matrices to multiple dimensions. An N-way tensor (or Nth-order tensor) has N indices, making it an N-dimensional array. For example, a vector is a first-order tensor. A matrix is a second-order tensor. Tensors of order three or higher are called higher-order tensors.

Tensor decomposition is an active area of research that has been frequently applied to health care data [69]. It has been found to be an efficient method for phenotyping EHRs [85].

Furthermore, it has a wide range of applications that extend beyond health data analysis. These applications include recommender systems [82] and signal processing [140]. Thus far, several tensor decomposition techniques have been developed [86], such as Tucker [152], PARAFAC2 [60], SPARTan [116], COPA [4], Sparse H-Tucker [115], TASTE [5]. The most popular technique is called canonical polyadic (CP) tensor factorization [59, 22], which is also known by 2 different names: canonical tensor decomposition (CANDECOMP) and parallel factor decomposition (PARAFAC). These names were introduced separately by Carroll and Chang [22] and Harshman [59], respectively. CP decomposition looks for the best low-rank approximation for the sum of squared errors [86]; in other words, it models interlinked data as a tensor and discovers phenotypes. It decomposes a tensor into a sum of component rank-one tensors as shown in Figure 4.1.

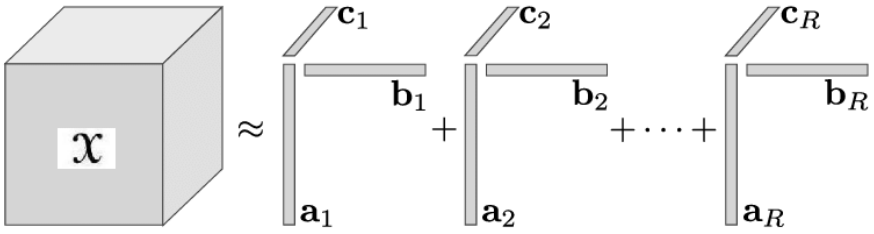


Figure 4.1: Canonical Polyadic (CP) decomposition.

Consider a third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, hence its CP decomposition is as follows:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r,$$

where the symbol "$\circ$" represents the vector outer product. R is a positive integer and $\mathbf{a}_r \in \mathbb{R}^I, \mathbf{b}_r \in \mathbb{R}^J$, and $\mathbf{c}_r \in \mathbb{R}^K$ for $r = 1, \ldots, R$.

Tensor's rank, R, is determined by the smallest number of "rank one" tensors in which the tensor can be decomposed [89], to put it another way, it is the least number of components in an exact CP decomposition. One should note that there exists a CP decomposition for any tensor as per the study of Bourbaki [16], and using a low-rank model is more affordable in analysis.

Recently, Hong et al. [71] introduced generalized canonical polyadic (GCP) decomposition that provides the option to incorporate alternative loss functions alongside squared errors. This decomposition is an efficient tool for dimensionality reduction of small-scale and dense tensors. However, Kolda and Hong [87] proposed stochastic gradient descent as a solution to address the difficulty of fitting generalized CP to large-scale tensors. GCP decomposition accommodates diverse data types, such as binary and count data, and finds application in various fields like social network analysis, neuroimaging, and environmental modeling. This makes it an ideal choice for modelling a vast and heterogeneous data set, such as the EHRs, and longitudinal data. Although CP is still considered a fundamental approach, GCP's versatility and broader range of applications make it an attractive choice for various tasks. Researchers often choose GCP based on their specific data and objectives. CP factorization and its generalization, GCP decomposition, are essential tools for tensor analysis. They lead to a factorized tensor that includes the most significant computational characteristics, and numerous studies demonstrate their exceptional performance in phenotyping EHRs [85]. Furthermore, privacy-enhancing technologies have been frequently applied to them in medical settings [155, 96].

4.2 Notations and Basic Definitions

We first describe the preliminaries and notations used in this chapter. Tabel 4.1 shows some basic symbols used for tensor factorization.

Notations	Descriptions
$\mathbb{N}_0$	$\{0\} \cup \mathbb{N}$
$\circ$	Outer Product
N	Number of dimensions (ways) of a tensor
R	Number of ranks
$\boldsymbol{\mathcal{X}}, \mathbf{X}, \mathbf{x}, x$	Tensor, matrix, vector, scalar
$\boldsymbol{\mathcal{X}}$	Observed tensor
$\boldsymbol{\mathcal{M}}$	Model tensor
$\mathbf{B}$, $\mathbf{C}$	Nonpatient factor matrices
$\mathbf{A}$	Patient factor matrix
$\mathbf{a}, \mathbf{b}, \mathbf{c}$	Column vectors as latent variables

Table 4.1: Symbols

The generalized canonical polyadic (GCP) decomposition approximates a N-way observed tensor $\boldsymbol{\mathcal{X}}$ of size $n_1 \times n_2 \times \ldots \times n_N$ by the sum of R rank-1 tensors as model $\boldsymbol{\mathcal{M}}$, where R is smaller or equal to the rank of tensor $\boldsymbol{\mathcal{X}}$, as shown in Figure 4.2 and presented as follows [71]:

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{M}} = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(N)} = [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!],$$

where $\mathbf{a}_r^k$ indicates the r^{th} column of $\mathbf{A}^{(k)}$ for all $k = 1, \ldots, N$ and $r = 1, \ldots, R$. $\mathbf{A}^{(k)}$ is the k-mode factor matrix of size $I_k \times R$, $k = 1, \ldots, N$, consisting of R latent components or phenotypes vectors, expressed as follows:

$$\mathbf{A}^{(k)} = \left[\mathbf{a}_1^{(k)} \cdots \mathbf{a}_R^{(k)}\right].$$

In addition, it is often convenient to express the decomposition with a positive weights vector of $\boldsymbol{\lambda}$ as follows:

$$\mathcal{X} \approx \mathcal{M} = [\![\boldsymbol{\lambda}; \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!] = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(N)}.$$

Assume we have a probability density function (PDF) or probability mass function (PMF) that gives the likelihood of each entry, as follows:

$$x_i \sim p\left(x_i \mid \theta_i\right) \quad \text{where} \quad f\left(\theta_i\right) = m_i,$$

where x_i is an observation of a random variable and $f(\cdot)$ is an invertible link function that relates the model parameter m_i to the corresponding natural parameter of the distribution, θ_i. We aim to find the model $\mathcal{M}$ that is the maximum likelihood estimate (MLE) over all entries. The conditional independence of observations implies that the overall likelihood is the product of the likelihoods. We next convert the maximizing problem to a minimization problem, which is a common approach in optimization. The GCP decomposition is then carried out by minimizing the negative log-likelihood over all entries, which is called a loss function. Meaning that finding factor matrices $\mathbf{A}^{(k)} \in \mathbb{R}^{I_k \times R}$ for $k = 1, \ldots, N$ with a given R such that solves the following optimization problem:

$$\min L(\mathcal{M}; \mathcal{X}) \equiv \sum_{i_N=1}^{n_N} \cdots \sum_{i_1=1}^{n_1} \ell\left(x_i, m_i\right), \quad \text{subject to} \quad \mathcal{M} = [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!],$$

where $x_i = x\left(i_1, \ldots, i_N\right), m_i = m\left(i_1, \ldots, i_N\right)$, and $\ell(x_i, m_i) \equiv -\log p\left(x_i \mid f^{-1}(m_i)\right)$ is the loss function. The form of data may dictate the choice of loss function from the below table [71, Table 1]. Thus, the loss function lets us adjust the model to data's nature.

Distribution	Data type	Link function	Loss function	Constraints
$\mathcal{N}(\mu, \sigma)$	Continuous	$m = \mu$	$(x-m)^2$	$x, m \in \mathbb{R}$
Gamma(k, σ)	Positive continuous	$m = k\sigma$	$x/(m+\epsilon) + \log(m+\epsilon)$	$x > 0, m \geq 0$
Rayleigh (θ)	Non-negative continuous	$m = \sqrt{\pi/2}\theta$	$2\log(m+\epsilon) + (\pi/4)(x/(m+\epsilon))^2$	$x > 0, m \geq 0$
Poisson (λ)	Count	$m = \lambda$	$m - x\log(m+\epsilon)$	$x \in \mathbb{N}_0, m \geq 0$
		$m = \log\lambda$	$e^m - xm$	$x \in \mathbb{N}_0, m \in \mathbb{R}$
Bernoulli (ρ)	Boolean	$m = \rho/(1-\rho)$	$\log(m+1) - x\log(m+\epsilon)$	$x \in \{0,1\}, m \geq 0$
		$m = \log(\rho/(1-\rho))$	$\log(1+e^m) - xm$	$x \in \{0,1\}, m \in \mathbb{R}$
NegBinom (r, ρ)	Count	$m = \rho/(1-\rho)$	$(r+x)\log(1+m) - x\log(m+\epsilon)$	$x \in \mathbb{N}_0, m \geq 0$

Table 4.2: Loss functions. ϵ is a numerical adjustment. Blue colour parameters are supposed to be constant.

In the following, we will discuss how different distributions and their corresponding probabilities naturally determine the elementwise loss function ℓ. Every distribution has its own standard notation for the generic parameter. We show that the L^2 loss function, $(x-m)^2$, is derived from the assumption that the data is Gaussian distributed. Consider the normal or Gaussian distribution, denoted by $\mathcal{N}(\mu, \sigma)$, where μ represents the mean and σ represents the standard deviation. We can assume that σ remains constant for all entries.

$$x_i \sim \mathcal{N}(\mu_i, \sigma), \quad \text{with} \quad \mu_i = m_i \quad \text{for all} \quad i \in \Omega,$$

where Ω is the set of indices for which the values of $\mathfrak{X}$ are known. Given the link function, the data's Gaussian distribution may be expressed as follows.

$$x_i = m_i + \epsilon_i, \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, \sigma) \quad \text{for all} \quad i \in \Omega.$$

The link function connecting the natural parameter μ_i and the model m_i is just the identity

function, $\mu_i = m_i$. The probability density function for the normal distribution $\mathcal{N}(\sigma, \mu)$ can be derived as:

$$p(x \mid \mu, \sigma) = e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}.$$

The loss function is then computed elementwise as:

$$\ell(x, m) = (x - m)^2/\left(2\sigma^2\right) + \frac{1}{2}\log\left(2\pi\sigma^2\right),$$

σ remains constant, thus it does not affect the optimization process. Therefore, we can eliminate those terms and simplify the equation to its standard form.

$$\ell(x, m) = (x - m)^2, \quad \text{for} \quad x, m \in \mathbb{R}.$$

Thus, we often refer to L^2 loss as Gaussian loss.

Count data can be modeled using the Poisson distribution. If the tensor values are counts, they can be modeled as a Poisson distribution. This distribution is often used to describe the number of events that occurred in a specific time window, such as the number of emails per month. The probability mass function (PMF) for a Poisson distribution with a mean value of λ is expressed as:

$$p(x \mid \lambda) = e^{-\lambda}\lambda^x/x!, \quad \text{for} \quad x \in \mathbb{N}_0.$$

Using the identity link function $f(\lambda) = \lambda$ while omitting constant terms, we obtain:

$$\ell(x, m) = m - x\log m, \quad \text{for} \quad x \in \mathbb{N}_0,\ m \geq 0.$$

There is another option for the link function, which is the logarithmic link. It is represented as $f(\lambda) = \log\lambda$. Then, the loss function is as follows:

$$\ell(x, m) = e^m - xm, \quad \text{for} \quad x \in \mathbb{N}_0,\ m \in \mathbb{R}.$$

The benefit of this loss function is that m is unconstrained.

Moreover, we can choose the loss function arbitrary, e.g., a robust loss function like Huber can be selected to minimize outliers' influence in noisy data such as EHRs. For instance, Gamma distribution is suitable for positive continuous data, and if the feature values are counts, they can be modeled as a Poisson or a Negative Binomial distribution. The robust loss function, Huber, which is less sensitive to outliers [73] and can be considered as a smooth approximation of an L^1 loss, is as follows:

$$\ell(x, m; \Delta) = \begin{cases} (x-m)^2 & \text{if } |x-m| \le \Delta \\ 2\Delta|x-m| - \Delta^2 & \text{otherwise.} \end{cases}$$

Modified Huber loss is used for binary classification problems with $x \in \{\pm 1\}$. Refer to [165] for the use of modified Huber loss in stochastic gradient descent algorithm. The modified Huber loss is:

$$\ell(x, m) = \begin{cases} \max(0, 1-xm)^2 & \text{for } xm \ge -1 \\ -4x\,m & \text{otherwise.} \end{cases}$$

Another choice is β-divergence, which has been popular in matrix and tensor factorization [51]. The formula is given only depending on x:

$$\ell(x, m; \beta) = \begin{cases} \frac{1}{\beta} m^\beta - \frac{1}{\beta - 1} x m^{\beta - 1} & \text{if } \beta \in \mathbb{R} \backslash \{0, 1\} \\ m - x \log m & \text{if } \beta = 1 \\ \frac{x}{m} + \log m & \text{if } \beta = 0. \end{cases}$$

β-divergence has shown better results than the other loss functions for the integer data type [71]. Furthermore, in a heterogeneous data set, a different loss function for each entry could

be easily defined as $\ell_i(x_i, m_i)$. It is worth noting that the loss function should reflect the intuitive concept of what is meant by "fit the data well" [153].

For simplicity, we consider a 3-way tensor scenario; however, our model generalizes to N modes as well. The GCP decomposes the observed longitudinal health data $\mathcal{X}$ into 3-factor matrices: a patient factor matrix $\mathbf{A}$ and 2 nonpatient factor matrices $\mathbf{B}$ and $\mathbf{C}$. For a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, the generalized CP decomposition with weight vector $\boldsymbol{\lambda} = \mathbf{1}$ is represented as (for simplicity, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ notations are used rather than $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)}$)

$$\mathcal{X} \approx \mathcal{M} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r,$$

where $R = \text{rank}(\mathcal{X})$, and $\mathbf{a}_r \in \mathbb{R}^I, \mathbf{b}_r \in \mathbb{R}^J, \mathbf{c}_r \in \mathbb{R}^K$ are the r^{th} column vectors within the factor matrices $\mathbf{A} \in \mathbb{R}^{I \times R}$, $\mathbf{B} \in \mathbb{R}^{J \times R}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$, respectively.

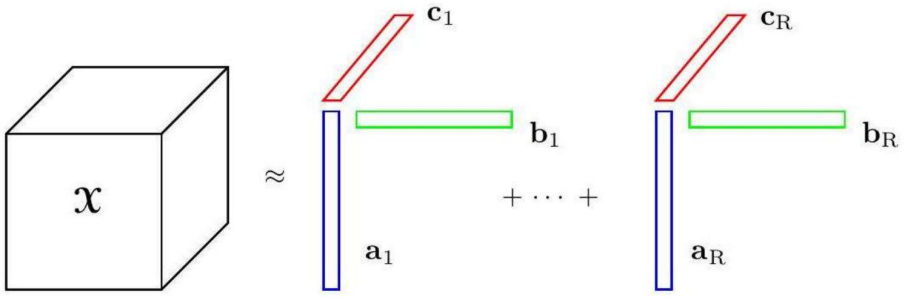


Figure 4.2: The generalized CP decomposition.

4.3 Method

In this section, we present the model that we used to generate synthetic longitudinal health data using generalized CP decomposition, along with 3 sampling and simulation techniques for the latent space of GCP. We also describe the data, the utility metrics, and the experimental details of our study [84]. There are a bunch of simple situations, such as when there is no unique or direct correspondence between the records in the synthetic data and the original data, where pre-existing SDG strategies have demonstrated low identity disclosure risks in evaluations [129, 48, 72, 126]. This effectively could decrease the kind of record linkage attacks that have challenged traditional disclosure control approaches, as attackers cannot easily match synthetic records to outside data [10].

First of all, it is important that our model satisfies the situation outlined above. Relying solely on the model tensor $\mathcal{M}$ for synthesis is an incorrect approach. This is because the elements of $\mathcal{M}$ directly approximate the elements of the actual tensor $\mathcal{X}$, with a one-to-one mapping and direct correspondence between the entries of the GCP model and the entries of the original data. We fix this by simulating and modeling a latent factor matrix of GCP decomposition linked to patients, which was inspired by studies [96, 139]. We focus on the patient factor matrix $\mathbf{A}$ in the model's latent space, which contains key phenotypes and information about patients. Still, if the model is overfitted to the original data, there is a possibility of generating a synthetic record that matches a real record [48], which can increase the risk of identity disclosure. Therefore, we must avoid overfitting the model in order to reduce the identity disclosure risk, which we will discuss shortly.

The patient factor matrix $\mathbf{A}$ is not a longitudinal data set and is a significantly smaller data set than the model tensor $\mathcal{M}$; hence its sampling and synthesis would be considerably simpler and more efficient. Thus, the aim will be to find the optimal sampling or synthesis technique for the patient factor matrix $\mathbf{A}$, which is denoted by $\hat{\mathbf{A}}$. As determined by our study [84], we may employ one of the methods listed below to create $\hat{\mathbf{A}}$. We then denote the synthetic version of $\mathcal{X}$ by $\hat{\mathcal{X}}$ and generate it as shown below and in Figure 4.3.

$$\hat{\mathcal{X}} \approx [\![\hat{\mathbf{A}}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \hat{\mathbf{a}}_r \circ \mathbf{b}_r \circ \mathbf{c}_r.$$

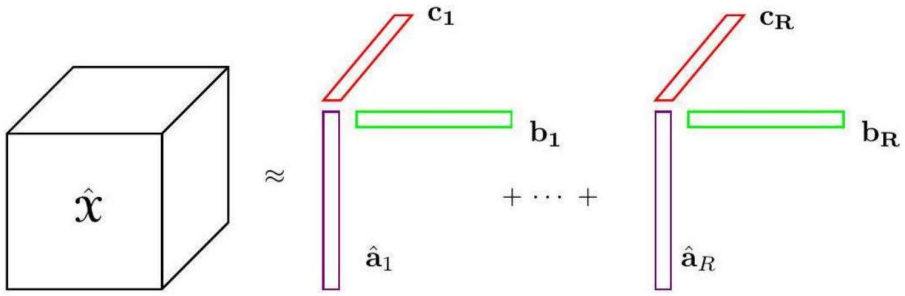


Figure 4.3: The generative model in terms of the generalized CP decomposition.

As previously stated, some studies [127, 157, 72] claim that fully synthetic data does not pose an elevated identity disclosure risk. However, when generating synthetic data, it is possible to overfit the synthesis model to the real data. This implies that the generated data will seem remarkably similar to the original data, resulting in a synthetic record being matched to a real individual [49], which poses a risk of identity disclosure. As a result, we implement the following heuristic, similar to the "elbow" rule, to determine the rank of the GCP. This will help in preventing overfitting the model, avoiding unintentional copying of rows from the original data, and eventually reducing the chance of identity disclosure.

We employed the elbow method to determine the rank of GCP decomposition, which has been used in similar studies on matrix factorization [58, 34]. The elbow method is widely used in cluster analysis and PCA to find the optimal number of clusters and components [149, 108]. The procedure involves plotting the explained variance or objective function versus the number of clusters or components. The ideal value is established by determining the point of observable change, known as the elbow. This technique finds a reasonable balance between accurately fitting the data and avoiding overfitting. Following the selection of the loss function, we attempted to find the rank R by doing several runs with different values of R, where $R \leq \min\{IJ, IK, JK\}$ and $\min\{IJ, IK, JK\}$ is a weak upper bound

on the maximum rank of a general third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$. We then selected the rank for which there were no significant changes in the objective function, the fit score "$1 - \| \mathcal{M} - \mathcal{X} \| \ / \|\mathcal{X}\|$" from that rank to higher ranks. This heuristic, which we also present as Algorithm 4.1, allows important features and phenotypes to be captured by the model and also avoids overfitting, which prevents the model from becoming a close approximation of the actual data.

Algorithm 4.1 Heuristic for choosing rank R for a 3-way tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ to avoid overfitting of the model

Input: $\epsilon > 0$, $1 \leq R_1 \leq \min\{IJ, IK, JK\}$ ▷ ϵ acts as the overfitting controller and R_1 is an integer

function FIND REASONABLE RANK(R)
$R_a \leftarrow R_1 + 1$
$R \leftarrow R_1$
$R_b \leftarrow R_1 - 1$
Fit(R)=Run GCP with rank R
f=The fit score of Fit(R)
Fit(R_b)=Run GCP with rank R_b
f_b=The fit score of Fit(R_b)
Fit(R_a)=Run GCP with rank R_a
f_a=The fit score of Fit(R_a)
$m \leftarrow 0$
$n \leftarrow 0$
while $m = 0$ **do**
Fit(R) ▷ Run GCP with rank R
Fit(R_b) ▷ Run GCP with rank R_b
Fit(R_a) ▷ Run GCP with rank R_a
if $|f_b - f| \leq \epsilon$ **then**
while $n = 0$ **do**
$R_a \leftarrow R$
$R \leftarrow R_b$
$R_b \leftarrow R - 1$
Fit(R) ▷ Run GCP with rank R
Fit(R_b) ▷ Run GCP with rank R_b
Fit(R_a) ▷ Run GCP with rank R_a
if $|f_b - f| > \epsilon$ **then**
$n \leftarrow 1$
$m \leftarrow 1$
end if
end while
else if $|f_a - f| \leq \epsilon$ **then**
$R \leftarrow R_a$
$m \leftarrow 1$
else
$R_b \leftarrow R_a$
$R \leftarrow \min(R_a + 1, \min\{IJ, IK, JK\})$
$R_a \leftarrow R + 1$
end if
end while
return R
end function

For addressing missing observations in the GCP model, Hong et al. [71] used an indicator for the missingness of observations by assigning weights 0 if an observation is missing or 1 if it is observed; this approach for handling missing data is essentially the same as the work of Acar et al. [2]. Besides, the missing structure of synthetic data must be similar to that of the original data. We assigned weights 0 or 1 to every element of the tensor $\boldsymbol{\mathcal{X}}, x_i$:

$$w_i = \begin{cases} 1 & \text{if } x_i \text{ is observed,} \\ 0 & \text{if } x_i \text{ is missing.} \end{cases}$$

Let $\boldsymbol{\mathcal{W}}$ be the missingness indicator or mask tensor made of weights w_i, we then treated it as an input tensor and decomposed it using GCP decomposition. Assume $\mathbf{A}_{\boldsymbol{w}} \in \mathbb{R}^{I \times \hat{R}}$ to be its patient factor matrix attained from the GCP decomposition with rank $\hat{R}$:

$$\boldsymbol{\mathcal{W}} \approx [\![\mathbf{A}_w, \mathbf{B}_w, \mathbf{C}_w]\!].$$

We combined $\mathbf{A}_{\boldsymbol{w}}$ and $\mathbf{A}$ to create a new patient factor matrix $\mathbf{A}_{\text{mix}} = [\mathbf{A}_{\boldsymbol{w}}, \mathbf{A}] \in \mathbb{R}^{I \times (R+\hat{R})}$, from which we simulated to obtain the synthesized form of $\hat{\mathbf{A}}_{\text{mix}} = \left[\hat{\mathbf{A}}_{\boldsymbol{w}}, \hat{\mathbf{A}}\right]$. Then, we restored $\hat{\mathbf{A}}_{\boldsymbol{w}}$ and $\hat{\mathbf{A}}$ to their source latent space in order to obtain the synthetic data and its missingness indicator tensor $\hat{\boldsymbol{\mathcal{W}}}$.

The electronic health records might be an irregular tensor, with patients having a varied number of clinical visits. However, for CP and generalized CP decompositions, the input must be a regular tensor. Therefore, we needed to convert the irregular observed tensor into the regular one. We started by finding the maximum number of visits among all patients. We then added extra missing visits at the end of patients' records if their total number of visits was less than the maximum, resulting in all patients having the same number of visits. We then performed the GCP decomposition and added the number of clinical visits as a new variable to the patient factor matrix $\mathbf{A}$ because this feature is not a longitudinal variable and can be directly added to the patient factor matrix. We then sampled or synthesized the $\mathbf{A}$ along with that variable, and obtained the synthetic number of records. Finally, we

modified the number of records in the post-processing. We also recommend applying the same to the baseline attributes. We can simulate different numbers of patients in synthetic data as well with this generative model.

Our generative model generates synthetic data through the following steps: GCP decomposition and synthesis or sampling of patient latent variables using the methods provided in the next subsections 4.3.1, 4.3.2, 4.3.3. Once the simulated patient latent variables have been inserted into the latent space of the GCP decomposition, we then convert the latent space back into a tensor. At last, after going through post-processing, we obtain the synthesized longitudinal data.

4.3.1 Sequential Decision Trees

We synthesized the patient factor matrix $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_R]$ using sequential decision trees presented in [50]. In data synthesis, sequential decision tree-based synthesizers outperform deep learning methods, such as generative adversarial networks (GANs) or recurrent neural networks (RNNs) when the data set is not huge and not longitudinal. Thus, it seems suitable for sampling and synthesizing the patient factor matrix $\mathbf{A}$. We discuss the sequential synthesis [50] that we applied on the patient factor matrix $\mathbf{A}$ in the following:

Assume that the variables in the patient factor matrix $\mathbf{A}$ are $\mathbf{a}_1, \ldots, \mathbf{a}_R$, and their corresponding synthesized versions are $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_R$. We fit $R-1$ models using a model M_j defined by $\mathbf{a}_{j+1} \sim f(\{\forall i \leq j : \mathbf{a}_i\})$, where M_j and $f(\cdot)$ denote a decision tree model and the tree model fitting function, respectively. We then sample $\hat{\mathbf{a}}_1$ from $\mathbf{a}_1$, and apply $\hat{\mathbf{a}}_{j+1} = M_j(\{\forall i \leq j : \hat{\mathbf{a}}_i\})$ to sample the remaining variables. In this method, classification and regression trees are applied depending on the kind of data. Thus, it doesn't matter if the variables are continuous or categorical. A greedy approach [17] is used to create classification and regression trees (CART) [128] by repeatedly separating variables until a stopping requirement is met. At each split, the variable that minimizes a loss function is chosen. The goal is to identify the best binary splits for each variable. The particle swarm method [15]

is used to optimize the variable sequence. To avoid overfitting, tree pruning is carried out using a cost-complexity criteria [17]. Supplementary Appendix B in [50] provides further details of this technique.

The following example is similar to the one in [50, Figure 1] and shows the sequential synthesis process for a data set with four variables, T through X. The fitting step builds four models, $\{M_1, M_2, M_3, M_4\}$.

Example 4.3.1. *Let's consider five variables: T, U, V, W, X. The generation is done in a sequential manner, so it is necessary to have a specific order. There are several factors that can be considered when selecting a sequence. The sequence $T \rightarrow X \rightarrow V \rightarrow U \rightarrow W$ is defined for our example. Prime notation marks the variables as synthesized. To illustrate, T' denotes the synthesized version of T. Fitting and synthesis are the two main components of the generating process. The procedure for sequential generation is as follows:*

Fitting:

In this phase, four decision tree models are constructed: M_1 to M_4.

- *Build a model M_1, the model takes as input T and the outcome as X ($X \sim T$)*
- *Build a model M_2, the model takes as input T and X and the outcome as V ($V \sim T+X$)*
- *Build a model M_3, the model takes as input T, X, and V and the outcome as U ($U \sim T + X + V$)*
- *Build a model M_4, the model takes as input T, X, V, and U and the outcome as W ($W \sim T + X + V + U$)*

Synthesis:

- *Sample from the distribution of the first variable T to get T'*
- *Synthesize X as $X' = M_1(T')$*

- *Synthesize* V *as* $V' = M_2\,(T', X')$
- *Synthesize* U *as* $U' = M_3\,(T', X', V')$
- *Synthesize* W *as* $W' = M_4\,(T', X', V', U')$

The four decision tree models $\{M_1, M_2, M_3,\ M_4\}$ *make up the overall generative model.*

The particle swarm optimization[15] is faster than a random search for an acceptable variables order [50]. We refer to Appendix A for the details of this optimization in sequential synthesis.

4.3.2 Hamiltonian Monte Carlo (HMC)

There are many MCMC algorithms for sampling from a probability distribution [19, 132]. Hamiltonian Monte Carlo (HMC), a Markov chain Monte Carlo (MCMC) algorithm, is a widely used approach that is generally considered state-of-the-art. Several famous software programs, including Stan [21] and Tensorflow [1], use it as their default sampler for challenging distributions.

Applying Hamiltonian dynamics in simulation was initially introduced by Alder and Wainwright [6] in physics field. The hybrid Monte Carlo method, which combines MCMC with Hamiltonian dynamics, was introduced by Duane et al. [43]. The method, which Neal [109] improved, is now referred to as Hamiltonian Monte Carlo in the statistical field. HMC uses first-order gradient information to prevent random walk behavior and sensitivity to correlated parameters, which are common in other MCMC methods. These characteristics enable it to converge to high-dimensional target distributions considerably faster than simpler approaches like random walk Metropolis or Gibbs sampler [70].

The goal of sampling is to draw from a target density $\pi(q)$ for parameters q. The HMC algorithm introduces auxiliary momentum variables, p, to the parameters of the target distribution, q, with the joint density:

$$\pi(p, q) = \pi(p \mid q)\pi(q),$$

the joint density defines a Hamiltonian:

$$\begin{aligned}\mathcal{H}(p, q) &\equiv -\log \pi(p, q) \\ &= -\log \pi(p \mid q) - \log \pi(q) \\ &\equiv K(p \mid q) + V(q),\end{aligned}$$

where $K(p \mid q) \equiv -\log \pi(p \mid q)$ is the kinetic energy, and $V(q) \equiv -\log \pi(q)$ is the potential energy. The potential energy is computed by the target distribution while the kinetic energy is specified by the implementation.

For $\mathbf{q}(t) \in \mathbb{R}^d$ and $\mathbf{p}(t) \in \mathbb{R}^d$ as variables that describe the evolution of a state vector and its momentum over time through Hamilton's equations:

$$\frac{d\mathbf{q}}{dt}(t) = \frac{\partial \mathcal{H}}{\partial \mathbf{p}}(\mathbf{p}(t), \mathbf{q}(t)), \quad \text{and} \quad \frac{d\mathbf{p}}{dt}(t) = -\frac{\partial \mathcal{H}}{\partial \mathbf{q}}(\mathbf{p}(t), \mathbf{q}(t)).$$

$$\mathcal{H}(\mathbf{p}, \mathbf{q}) = K(\mathbf{p} \mid \mathbf{q}) + V(\mathbf{q}).$$

We picked Hamiltonian Monte Carlo (HMC) as another suitable latent-space sampler for our generative model [84] because all latent variables are continuous, and it also has advantages over other Markov chain Monte Carlo algorithms.

The distribution of patient latent variables depends on the kind of loss function used in the GCP decomposition. We employ this method on the latent space generated by Gaussian loss in the GCP decomposition. After pre-processing and zero-mean normalization of patient latent variables, we assume that the patient factor matrix variables (patient latent variables) $\mathbf{a}_1, \ldots, \mathbf{a}_R$ are generated from a multivariate Gaussian distribution with a known mean vector $\boldsymbol{\mu}$ of $\mu_{\mathbf{a}_1}, \ldots, \mu_{\mathbf{a}_R}$ (which are set to zero) and an unknown covariance matrix $\boldsymbol{\Sigma}$.

The covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as follows: $\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{C}\boldsymbol{\sigma}$, where $\boldsymbol{C}$ is a correlation matrix and $\boldsymbol{\sigma} = \sqrt{\text{diag}(\boldsymbol{\Sigma})}$. Diagonal entries of the covariance matrix $\boldsymbol{\Sigma}$ are denoted as $\Sigma_{i,i}$, representing the variance of latent variables. On the other hand, the off-diagonal entries, denoted as $\Sigma_{i,j}$, indicate the covariance between variables. Further, let denote the off-diagonal entries of $\boldsymbol{C}$ by $C_{i,j}$. Hence

$$\sigma_i = \sqrt{\Sigma_{i,i}}, \quad \text{and} \quad C_{i,j} = \frac{\Sigma_{i,j}}{\sigma_i\,\sigma_j}.$$

A suitable prior for standard deviations, σ_i, $i = 1, \ldots, R$ would be a weakly-informative prior like the half-Cauchy distribution with the non-negative support and small scale s. The half-Cauchy is a sensible default prior for scale parameters in hierarchical models [122, 100, 53]. Additionally, using weakly informative priors in MCMC algorithms has practical advantages. Further, it is recommended to give the correlation matrix $\boldsymbol{C}$ a Lewandowski-Kurowicka-Joe (LKJ) distribution prior [156] with shape $\eta > 0$, where $\text{LKJ}(\boldsymbol{C} \mid \eta) \propto \det(\boldsymbol{C})^{(\eta-1)}$. The LKJ distribution is a probability distribution over positive definite symmetric matrices with unit diagonals. Hence, we apply the following model:

$$\begin{aligned}
&\mathbf{a}_1, \ldots, \mathbf{a}_R \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where } \quad \boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{C}\boldsymbol{\sigma}, \\
&\qquad \sigma_i \sim \text{Cauchy}(0, s) \text{ constrained by } \sigma_i > 0, \qquad \text{for } i = 1, \ldots, R \\
&\qquad \boldsymbol{C} \sim \text{LKJ}(\eta).
\end{aligned}$$

We then implement the model using Hamiltonian Monte Carlo (HMC). We apply the No-U-Turn Sampler (NUTS) [70], an extension of the Hamiltonian Monte Carlo (HMC) algorithm, with the auxiliary distribution $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$. We allow mass matrix $\mathbf{M}$ (more formally, Euclidean metric) to be estimated from warmup draws and restricted to a diagonal matrix.

Furthermore, NUTS eliminates the need to specify the number of steps and step size parameters while still maintaining (and sometimes enhancing) HMC's capability to efficiently

generate independent samples. We employ an approach that automatically adjusts the step size shared by NUTS and HMC. This is done through an adaptation of the dual averaging algorithm of Nesterov [111], allowing NUTS to be run without any manual tuning. The algorithm of NUTS-HMC with dual averaging is provided in Algorithm 6 in [70]. We implement this algorithm using Stan, a powerful probabilistic programming language for building statistical models [21] and state-of-the-art automatic differentiation techniques to compute the derivatives in HMC without user input. The Stan model block is provided in Appendix B. Stan, similar to other HMC implementations, uses the leapfrog integrator, a method for numerical integration that is meant to keep Hamiltonian systems of equations stable.

4.3.3 Copula

A copula is a Latin phrase that means link. Copula has recently been used in a variety of fields, including econometric modeling and quantitative risk management, because of its ability to capture the core of multivariate data distributions and their connections. Sklar [142] developed this notion in statistical modeling in 1959.

Theorem 4.3.2 ([142], Sklar's theorem)**.** *Let* $(X_1, \ldots, X_j, \ldots, X_d)$ *be a* d*-dimensional random vector with joint distribution function* H *and marginal distribution functions* $F_i, i = 1, \ldots, d$*. Then there exists a* d*-copula* $C : [0,1]^d \to [0,1]$*, such that for all* x *in* $\mathbb{R}^d$*, the joint distribution function can be expressed as:*

$$H(x_1, \ldots, x_j, \ldots, x_d) = C(F_1(x_1), \ldots, F_j(x_j), \ldots, F_d(x_d)),$$

with associated density function h*, expressed by the multiplication of the copula density function* c *and marginal densities* $f_i, i = 1, \ldots, d$*:*

$$h(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \times \prod_{i=1}^{d} f_i(x_i).$$

Copula-based synthetic data generation has gained significant attention in recent years. Unlike other generative models like generative adversarial networks, the copula is an appropriate choice for synthesizing non-longitudinal and small data sets [11]. Copula models are useful sampling approaches for describing dependencies and marginal distributions.

Therefore, we chose the Gaussian copula [92, 114] as one of the techniques for simulating and synthesizing the patient factor matrix $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_R]$. Below, we provide the details of the copula used in the generative model of our study [84], which is also presented as Algorithm 4.2 with details.

Assuming that $\mathbf{a}_1, \ldots, \mathbf{a}_R$ are latent variables with marginal distributions $F_1, \ldots, F_R$ (depending on the choice of the objective function in GCP), and covariances $\mathrm{Cov}(\mathbf{a}_i, \mathbf{a}_j)$, $i \neq j$. Then, using a Gaussian copula:

1. Generate variables $\mathbf{t}_1, \ldots, \mathbf{t}_R$ from a multivariate normal distribution with means all equal to 0, variances all equal to 1, and $\mathrm{Cov}(\mathbf{t}_i, \mathbf{t}_j) = \mathrm{Cov}(\mathbf{a}_i, \mathbf{a}_j), \forall i \neq j$.

2. Denote the cumulative distribution function of standard normal distribution by Φ. Generate the uniform variables such that $\mathbf{u}_i = \Phi(\mathbf{t}_i), i = 1, \ldots, R$; the Gaussian copula is the joint distribution of $\mathbf{u}_1, \ldots, \mathbf{u}_R$.

3. Generate $\hat{\mathbf{a}}_i = F_i^{-1}(\mathbf{u}_i), \forall\, i = 1, \ldots, R$, where F_i^{-1} denotes the inverse CDF of the marginal distribution of $\mathbf{a}_i$.

We used parametric and nonparametric marginals F_i. For the nonparametric marginals, we used the empirical CDF. We employed Gamma, Beta, and truncated Gaussian distributions as parametric marginals. The choice of the objective function in Section 4.2 determines which parametric marginal distribution should be used. The analysis showed that the empirical CDF marginals outperformed parametric ones.

The sampling will never be exact, and because we are sampling in the latent space, the correlation of the synthetic tensor might not be what we want. Therefore, we selected a

sample such that the Frobenius norm (the Euclidean norm generalised to matrices) of the difference between the correlation matrices of the original and synthetic data is less than or equal to a threshold ϵ, which can be viewed as an optimization. Finally, we produce the synthetic patient factor matrix from the obtained samples $\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_R]$. We refer to Algorithm 4.2 for a greater overview of our approach.

Algorithm 4.2 Copula sampling from latent variables in $\mathbf{A}$

Input: Patient factor matrix $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_R]$ as a data set with R latent variables, ϵ a threshold for the optimization, $Corr_{\text{Original}}$ the correlation matrix of the original data set.

function SIMULATED PATIENT FACTOR MATRIX($\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_R]$)

$\mathbf{a}_i = i$th latent variable ▷ $\mathbf{a}_i$ is the ith column of $\mathbf{A}$

$R =$ number of variables in $\mathbf{A}$

$\mathbf{\Sigma} = R \times R$ covariance matrix of latent variables $\mathbf{a}_1 \cdots \mathbf{a}_R$

$\Phi =$ cumulative distribution function (CDF) of a standard normal distribution

$F_i^{-1} =$ inverse CDF of the marginal distribution of $\mathbf{a}_i$ ▷ choose F_i as the nonparametric marginals using the empirical distribution and parametric marginals using distributions: Gamma, Beta, and Truncated Gaussian distributions

$m = 1$

while $m \neq 0$ **do**

Draw $\mathbf{t}_1, \cdots, \mathbf{t}_R \sim N(\mathbf{0}, \mathbf{\Sigma})$, where $N(\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes an R-dimensional Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$

for $i \in 1 \ldots R$ **do**

$\mathbf{u}_i = \Phi(\mathbf{t}_i)$

$\hat{\mathbf{a}}_i = F_i^{-1}(\mathbf{u}_i)$ ▷ This implies that variables of $\hat{\mathbf{a}}_i$ follows the desired distribution

end for

if $\|Corr_{\text{Original}} - Corr_{\text{Synthetic}}\|_F \leq \epsilon$ **then** ▷ Frobenius norm of the difference between the correlation matrices of the original and synthetic data set made of patient factor matrix of $[\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_R]$ needs to be less than or equal to a threshold ϵ so that we can choose that sample from Gaussian copula

$m = 0$

end if

end while

return $\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_R]$

end function

4.4 Data Utility

In this section, we describe how we evaluate the utility of the synthetic data set. The analysis of synthetic data should provide similar statistical inferences and conclusions to those obtained from actual data. Therefore, we evaluate our proposed model on the utility aspect of the generated data. We assess the ability of the model in terms of dependency structure and marginal fitting (univariate distribution similarity) [48, 57], using the following metrics:

- We may use bivariate metrics to validate that the same underlying relationship between variables exists in the synthetic data. The pairwise correlation difference evaluates the difference in correlation between actual and synthetic data.

  Assume X_R and X_S represent the real and synthetic data sets, respectively. We compute the absolute difference between the pairwise correlation in the original and synthetic data as follows:

  $$Corr_{\mathrm{X}_R,\mathrm{X}_S}(i,j) = |Corr_{\mathrm{X}_R}(i,j) - Corr_{\mathrm{X}_S}(i,j)| \quad \text{for each pair of variables } (i,j),$$

  where $Corr_{\mathrm{X}_R}(i,j)$ and $Corr_{\mathrm{X}_S}(i,j)$ denote the correlation between each pair of variables (i,j) in real and synthetic data sets, respectively. This yields a distinct value of $Corr_{\mathrm{X}_R,\mathrm{X}_S}(i,j)$ for each pair of variables, which is further analyzed using visualization tools like boxplot. To obtain optimal synthetic data, the median of $Corr_{\mathrm{X}_R,\mathrm{X}_S}(i,j)$ for every possible pair of variables (i,j) should be close to zero with minimal variation.

- The Hellinger distance between the synthetic and original variables indicates whether they are drawn from the same distribution. The Hellinger distance is a metric in the range of 0 to 1, where 0 indicates no difference between the distributions. This makes it easier to interpret. For a high-utility synthetic data set, the median of Hellinger distances for all variables should be close to zero, and the variation should be small.

This will indicate that the synthetic data properly reflects the distribution of each variable [50].

Consider the probability distributions P and Q over the same domain $\mathcal{X}$, the Hellinger distance is defined as:

$$H(P,Q) = \sqrt{1 - BC(P,Q)},$$

where $BC(P,Q)$ is the Bhattacharyya coefficient [14], and it can be computed for discrete probability distributions as follows:

$$BC(P,Q) = \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)},$$

where $P(x)$ and $Q(x)$ represent the probabilities of occurrence for observation x.

For continuous probability distributions, with $P(dx) = p(x)dx$ and $Q(dx) = q(x)dx$ where $p(x)$ and $q(x)$ are the probability density functions, the Bhattacharyya coefficient is defined as:

$$BC(P,Q) = \int_{\mathcal{X}} \sqrt{p(x)q(x)}dx.$$

- The root mean square difference between the real and synthetic correlations, denoted as RMSDC, is the root mean square difference between the correlations of the original variables and those of the associated synthetic variables. It is used to measure how well the dependency structure is captured. Assume n variables of $\{1, \ldots, n\}$ in the real data set X_R that we want to synthesize as synthetic data set X_S. We indicate $M = n(n-1)/2$ as the total number of possible distinct pairs of variables. Then, RMSDC is computed as:

$$\text{RMSDC} = \sqrt{\frac{1}{M} \sum_{i<j} \sum_{j} |Corr_{\mathrm{X}_R}(i,j) - Corr_{\mathrm{X}_S}(i,j)|^2},$$

where $Corr_{\mathrm{X}_R}(i,j)$ and $Corr_{\mathrm{X}_S}(i,j)$ denote the correlation between each pair of variables (i,j) in real and synthetic data sets, respectively. A lower RMSDC indicates a better capture of the dependency structure.

- Descriptive statistics.

4.5 Experimental Details

Initially, we arranged the study into trials on continuous data sets and experiments on categorical data. We imputed the continuous data set with missing observations using the generalized CP decomposition. This allowed us to obtain a data set for our analysis that had no missing values. We evaluated 3 simulation and sampling strategies described in Section 4.3 in our model on the GCP's patient factor matrix.

The GCP decomposition was conducted using various loss functions depending on the kind of variables in our trials. These included Gamma, β-divergence (similar to Gamma), and Gaussian with non-negativity constraints. Despite the continuous data set being non-negative, the Gaussian loss (L2 loss) outperformed the others for all 3 approaches. This is because the dependency structure and univariate distributions of the original variables were significantly better preserved in the generated synthetic data. Refer to Table 4.2 for the choices of the loss function.

The generalized canonical polyadic (GCP) tensor decomposition was conducted using the algorithm of Hong et al. [71], which is implemented in the Tensor Toolbox for MATLAB [9]. There is also a C++ programming software for generalized CP decompositions developed by Sandia National Laboratories [117], known as Genten, which is accessible in [118].

We applied sequential trees to validate the generative model on the data set with missing and irregular clinical visits. The categorical data trial was accomplished by GCP decomposition using the Poisson log link and Gaussian losses.

4.5.1 Data

For analysis, we used the MIMIC-III (version 1.4) data set [79], a publicly available anonymized EHRs dataset of intensive care unit patients. The continuous data set that we derived from the MIMIC-III data and employed to evaluate the performance of our proposed model was a 3-way tensor of laboratory measurements for patients within the hospital who had 36

clinical visits. The resulting data set was a tensor made up of 226 patients, 4 laboratory tests (creatinine, potassium, sodium, hematocrit), and 36 clinical visits, with 21% of the data missing. The categorical data set used in our analysis was a tensor consisting of 246 patients, 2 categorical features (admission type, admission location), and 5 clinical visits, with no missing records.

4.6 Results and Analysis

In this section, we present the findings and analysis of our experiments on generating synthetic longitudinal health data from our study [84]. We show that our method is capable of handling different data structures and scenarios. We conducted numerous experiments and considered the following structures for both the original data and the synthetic data in order to validate our model:

- Synthetic data for dense original data with continuous variables
- Synthetic data with a varying number of patients compared to the original data
- Synthetic data for original data with missing observations
- Synthetic data for original data with irregular clinical visits
- Synthetic data for original data with categorical variables

4.6.1 Experiments on Continuous Dense Data

According to the evaluations, β-divergence is not a suitable objective function for synthesizing continuous data in EHRs with patient factor matrix simulation using sequential trees, or HMC with the multivariate Gaussian distribution model. We refer to Appendix C for the findings.

We learned through several analyses that standardizing data and using Gaussian loss improves the results significantly. In the following, we present the outcomes of synthesizing a dense continuous data set that contains 226 patients, 4 laboratory test variables, and 5 clinical visits. In the preceding sections, we described how we obtained the data set for our study. The model used GCP decomposition with Gaussian loss and $R = 20$. It can be observed that synthetic data sets generated by all 3 patient factor matrix simulation

methods have comparable dependency structures and marginal distributions to the real one. The outcomes of copula, sequential trees, and the HMC can be found here.

According to Figure 4.11(B) and Figure 4.5, the synthetic data generated from copula using empirical CDF marginals almost preserves the dependency structure and distribution of the original variables. The box plots in Figure 4.4 represent the variation of Hellinger distance and Pearson correlation between variables in both the synthetic and original data sets. Further, RMSDC = 0.04 was obtained.

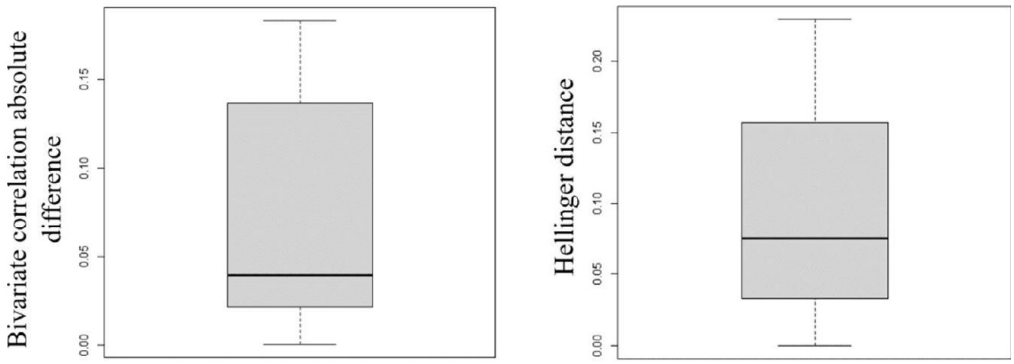


Figure 4.4: The box plots display the variation of the Hellinger distance and Pearson correlation between the original variables and the synthetic variables generated by copula. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance between the original variables and the synthetic variables. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

The following plot displays the different tensor modes, illustrating how the synthetic data generated from the copula with empirical CDF marginals almost preserves the structure and distribution of the original variables in trials on continuous dense data. The figure shows

that the synthetic data generated from the copula with empirical CDF marginals effectively maintains the distribution of the original variables in trials on continuous dense data, and the structure of the original data in different modes was preserved upon synthesis.

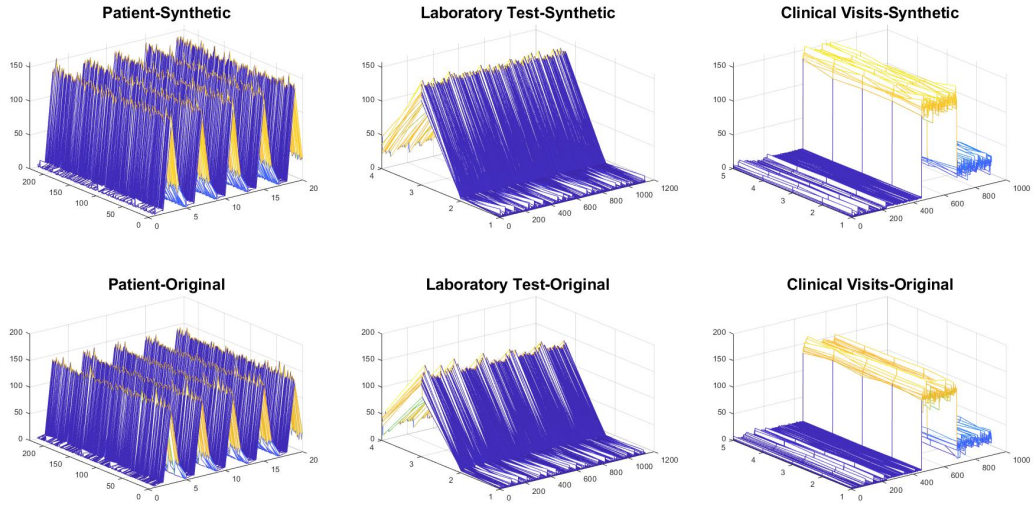


Figure 4.5: The different modes (Patients, Laboratory tests, and Clinical visits) of copula's generated data and the original data.

According to the summaries presented in Tables 4.3 and 4.4, the maximum value of the variables is slightly lower than that of the original.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	2.23	110.6	13.64
1st Q	0.52	3.77	134.5	27.97
Median	1.08	4.17	138	31.64
Mean	1.67	4.22	138	31.9
3rd Q	2.21	4.62	141.9	35.75
Max	14.1	7.32	157.5	48.49

Table 4.3: A summary of the copula's synthetic variables.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0.2	2.5	109	9.2
1st Q	0.7	3.8	135	28.1
Median	1	4.1	138	31.6
Mean	1.64	4.23	138.4	32.08
3rd Q	1.6	4.51	141.6	35.7
Max	16.2	10	170	52.6

Table 4.4: A summary of the original variables.

The following are the outcomes of sampling the patient factor matrix of the previously mentioned GCP decomposition using the sequential trees approach. According to Figure 4.6, the structure of the original data was preserved in different modes upon synthesis.

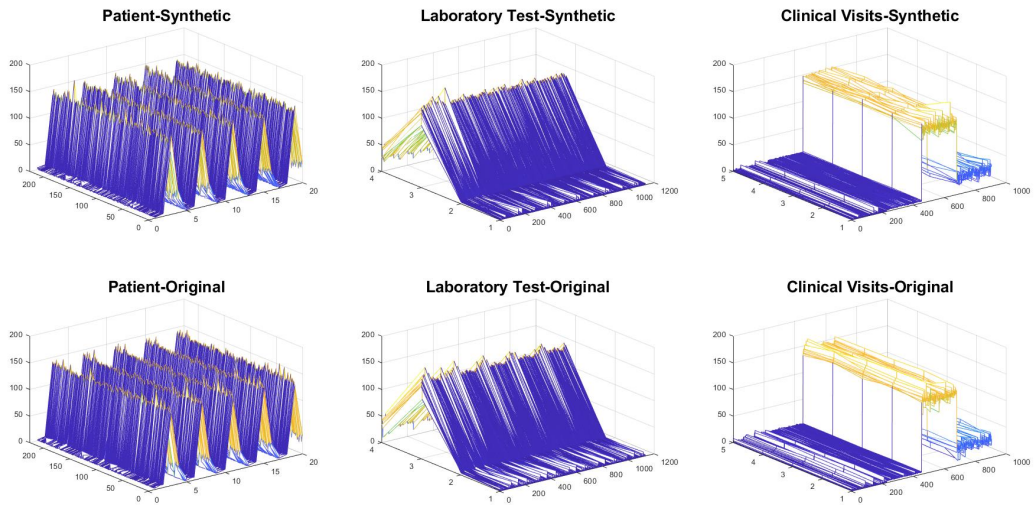
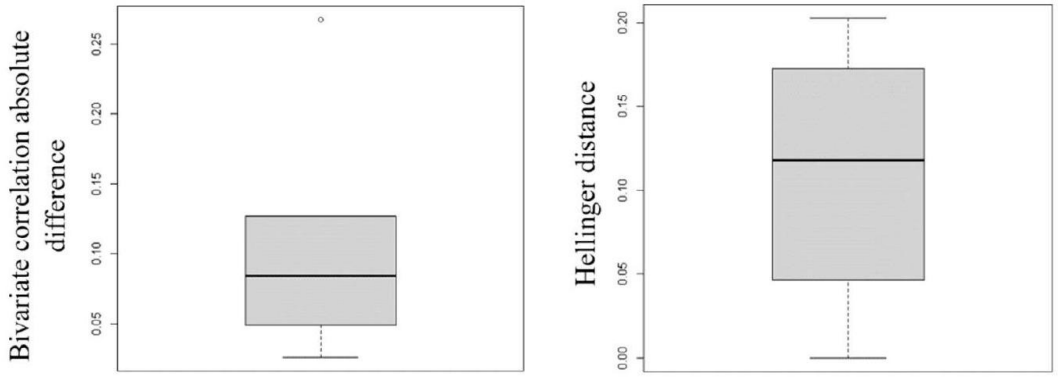


Figure 4.6: The different modes (Patients, Laboratory tests, and Clinical visits) of the sequential trees' generated data and the original data.

According to the above figure, the structure of the original data in different modes was preserved upon synthesis.

Figure 4.11(C) indicates that the Pearson correlations and univariate distributions are similar between the synthetic and original data sets shown in Figure 4.11(A). Figure 4.7 of the variation of the Hellinger distance also shows that synthetic variables from sequential decision trees are derived from a similar distribution as the original variables. However, the copula performed slightly better in capturing the correlations between the variables. The RMSDC computed for this experiment was 0.078.



Laboratory tests (A)

Laboratory tests (B)

Figure 4.7: The box plots display the variation of the Hellinger distance and Pearson correlation between the original variables and the synthetic variables generated by sequential decision trees. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance between the original variables and the synthetic variables. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

The summary in Tabel 4.5 shows that the range of variables has significantly improved compared to the previous copula analysis, refer to Table 4.4 for the summary of the original data.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	1.88	116.4	10.09
1st Q	0.67	3.7	134.2	27.01
Median	1.25	4.18	138.8	31.68
Mean	1.53	4.23	138.7	31.87
3rd Q	1.86	4.71	143.3	36.38
Max	14.84	8.16	177.4	54.8

Table 4.5: A summary of the sequential decision trees' synthetic variables.

The following is the result of the MCMC method using the Hamiltonian Monte Carlo algorithm. If the distribution of the Hamiltonian Monte Carlo (HMC) model is well specified, the resulting outcome will be more effective.

Figure 4.8 indicates that the structure of the synthetic data in different modes is comparable to that of the original data.

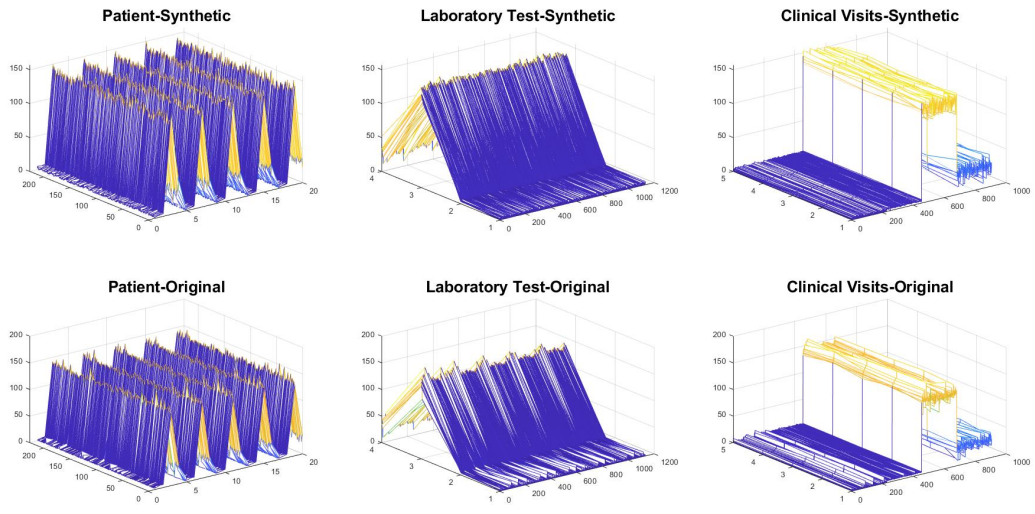


Figure 4.8: The different modes (Patients, Laboratory tests, and Clinical visits) of HMC's generated data and the original data.

Furthermore, the obtained RMSDC was 0.071. We expected the HMC to perform well on the Gaussian latent space, which is defined by our Gaussian model. Defining a proper model distribution for the HMC would significantly enhance the findings. Figure 4.9 shows that HMC was slightly better at capturing the correlations between variables.

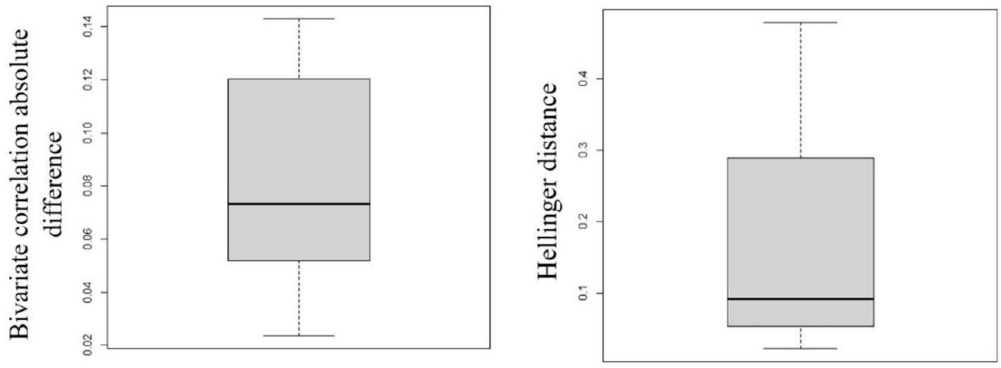


Figure 4.9: The box plots display the variation of the Hellinger distance and Pearson correlation between the original variables and HMC's synthetic variables. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance between the original variables and the synthetic variables. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

Based on Table 4.6, the summary of HMC synthetic variables is similar to that of the other methods. The maximum values of the variables dropped in the synthetic data compared to the actual data in Table 4.4.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	0.66	117.8	7.22
1st Q	0.6	3.74	134.5	27.85
Median	1.85	4.19	138.6	32.18
Mean	2.01	4.23	138.5	31.93
3rd Q	3.1	4.7	142.8	36.47
Max	7.24	6.78	156.9	51.63

Table 4.6: A summary of the HMC's synthetic variables.

We present Figures 4.11 and 4.10 for an easier comparison of the three sampling techniques on GCP decomposition with Gaussian loss. In Figure 4.11, we can observe the dependency structure and univariate distribution of both synthetic and original variables simultaneously.

Figure 4.10 demonstrates that all three synthetic data sets have similar statistical properties in terms of dependency and univariate distributions.

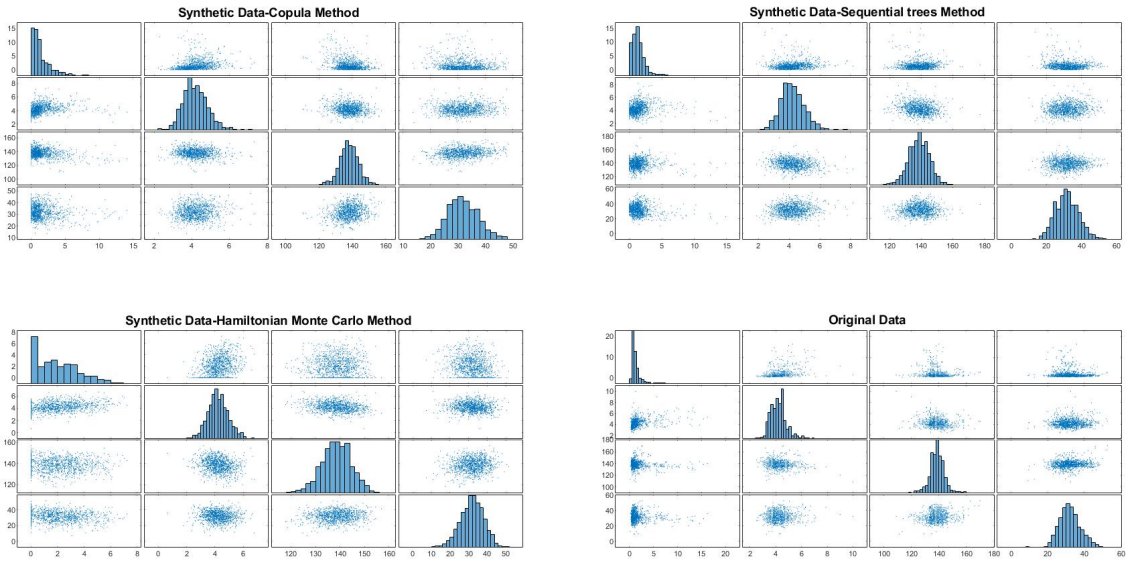


Figure 4.10: The distribution and scatter plots of the original variables and synthetic variables generated using copula, sequential trees, and HMC in experiments on continuous dense data.

The findings and figures demonstrate that all 3 synthetic data sets have similar statistical properties in terms of dependency and univariate distributions. However, the copula and sequential trees performed slightly better than the MCMC technique when the HMC algorithm was used. In brief, we need to define a proper distribution that corresponds to the latent space in order to obtain decent results through HMC sampling. The Gaussian loss is the ideal loss function for simulating from the patient factor matrix using sequential trees, but the observed data must be standardized beforehand. Further analysis, which we have not included in this thesis, has shown that it is preferable to use empirical CDF marginals instead of parametric ones when sampling the patient factor matrix by copula. The outcomes of generating synthetic data using β-loss in GCP decomposition can be found in C.

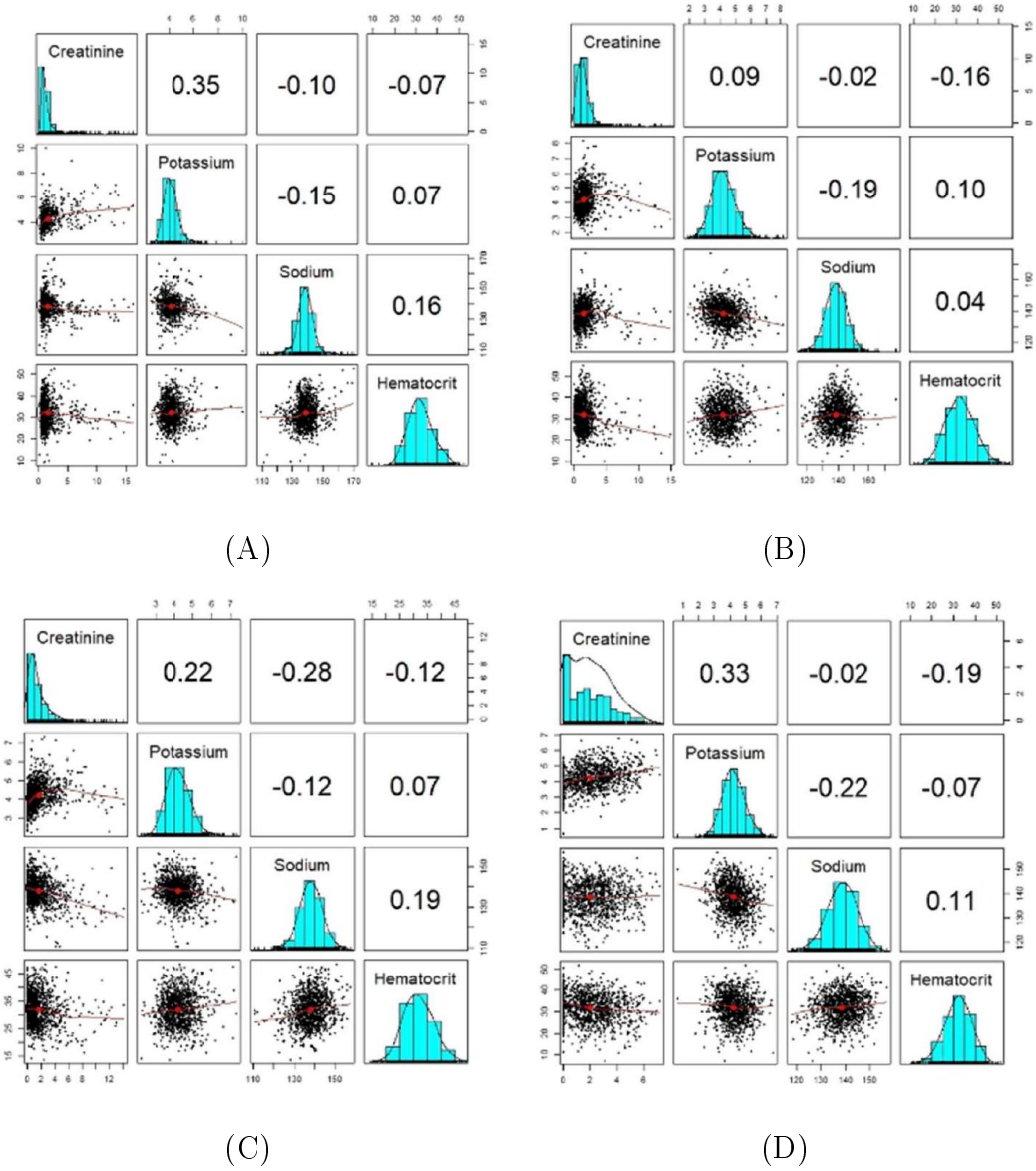


Figure 4.11: The plots show the correlation and distribution of the original and synthetic variables. A correlation matrix displaying bivariate scatter plots of the adjacent variables below the diagonal, histograms of the data distribution of the respective variables on the diagonal, and the Pearson correlation above the diagonal. Ellipses specify the direction of the correlation. The information regarding the relationship between two selected variables is always perpendicular to each other. (A) is the plot of the original variables. (B) is the plot of synthetic variables generated by sequential decision trees. (C) is the plot of synthetic variables generated by copula. (D) is the plot of synthetic variables generated by HMC.

The copula approach and sequential decision trees showed somewhat similar outcomes and demonstrated slightly better performance compared to Hamiltonian Monte Carlo (HMC) in these trials.

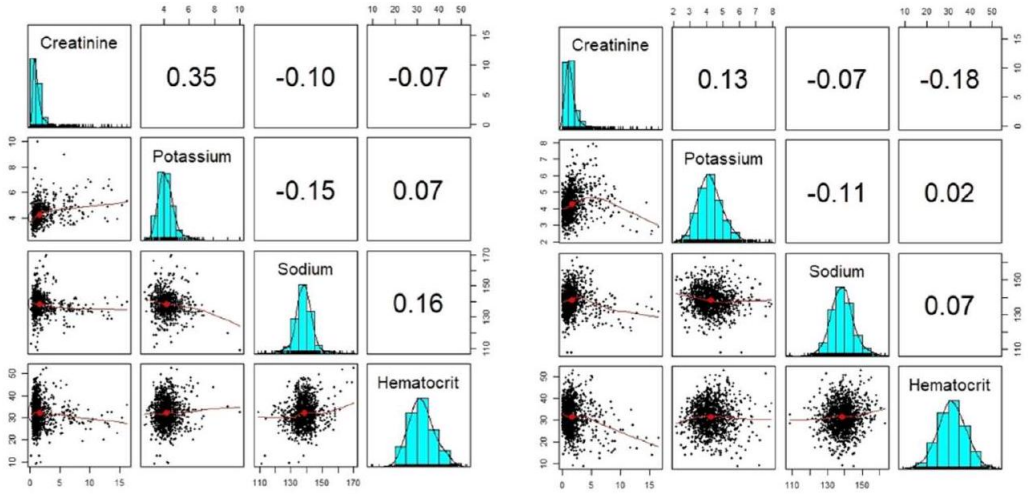
In the next section, we will provide the results of generating synthetic data with an additional number of patients compared to the original data set.

4.6.2 Experiments on Continuous Data with a Different Number of Patients in Synthetic Data Compared to Original Data

We can generate different numbers of patients in the synthetic data using all 3 patient factor matrix simulations. However, we only applied the sequential trees approach, and the results are as follows: the findings indicate that the dependency and univariate structure of the original variables are well maintained in the synthetic data.

The following are the outcomes of generating 250 patients from the patient factor matrix using sequential trees. The patient factor matrix is derived from the GCP decomposition in the previous experiment. The original data set is dense, consisting of 226 patients, and is the same data set used in the previous experiment.

Figure 4.12 indicates that the Pearson correlations and univariate distributions are similar between the synthetic and original data sets. Figure 4.13 illustrates that the synthetic variables created by sequential decision trees are derived from a distribution similar to that of the original variables. The RMSDC computed for this experiment was 0.066.



(A) (B)

Figure 4.12: The plots show the correlation and distribution of sequential decision trees' synthetic variables, as well as the original variables. A correlation matrix displaying bivariate scatter plots of the adjacent variables below the diagonal, histograms of the data distribution of the respective variables on the diagonal, and the Pearson correlation above the diagonal. Ellipses specify the direction of the correlation. The information regarding the relationship between two selected variables is always perpendicular to each other. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

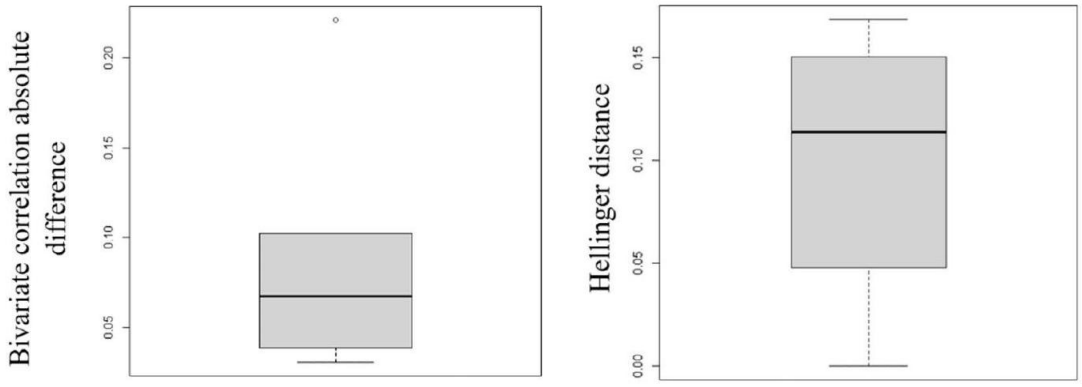


Figure 4.13: The box plots display the variation of the Hellinger distance and Pearson correlation between the original variables and sequential decision trees' synthetic variables. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance for all variables between the original and synthetic data sets. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

The structure of the generated data in different modes is presented in Figure 4.14, which shows that the structure of the synthetic data in different modes is comparable to the original one.

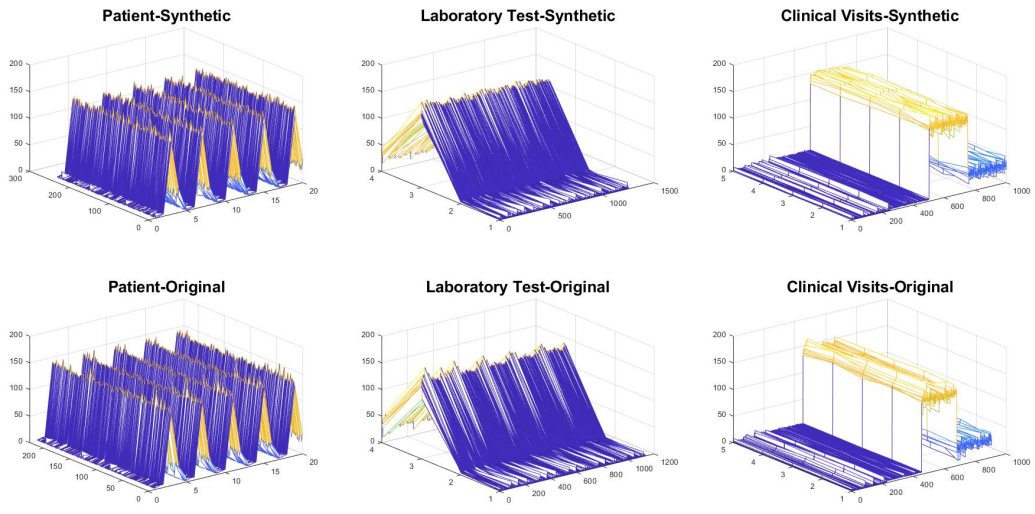


Figure 4.14: The different modes (Patients, Laboratory tests, and Clinical visits) of sequential trees' generated data and the original data.

In this scenario, the summary in Table 4.7 also demonstrates that the ranges of the synthetic variables are comparable to those of the original ones in Table 4.4.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	2.16	108.3	8.97
1st Q	0.76	3.71	134.4	27.37
Median	1.22	4.18	138.4	31.44
Mean	1.67	4.26	138.4	31.56
3rd Q	1.88	4.72	142.2	35.87
Max	16.42	7.94	163.1	53.04

Table 4.7: A summary of sequential decision trees' synthetic variables.

Based on the results of this section, we are optimistic that our model may perform even

better when generating larger data sets.

4.6.3 Experiments on Continuous Data with Missing Observations

The following is a trial on the continuous data set derived from the MIMIC-III data set, without imputation. As mentioned earlier, the data set consists of 226 patients, 4 laboratory tests, and 36 clinical visits, with 21% of the observations missing. We have previously described the approach for synthesizing this sort of data. In this experiment, we attempted to sample the patient factor matrix using sequential trees and generate the same sample size of 226 patients as in the original data. We performed GCP factorization with Gaussian loss and $R = 100$. Then, we applied the CP decomposition via an alternating least square (CP-ALS decomposition) to the missing tensor with $R = 50$, where the loss function was also Gaussian. The structure of the synthetic and original data sets in different modes is similar, as shown in Figure 4.15. Furthermore, Figures 4.16 and 4.17 below indicate that the dependency structure is preserved, and the marginal fitting is quite comparable in both the synthetic and original data sets. This demonstrates that our proposed model for EHRs synthesis performs well.

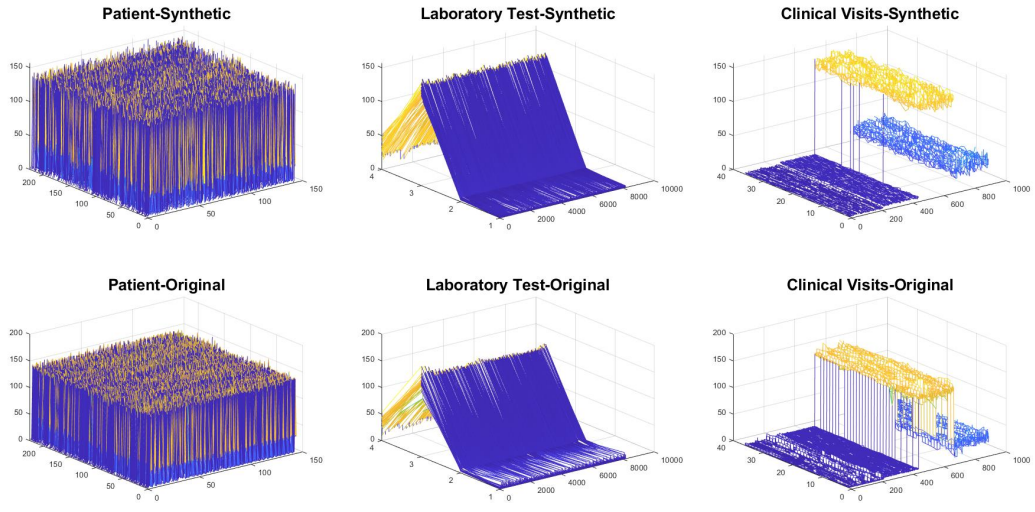
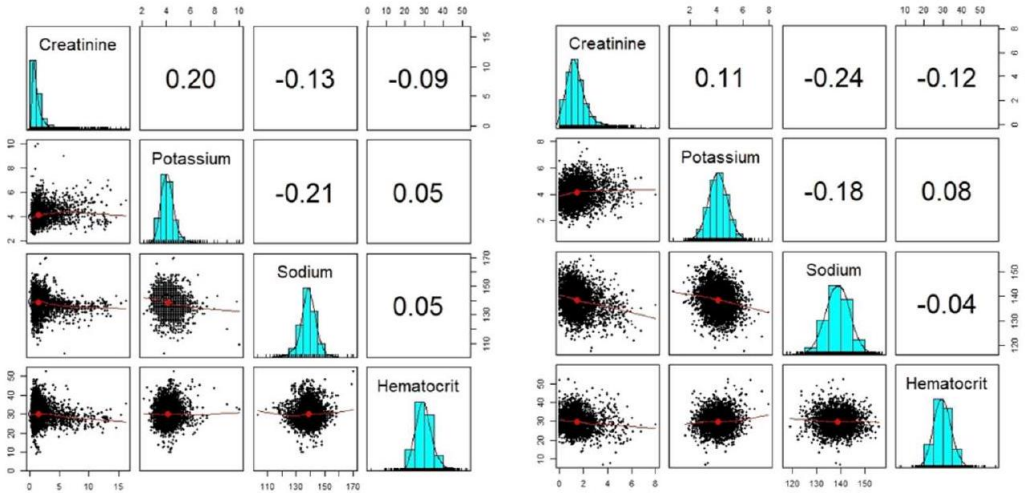


Figure 4.15: The different modes (Patients, Laboratory tests, and Clinical visits) of the original and sequential decision trees' synthetic data.

The structure of synthetic and original data sets in different modes are similar as shown here.



(A) (B)

Figure 4.16: The plots show the correlation and distribution of sequential decision trees' synthetic variables, as well as the original variables. A correlation matrix displaying bivariate scatter plots of the adjacent variables below the diagonal, histograms of the data distribution of the respective variables on the diagonal, and the Pearson correlation above the diagonal. Ellipses specify the direction of the correlation. The information regarding the relationship between two selected variables is always perpendicular to each other. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

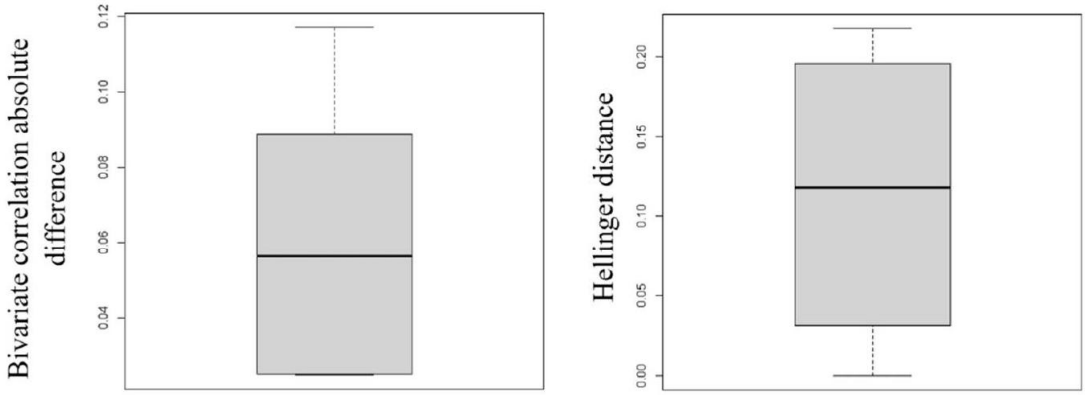
According to Tables 4.8 and 4.9, the missing percentages in the synthetic and actual data, are 46% and 21%, respectively. It seems to be almost double. As the missing tensor is binary, we can try to factorize it in order to determine if there are any possible improvements in the outcomes.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	0.71	117.5	7.07
1st Q	0.85	3.64	135.3	26.58
Median	1.3	4.12	138.6	29.64
Mean	1.47	4.13	138.6	29.82
3rd Q	1.87	4.62	142	32.77
Max	7.97	7.94	156.8	57.9
#NA's	3893	3636	3733	3661

Table 4.8: A summary of the sequential decision trees' synthetic variables.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	2.1	103	2
1st Q	0.6	3.7	135.8	26.9
Median	1	4	139	29.6
Mean	1.52	4.1	138.7	29.86
3rd Q	1.6	4.4	142	32.5
Max	16.2	10	170	52.6
#NA's	2050	1532	1752	1515

Table 4.9: A summary of the original variables.



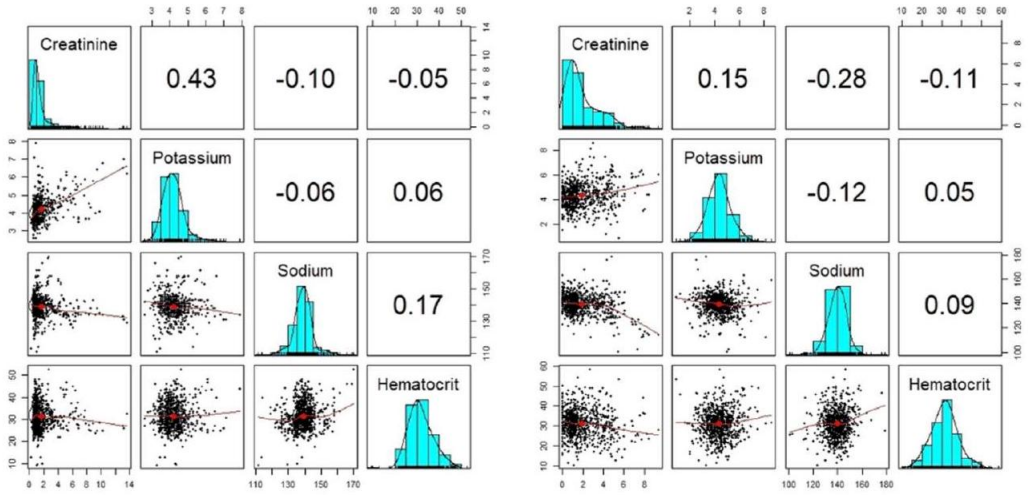
Laboratory tests (A)

Laboratory tests (B)

Figure 4.17: The box plots display the variation of the Hellinger distance and Pearson correlation between the original variables and sequential decision trees' synthetic variables. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance for all variables between the original and synthetic data sets. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

4.6.4 Experiments on Continuous Data with Irregular Clinical Visits

To create irregularity in clinical visits, a subset of the continuous data was created by choosing the first 10 observations. Those outcomes are presented in this section, and we have previously elaborated upon the method used for addressing this particular scenario. The GCP decomposition was performed with Gaussian loss and $R = 30$, resulting in an MSE of approximately 0.004. The experiment was conducted by employing the sequential decision trees approach to sample the patient factor matrix. The RMSDC was computed as 0.14.



(A) (B)

Figure 4.18: The plots show the correlation and distribution of variables generated by sequential trees and the original ones. A correlation matrix displaying bivariate scatter plots of the adjacent variables below the diagonal, histograms of the data distribution of the respective variables on the diagonal, and the Kendall correlation above the diagonal. Ellipses specify the direction of the correlation. The information regarding the relationship between two selected variables is always perpendicular to each other. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

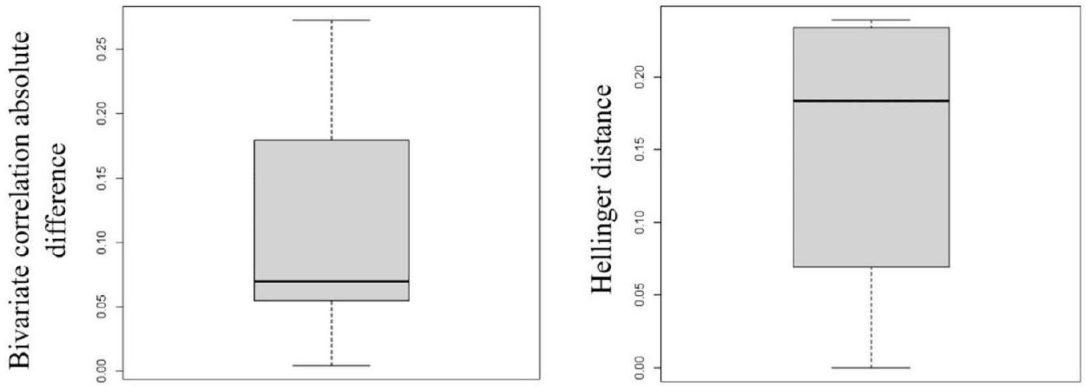


Figure 4.19: The box plots display the variation of the Hellinger distance and Kendall correlation between the original variables and sequential decision trees' synthetic variables. (A) is the box plot of the absolute difference in correlations between all variable pairs in the real and synthetic data. (B) is the box plot of the Hellinger distance for all variables between the original and synthetic data sets. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

Upon analyzing the results in Figures 4.18 and 4.19, we found that the process of generating synthetic data generally maintained the bivariate relationships and univariate distributions in the data.

When analyzing Tables 4.10 and 4.11, it was found that the provided descriptive statistics have remained comparable throughout the process of generating synthetic data.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
min	0	0.87	100.3	10.62
1st Q	0.76	3.73	134.9	27.02
Median	1.31	4.31	139.8	31.43
Mean	1.9	4.34	139.5	31.36
3rd Q	2.82	4.88	144.6	35.21
Max	9.35	8.59	178.9	58.15

Table 4.10: A summary of the sequential decision trees' synthetic variables.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0.2	2.6	111.2	9.2
1st Q	0.76	3.8	135.6	27.83
Median	1.07	4.15	139	31
Mean	1.6	4.21	139	31.61
3rd Q	1.7	4.5	142	35.1
Max	13.6	7.9	170	52.6

Table 4.11: A summary of the original variables.

4.6.5 Experiments on Categorical Data

The categorical data contains 2 variables: "Admission Type" and "Admission Location." The GCP decomposition was implemented using 2 different loss functions: the Poisson log link, which we discuss its results in Appendix D, and the Gaussian loss function. Initially, a series of postprocessing steps were conducted. The outcomes of the synthesis, which apply the Gaussian loss function and use Hamiltonian Monte Carlo (HMC), are presented in the

following. In this experiment, $R = 10$ was obtained. The structure of the generated data in different modes is shown in Figure 4.20.

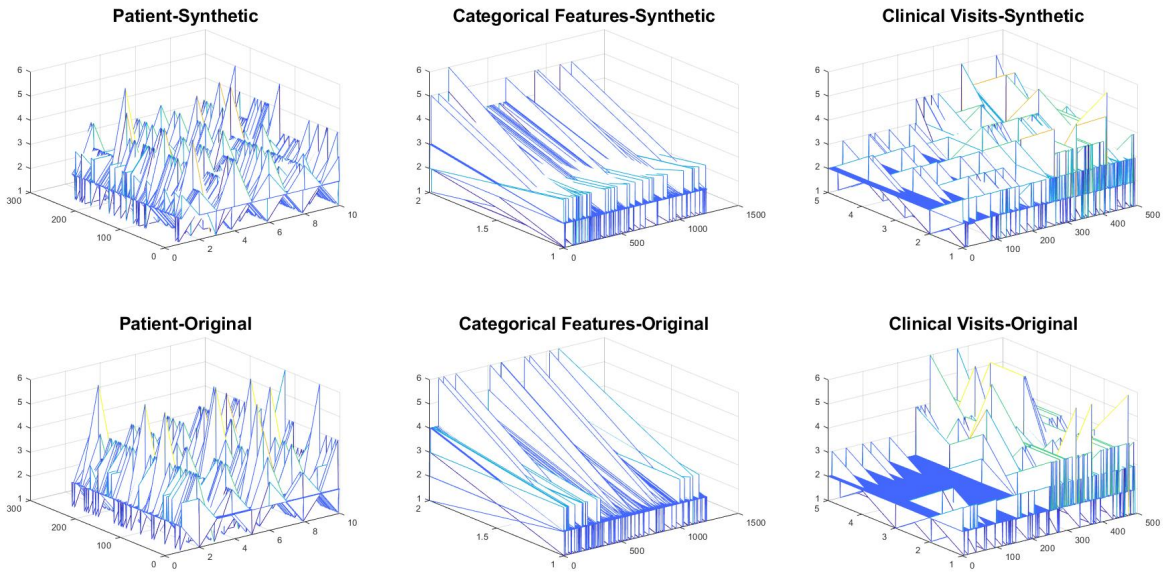
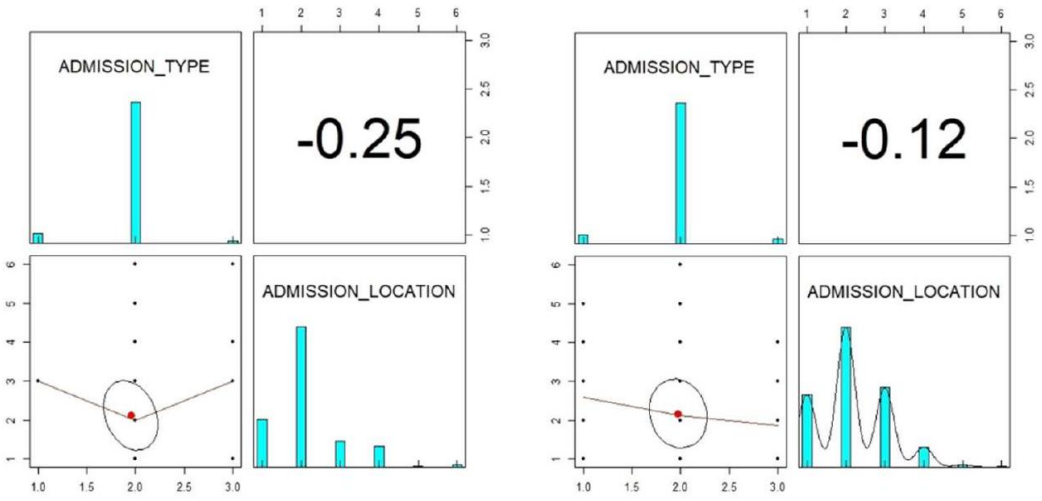


Figure 4.20: The different modes (Patients, Categorical features, and Clinical visits) of the HMC's generated categorical data are shown.

The Hellinger distance and Kendall correlation were calculated for the categorical variables as Figures 4.21 and 4.22.



(A) (B)

Figure 4.21: The plots show the Kendall correlation and distribution of the HMC's synthetic variables, as well as the original variables. A correlation matrix displaying bivariate scatter plots of the adjacent variables below the diagonal, a bar chart of the data distribution of the respective variables on the diagonal, and the Kendall correlation above the diagonal. Ellipses specify the direction of the correlation. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

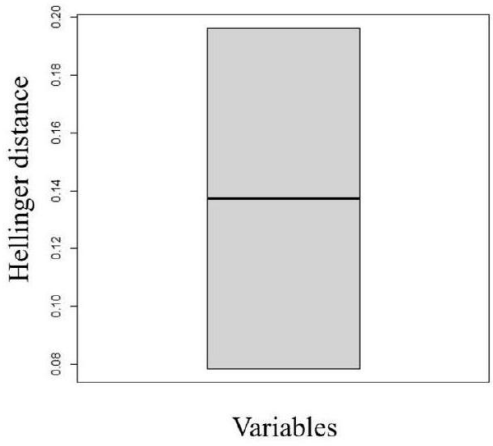


Figure 4.22: The box plot shows the variation of Hellinger distance for all variables between the original and HMC's synthetic data sets. This shows the similarity of the univariate distributions between the real and synthetic data. This is a value between 0 and 1, with lower values indicating similarity between the univariate distributions of the real and synthetic variables.

All the findings demonstrate that the generative model is applicable to any sort of variable.

Chapter 5

Conclusions and Future Works

5.1 Conclusions

Markov chain Monte Carlo (MCMC) algorithms are widely used simulation algorithms. One of their key contributions is revolutionizing Bayesian inference, allowing users to sample from the posterior distributions of complex statistical models. This is achieved by defining a Markov chain that converges to the desired probability distribution. For effective sampling, the relevant Markov chain must quickly converge to its stationary state. As a result, understanding the time it takes for Markov chains to reach the stationary state, known as mixing time, is essential for the algorithm's proper implementation. Therefore, recent mathematical research has focused on the convergence rates of Markov chains and bounding mixing times. This research has numerous applications, the most important of which are Markov chain Monte Carlo methods.

This thesis discussed commonly used techniques for estimating mixing time in discrete and continuous state spaces. There have been numerous studies for bounding the convergence rate of discrete-state Markov chains, and their extensive collection has inspired studies of Markov chains' mixing times in continuous state spaces. This includes spectral and pro-

file methods, comparison approaches, decomposition techniques, coupling, and geometric bounds. We also examined the precision of the spectral profile bound on the L^2 mixing time for continuous-state Markov chains. Several studies have been conducted to investigate the sensitivity of bounds on mixing times, and Kozma's research [88] inspired our study [83]. Analyzing the sensitivity of the bound on mixing time reveals how sharp the bound is and how much information will be lost if the method is used for bounding. The perturbation studies [120, 110] are closely related to the problem of investigating the sensitivity of Markov chains to minor changes.

There are various MCMC algorithms for sampling from target distributions [131, 132]. The Hamiltonian Monte Carlo (HMC) is widely used in practice and considered state-of-the-art [44]. We proposed a generative model [84] that uses techniques such as HMC, copula, and sequential decision trees to sample from the latent space of state-of-the-art generalized canonical polyadic (GCP) tensor factorization. Tensor decompositions are becoming increasingly popular due to their interpretability and adaptability in high-dimensional and heterogeneous data sets. The model appears to be a useful tool for generating synthetic longitudinal health data. We could overcome the challenges of synthesizing massive longitudinal health data by synthesizing a much smaller non-longitudinal data set instead. As a result, our generative model has the potential to produce valuable synthetic data across a wide range of fields and research domains.

5.2 Research Extensions

The results in Chapter 3, which contributed to our study [83], were derived from the assumption of exactly half-lazy and Hilbert-Schmidt Markov chains. As a future research, Lemma 11 in Chen et al. [26], which is Lemma 2.4.15 in this thesis, can be expanded to include a wider range of Markov chains in our Theorem 3.3.5. We provided Lemma 3.3.6, which can be used to extend Theorem 3.3.5 to include half-lazy 3.2.1 and smooth Markov chains 2.2.7. The lemma serves as a foundation for future research. Extending Theorem 3.3.5 to beyond half-lazy chains poses a challenge due to the fact that the stationary measures of the lazy and non-lazy chains P and $\tilde{P}$, respectively, may not be the same. By using Lemma 2.4.30 in Chapter 2, one can find the link between the spectral gaps of $\tilde{P}$ and P. However, determining a lower bound for the ratio of densities as Lemma 3.3.4 in this context will be difficult since the stationary distributions are not identical. We have analyzed the behavior of the stationary distributions in a simple toy problem, and interestingly, as the number of steps approaches infinity, the stationary measures converge towards each other. Another possible research direction is to modify Assumption 3.3.3 to include non-Hilbert-Schmidt Markov chains. To address this, we recommend that interested readers utilize Lemma 2.1 of [36] and then build upon our lemma 3.3.4 in this particular context. As another direction for future research, it would be worthwhile to investigate the precision of the conductance profile bound for L^2 mixing time proposed by Chen et al. [26] in light of the work by Addario-Berry and Roberts [3] on the robustness of the mixing time bounds using bottleneck sequences. This could extend the study from discrete to continuous state spaces.

A future study on the generative model provided in Chapter 4 could involve conducting a more rigorous comparison between the original and synthetic data sets to evaluate both the generative model and the superior sampling approaches. Another possible future study would be to employ a large data set for this model and examine different simulation techniques, such as other MCMC algorithms, generative adversarial network (GAN), or recurrent neural network (RNN) models. We focused on longitudinal health data in our model. However,

there are also other types of longitudinal data, such as transactions in financial data sets, that occur over time. We believe that our model could perform well on various types of longitudinal data sets. It would be interesting and valuable to carry out a future study to assess the effectiveness and feasibility of this model on different types of longitudinal data, such as financial data. We can also look at the HMC sampling approach and see if we can improve its results on discrete and continuous variables by defining more appropriate distributions, such as Tweedie.

We reduced the risk of identity disclosure in our model by avoiding overfitting. However, we did not measure the risk of privacy disclosure in the model. There are different types of privacy risks, including identity disclosure, which refers to the risk of correctly matching a synthetic record to a real record. Another type of privacy risk is attribution risk [46], which occurs when an adversary learns that a certain individual has a particular characteristic, as well as membership disclosure, which combines the two [47]. It is beneficial to evaluate the privacy risk, especially the model's membership disclosure.

Appendix A

Particle Swarm Optimization

The particle swarm technique [15] is employed to optimize variable ordering in sequential synthesis [50]. This approach does not need the objective function to be continuous; instead, it finds the global optimum using a search heuristic. For the objective function, the utility metric (distinguishability) is computed, and a hinge loss function is employed and minimized [134]. The hinge loss sets the utility metric (distinguishability) to zero when it falls below a certain threshold. For instance, there is no loss if the distinguishability is smaller than 0.05. The threshold ensures that the generated trees don't overfit the data. The loss as an optimization is:

$$\text{Hing-loss} = \max(0, d - 0.05),$$

where d denotes the distinguishability metric.

This distinguishability is based on propensity scores. The real and synthetic data sets are combined, and each record is labeled as either real or synthetic. To distinguish between actual and synthetic records, a binary classification model is created with the original variables as predictors and the binary indicator variable as the outcome. The propensity score (predicted probability) is determined using 10-fold cross-validation and the generalized boosted models

are used as the classification approach [50]. The distinguishability score is calculated as the mean square difference between the predicted probability and 0.5, the threshold at which the two data sets cannot be distinguished.

$$d = \frac{1}{N} \sum_{i=1}^{N} (p_i - 0.5)^2 ,$$

where N is the size of the synthetic data set and p_i is the propensity score for observation i. The classifier may distinguish between entirely distinct data sets. In this situation, the propensity score will be either 0 or 1, with a d value around 0.25.

Appendix B

The Model Block of Stan for Hamiltonian Monte Carlo

We used Stan for implementing HMC. The following is the model block of Stan. In this block, x and $x_$sim represent the patient factor matrix variables $\mathbf{a}_i$ and the corresponding simulations $\hat{\mathbf{a}}_i$, respectively. M indicates the number of patients.

```
data {
        int<lower=0> M;
        vector[R] x[M];
}

transformed data {
        vector[R] mu = rep_vector(0, R);
}
parameters {
        cholesky_factor_corr[R] chol;
        vector<lower=0>[R] sigma;
}

transformed parameters {
        matrix[R, R] chol_cov = diag_pre_multiply(sigma, chol);
}

model {
        chol ~ lkj_corr_cholesky(1);
        sigma ~ cauchy(0, 5);
        x ~ multi_normal_cholesky(mu, chol_cov);
}

generated quantities {
        vector[R] x_sim[M];
        for(i in 1:M) {
                x_sim[i] = multi_normal_cholesky_rng(mu, chol_cov);
        }
}
```

Appendix C

Continuous SDG Using β-loss

In the following, we present the outcomes of synthetic continuous data generated by GCP using β-divergence with $\beta = 0.75$, $R = 15$, where the fit score and MSE were about .977 and 2.5, respectively. The data set used in this experiment consists of 226 patients, 4 laboratory tests, and 36 clinical visits. It is the imputed version of the continuous data set derived from the MIMIC data set. As the MSE is not too small it was expected that the result would not be outstanding. However, copula and sequential trees performed better than HMC. As can be observed, all three recommended methods of patient factor matrix sampling have a much greater correlation than the real one. Copula, sequential trees, and the HMC results can be found in the following, respectively.

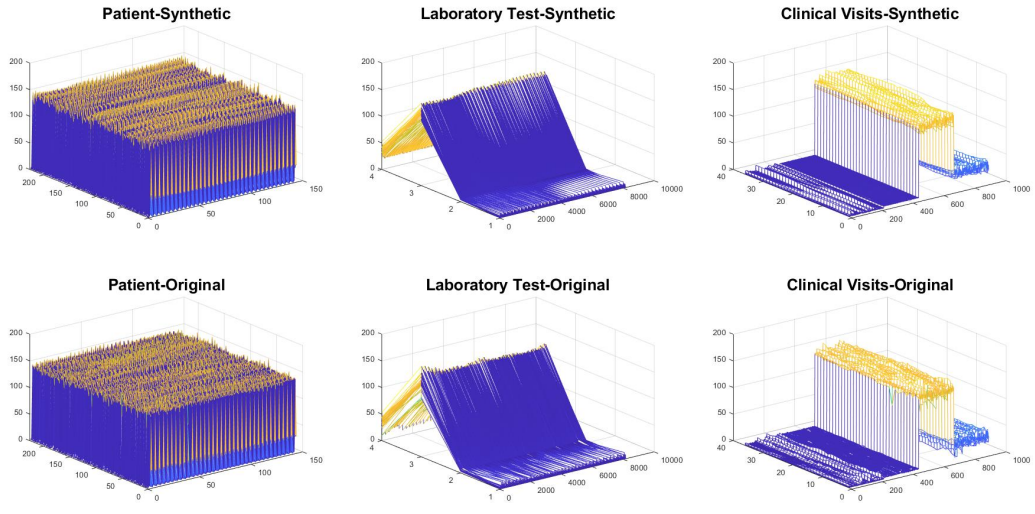
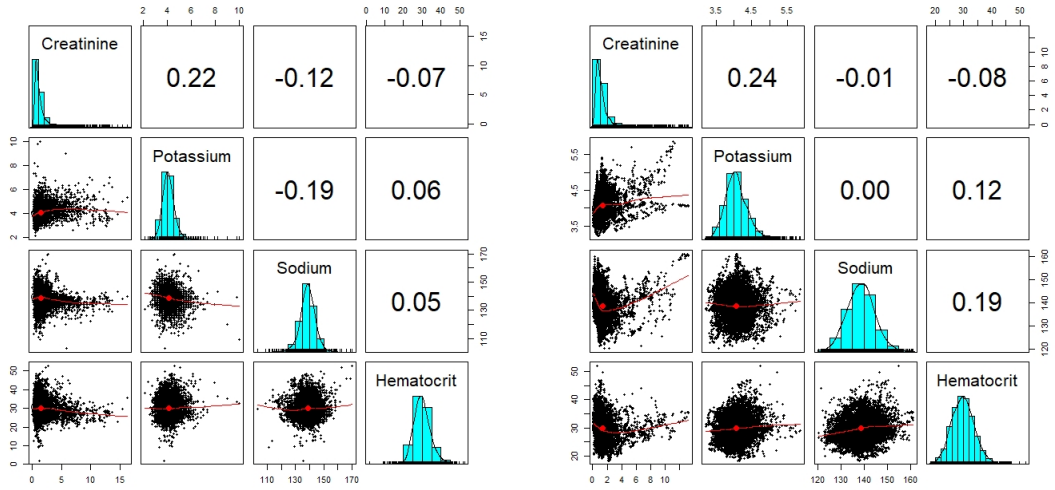


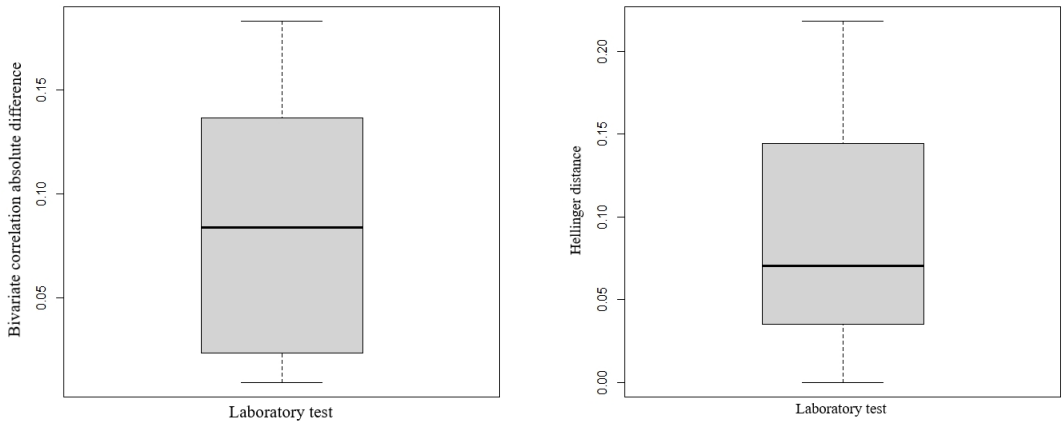
Figure C.1: The different modes (Patients, Laboratory tests, and Clinical visits) of copula's generated data and the original data are shown.



(A) (B)

Figure C.2: The plots display the correlation and distribution of all variables in the generated data by copula and the original data set. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

According to Figures C.2 and C.3, the dependency structure and distribution of the original variables are almost preserved in the synthetic data generated by copula using empirical CDF marginals. Figure C.1 shows the different modes on the original and synthetic tensor. The box plots for the variation of Hellinger distance and Pearson correlation between variables in the synthetic and original data in Figure C.3 also represent the same. RMSDC=0.104 was obtained.



(A) (B)

Figure C.3: (A) The box plot of Pearson correlation difference between the original and copula's synthetic variables. (B) The box plot of the variation of Hellinger distance.

According to the summaries in Tables C.1 and C.2, the minimum of variables "Sodium" and "Hematocrit" are somewhat higher than the original.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0.09	3.21	120.2	18.33
1st Q	0.64	3.86	134.5	26.99
Median	0.98	4.04	138.6	29.71
Mean	1.37	4.07	138.6	29.83
3rd Q	1.52	4.25	142.5	32.42
Max	13.15	5.87	161.2	52

Table C.1: A summary of the copula's synthetic variables.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	2.1	103	2
1st Q	0.62	3.74	135.4	27
Median	0.99	4.03	139	29.7
Mean	1.47	4.09	138.6	30.05
3rd Q	1.6	4.4	142	32.7
Max	16.2	10	170	52.6

Table C.2: A summary of the original variables.

Here are the outcomes of sampling the patient factor matrix of the previously mentioned GCP decomposition using sequential trees approaches. Figure C.4 shows the modes of the original and synthetic tensors.

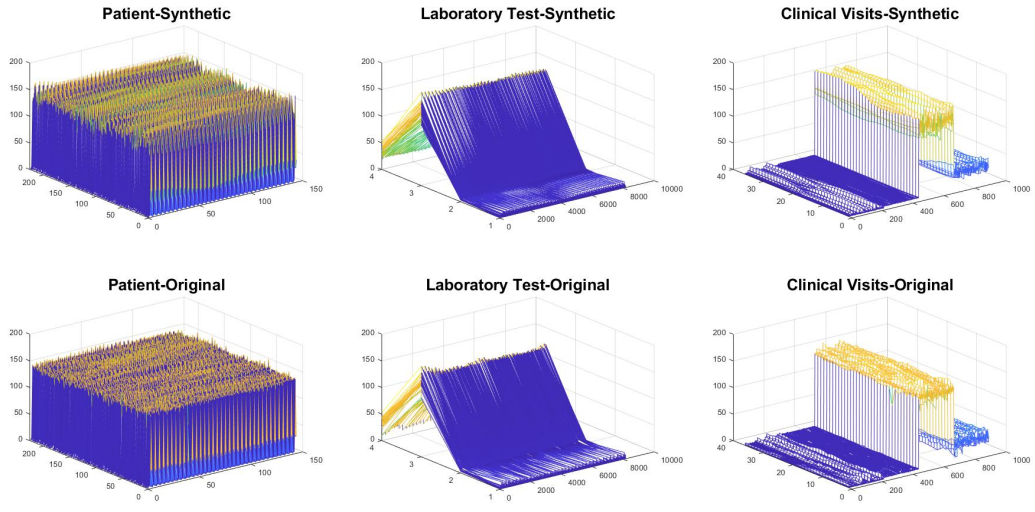
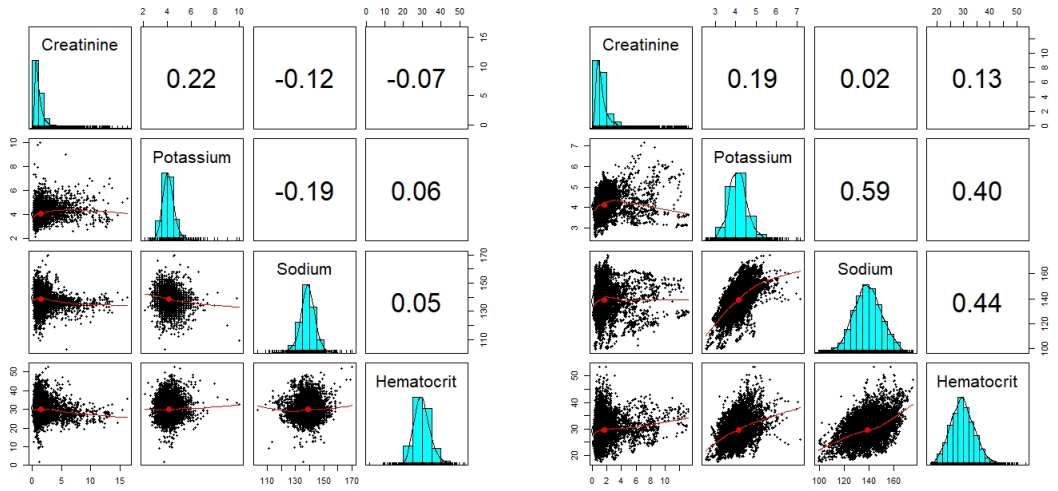


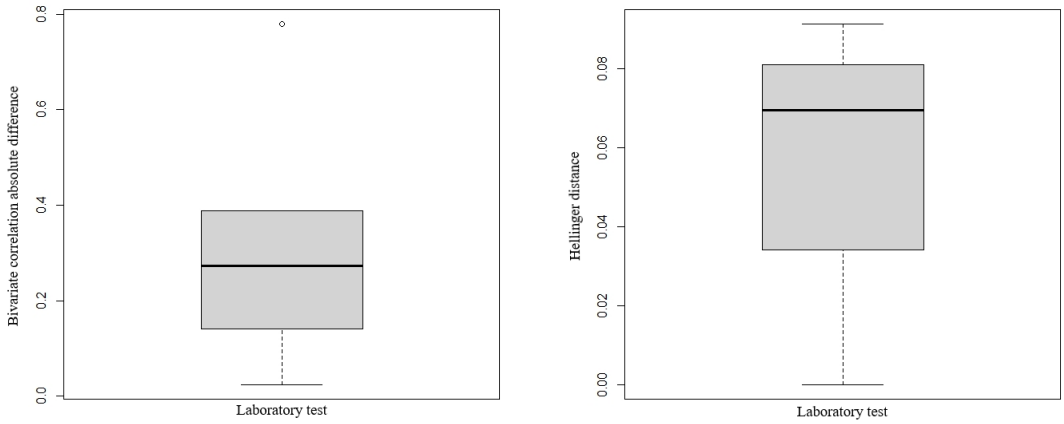
Figure C.4: The different modes (Patients, Laboratory tests, and Clinical visits) of the sequential trees' generated data are shown.

Figure C.5 indicates that the univariate distributions are similar for each variable in synthetic and original data sets. Figure C.6 of the variation of Hellinger distance also shows that sequential decision trees' synthetic variables are derived from the similar distribution as the original variables. However, the copula performed better in capturing the correlations between variables. The computed RMSDC for this experiment was 0.237.



(A) (B)

Figure C.5: The plot shows the correlation and distribution of the original data and the data generated by sequential decision trees. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.



(A) (B)

Figure C.6: (A) The box plot of Pearson correlation difference between the original and sequential decision trees' synthetic variables. (B) The box plot of the variation of Hellinger distance.

The below summary in Table C.3 displays that the range of variable "Sodium" has been significantly improved with respect to the previous analysis using copula, refer to Table C.2 for the summary of the original data set.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0.15	2.54	99.37	17.71
1st Q	0.74	3.8	131.21	26.5
Median	1.1	4.1	138.75	29.57
Mean	1.65	4.13	138.82	29.79
3rd Q	1.77	4.4	146.33	32.73
Max	13.15	7.18	175.29	53.31

Table C.3: A summary of the sequential decision trees' synthetic variables.

Figures C.7, C.8 and C.9 show the results of the Hamiltonian Monte Carlo performance on the data set. If the distribution of the HMC model is properly defined, the outcome would be quite satisfactory.

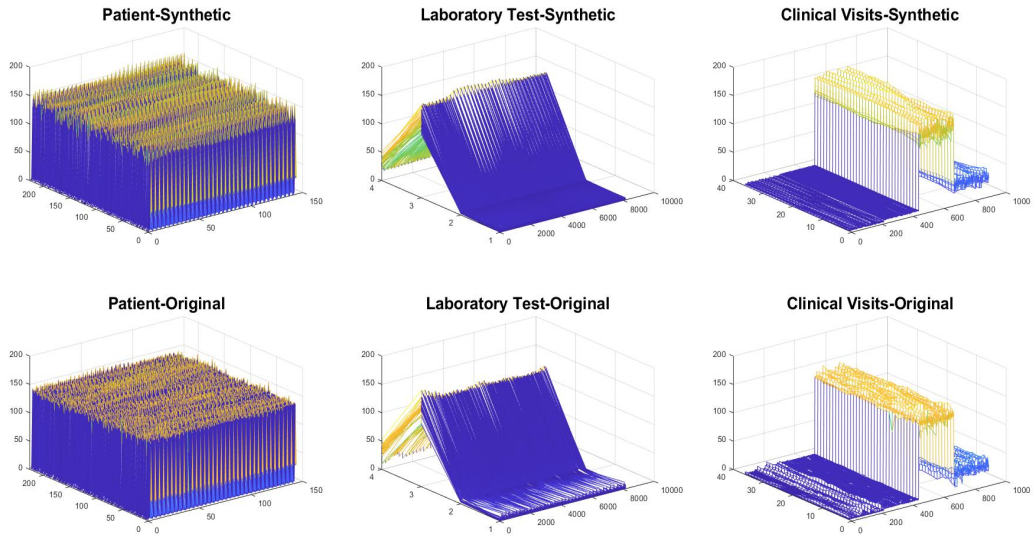
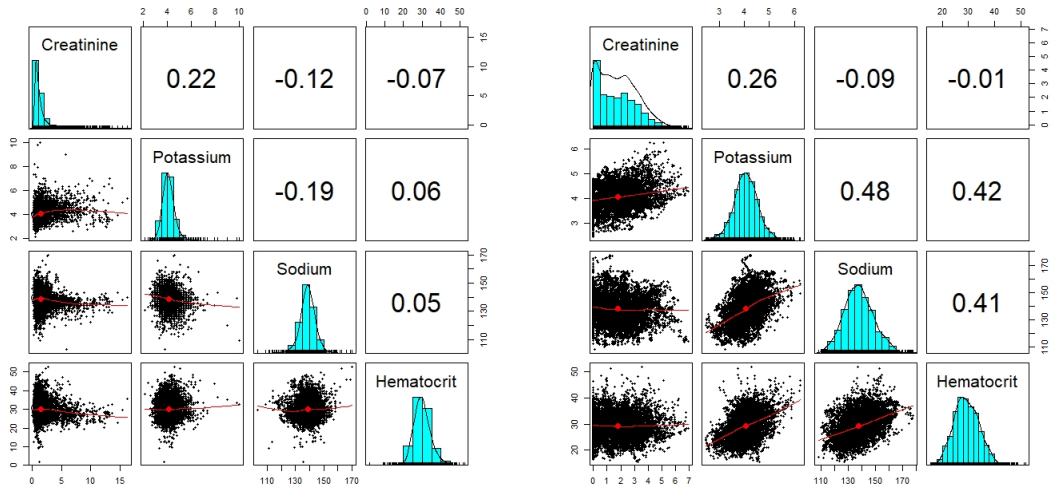


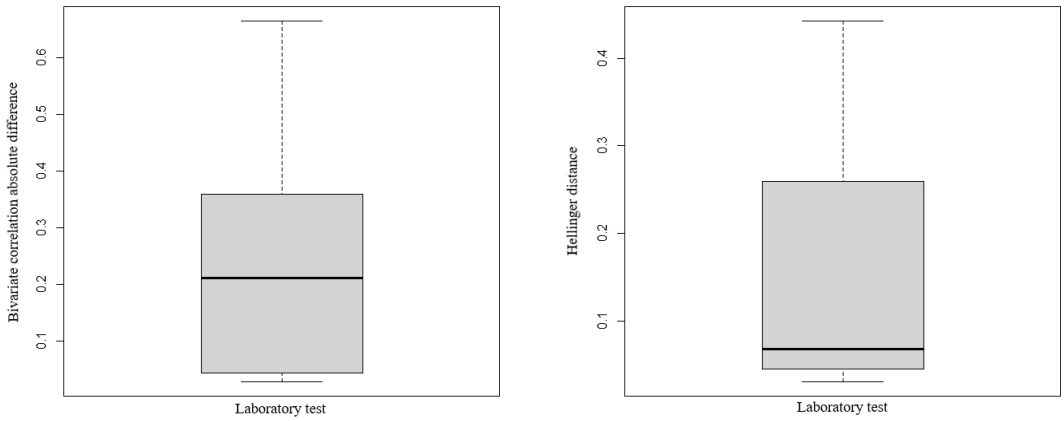
Figure C.7: The different modes (Patients, Laboratory tests, and Clinical visits) of HMC's generated data are shown.

We did not expect HMC to perform well here since the β-divergence loss causes a non-Gaussian latent space, and we won't get a good result even when standardizing the latent space. On the other hand, defining a proper model distribution for the HMC would considerably enhance the findings.



(A) (B)

Figure C.8: The plot shows the correlation and distribution of the original data and the data generated by the HMC. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.



(A) (B)

Figure C.9: (A) The box plot of Pearson correlation difference between the original and HMC's synthetic variables. (B) The box plot of the variation of Hellinger distance.

The comparison of the summary in Table C.4 with Table C.2 reveals that the HMC performed poorly in this particular scenario.

Metric	Variables			
	Creatinine	Potassium	Sodium	Hematocrit
Min	0	2.54	108.3	15.48
1st Q	0.58	3.75	130.6	25.91
Median	1.65	4.06	137.5	29.04
Mean	1.79	4.06	138	29.3
3rd Q	2.78	4.37	145	32.62
Max	6.9	6.25	177.6	51.91

Table C.4: A summary of the HMC's synthetic variables.

At last, we provide Figure C.10 to make it easier comparing the three sampling techniques on the GCP decomposition with β-divergence loss.

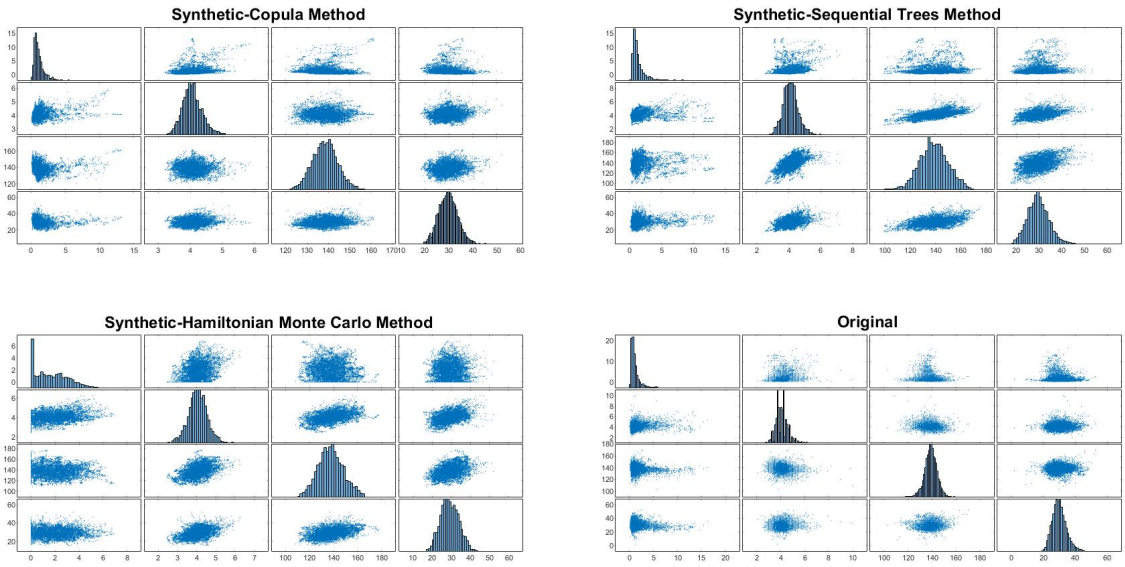


Figure C.10: The distribution and scatter plots of the original data set and synthetic data generated using copula, the sequential trees, and HMC.

Appendix D

Categorical SDG Using Poisson Log Link

In the following, we present the outcomes of Generating Synthetic Categorical Data Using Poisson log link. Figures D.1 and D.2 show the results from generating categorical data using the GCP decomposition with a Poisson log link using the HMC for simulation of the patient factor matrix.

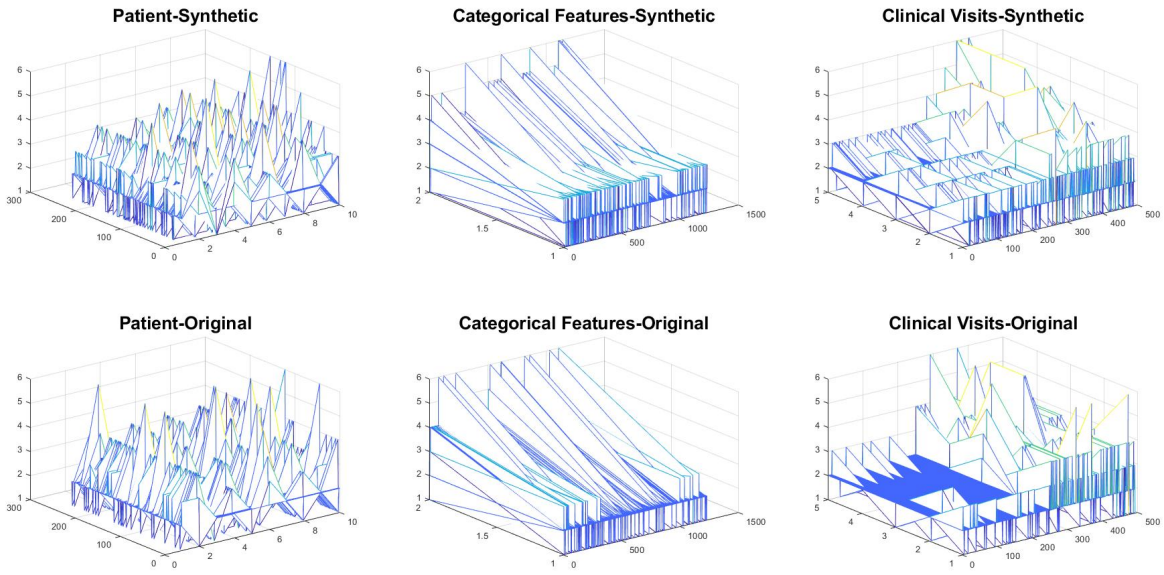
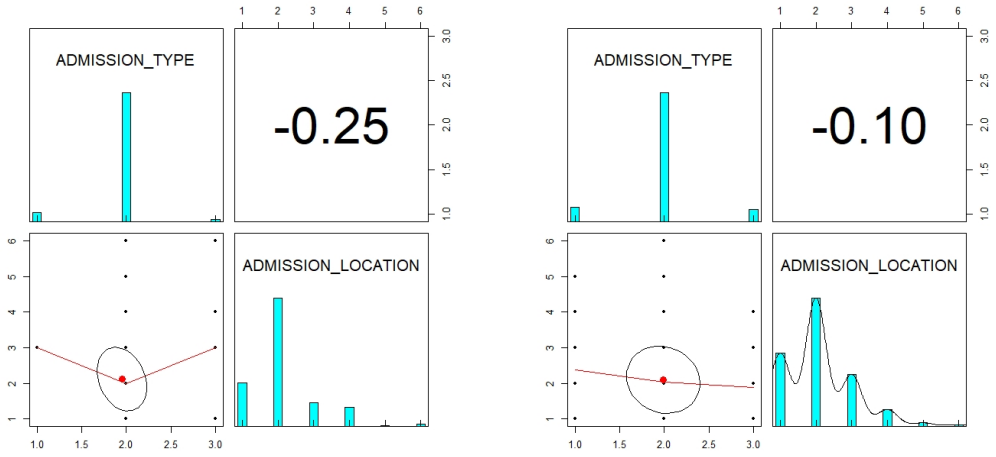


Figure D.1: The different modes (Patients, Categorical features, and Clinical visits) of the HMC's generated categorical data are shown.



(A) (B)

Figure D.2: The correlation and distribution plot illustration of the HMC's generated data. (A) is the plot of the original variables. (B) is the plot of the synthetic variables.

The Hellinger distance and Kendall correlation were computed for the categorical variables as shown in Figure D.3. All the results indicate that the generative model can be applied to any type of variable.

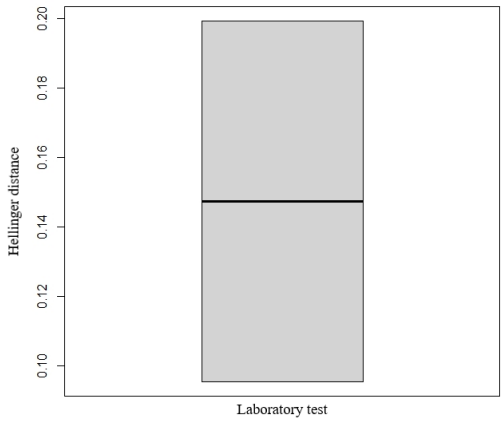


Figure D.3: The box plot shows the variation of Hellinger distance between the original and the HMC's synthetic categorical variables.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.

[3] Louigi Addario-Berry and Matthew I Roberts. Mixing time bounds via bottleneck sequences. *Journal of Statistical Physics*, 173:845–871, 2018.

[4] Ardavan Afshar, Ioakeim Perros, Evangelos E Papalexakis, Elizabeth Searles, Joyce Ho, and Jimeng Sun. COPA: Constrained PARAFAC2 for sparse & large datasets. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 793–802, 2018.

[5] Ardavan Afshar, Ioakeim Perros, Haesun Park, Christopher deFilippi, Xiaowei Yan, Walter Stewart, Joyce Ho, and Jimeng Sun. TASTE: temporal and static tensor factorization for phenotyping electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 193–203, 2020.

[6] Berni J Alder and Thomas Everett Wainwright. Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

[7] David Aldous and James Allen Fill. Reversible Markov Chains and Random Walks on Graphs, 2002. Unfinished monograph, recompiled 2014, available at https://www.stat.berkeley.edu/users/aldous/RWG/Book_Ralph/book.html.

[8] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[9] Bader, Brett W and Kolda, Tamara G and others. Tensor Toolbox for MATLAB, Version 3.1, June 2019. https://www.tensortoolbox.org/.

[10] Andrés F Barrientos, Alexander Bolton, Tom Balmat, Jerome P Reiter, John M de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. Providing access to confidential research data through synthesis and verification: An application to data on employees of the US federal government. 2018.

[11] Fodil Benali, Damien Bodénès, Nicolas Labroche, and Cyril de Runz. MTCopula: Synthetic Complex Data Generation Using Copula. In *23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP)*, pages 51–60, 2021.

[12] Nathanaël Berestycki. Mixing Times of Markov Chains: Techniques and Examples. *Alea-Latin American Journal of Probability and Mathematical Statistics*, 2016.

[13] Michael Betancourt. The convergence of Markov chain Monte Carlo methods: from the Metropolis method to Hamiltonian Monte Carlo. *Annalen der Physik*, 531(3):1700214, 2019.

[14] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[15] Mohammad Reza Bonyadi and Zbigniew Michalewicz. Particle swarm optimization for single objective continuous space problems: a review. *Evolutionary computation*, 25(1):1–54, 2017.

[16] Nicolas Bourbaki. *Elements of Mathematics: Commutative Algebra*. Springer Berlin, 1989.

[17] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984.

[18] Pierre Brémaud. Markov chains, volume 31 of Texts in Applied Mathematics. In *Gibbs fields, Monte Carlo simulation, and queues*. Springer, 1999.

[19] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[20] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in Markov chains. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 223–231. IEEE, 1997.

[21] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[22] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[23] Saptarshi Chakraborty and Kshitij Khare. Convergence properties of Gibbs samplers for Bayesian probit regression with proper priors. *Electronic Journal of Statistics*, 11(1):177–210, 2017.

[24] Jeff Cheeger. A lower bound for the smallest eigenvalue of the laplacian. *Problems in analysis*, 625(195-199):110, 1970.

[25] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022.

[26] Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *Journal of Machine Learning Research*, 21(92):1–71, 2020.

[27] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 149–160, 2021.

[28] Robert E Colgana, David E Gutierrezb, Jugesh Sundramc, and Gnana Bhaskar Tenalid. Analysis of medical data using dimensionality reduction techniques. In *Proceedings of the Conference: AMALTHEA–2013. Melbourne, FL: Florida Institute of Technology*, volume 10, pages 2270–1762, 2013.

[29] John B Conway. *A course in functional analysis*, volume 96. Springer, 2019.

[30] Thierry Coulhon. Ultracontractivity and nash type inequalities. *Journal of functional analysis*, 141(2):510–539, 1996.

[31] Thierry Coulhon, Alexander Grigor'yan, and Christophe Pittet. A geometric approach to on-diagonal heat kernel lower bounds on groups. In *Annales de l'institut Fourier*, volume 51, pages 1763–1827, 2001.

[32] Thierry Coulhon and Alexander Grigor'yan. On-diagonal lower bounds for heat kernels and Markov chains. 1997.

[33] Bryant Davis and James P Hobert. Approximating the Spectral Gap of the Pólya-Gamma Gibbs Sampler. *arXiv preprint arXiv:2104.13419*, 2021.

[34] Zachary J DeBruine, Karsten Melcher, and Timothy J Triche Jr. Fast and robust non-negative matrix factorization for single-cell experiments. *BioRxiv*, pages 2021–09, 2021.

[35] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, 2009.

[36] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23(2):151–178, 2008.

[37] Persi Diaconis and Laurent Saloff-Coste. Comparison techniques for random walk on finite groups. *The Annals of Probability*, pages 2131–2156, 1993.

[38] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible Markov chains. *The Annals of Applied Probability*, pages 696–730, 1993.

[39] Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695 – 750, 1996.

[40] Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability*, pages 36–61, 1991.

[41] Wolfang Doeblin. Exposé de la théorie des chaınes simples constantes de Markov á un nombre fini détats. *Mathématics of the Inter-Balkan Union*, 2(77-105):78–80, 1938.

[42] Jörg Drechsler and Jerome P Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243, 2011.

[43] Duane, Simon and Kennedy, Anthony D and Pendleton, Brian J and Roweth, Duncan. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

[44] Raaz Dwivedi. *Principled Statistical Approaches For Sampling and Inference in High Dimensions*. University of California, Berkeley, 2021.

[45] Martin Dyer, Leslie Ann Goldberg, Mark Jerrum, and Russell Martin. Markov chain comparison. *Probability Surveys*, 3:89–111, 2006.

[46] Khaled El Emam, Lucy Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *Journal of medical Internet research*, 22(11):e23139, 2020.

[47] Khaled El Emam, Lucy Mosquera, and Xi Fang. Validating a membership disclosure metric for synthetic health data. *JAMIA open*, 5(4):ooac083, 2022.

[48] Khaled El Emam, Lucy Mosquera, and Richard Hoptroff. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O'Reilly Media, 2020.

[49] Mark Elliot. Final report on the disclosure risk associated with the synthetic data produced by the sylls team. *Report 2015*, 2, 2015.

[50] Khaled El Emam, Lucy Mosquera, and Chaoyi Zheng. Optimizing the synthesis of clinical trial data using sequential trees. *Journal of the American Medical Informatics Association*, 28(1):3–13, 2021.

[51] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[52] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[53] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[54] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[55] Walter R Gilks and Pascal Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

[56] Sharad Goel, Ravi Montenegro, and Prasad Tetali. Mixing Time Bounds via the Spectral Profile. *Electronic Journal of Probability*, 11:1–26, 2006.

[57] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20:1–40, 2020.

[58] Emine Guven. Decision of the Optimal Rank of a Nonnegative Matrix Factorization Model for Gene Expression Data Sets Utilizing the Unit Invariant Knee Method: Development and Evaluation of the Elbow Method for Rank Selection. *JMIR Bioinformatics and Biotechnology*, 4(1):e43665, 2023.

[59] Richard A Harshman. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA working papers in phonetics*, 16(1):84, 1970.

[60] Richard A Harshman. PARAFAC2: Mathematical and technical notes. *UCLA working papers in phonetics*, 22(3044):122215, 1972.

[61] Abdelaali Hassaine, Dexter Canoy, Jose Roberto Ayala Solares, Yajie Zhu, Shishir Rao, Yikuan Li, Mariagrazia Zottoli, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Learning multimorbidity patterns from electronic health records using non-negative matrix factorisation. *Journal of Biomedical Informatics*, 112:103606, 2020.

[62] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[63] Gilbert Helmberg. *Introduction to Spectral Theory in Hilbert Space: North-Holland Series in Applied Mathematics and Mechanics*. Elsevier, 2014.

[64] Jonathan Hermon. On sensitivity of uniform mixing times. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 54(1):234 – 248, 2018.

[65] Jonathan Hermon and Gady Kozma. Sensitivity of mixing times of Cayley graphs. *Canadian Journal of Mathematics*, page 1–32, 2023.

[66] Jonathan Hermon and Yuval Peres. On sensitivity of mixing times and cutoff. *Electronic Journal of Probability*, 23, Jan 2018.

[67] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[68] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[69] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124, 2014.

[70] Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[71] David Hong, Tamara G Kolda, and Jed A Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Review*, 62(1):133–163, 2020.

[72] Jingchen Hu, Jerome P Reiter, and Quanli Wang. Disclosure risk evaluation for fully synthetic categorical data. In *International conference on privacy in statistical databases*, pages 185–199. Springer, 2014.

[73] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[74] Mark Jerrum. Mathematical foundations of the Markov chain Monte Carlo method. In *Probabilistic methods for algorithmic discrete mathematics*, pages 116–165. Springer, 1998.

[75] Mark Jerrum and Alistair Sinclair. Conductance and the rapid mixing property for Markov chains: the approximation of permanent resolved. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 235–244, 1988.

[76] Mark Jerrum and Alistair Sinclair. Approximating the permanent. *SIAM journal on computing*, 18(6):1149–1178, 1989.

[77] Mark Jerrum and Alistair Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. *Approximation Algorithms for NP-hard problems, PWS Publishing*, 1996.

[78] Mark Jerrum, Jung-Bae Son, Prasad Tetali, and Eric Vigoda. Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains. *The Annals of Applied Probability*, 14(4):1741–1765, 2004.

[79] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016.

[80] Galin L Jones and James P Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001.

[81] Sagar Kale. Eigenvalues and mixing time. *University of Darthmouth*, 2013.

[82] Alexandros Karatzoglou, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86, 2010.

[83] Elnaz Karimian Sichani and Aaron Smith. On the Precision of the Spectral Profile Bound for the Mixing Time of Continuous State Markov Chains. *arXiv preprint arXiv:2407.00749*, 2024.

[84] Elnaz Karimian Sichani, Aaron Smith, Khaled El Emam, and Lucy Mosquera. Creating High-Quality Synthetic Health Data: Framework for Model Development and Validation. *JMIR Formative Research*, 8(1):e53241, 2024.

[85] Yejin Kim, Jimeng Sun, Hwanjo Yu, and Xiaoqian Jiang. Federated tensor factorization for computational phenotyping. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 887–895, 2017.

[86] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[87] Tamara G Kolda and David Hong. Stochastic gradients for large-scale tensor decomposition. *SIAM Journal on Mathematics of Data Science*, 2(4):1066–1095, 2020.

[88] Gady Kozma. On the precision of the spectral profile. *Alea*, 3:321–329, 2007.

[89] Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

[90] Gregory F Lawler and Alan D Sokal. Bounds on the L^2 spectrum for Markov chains and Markov processes: a generalization of Cheeger's inequality. *Transactions of the American mathematical society*, 309(2):557–580, 1988.

[91] David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., United States of America, 2017.

[92] Zheng Li, Yue Zhao, and Jialin Fu. SYNC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 571–578. IEEE, 2020.

[93] Torgny Lindvall. Lectures on the coupling method, Corrected reprint of the 1992 original, 2002.

[94] László Lovász and Ravi Kannan. Faster mixing via average conductance. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 282–287, 1999.

[95] László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random structures & algorithms*, 4(4):359–412, 1993.

[96] Jing Ma, Qiuchen Zhang, Jian Lou, Joyce C Ho, Li Xiong, and Xiaoqian Jiang. Privacy-preserving tensor factorization for collaborative health data analysis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1291–1300, 2019.

[97] Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, pages 581–606, 2002.

[98] Oren Mangoubi, Natesh S Pillai, and Aaron Smith. Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? *arXiv preprint arXiv:1808.03230*, 2018.

[99] Russell A Martin and Dana Randall. Sampling adsorbing staircase walks using a new Markov chain decomposition method. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 492–502. IEEE, 2000.

[100] Yuzo Maruyama and Takeru Matsuda. Minimaxity under the half-Cauchy prior. *arXiv preprint arXiv:2406.08892*, 2024.

[101] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[102] Antonietta Mira. *Ordering, slicing and splitting Monte Carlo Markov chains*. University of Minnesota, 1998.

[103] Sarah Miracle, Amanda Pascoe Streib, and Noah Streib. Iterated decomposition of biased permutations via new bounds on the spectral gap of Markov chains. *arXiv preprint arXiv:1910.05184*, 2019.

[104] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[105] Ravi R Montenegro and Prasad Tetali. *Mathematical aspects of mixing times in Markov chains*. Now Publishers Inc, 2006.

[106] Ben Morris and Yuval Peres. Evolving sets, mixing and heat kernel bounds. *Probability Theory and Related Fields*, 133(2):245–266, 2005.

[107] Morten Mørup. Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):24–40, 2011.

[108] Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Astuti Feriani Tarigan. Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In *Journal of Physics: Conference Series*, volume 1361, page 012015. IOP Publishing, 2019.

[109] Radford M Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, 1994.

[110] Jeffrey Negrea and Jeffrey S Rosenthal. Approximations of geometrically ergodic reversible Markov chains. *Advances in Applied Probability*, 53(4):981–1022, 2021.

[111] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[112] James R Norris. *Markov chains*. Cambridge university press, 1998.

[113] Subahdip Pal, Kshitij Khare, and James P Hobert. Trace class Markov chains for Bayesian inference with generalized double Pareto shrinkage priors. *Scandinavian Journal of Statistics*, 44(2):307–323, 2017.

[114] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410. IEEE, 2016.

[115] Ioakeim Perros, Robert Chen, Richard Vuduc, and Jimeng Sun. Sparse hierarchical tucker factorization and its application to healthcare. In *2015 IEEE International Conference on Data Mining*, pages 943–948. IEEE, 2015.

[116] Ioakeim Perros, Evangelos E Papalexakis, Fei Wang, Richard Vuduc, Elizabeth Searles, Michael Thompson, and Jimeng Sun. SPARTan: Scalable PARAFAC2 for large & sparse data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 375–384, 2017.

[117] Eric T Phipps and Tamara G Kolda. Software for sparse tensor decomposition on emerging computing architectures. *SIAM Journal on Scientific Computing*, 41(3):C269–C290, 2019.

[118] Eric T Phipps, Tamara G Kolda, Daniel Dunlavy, Grey Ballard, and Todd Plantenga. Genten: software for generalized tensor decompositions v. 1.0. 0. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2017.

[119] Natesh S Pillai and Aaron Smith. Elementary bounds on mixing times for decomposable Markov chains. *Stochastic Processes and their Applications*, 127(9):3068–3109, 2017.

[120] Natesh S Pillai and Aaron Smith. On the mixing time of Kac's walk and other high-dimensional Gibbs samplers with constraints. *The Annals of Probability*, 46(4):2345 – 2399, 2018.

[121] JW Pitman. On coupling of Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(4):315–322, 1976.

[122] Nicholas G. Polson and James G. Scott. On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, 7(4):887 – 902, 2012.

[123] Qian Qin. Spectral properties of Markov operators in Markov chain Monte Carlo. 2017.

[124] Qian Qin, James P Hobert, and Kshitij Khare. Estimating the spectral gap of a trace-class Markov operator. *Electronic Journal of Statistics*, 13(1):1790–1822, 2019.

[125] Jean-Francois Rajotte, Robert Bergen, David L Buckeridge, Khaled El Emam, Raymond Ng, and Elissa Strome. Synthetic data as an enabler for machine learning applications in medicine. *Iscience*, 25(11), 2022.

[126] Jerome P Reiter. New approaches to data dissemination: a glimpse into the future (?). *Chance*, 17(3):11–15, 2004.

[127] Jerome P Reiter. Releasing Multiply Imputed, Synthetic Public use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(1):185–205, 12 2004.

[128] Jerome P Reiter. Using CART to generate partially synthetic public use microdata. *Journal of official statistics*, 21(3):441, 2005.

[129] Jerome P Reiter. Synthetic Data: A Look Back and A Look Forward. *Trans. Data Priv.*, 16(1):15–24, 2023.

[130] Gareth O Roberts and Jeffrey S Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electronic Communications in Probability*, 2(none):13 – 25, 1997.

[131] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

[132] Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. 2004.

[133] Peter N Robinson, Christopher J Mungall, and Melissa Haendel. Capturing phenotypes for precision medicine. *Molecular Case Studies*, 1(1):a000372, 2015.

[134] Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural computation*, 16(5):1063–1076, 2004.

[135] Jeffrey S Rosenthal. Convergence rates for Markov chains. *Siam Review*, 37(3):387–405, 1995.

[136] Jeffrey S Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566, 1995.

[137] Jeffrey S Rosenthal. Faithful couplings of Markov chains: now equals forever. *Advances in Applied Mathematics*, 18(3):372–381, 1997.

[138] Nicolas Ruiz, Krishnamurty Muralidhar, and Josep Domingo-Ferrer. On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, pages 59–74. Springer, 2018.

[139] Mikkel N Schmidt and Shakir Mohamed. Probabilistic non-negative tensor factorization using Markov chain Monte Carlo. In *2009 17th European Signal Processing Conference*, pages 1918–1922. IEEE, 2009.

[140] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[141] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

[142] M Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pages 229–231, 1959.

[143] Aaron Smith. Analysis of convergence rates of some Gibbs samplers on continuous state spaces. *Stochastic Processes and their Applications*, 123(10):3861–3876, 2013.

[144] Aaron Smith. Comparison theory for Markov chains on different state spaces and application to random walk on derangements. *Journal of Theoretical Probability*, 28:1406–1430, 2015.

[145] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[146] Perla Sousi. Mixing times of markov chains. *Preprint*, 2020.

[147] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. Challenges of differentially private prediction in healthcare settings. In *Proceedings of the IJCAI 2021 Workshop on AI for Social Good*, 2021.

[148] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[149] Muhammad Ali Syakur, B Khusnul Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 336(1):012017, apr 2018.

[150] Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Probability and its Applications (New York). Springer, 2000.

[151] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[152] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[153] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.

[154] Santosh Vempala. Geometric random walks: a survey. *Combinatorial and computational geometry*, 52(573-612):2, 2005.

[155] Yining Wang and Animashree Anandkumar. Online and differentially-private tensor decomposition. *arXiv preprint arXiv:1606.06237*, 2016.

[156] Zhenxun Wang, Yunan Wu, and Haitao Chu. On equivalence of the lkj distribution and the restricted wishart distribution. *arXiv preprint arXiv:1809.04746*, 2018.

[157] Lan Wei and Jerome P Reiter. Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Statistical Journal of the IAOS*, 32(1):93–108, 2016.

[158] Keru Wu, Scott Schmidler, and Yuansi Chen. Minimax mixing time of the Metropolis-adjusted Langevin algorithm for log-concave sampling. *The Journal of Machine Learning Research*, 23(1):12348–12410, 2022.

[159] Chao Yan, Ziqi Zhang, Steve Nyemba, and Zhuohang Li. Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutoria. *JMIR AI*, 3:e52615, 2024.

[160] Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: a survey. *PeerJ Computer Science*, 10:e1905, 2024.

[161] Wai Kong Yuen. Applications of geometric bounds to the convergence rate of Markov chains on R^n. *Stochastic processes and their applications*, 87(1):1–23, 2000.

[162] Wai Kong Yuen. *Application of Geometric Bounds to Convergence Rates of Markov Chains and Markov Processes on R^n*. National Library of Canada, 2002.

[163] Wai Kong Yuen. Generalization of discrete-time geometric bounds to convergence rate of Markov processes on R^n. *Stochastic models*, 18(2):301–331, 2002.

[164] Tobias Zahner, Tobias Lochbühler, Grégoire Mariethoz, and Niklas Linde. Image synthesis with graph cuts: a fast model proposal mechanism in probabilistic inversion. *Geophysical Journal International*, 204(2):1179–1190, 2016.

[165] Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, page 116, 2004.

[166] Bumeng Zhuo and Chao Gao. Mixing Time of Metropolis-Hastings for Bayesian Community Detection. *Journal of Machine Learning Research*, 22(10):1–89, 2021.