CONDITIONS FOR DETERMINISTIC LIMITS OF MARKOV JUMP PROCESSES: THE KURTZ THEOREM IN CHEMISTRY

by

Ada Sedova

A Thesis

Submitted to the University at Albany, State University of New York

in Partial Fulfillment of

the Requirements for the Degree of

Master of Arts

College of Arts and Sciences

Department of Mathematics and Statistics

2015

Abstract

A theorem by Kurtz on convergence of Markov jump processes is presented as it relates to the use of the chemical master equation. Necessary mathematical background in the theory of stochastic processes is developed, as well as requirements of the mathematical model necessitated by results in the physical sciences. Applicability and usefulness of the master equation for this type of combinatorial model in chemistry is discussed, as well as analytical connections and modern applications in multiple research fields.

Acknowledgements

I would like to thank Dr. Martin Hildebrand at UAlbany Mathematics for encouragement and patience with me as I worked on graduate degrees in two departments, and also for his unwavering sense of humor that always finds a way to make light of stressful situations. I would also like to thank Dr. Karin Reinhold for support and understanding, as well as Dr. N.K. Banavali for allowing me to take breaks from my PhD research to work on this thesis. Finally, I am grateful for the existence of many delightful texts listed in the bibliography which allowed me to learn much of this theory through self-study.

Contents

Chapter 1

Introduction

My work in chemical physics has involved frequent interactions with the "master equation" for the probability, as a function of time, describing the states of a chemical system. This equation is usually presented together with an intuitive apology for its attendance in the description. The master equation seemed to me, for some time, to have an origin and derivation as mysterious as its pervading usage. During my work on a particular computational model of intermediate states involved in a RNA folding process, I was told that the probabilities derived from solutions of the master equation corresponded to the macroscopic kinetic rates derived from experiments. In other words, I was told that the values of the kinetic rate constants and the constants in the probabilistic equations were identical. How such a correspondence could be possible was not explained to me, and I soon discovered that the reason behind the use of this "trick" was not entirely clear to most workers in chemistry and the physical sciences.

Finally, I happened upon the explanation: that the probabilistic solutions to the master equations actually converge to the macroscopic kinetic equations with rates equal to the constants in the stochastic formulation, in the "thermodynamic limit." Why and how this occurred was not clear to me. However, I was directed to a three-page paper found in a 1972 issue of the *Journal of Chemical Physics*, written by a mathematician. In it, Thomas G.

Kurtz [36] explains to the scientific community that, as had been hypothesized, a certain type of combinatorial, "mesoscopic" stochastic model of chemical reactions— involving systems of stochastic linear differential equations with coefficients set equal to the rates derived from experiment— did, in fact, converge in mean to the solution of the corresponding deterministic model for the same reactions— in a limit where the volume of the system was made very large, but the density was kept constant (as long as the sequence of initial densities of the stochastic processes converged to the initial conditions of the deterministic model). This limit, with the number of members of some chemical population and the system volume going to infinity, while the density, (or ratio of particle number to volume) is kept constant, is known as the thermodynamic limit.

Apart from a handful of theoretical physical scientists whose work involves extensive use of advanced mathematical methods [7, 10, 15, 28], I have not found mention of Kurtz's theorem. Neither have I found mention of the requirement for a proof of the correspondence between a stochastic description of chemical reactions and the deterministic one. It seems that the use of macroscopic rates, in the stochastic description which counts molecule numbers, has become an accepted experimental fact [28]: in certain chemistry textbooks the means of stochastic processes are presented as identical to the deterministic solutions, with no mention of the thermodynamic limit [47]. Through my research on this topic, I have tried to bridge my disconnect, and I have discovered that several decades ago, this question was a new and exciting theoretical problem whose details have since become forgotten, as the methods have become routine techniques. However, the use of such descriptions, and their mathematical subtleties, are far from straight-forward. It is important to understand the applicability and the limitations of this, as any, mathematical model, whether one approaches from the viewpoint of an applied mathematician or that of a physical scientist.

As early as the 1930s, scientists had attempted to apply the theory of stochastic processes to model molecular events [14]. From the 1950s through the late 1970s, chemists and physicists became well-versed in a type of stochastic process known as the Markovian birth-

death description, and were fervently attempting to apply it, to model chemical reactions—in order to gain more insight about the fluctuations around the deterministic curves seen in experiments [39]. These were assumed to originate from the actual molecular nature of the reactions, which the deterministic differential equations, with their continuous changes in concentration, failed to address [7, 47]. Connections between microscopic fluctuations and macroscopic transport processes had become apparent via work relating to (and extending) Einstein and Smoluchowski's descriptions of Brownian motion, and that of Langevin [15]. Based on some preliminary results modeling unimolecular reactions with discrete-state, continuous-time Markov chains and solving the associated Kolmogorov equations, it had been found that the means of the stochastic processes were *identical* to the solutions of the deterministic equations modeling the same processes, when the deterministic rates were used as the stochastic intensities. This result was similar to a now-classical correspondence found between the simple, deterministic, linear-rate, Malthusian population growth model and a stochastic description of population growth applied to genetic lineage by Yule [34].

Unfortunately, for more complex reactions this did not turn out to be true. However, as the particle number was allowed to grow large, while maintaining constant density, the stochastic means seemed to converge appropriately [39]. I ultimately found a series of articles [26], [27], (and references within), written around the same time as the Kurtz article, by a mathematical physicist named Joel Keizer, who was interested in the same correspondence but from the Fokker-Planck-type diffusion description. Kurtz's theorem was important to Keizer, because if the diffusion model was found to converge to the deterministic model, and the Markov-chain model converged to both the diffusion model and the deterministic model, the theories were consistent. Furthermore, using a synthesis of the results, Keizer was able to estimate the appropriate time scales for the correct application of each of the various descriptions, a highly important requirement for success in mathematical modeling of a physical process.

The article by Kurtz in the *Journal of Chemical Physics* did not include mathematical

details. Upon following its references to the *Journal of Applied Probability*, where the actual proof of the Kurtz theorem could be found [34, 35], I was a bit deterred by what seemed like a mathematical theory very distant from the probability theory I had been exposed to. I then found a piece by Keizer in a collection written in tribute to Marc Kac, with whom Keizer had worked, in which he formulated another proof of the Kurtz theorem using a model from chemical physics and a theorem of Kac's [28]. Although this proof was less transparent to me than what I had read in Kurtz's mathematical paper, Keizer's introduction gave me some encouragement:

> "In 1972 a short article by Kurtz [1] appeared in the *Journal of Chemical Physics* which explained to chemists the relevance of some of his work on limit theorems for stochastic processes. No proofs appear in that paper and for good reasons. First, Kurtz had previously published the proofs in the mathematics literature [2], and second the proofs make heavy use of martingale theorems, which are not standard fare for physicists and chemists. Kurtz's basic results state for certain birth and death processes that the averages, conditioned on a precise knowledge on the initial variables, satisfy simple differential equations when the size of the system becomes large. Furthermore, the deviations from the conditional average are nonstationary, Gaussian diffusion processes. For the birth and death processes used to model chemical reactions [1, 3], Kurtz's results imply that the phenomenological mass action laws are satisfied by the conditionally averaged concentration variables, a fact that had been anticipated in earlier work [4]."
>
> *(Here Keizer's citations 1,2,3, and 4 translate to* [36], [34, 35], [39], *and* [42], *respectively).*

Keizer describes Kac's reaction to the Kurtz paper as such: "The proof of Kurtz's theorem is, at best, difficult, and even Kac, after examining the proof, remarked that if such results appear in the *Journal of Applied Probability*, what must the *Journal of* ***Pure*** *Probability* be like!" [28]

The paper by Kurtz introduced me not only to the theory of stochastic processes, but to a deep analytical extension of this theory using methods from functional analysis and operator theory. These results can only be attained using a rigorous measure-theoretic description of probability spaces, and create a connection between stochastic differential equations and branches of potential theory and the theory of dynamical systems, including the Abstract Cauchy Problem. I discovered a rich mathematical field developed by the work of Dynkin, Skorokhod, Feller, Doob, Hille, Yosida, and others. It is interesting that to prove such a commonly accepted experimental idea requires such a detailed mathematical description and such advanced methods. On the other hand, these same areas of mathematics are the foundations for a number of other methods in physics, including the perturbation theory and semigroup methods used quantum mechanics, and the mathematical formalism used to interpret all spectroscopic experimental data [10, 15].

Does the knowledge of the details of Kurtz's proof change the way I choose to model physical phenomena? Possibly. For the computational physical scientist, having a deeper understanding of the mathematical theory, including the limitations and implications of its use in modeling, can only prove beneficial. As Karlin wrote in the introduction his classic text [24], one of his desires was "to make the student who is more concerned with application aware of the relevance and importance of the mathematical subtleties underlying stochastic processes." It is possible that fundamental errors may result from the improper application and incomplete understanding of a mathematical model.

The successful application of mathematical models to study physical phenomena relies in large part on effective communication between the mathematician and the physical scientist. It is important for the physical scientist to have awareness about the complexities and subtleties relating to the mathematical theory being used. Not only will results be easier to interpret if they do, or do not, agree with experiment, but awareness about the existence of multiple fields in mathematics as well as their connections to each other may prove to be useful at some point in other applications, and will surely improve communication with

the mathematical community. Alternately, from an applied mathematics perspective, it is crucial to understand the phenomena being modeled as deeply as possible, and the results of making certain assumptions and approximations.

For this reason, I have chosen to approach the writing of this text in the following way: I include textual explanations of essential mathematical results and limit the level of rigor for the more advanced topics to that of an introductory survey— one of my aims is to fill in the gap in knowledge that may exist in the scientific community about deeper results in stochastic processes, as well as to provide the mathematical community with an overview of a complex and continually developing modeling problem. Therefore I present some in-depth discussions about the use of Markov models in chemical kinetics and the limits of their validity. Additionally, it is important to understand how the current model relates to other models being used to describe the same phenomena and the differences in the applications, both from a theoretical and practical standpoint. Consequently I shall briefly present other methods and representations that relate to the chemical master equation technique.

Mathematics has frequently been both guided and inspired by its connection and relationship with physical theories [12], and as Feller wrote, "many parts of the purest mathematics owe their origin to physical problems." The equations resulting from the use of Markov models for biology, physics, and chemistry, such as the Kolmogorov equations, offered mathematics "integrodifferential equations of a type never studied before,"[12]. Therefore, even if the model is shown to be less accurate than expected, its associated methods offer stimulation for the development of new theories, which can in turn be used to explore other physical systems.

> "As for practical usefulness, it should be borne in mind that for a mathematical theory to be applicable it is by no means necessary that it be able to provide accurate models of observed phenomena. Very often in applications the constructive role of mathematical theories is less important than the economy of thought and experimentation resulting from the ease with which qualitatively reasonable

working hypotheses can be eliminated by mathematical arguments. Perhaps even more important is the constant interpretations of observations in the light of theory and of theory in the light of observations; in this way mathematical theory can become an indispensable guide not only to a better understanding, but even to a proper formulation of scientific problems." *William Feller* [12].

Chapter 2

Mathematical Basics

2.1 Preliminaries

In this thesis I assume the reader is familiar with basic, non-measure-theoretic probability theory, including definitions such as event, state space, random variable, probability density function, cumulative distribution function, expectation, etc. However, in this section I review some basic results and definitions that will be necessary for the following sections. This section follows presentations of Durrett, [8], Ross, [44], and Karlin, [24].

Definition. The exponential distribution with parameter λ:

The exponential (λ) distribution has probability density function (p.d.f.):

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & otherwise, \end{cases} \tag{2.1}$$

and cumulative distribution function (c.d.f):

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & otherwise. \end{cases} \tag{2.2}$$

Definition. The Poisson distribution with parameter λ:

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k = 0, 1, \ldots \tag{2.3}$$

Proposition. The sum of two independent Poisson random variables is a Poisson random variable.

Proof: The moment generating function for the Poisson distribution is:

$$\phi_x(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}e^{tk} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda(1-e^t)} \tag{2.4}$$

If X and Y are two independent random variables, and $X \sim \text{Poisson}(\lambda), Y \sim \text{Poisson}(\mu)$, then

$$\phi_{x+y} = \phi_x\phi_y = e^{-\lambda(1-e^t)}e^{-u(1-e^t)} = e^{-(\lambda+u)(1-e^t)} \longrightarrow X + Y \sim \text{Poisson}(\lambda + u) \quad \square \tag{2.5}$$

Definition. The gamma (n, λ) distribution:

$$\Gamma(n, \lambda) = \begin{cases} \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!} = \frac{e^{-\lambda t}\lambda^n t^{n-1}}{\Gamma(n)} & t \geqslant 0 \\ 0 & otherwise \end{cases} \tag{2.6}$$

Proposition. The sum of n independent exponential (λ) distributions is a gamma(n, λ) distribution: If $X_1, X_2, \ldots, X_n$ are independent exponential(λ) distributions, then if $Y_n = X_1 + X_2 + \cdots + X_n$,

$$\mathbb{P}(Y_n = t) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}, \text{ for } t \geqslant 0, \ 0 \text{ otherwise.} \tag{2.7}$$

In other words, Y_n is gamma(n, λ).

Proof:

(By induction). We see for $n = 1$, this is true. Also, for $n = 2$, we can compute the probability distribution for an exponential waiting time of s, plus another exponential waiting time of

$t - s$, for a total waiting time of t:

$$\int_0^t \lambda e^{-\lambda s} \lambda e^{-\lambda(t-s)} ds = \lambda e^{-\lambda t} \cdot \lambda t, \tag{2.8}$$

which is gamma$(2, \lambda)$. Now, assuming the statement is true for n, then for $n+1$, we have:

$$\begin{gathered}\int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} \lambda e^{-\lambda(t-s)} ds = \int_0^t \lambda^2 e^{-\lambda t} \frac{(\lambda s)^{n-1}}{(n-1)!} ds = \lambda e^{-\lambda t} \lambda^n \left[\frac{s^n}{n!}\right]_0^t = \\ \frac{\lambda e^{-\lambda t} \lambda^n t^n}{n!} = \Gamma(n+1, \lambda) \quad \square\end{gathered} \tag{2.9}$$

Definition. The probability generating function, or simply the **generating function**: If X takes on only nonnegative integer values, we can use a different function (besides the characteristic function or the moment-generating function) to obtain the moments. For a nonnegative, integer-valued random variable, define

$$g_X(s) = \mathrm{E}[s^X] = \sum_{k=0}^{\infty} s^k \mathbb{P}(s = k) \tag{2.10}$$

Differentiating the generating function and setting $s = 1$, we obtain the factorial moments:

$$g_X(s)' \Big|_{s=1} = \sum_{k=1}^{\infty} k s^{k-1} \mathbb{P}(s = k) \Big|_{s=1} = \mathrm{E}\left[k\right], \tag{2.11}$$

$$g_X(s)'' \Big|_{s=1} = \sum_{k=2}^{\infty} k(k-1) s^{k-2} \mathbb{P}(s = k) \Big|_{s=1} = \mathrm{E}\left[k(k-1)\right], \text{ etc.} \tag{2.12}$$

Properties:

$g(s)$ determines the distribution function uniquely, and for independent nonnegative integer-valued random variables,

$$g_{X+Y}(s) = g_X(s) g_Y(s) \tag{2.13}$$

Definition. The Laplace transform: For nonnegative random variables we can also use the Laplace transform instead of the characteristic function:

$$\begin{aligned} \psi_X(s) &= \sum_{n=0}^{\infty} e^{-s\lambda_n} \mathbb{P}(X = \lambda_n), \text{ or} \\ \psi_X(s) &= \int_0^{\infty} e^{-sX} p_X(x) dx \end{aligned} \tag{2.14}$$

2.2 Stochastic Processes- An Introduction

A stochastic process is a sequence of events with associated probabilities. We say the stochastic process $X_t, t \in T$, is a collection of random variables such that for each t, X_t is a random variable. The index is often interpreted as time and, as a result, we can say X_t is the state of the process at time t. T is the index set of the process. Thus the process can be thought of as moving forward in time, through a series of time steps. A state space is the set of possible states or values for X_t. For a stochastic process, both the state space and the index set can be defined to be discrete or continuous. By discrete we mean finite or countable.

Stochastic processes can be differentiated by state space, index parameter, and the type of dependence that exists between the random variables. Therefore we can have a discrete-time, discrete-space process, a discrete-state continuous-time process, etc. In this paper I will focus primarily on discrete-state stochastic processes that occur in either discrete time or continuous time.

Dependence relations among variables are defined by stating the joint distribution functions for every finite family of random variables of the process. We can also talk about the difference in values between steps of the process, or its increments, which are also random variables. Several types of relations between the increments will prove to be important:

Definition. Independent increments: If the random variables

$$X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, \ldots, X_{t_n} - X_{t_n - 1} \tag{2.15}$$

are independent for all choices of $t_1, \ldots, t_n$, satisfying $t_1 < t_2, \ldots, < t_n$, we say X_t is a process with independent increments.

Definition. Stationary increments: X_t is a stochastic process with stationary increments if the distribution of $X_{t_1+h} - X_{t_1}$ is equal to the distribution of $X_{t_2+h} - X_{t_2}$ for all $t_2, t_1,$ and h.

The simplest type of stochastic process would have every time step independent of every other one. Because this type of process would be uninteresting and not very useful for applications, we would like the variables to have some type of memory about their past. This is because in many physical processes, such as the motion of a particle in some medium, the particles past history (in terms of velocity, position, momentum, etc.) has some influence on its future trajectory. However, taking account of long periods of past history proves to be highly difficult and cumbersome mathematically. Therefore, we define a type of process that has no memory of its past, but its *current state* and its *next-step transition* are dependent, probabilistically. Thus the process is defined via a conditional probability:

Definition. Markov Processes: A process is *Markovian* if

$$\begin{aligned} \mathbb{P}\{a < X_t \leqslant b | X_{t_1} = x_1, X_{t_2} = x_2, \ldots, X_{t_n} = x_n\} = \mathbb{P}\{a < X_t \leqslant b | X_{t_n} = x_n\}, \\ \text{for } t_1 < t_2 < \cdots < t_n < t. \end{aligned} \tag{2.16}$$

A realization, or sample function, of a stochastic process X_t is an assignment to each t some value of X_t. Here I will call a Markov process with a finite or denumerable state space a *Markov chain*, whether the time index is discrete or continuous. A Markov process for which all sample functions are continuous is called a *diffusion process*.

I present discrete-time Markov chains, continuous-time Markov chains, and continuous state space, continuous time Markov processes, but I will focus on continuous time Markov

chains, which are a type of "jump process." Discrete state space, continuous time processes are the types most commonly found in scientific applications, because the description is simpler, although the solutions of associated differential equations may be more difficult to compute [17]. For computational applications, discrete time chains are also common, and for those there is an established theory that provides many methods for performing calculations [8].

2.3 Markov Chains: Discrete Time

In this section a brief survey of essential definitions and results from the theory of discrete-time Markov chains is presented, for finite and infinite, but countable, state spaces. Although this thesis will focus on continuous-time Markov chains and their use in modeling chemical reactions, certain key results from the theory of discrete-time chains have analogies in, or offer insights into, the continuous-time theory. Additionally, for computer simulations of Markov processes, the time domain must necessarily be discrete, even if the process being simulated is a continuous-time process, in which case the simulation must approximate the continuous-time domain. Therefore, an understanding of the discrete-time process is important. This section follows the expositions of Karlin [24], Durrett [8], Anderson [1], Lawler [37], and Ross [44]. Proofs of some theorems and lemmas and further details can be found there.

A discrete-time Markov chain, X_n, has a finite or countable state space and also $T = \{0, 1, 2, \dots\}$. As the time index advances by one, the process transitions to the next state (which may possibly be the same state as current state). The position in time may or may not be important to the process. If the probability of transition from a state j to a state k depends on the current value of the time index, we call the *process* nonstationary. If the same probability does not depend on the position in time but only the length of the time interval, we call the process stationary. Therefore for the one-step transition probability of a nonstationary Markov chain, we have $P_n(i, j)$, where n is the time index, and i and j are the

events e_i and e_j. A stationary Markov process is also called a time-homogeneous process, and is completely determined by

$$P(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i) \text{ for any } n, \tag{2.17}$$

and the initial distribution or state.

In this paper I will only consider stationary processes. The above equation is the *Markov property* for discrete time Markov chains. It states that the conditional probability that the process is in state j one time step after it was in state i, given the entire history of the process from the initial state, is the same as the conditional probability given only the previous state and time. In other words, the process has a memoryless property.

For discrete time Markov chains, the transition probabilities can be arranged in matrix form, where the current state is represented by the row, the next state by the column, and the entry (i,j) is the probability of transitioning from state i to state j. This is known as the *transition matrix*. The transition matrix must have all elements $\geqslant 0$, and the row sums must equal 1, because the probability of transitioning to one of any of the possible states must be a sure event. An *absorbing state* is a state for which the probability of transitioning back into the same state on the next step (i.e. staying in the current state) is 1, and thus the probability of transitioning into any other state is 0.

The probability $P^m(i,j) = \mathbb{P}(X_{n+m} = j | X_n = i)$ is the probability that, given the value of X was i at time n, it reaches a value of j, m steps later. To offer insight about what this probability may be, we consider a three-state chain, and ask what is the probability that we are at state 1 on day $n+1$, given we were in state 2 on day $n-1$?

$$\mathbb{P}(X_{n+1} = 1 | X_{n-1} = 2) = \sum_{k=1}^{3} \mathbb{P}(X_{n+1} = 1 | X_{n-1} = 2) = \sum_{k=1}^{3} P(2,k)P(k,1) \tag{2.18}$$

In words, we must sum over the probabilities of all possible intermediate states in the transition from state 2 to state 1. Likewise, for the transition from state i to state j in $m+n$ steps, we can split the path into two pieces, and due to the Markov property, the probabilities of the individual pieces are independent. Thus we have the Chapman-Kolmogorov equation:

$$P^{m+n}(i,j) = \sum_k P^m(i,k)P^n(k,j) \tag{2.19}$$

The above equation is also the equation for matrix multiplication for a matrix P^m by a matrix P^n. Therefore, since we can write the transition probabilities in matrix form, we see that the Chapman-Kolmogorov equations imply that the m-step transition probability is the m^{th} power of the one-step transition matrix $\mathbf{P}$.

2.3.1 Long-term behavior of the discrete-time Markov chain

Now that we see that multiple steps of the process are expressible as powers of the transition matrix, we can suppose that there may be some results relating to long term behavior of the chain, after many steps, and that this behavior may be related to results from linear algebra. In fact, this is true. For some finite transition matrices, a simple computer calculation can show that powers of the matrix can rapidly converge to a matrix whose rows are identical. This implies that there may exist a limiting distribution, such that after some time, the probabilities of transitioning into a particular state from any other state are the same, and are independent of the starting distribution. These probabilities can also be seen as a limiting fraction of time spent in each state, or as a fixed distribution of the states. For these matrices, the limiting *distribution* can be thought of as "stationary." However, we will see that the term "stationary distribution" will be defined in a different way, and will apply to both cases where the matrix powers converge, and when they do not. Studying limit behaviors of the transition matrices can help to answer some important questions, such as: if there is an absorbing state, what is the probability that ultimately, the process will be absorbed or will

avoid absorption? More generally, what fractions of the states will ultimately be populated if the process stabilizes over long time scales? To answer these questions, it is necessary to classify states and sets of states based on their potential for recurrence and their interaction with each other.

Definition. Classification of States of a Markov chain:

Accessible $\longrightarrow$: A state y is accessible from a state x, $x \longrightarrow y$ if there is a positive probability of transitioning from state x to state y, directly or via any series of transitions through other states: $P^n(x, y) > 0$ for some $n \geqslant 0$.

Communicates $\longleftrightarrow$: Two states communicate if each is accessible to the other. We write $x \longleftrightarrow y$.

Communication, or $\longleftrightarrow$, has the following properties: reflexivity, symmetry, and transitivity. Thus it is an equivalence relation and partitions the state space into equivalence classes. The space is called **irreducible** if the equivalence relation induces only one class.

Definition. Time of first return: Let $T_y = \min\{n \geqslant 1 : X_n = y\}$ be the first return to a state, *after having been there once before at time* $n = 0$. Then the probability that a process that starts at y ever returns to y can be written as $f_y = \mathbb{P}(T_y < \infty | X_0 = y)$.

Definition. Stopping time: T is a stopping time if the occurrence of an event at $\{T = n\}$ is completely determinable by the values of the process up to that time. In others words, if we had to decide to stop at time n, based on some criteria, we would know it was time to stop when $\{T = n\}$. The time of a first return to a state, T_y, is a stopping time. However, the time of *last* return to a state is not a stopping time, because we do not know if the current state is the last time the state will be visited, based only on the past history of events. NOTE: T is a random variable.

Definition. Strong Markov Property (for discrete-time processes): Let T be a stopping time. Then given $T = n$ and $X_T = y$, X_{T+k} for $k \geqslant 0$ is also a Markov chain with initial state y.

As a result of the Strong Markov Property, the probability of a *second* return to y is $(f_y)^2$, and thus the probability of an n^{th} return is $(f_y)^n$. Because of this, we see that there are two possibilities for $(f_y)^n$ as $n \longrightarrow \infty$. If $f_y < 1$, then $(f_y)^n \longrightarrow 0$ as $n \longrightarrow \infty$. If $f_y = 1$, then $(f_y)^n = 1$ for all n, so the chain returns to y infinitely many times.

Thus we have the following definitions:

Definition. Transient state: A state is transient if $f_y < 1$. The probability of returning to it eventually goes to zero: $(f_y)^n \longrightarrow 0$ as $n \longrightarrow \infty$.

Definition. Recurrent state: A state is recurrent if $f_y = 1$. The probability of the state's repeated occurrence in the chain, $(f_y)^n$, also equals 1. The state will continually be returned to as the process proceeds in time.

We can now express the concept of accessibility more formally, using the Strong Markov Property:

Definition. Accessible (second definition): A state x is accessible to a state y if

$$\mathbb{P}(T_y < \infty | X_0 = x) > 0 \tag{2.20}$$

Using the above, we can now derive the following

Theorem. If y is accessible from x, but x is not accessible from y, then x is transient. $\square$

The sets of states of a Markov chain, due to the fact that communication of states defines an equivalence relation, can be partitioned into disjoint sets and classified in the following way:

Definition. Classification of Sets of the State Space:

Closed set- a set of states is closed if it is impossible to ever leave the set:

$$\text{If } A \text{ is closed and } i \in A \text{ and } j \notin A, \text{ then } P(i,j) = 0. \tag{2.21}$$

Irreducible set- a set S is irreducible if, whenever $i, j \in S$, then j is accessible from i.

Lemma. If x is recurrent and $x \longrightarrow y$, then y is recurrent. □

Lemma. In a finite, closed set there must be at least one recurrent state. □

Theorem. If S is a finite, closed, irreducible set, then all states in S are recurrent.

□

Theorem. Any finite state space of a discrete-time Markov chain can be partitioned into a disjoint union $T \cup R_1 \cup \cdots \cup R_k$, where T is a set of all the transient states, and the R_i are closed, irreducible sets of recurrent states. □

Let $P_x(A) = \mathbb{P}(A|X_0 = x)$ be the probability of A given that the initial state was x, and E_x be the expectation with respect to this initial state. Let $N(y)$ be the total number of visits to y, *after* time $n = 0$, i.e. for times $n \geqslant 1$. Thus if $X_0 = y,\ N(y)$ counts returns to y.

Because of the Strong Markov Property, the probability that we make k visits to y, when we started at x, equals $P_x(T_y < \infty)(P_y(T_y < \infty))^{k-1}$. Let $i_N(k) = 1$ if $N \geqslant k$, and 0 otherwise. Then since $N = \sum_{k=1}^{\infty} i_N(k)$, we have that $\mathrm{E}(i_N(k)) = \mathbb{P}(N \geqslant k)$, and thus

$$\mathrm{E}N = \sum_{k=1}^{\infty} \mathrm{E}(i_N(k)) = \sum_{k=1}^{\infty} \mathbb{P}(N \geqslant k). \tag{2.22}$$

But $\mathbb{P}(N \geqslant k) = P_x(T_y < \infty)(P_y(T_y < \infty))^{k-1}$, thus we have

$$\begin{aligned} \mathrm{E}_x N(y) = \sum_{k=1}^{\infty} P_x(T_y < \infty)(P_y(T_y < \infty))^{k-1} = P_x(T_y < \infty) \sum_{k=1}^{\infty} (P_y(T_y < \infty))^{k-1} \\ = \frac{P_x(T_y < \infty)}{1 - P_y(T_y < \infty)} \end{aligned} \tag{2.23}$$

From the above we can see that $\mathrm{E}_y N(y) = \infty$ if and only if $P_y(T_y < \infty) = 1$, which is the definition of recurrence. Furthermore, noting that $N(y) = \sum_{n=1}^{\infty} 1_{\{X_n = y\}}$, we see that $\mathrm{E}_x N(y) = \sum_{n=1}^{\infty} P_x(X_n = y) = \sum_{n=1}^{\infty} P^n(x, y)$. Thus we have proved the following theorem:

Theorem. y is recurrent if and only if

$$\sum_{n=1}^{\infty} P^n(y, y) = \mathrm{E}_y N(y) = \infty \quad \square \tag{2.24}$$

Thus we see that a state is either recurrent or transient— there is no in-between, and if it is recurrent, the number of times we visit that state will be infinite with probability one. If a state is transient, then it will NOT be visited infinitely, with probability one, and it will eventually never be visited again. Furthermore, the number of times it is visited has a finite expectation. We also see that if a chain has absorbing states which are accessible from other states, then the chain will eventually stay in one of those states, and the states that communicate with it are necessarily transient and will eventually be unpopulated.

Periodicity and limiting behavior

In a discrete-time Markov chain, we have the possibility that a particular set of states can recur in regular alternation for all time, thus preventing the convergence to a limiting distribution. For instance, $P^n(x, x)$ may equal zero for n odd, or unless n is a multiple of some number. A classic example is the Ehrenfest chain, where we have two urns containing a total of N balls. We pick one ball at random and transfer it into the other urn. The states consist of the number of balls at time n in only one chosen urn, say urn 1. The transition

matrix for $N = 3$ is then:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{2.25}$$

From this we see that if we go from one to zero, we cannot go back to zero until two steps later. In general, the zeros will persist infinitely, and will prevent convergence of $P^n(x, y)$ as $n \longrightarrow \infty$. The behavior of the chain is cyclical. Using this idea we can formulate the following definition:

Definition. The **period** of a state is the greatest common divisor of all n such that $P^n(x, x) > 0$.

A chain with all states having period 1 is called aperiodic. Obviously, if $P(x, x) > 0$, then x has period 1. However, a state whose one-step transition probability to itself is zero may still have period 1, for example, if it is possible to go from x back to x in either 5 steps or in 6 steps.

Lemma. If x and y communicate, then x and y have the same period. □

From this we can see that if all states communicate with each other in a chain, in other words, if the chain is irreducible, and one state has period 1, then all states have period 1, and we can say that the *chain* is aperiodic. It seems that for an irreducible, aperiodic chain, convergence to some limiting distribution is possible. From our initial experiment with a converging matrix, we define:

Definition. A **stationary distribution** is a solution of the matrix equation $\pi\mathbf{P} = \pi$, where $\mathbf{P}$ is the transition matrix, and π is a row vector.

Theorem. If the state space is finite, then there is at least one stationary distribution.

□

In the introductory discussion, π would be one of the identical rows in our converging matrix. Note, however, that this definition does not mention any convergence. We guess that if $P^n(x, y)$ converges as $n \longrightarrow \infty$ it must be to $\pi(y)$. Interestingly, even if convergence cannot occur due to periodicity, in a *finite* state space, we still have a solution to the equation $\pi\mathbf{P} = \pi$. We can see that π is a left eigenvector of the matrix $\mathbf{P}$, with eigenvalue 1. If the chain is reducible with r recurrence classes, then we will have r different solutions to the equation $\pi\mathbf{P} = \pi$; eigenvalue 1 will have multiplicity r. If the chain is irreducible but periodic with period d, $\mathbf{P}$ will have d eigenvalues with absolute value 1: z with $z^d = 1$, z complex. Each of these is a simple eigenvalue, and corresponding to the simple eigenvalue of 1 is a unique solution to $\pi\mathbf{P} = \pi$.

Theorem. If P is irreducible, then the stationary distribution is unique. □

Therefore, if the chain is irreducible and periodic, we still have a unique stationary distribution. This distribution gives us a limiting percentage of time spent in each state. This can be shown using a strong law of large numbers for Markov chains:

Theorem. Suppose P is irreducible and has stationary distribution π. Let $r(x)$ be some function of state x such that $\sum_x \pi(x)|r(x)| < \infty$. Then

$$\lim_{n\longrightarrow\infty} \frac{1}{n} \sum_{k=1}^{n} r(X_k) = \sum_x \pi(x) r(x), \tag{2.26}$$

Putting $r(y) = 1$ and $r(x) = 0$ for $x \neq y$, $\sum_{k=1}^{n} r(X_k)$ gives the sum of the number of times we reach state y by time n, or $N_n(y)$. Thus we have

$$\lim_{n\longrightarrow\infty} \frac{N_n(y)}{n} \longrightarrow \pi(y) \tag{2.27}$$

showing that π is the limiting fraction of time spent in the state y. □

Finally, if P is neither reducible nor periodic, we have convergence of P^n to the stationary distribution π as $n \longrightarrow \infty$:

Theorem. If P is irreducible, aperiodic, and has stationary distribution π, then $P^n(x, y) \longrightarrow \pi(y)$ as $n \longrightarrow \infty$. $\square$

2.3.2 Infinite state space considerations

Several results above were limited to the case of a finite state space. For a Markov chain in discrete time but with countable but infinite state space, additional definitions are necessary. The classic example of difficulties that may arise in the case of the infinite state space is the one-dimensional random walk with a one-sided reflection: a particle can make one step to the right with probability p, one step to the left with probability $1 - p$, but if it reaches zero, it cannot go further to the left so it either stays at zero with probability $1 - p$ or moves back right. For this chain we have three separate types of limiting behavior, in the case $p > \frac{1}{2}, p < \frac{1}{2},$ or $p = \frac{1}{2}$. A stationary distribution exists only for the case $p < \frac{1}{2}$. However, it can be shown that for $p = \frac{1}{2}$, the probability, starting from any point $x \geqslant 0$, that we return to 0 in a finite amount of time, is 1. On the other hand, when we try to calculate the expectation time for our first return to 0, we obtain infinity: $\mathbb{P}(T_0 < \infty | X_0) = 1$, but $\mathrm{E}(T_0 | X_0 = x) = \infty$.

A similar situation occurs in a population model called binary branching, where the states of the Markov chain is the population number at time n. Here it can be shown that, for the case where the mean number of offspring per individual is 1, the probability for eventual extinction as $n \longrightarrow \infty$ is 1, but the *expected time* for extinction is infinite. Thus we must create two additional definitions:

Definition. A state x is called **positive recurrent** if $\mathrm{E}(T_0 | X_0 = x) < \infty$.

Definition. A state x is called **null recurrent** if it is recurrent but not positive recurrent:

$\mathbb{P}(T_0 < \infty | X_0 = x) = 1$, but $\mathrm{E}(T_0 | X_0 = x) = \infty$

The definition for a transient state is the same as in the case of a finite state space.

We now can state the following theorems:

Theorem. For an irreducible discrete-time Markov chain the following are equivalent:

1) Some state is positive recurrent

2) All states are positive recurrent

3) There is a stationary distribution π.

Theorem. If P is irreducible, aperiodic, and all states are *positive recurrent*, then

$$\lim_{n \longrightarrow \infty} P^n(x, y) = \pi(y) \tag{2.28}$$

where π is the unique nonnegative solution of $\pi P = \pi$, with $\sum_i \pi(i) = 1$.

2.4 Markov chains in continuous time

This section follows the expositions of Durrett [8], Karlin [24], Feller [13], and Norris [41], with additional details from Barucha-Reid [5], Vrbik [54], and Bartholomay [2].

2.4.1 Introduction to continuous time Markov chains:

Continuous time Markov chains are Markov processes with a discrete state space but a continuous time parameter. Because the possible transition times are no longer denumerable, the Markov property takes on a slightly different form:

Definition. $X_t, t \geqslant 0$, is Markovian if, for any $0 \leqslant s_0 < s_1 < \cdots < s_n < s$,

$$\mathbb{P}(X_{t+s} = j | X_s = i, X_{s_n} = i_n, \ldots, X_{s_0} = i_o) = \mathbb{P}(X_{t+s} = j | X_s = i). \tag{2.29}$$

Note that once again we assume the process is time-homogeneous, in that the transition probabilities only depend on the time difference $(t+s)-s = t$ and not the value of $t+s$ or s.

The notation for the transition probabilities now contains an index for the time increment, t:

$$p_t(i,j) = \mathbb{P}(X_t = j | X_0 = i). \tag{2.30}$$

The continuous-time version of the Chapman-Kolmogorov (CK) equation is:

$$\sum_k p_s(i,k) p_t(k,j) = p_{s+t}(i,j). \tag{2.31}$$

In discrete and continuous time, it is possible to construct stochastic processes which satisfy the Chapman-Kolmogorov equations, but are not Markovian. Thus, although it was conjectured for some time that the Markov property could be defined by the CK equations, this is not the case. However, if a process is Markovian and satisfies the CK equations, then it is a unique *Markovian* solution to these equations, and the Markovian process is uniquely and completely determined by its transition probabilities and the initial condition.

For the continuous-time Markov chain, because transitions can happen at any time on the continuum and not in a regularly-spaced way as in the discrete-time case, it makes sense to talk about an average rate of transitions per unit time. If the unit of time is allowed to go to zero, under appropriate assumptions for the process, we are able to define an instantaneous jump rate.

2.4.2 Exponential distribution of waiting times

Let T_i be the waiting time until transition into the next state. This is also sometimes called the "holding time." By the Markov property, T_i must be memoryless and thus must be exponentially distributed.

Proposition. Memoryless property of the exponential distribution:

$$\begin{aligned} \mathbb{P}(X > t) &= 1 - \mathbb{P}(X \leqslant t) = 1 - F(t) = e^{-\lambda t}, \\ \mathbb{P}(T > t + s | T > t) &= \frac{\mathbb{P}(T > t + s)}{\mathbb{P}(T > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(T > s). \quad \square \end{aligned} \tag{2.32}$$

We have shown that the exponential distribution has the memoryless property, but we must show that it is the only such function with this property. This proof is not as easy, and a completely rigorous version requires some advanced methods.

Theorem. The exponential equation: This is the functional equation $u(s+t) = u(s)u(t)$. It has been shown that monotone solutions of this equation are necessarily of the form $e^{-\lambda t}$. According to Feller, "This is a logarithmic variant of the famous Hamel equation $f(t + s) = f(t) + f(s)$. [It can be proved that] its solutions are either of the form at or else unbounded in every interval," [13]. If we are allowed to assume that u is differentiable, we can show the following:

Proof:

Let $u(s) = 1 - F(s)$, where $F(s)$ is the c.d.f. of some distribution. Also, assume $u(s)$ is differentiable. Then, if

$$\begin{aligned} &u(s + t) = u(s)u(t), \\ &u'(s + t) = \frac{\partial}{\partial s}[u(s + t)] = u'(s)u(t), \text{ and} \\ &u'(s + t) = \frac{\partial}{\partial t}[u(s + t)] = u(s)u'(t), \end{aligned} \tag{2.33}$$

Thus $u'(s) = au(s)$, where $a = \frac{u'(t_0)}{u(t_0)}$, for some t_0 such that $u(t_0) \neq 0$. The solution to this differential equation is Ae^{-as}. Because $u(0) = 1 - F(0) = 1,\ A = 1$. Because $u(s)$ can be $< 1, a$ is a negative number. Thus $u(s) = e^{-\lambda s}$ for $\lambda > 0$. $\square$

Now we can provide another description of the continuous time Markov chain using the property of exponentially distributed waiting times.

Definition. Alternate definition for a continuous time, discrete state Markov process: A continuous time, discrete state stochastic process is Markovian if, with each transition into state i,

1) the waiting time before transition into a different state is *exponentially distributed*,

2) upon leaving state i, it enters state j with probability $P(i,j)$, where as in the discrete time version, $\sum_j P(i,j) = 1$, and additionally, $P(i,i) = 0$,

3) the waiting times and the transition probabilities are independent random variables.

The following type of process will be helpful for use in the next section:

Definition. Counting process: A stochastic process $\{N(t), t \geqslant 0\}$ is a counting process if $N(t)$ is a number of events that occur by time t. Formally, $N(t)$ must satisfy:

1) $N(t) \geqslant 0$,

2) $N(t)$ is integer-valued,

3) If $s < t,\ N(s) \leqslant N(t)$,

4) For $s < t,\ N(t) - N(s)$ is the number of events that occur in the interval $(s,t]$.

A counting process has independent increments if the number of events that occur in disjoint time intervals are independent. It has stationary increments if the distribution of the number of events that occur in any interval of time depends only on the length of the time interval.

2.4.3 Poisson Process

The Poisson process is the simplest continuous time Markov chain, which is what makes it the classical starting point to examine these types of processes. Yet it has some interesting properties and numerous applications in the physical sciences as a descriptive and predictive model. For the Poisson process, the only possible transition from a state n is to state

$n+1$, and the rate of transitions is constant. Phenomena modeled by Poisson processes include rates of chromosome breakages and disintegrations of radioactive particles [13, 24]. Such physical processes involve the occurrence of relatively rare events: for instance, for radioactive disintegration, the Poisson model is more accurate if the observation time is brief compared to the half life of the radioactive particles. Furthermore, successfully modeled phenomena are those for which we can assume that the averages of all forces affecting the physical process remain constant [13].

Consider the zero term of the Poisson distribution with parameter λt. For fixed t, we have our basic discrete Poisson distribution, which can also can be understood as the probability that no event occurs within a fixed time interval of length t. But if we allow t to be a continuous variable, the same expression can also be seen as the probability that the waiting time for the first event exceeds t, so that now a discrete probability distribution becomes a continuous random process.

Consider this latter definition of the Poisson function, and let the k^{th} term in the expression $e^{-\lambda t}\frac{(\lambda t)^k}{k!}$ be the probability that k events have occurred by time t. Each state k in the process is defined by the expression "k events have occurred," thus with each new event, we have a "jump" to the next state along the time continuum. Define $P_n(t)$ as the number of events occurring by time t, and thus let $P_0(h)$ be the probability that no event, or jump, occurs by time h (thus the waiting time exceeds h). Partition a time interval of unit length into N subintervals of length $h = \frac{1}{N}$. The probability of a jump within any one of these subintervals is $1 - P_0(h)$, and the expected number of subintervals containing a jump must thus be $N(1 - P_0(h)) =$

$$\frac{1 - P_0(h)}{h} \tag{2.34}$$

We imagine that as $h \longrightarrow 0$, this expression should converge to the expected number of jumps within any time interval of unit length, in other words the "rate" of jumping:

$$\frac{1 - P_0(h)}{h} \longrightarrow \lambda \text{ as } h \longrightarrow 0. \tag{2.35}$$

This was an intuitive introduction to the idea of a Poisson process. Formally, we can define the Poisson process in a number of ways; here I show three equivalent definitions which each illuminate a different aspect of the process.

Definition. Poisson Process I: Let $t_1, t_2, \ldots$ be independent exponential (λ) random variables, and let $T_n = t_1 + \cdots + t_n,\ n \geqslant 1$, and $T_0 = 0$. Then a **Poisson process** $N(s)$ is a counting process, such that

$$N(s) = \max\{n : T_n \leqslant s\}. \tag{2.36}$$

From the above, we can derive the functional form of the distribution:

$$\begin{gathered}
\mathbb{P}(t_{n+1} > s - t) = e^{-\lambda(s-t)} \\
\mathbb{P}(T_{n+1} > s | T_n = t) = \mathbb{P}(t_{n+1} > s - t) = e^{-\lambda(s-t)} \\
\mathbb{P}(N(s) = n) = \int_0^s \mathbb{P}(T_n = t)\mathbb{P}(T_{n+1} > s | T_n = t)dt \ = \\
\int_0^s \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda(s-t)} dt
\end{gathered} \tag{2.37}$$

(because the sum of n exponential (λ) distributions is a gamma (n, λ) distribution, from section 1),

$$= \frac{(\lambda t)^n}{(n-1)!} e^{-\lambda s} \int_0^s t^{n-1} dt \ = \frac{(\lambda s)^n}{(n)!} e^{-\lambda s} \quad \square \tag{2.38}$$

Before we give the second definition of a Poisson Process, we must prove the following

Lemma. If $N(s)$ is a Poisson process, as defined above, then $N(t+s) - N(s),\ t \geqslant 0$, is a rate λ Poisson process independent of $N(r), 0 \leqslant r \leqslant s$. Additionally, $N(t)$ has independent increments.

Proof:

By the "sum of Poisson distributions property" proved in section 1, $N(t+s) - N(s)$ is Poisson distributed. Because of the memoryless property of the exponential distribution, and by the

properties of a counting process, the increments are independent. $\square$

Conversely, if $N(t+s)-N(s)$ is Poisson (λ) distributed, and has independent increments, they must be exponentially distributed, thus $N(s)$ is also Poisson (λ).

Thus we have established an equivalent definition:

Definition. Poisson Process II: A counting process $N(s)$ is a Poisson process if and only if:

1) $N(0) = 0$,

2) $N(t+s) - N(s)$ is Poisson (λt),

3) $N(t)$ has independent increments.

We now derive the third definition of the Poisson Process. Assume we have a counting process $N(s)$, and let $P_m(t)$ be the probability that the number of events occurring by time t is m. Assume also that the process is time-homogeneous and has independent increments. Furthermore, assume that the probability of one event occurring in a time h, as $h \longrightarrow 0$, is $\lambda h + o(h)$, (where a function f is said to be $o(h)$ if $\frac{f(h)}{h} \to 0$ as $h \to 0$), and the probability of two or more events occurring in time h is $o(h)$. Thus we have the $\sum_{m=2}^{\infty} P_m(h) \longrightarrow 0$ as $h \longrightarrow 0$ as well.

Therefore, for $m = 0$, we can write:

$$\begin{aligned} P_0(t+h) &= P_0(t)P_0(h) \text{ (by independent increments)}, \\ &= P_0(t)\left(1 - P_1(h) - \sum_{m=2}^{\infty} P_m(h)\right), \\ P_0(t+h) - P_0(t) &= -P_0(t)P_1(h) - P_0(t)\left(\sum_{m=2}^{\infty} P_m(h)\right) \\ \implies \frac{P_0(t+h) - P_0(t)}{h} &= \frac{-P_0(t)P_1(h)}{h} - \frac{P_0(t)\left(\sum_{m=2}^{\infty} P_m(h)\right)}{h} \end{aligned} \tag{2.39}$$

Now taking the limit as $h \longrightarrow 0$, we have $P'(t) = -\lambda P_0(t)$, which is a differential equation whose solution is $P_0(t) = Ce^{-\lambda t}$. Since $P_0(0) = 1, C = 1$, so $P_0(t) = e^{-\lambda t}$. We must now

solve for the remaining $P_m(t)$.

$$\begin{aligned} &P_m(t+h) = P_m(t)P_0(h) + P_{m-1}(t)P_1(h) + \sum_{i=2}^{m} P_{m-i}(t)P_i(h) \\ &\text{Note : } \sum_{i=2}^{m} P_{m-i}(t)P_i(h) \ \leqslant \sum_{i=2}^{m} P_i(h) = o(h), \end{aligned} \tag{2.40}$$

(because probabilities are $\leqslant 1$). Thus $\sum_{i=2}^{m} P_{m-i}(t)P_i(h) = o(h)$. Also, $P_0(h) = 1 - P_1(h) + o(h)$. So we have:

$$\begin{aligned} &P_m(t+h) = P_m(t)[1 - P_1(h) + o(h)] + P_{m-1}(t)P_1(h) + o(h) \\ &\implies P_m(t+h) - P_m(t) = -P_m(t)P_1(h) + P_m(t)o(h) + P_{m-1}(t)P_1(h) + o(h) \end{aligned} \tag{2.41}$$

Again, by dividing by h and taking $\lim h \longrightarrow 0$, we have

$$P'_m(t) = -\lambda P_m(t) + \lambda P_{m-1}(t) \tag{2.42}$$

We do not have to check differentiability conditions for the $P_m(t)$ because by definition, $o(h)$ does not depend on t. Note also that equation (2.41) says: the probability that we have m events by time $t+h$ is equal to the probability that, either we had m events at time t and no events occurred in the additional time interval h, or, that we had $m-1$ events happen by time t, and one event occurred in the next interval of time length h, or that multiple events happened between t and $t+h$.

We now need to solve the differential equation (2.43), with initial conditions $P_m(0) = 0$ for $m = 1, 2, \ldots$. Let $Q_m(t) = P_m(t)e^{\lambda t}$. Then $Q'_m(t) = P'_m(t)e^{\lambda t} + P_m(t)\lambda e^{\lambda t} =$

$$\begin{aligned} &[-\lambda P_m(t) + \lambda P_{m-1}(t)]e^{\lambda t} + \lambda Q_m(t) \\ &= -\lambda P_m(t)e^{\lambda t} + \lambda P_{m-1}(t)e^{\lambda t} + \lambda Q_m(t) \\ &= \lambda P_{m-1}(t)e^{\lambda t} = \lambda Q_{m-1}(t) \end{aligned} \tag{2.43}$$

So $Q'_m(t) = \lambda Q_{m-1}(t)$, with $Q_m(t) \equiv 1$. The initial conditions are $Q_m(0) = 0$ for $m = 1, 2, \ldots$. Thus $Q'_1(t) = \lambda Q_0(t),\ Q_1(t) = \lambda t + c_1$, and with $c_1 = 0$, by the initial conditions, gives $Q_1(t) = \lambda t$. Therefore,

$$P_1(t) = Q_1(t)e^{-\lambda t} = \lambda e^{-\lambda t}. \tag{2.44}$$

Likewise, $Q'_2(t) = \lambda Q'_1(t)$, so $Q_2(t) = \frac{\lambda^2 t^2}{2}$, after finding that $c_2 = 0$. So

$$\begin{aligned} P_2(t) &= e^{-\lambda t}\frac{\lambda^2 t^2}{2}, \\ Q_3(t) &= \frac{\lambda^3 t^3}{3!} \text{ and } P_3(t) = e^{-\lambda t}\frac{\lambda^3 t^3}{3!}, \\ \ldots Q_m(t) &= \frac{\lambda^m t^m}{m!}, \text{ and } P_m(t) = e^{-\lambda t}\frac{\lambda^m t^m}{m!}, \end{aligned} \tag{2.45}$$

which is the Poisson distribution. □

We have thus established a third definition for the Poisson Process:

Definition. Poisson Process III

The counting process $N(t), t \geqslant 0$, is a Poisson Process with rate $\lambda, \lambda \geqslant 0$, if

1) $N(0) = 0$,

2) The process has stationary, independent increments,

3) $\mathbb{P}(N(h) = 1) = \lambda h + o(h)$,

4) $\mathbb{P}(N(h) \geqslant 2) = o(h)$

(the remaining parts of the proof of equivalence to the other two definitions follow trivially).

The solution to the above differential equations can also be obtained with the use of Laplace transform methods. I demonstrate this here, because this type of technique was used often in solving more difficult such equations for specific models in biology and chemistry.

We begin identically to the above with the solution $P_0(t) = e^{-\lambda t}$. Then, let $\Psi_k(s) =$

$\mathcal{L}\{P_k(t)\}$ be the Laplace transform of $P_k(t)$. The differential equations for $P_k(t)$ become:

$$\begin{aligned} s\Psi_k(s) &= -\lambda\Psi_k(s) + \lambda\Psi_{k-1}(s) \\ \Psi_k(s) &= \frac{\lambda\Psi_{k-1}(s)}{s+\lambda} \\ \implies \Psi_1(s) &= \frac{\lambda}{(s+\lambda)^2},\ \Psi_2(s) = \frac{\lambda^2}{(s+\lambda)^3} \\ \ldots \Psi_k(s) &= \frac{\lambda^k}{(s+\lambda)^{k+1}} \end{aligned} \tag{2.46}$$

Note, for $\Psi_0(s)$ we also have $\frac{1}{s+h}$. The inverse Laplace transform of $\frac{n!}{(s-a)^{n+1}} = t^n e^{at}$. Thus the inverse transform of $\frac{\lambda^k}{(s+\lambda)^{k+1}}$,

$$\mathcal{L}^{-1}\left[\frac{\lambda^k}{(s+\lambda)^{k+1}}\right] = \frac{\lambda^k t^k}{k!}e^{-\lambda t}. \tag{2.47}$$

We can see that this saves a bit of writing!

2.4.4 Pure birth process

We slightly generalize the Poisson process to obtain the pure birth process. The rate is now no longer constant, but is assumed to be a function of n. Thus we allow the probability of an event occurring at a given time to depend upon the number of events which have already occurred.

Let $\{\lambda_k\}$ be a sequence of positive numbers. Assume:

1. $\mathbb{P}(X_{t+h} - X_t = 1 | X_t = k) = \lambda_k h + o(h)$
2. $\mathbb{P}(X_{t+h} - X_t = 0 | X_t = k) = 1 - \lambda_k h + o(h)$
3. $\mathbb{P}(X_{t+h} - X_t < 0 | X_t = k) = 0$
4. $\mathbb{P}(X_{t+h} - X_t \geqslant 2 | X_t = k) = o(h)$.

We can also assume (optional):

5. $X(0) = 0$. In this case we are counting the number of births, not the population, and the definition is closer to the Poisson process.

Note that by the description, $\lim_{n\to\infty} \frac{o(h)}{h} = 0$ *uniformly* in $t \geqslant 0$, since $o(h)$ does not depend on t. This property will be shown to be important in a rigorous description of the more general process.

A differentiation procedure similar to the one used for the Poisson process can now be performed to attain another set of differential equations for the transition probabilities [24]:

Set

$$P_n(t) = \mathbb{P}(X_t = n | X_0 = 0). \tag{2.48}$$

Then we have

$$\begin{aligned} P_0'(t) &= -\lambda P_0(t), \text{ for } n = 0, \\ P_n'(t) &= -\lambda_n P_n(t) + \lambda_{n-1} P_{n-1}(t), \text{ for } n \geqslant 1, \end{aligned} \tag{2.49}$$

with boundary conditions $P_0(0) = 1, P_n(0) = 0$. The above equations are obtained by the following:

$$\begin{aligned} P_n(t+h) &= P_n(t)[1 - \lambda_n h + o(h)] + P_{n-1}(t)[\lambda_{n-1}h + o(h)] + \sum_{k=2}^{n-2} P_k(t) o(h) \\ &= P_n(t) - \lambda_n h P_n(t) + P_n(t) o(h) + P_{n-1}(t)\lambda_{n-1} h + P_{n-1}(o(h)) + o(h) \end{aligned} \tag{2.50}$$

$$\implies \frac{P_n(t+h) - P_n(t)}{h} = \frac{-\lambda_n h P_n(t)}{h} + \frac{P_n(t)(o(h))}{h} + \frac{P_{n-1}(t)\lambda_{n-1}h}{h} + \frac{P_{n-1}(o(h))}{h} + \frac{o(h)}{h}. \tag{2.51}$$

Taking $\lim h \longrightarrow 0$ gives the equations (2.50).

The Yule Process

The Yule process, or Yule-Furry process, as it is sometimes called, was a description used by Udny Yule in 1924 to analyze creation of new species via random genetic mutations [56]. It was also used by W. H. Furry in 1937 to model phenomena associated with cosmic rays [14]. In both cases the descriptions are crude, because the required assumptions neglect

important details of the processes. For the species case, the assumption that each species (member) has the same probability as any other species of undergoing a genetic mutation that leads to diversification neglects large variations due to species sizes, as well as genome sizes. Additionally, the chance of extinction has been neglected. In both cases, the assumption of no interaction among members that results in the independence of each members reproduction is relatively unrealistic [13].

Yule process: The rates are defined to be *linear* functions of n: $\lambda_n = n\lambda$. Let n_0 be the initial population. We have:

$$\begin{aligned} P_n'(t) &= -\lambda n P_n(t), \text{ for } n = n_0, \\ P_n'(t) &= -\lambda n P_n(t) + \lambda(n-1)P_{n-1}(t), \text{ for } n > n_0 \end{aligned} \tag{2.52}$$

Note that now we have not used the optional condition that $X_0 = 0$, because we want to count population numbers, not numbers of births. The initial conditions are: $P_n(0) = 1$ for $n = n_0, 0$ for $n \neq n_0$. Consider first the case where $n_0 = 1$. Then $P_1(t) = c_1 e^{-\lambda t}$, and by the initial conditions, $c_1 = 1$. We can solve for the remaining $P_n(t)$ inductively, and we obtain

$$P_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}, \text{ for } n = 1, 2, \ldots \tag{2.53}$$

To check this, we can use the generating function; $g(s,t) = \sum_{n=0}^{\infty} P_n(t)s^n$. Then, using the second differential equation in (2.53), we have

$$\begin{aligned} \frac{\partial g}{\partial t} &= -\lambda s \sum_{n=0}^{\infty} n s^{(n-1)} P_n(t) + \lambda \sum_{k=0}^{\infty} (k-1)s^k P_{k-1}(t) \\ &= -\lambda s \sum_{n=1}^{\infty} n s^{(n-1)} P_n(t) + \lambda \sum_{k=2}^{\infty} (k-1)s^k P_n(t), \end{aligned} \tag{2.54}$$

(because the zeroth term is zero in the left sum and the first two terms of the right sum are

zero),

$$= -\lambda s \sum_{n=1}^{\infty} n s^{(n-1)} P_n(t) + \lambda \sum_{n=1}^{\infty} n s^{(n+1)} P_n(t), \tag{2.55}$$

(after setting $n = k - 1$),

$$= -\lambda s \sum_{n=1}^{\infty} n s^{(n-1)} P_n(t) + \lambda s^2 \sum_{n=1}^{\infty} n s^{(n-1)} P_n(t), \tag{2.56}$$

and

$$\frac{\partial g}{\partial s} = \sum_{n=1}^{\infty} n s^{(n-1)} P_n(t). \tag{2.57}$$

Thus

$$\frac{\partial g}{\partial t} = \lambda s(s-1)\frac{\partial g}{\partial s}. \tag{2.58}$$

The general solution [5] of this p.d.e. is

$$g(s,t) = f[(1 - \frac{1}{s})e^{\lambda t}], \tag{2.59}$$

where f is an arbitrary function. At time $t = 0$, $P_n(t)$ is nonzero only for $n = 1$, so $g(s, 0) = s$, and thus $s = f(1 - \frac{1}{s})$. Consequently, $f(\xi) = \frac{1}{1-\xi}$, and

$$g(s,t) = \frac{se^{-\lambda t}}{1 - (1 - e^{-\lambda t)}s}. \tag{2.60}$$

We can rewrite the above using an infinite sum:

$$g(s,t) = se^{-\lambda t} \sum_{k=0}^{\infty} (1 - e^{-\lambda t})^k s^k = e^{-\lambda t} \sum_{k=0}^{\infty} (1 - e^{-\lambda t})^k s^{(k+1)}. \tag{2.61}$$

Setting $n = k + 1$, and noting that $P_0(t) = 0$ for all t, we have

$$g(s,t) = e^{-\lambda t} \sum_{n=1}^{\infty} (1 - e^{-\lambda t})^{(n-1)} s^n \Longrightarrow P_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{(n-1)}. \quad \square \tag{2.62}$$

Now we find the solution for the general case, where $n_0 = N$ is arbitrary (but greater than 1). Let these probabilities be called $P_{n_N}(t)$. Since we have assumed that each member of the population reproduces independently, we have the equivalent of N separate Yule processes, each with an initial population of 1. By properties of the generating function, $g_N(s)$, the generating function for P_{n_N}, equals $[g(s)]^N$, where $g(s)$ was the generating function for the Yule process above. Therefore

$$g_N(s) = \left[\frac{se^{-\lambda t}}{1-(1-e^{-\lambda t})s}\right]^N = (se^{-\lambda t})^N \sum_{m=0}^{\infty} \binom{m+N-1}{m} (1-e^{-\lambda t})^m s^m, \tag{2.63}$$

the negative binomial series. Putting $n = N + m$, we have

$$g_N(s) = \sum_{n=N}^{\infty} \binom{n-1}{n-N} (e^{-\lambda t})^N (1-e^{-\lambda t})^{n-N} s^n \tag{2.64}$$

From this we can immediately see $P_{n_N}(t)$ as the coefficient of s^n in g_N, i.e.

$$P_{n_N}(t) = \binom{n-1}{n-N} (e^{-\lambda t})^N (1-e^{-\lambda t})^{n-N}. \quad \square \tag{2.65}$$

Existence and uniqueness of the pure birth process:

For these processes, the initial conditions uniquely determine the solution. However, for some choices of λ, the sum of total probability is less than 1, so existence of a properly normalized probability function fails. Thus we have a result known as the

Divergent birth process- With rapidly increasing λ, it may happen that

$$\sum_n P_n(t) < 1 \tag{2.66}$$

For the linear rates in the Yule process, this was not a problem. But for nonlinearly increasing rates, we find that the probabilities may not sum to one. This can be understood as indicating that it is possible for an infinite number of transitions (jumps) to occur in a

finite time. If the rates increase too fast in time, this type of explosion is possible. It can be shown that:

Theorem. In order that $\sum_n P_n(t) = 1$ for all t, it is necessary and sufficient that the series $\sum_n \lambda_n^{-1}$ diverges.

This theorem is proved rigorously in Feller [13], but the following gives an informal proof based on probabilistic reasoning:

We begin by defining the expectation that the first waiting time for a pure birth process exceeds t:

$$\mathrm{E}(T_0) = \int_0^\infty t e^{-\lambda_0 t} \lambda_0 dt = \frac{1}{\lambda_0} \tag{2.67}$$

Then by the Strong Markov Property, $\mathrm{E}(T_j) = \frac{1}{\lambda_j}$, and thus $\lambda_0^{-1} + \lambda_1^{-1} + \cdots + \lambda_n^{-1}$ is the expected time it takes the system to pass through the states $s_0, s_1, \ldots, s_n$. Consequently, $\sum_n \lambda_n^{-1}$ is the expected time it takes the system to pass through all of the (infinite number of) states. If $\sum_n \lambda_n^{-1}$ converges, then the expected time to pass through an infinite number of states is finite. Therefore, since $1 - \sum_n P_n(t)$, which is the probability that we have gone through all states by time t, is greater than zero, $\sum_n P_n(t)$ must be < 1. $\square$

2.4.5 Birth and death processes

We now allow the system to move one step in either direction: the addition of a member, or the removal of one.

NOTE: For most of this section, I revert to the original notation introduced in section 2.4.1, because now we would like to explicitly list the time, as well as both the current state and the one being jumped to: $p_t(i, j)$ represents a jump from state i to state j in time t.

Definition. Postulates for the birth-death process:

1) X_t is Markov on states $1, 2, \ldots$ and $p_t(i, j) = \mathbb{P}(X_{(s+t)} = j | X_s = i)$ is time-homogeneous.

2) $p_h(i, i+1) = \lambda_i h + o(h),\ i \geqslant 0$.

3) $p_h(i, i-1) = \mu_i h + o(h),\ i \geqslant 0$.

4) $p_h(i,i) = 1 - (\lambda_i + \mu_i)h + o(h), i \geqslant 0.$

5) $p_0(i,j) = \delta_{ij}$.

6) $\mu_0 = 0, \lambda_0 > 0;\ \mu_i, \lambda_i > 0$ for $i = 1, 2, \ldots$.

7) $p_h(i,j) = o(h)$ for $j \neq i-1, i, i+1$.

We now have:

$$\begin{aligned} p_{t+h}(i,j) &= \sum_{k=0}^{\infty} p_h(i,k)p_t(k,j) = p_h(i,i-1)p_t(i-1,j) + \\ & p_h(i,i)p_t(i,j) + p_h(i,i+1)p_t(i+1,j) + \sum_{k \neq i-1,i,i+1} p_h(i,k)p_t(k,j). \end{aligned} \tag{2.68}$$

Again the last sum is $o(h)$, and we obtain

$$\begin{aligned} &\lim_{h\to 0} \frac{p_{t+h}(i,j) - p_t(i,j)}{h} = \\ &\lim_{h\to 0} \left[\mu_i h p_t(i-1,j) - (\lambda_i + \mu_i)h p_t(i,j) + \lambda_i h p_t(i+1,j)\right] + 0. \end{aligned} \tag{2.69}$$

Thus

$$\begin{aligned} p_t'(0,j) &= -\lambda_0 p_t(0,j) + \lambda_0 p_t(1,j), \\ p_t'(i,j) &= \mu_i p_t(i-1,j) - (\lambda_i + \mu_i)p_t(i,j) + \lambda_i p_t(i+1,j),\ i \geqslant 1. \end{aligned} \tag{2.70}$$

Note that here the initial state is variable, and the final state, j, is fixed. We chose to split the time interval $(0, t+h]$ into $(0, h]$ and $(h, t+h]$. We are, in a sense, observing the process in retrospect. This is an example of what is called the *backward Kolmogorov equation*. We could also have chosen the subintervals $(0,t],\ (t, t+h]$; as we did with the Poisson process and the pure birth process. However, because the chain can now move in two directions, we cannot deduce from the postulates that the last summation, $\sum_{k\neq j-1,j,j+1} p_t(i,k)p_h(k,j)$ would be $o(h)$. It turns out that we must impose a uniformity condition on the probabilities in order to obtain a similar set of differential equations. One example of a sufficient condition is that, for $k \neq j-1, j, j+1$, $\frac{p_h(k,j)}{h} = o(1)$, where $o(1)$ tends to zero and is *uniformly bounded* for fixed j, as $h \to 0$, [24].

In this case, we obtain:

$$\begin{aligned} p_t'(i,0) &= -\lambda_0 p_t(i,0) + \mu_1 p_t(i,1), \\ p_t'(i,j) &= \lambda_{j-1} p_t(i,j-1) - (\lambda_j + \mu_j) p_t(i,j) + \mu_{j+1} p_t(i,j+1), \ j \geqslant 1. \end{aligned} \tag{2.71}$$

These are the *forward Kolmogorov equations* for the birth-death process.

Existence and uniqueness for the birth-death process:

The problem of existence and uniqueness for the birth-death process already becomes a difficult one. In the Poisson process and the pure birth process the systems of differential equations were recurrence relations, so existence and uniqueness depended completely upon the initial conditions as well as the choice of properly bounded rates. For the birth-death process, this is not so. The solutions for all of the $P_n(t) = \mathbb{P}(X_t = n)$ must be found simultaneously. Consequently, existence and uniqueness proofs require the use of advanced methods, [13], which reveal the following:

Theorem. For any coefficients $\lambda_n \geqslant 0, \mu_n \geqslant 0$, there always exists a positive solution $\{P_n(t)\}$ of the forward equations such that $\sum_n P_n(t) \leqslant 1$. If the coefficients are bounded or increase sufficiently slowly, the solution is unique and satisfies the regularity condition $\sum_n P_n(t) = 1$. However, it is possible to find coefficients such that $\sum_n P_n(t) < 1$ and such that there are an infinite number of solutions, [13]. $\square$

Limit theorems for the birth-death process:

In all cases, $\lim_{t\to\infty} P_n(t) = \pi_n$, for the forward equations, exist, and are independent of the initial conditions. They satisfy the equations obtained by setting the derivatives on the left hand sides of equations (2.71) and (2.72) equal to zero [13]. The limiting probabilities are related to the properties of the underlying discrete-time Markov chain. For the case of linear rates, we have a result similar to the reflecting random walk: $\pi_0 = \lim_{t\to\infty} P_0(t)$, the probability of ultimate extinction, is 1 if $\lambda \leqslant \mu$, and $(\frac{\mu}{\lambda})^i$ if $\lambda > \mu$, where i is the initial state.

Often the explicit solutions of the birth-death differential equations are too difficult to find, and if what is needed is actually just the mean and variance of the process, this can be calculated directly from starting equations with more ease, using methods such as generating functions, and Laplace transforms.

The forward and backward equations are not independent of each other: a solution of the backward equations with the given initial conditions is automatically a solution of the forward equations, except when the solution is not unique. This can occur for reasons similar to those that lead to the case of a divergent pure birth process.

2.4.6 General continuous-time Markov chains

We can generalize Markov processes with discrete states in continuous time to those for which we now assume a transition from a particular state to any other state is possible. (Another possible generalization is to drop the time-homogeneous assumption, although I will not elaborate on that here). The corresponding forward and backward equations can be derived in the same way, and under "ordinary circumstances" each system of equations uniquely determines the transition probabilities of the process.

Assuming differentiability of the transition probabilities, we find that the transition probabilities can be determined by their derivatives at zero:

$$q(i,j) = \lim_{h \to 0} \frac{p_h(i,j)}{h} \text{ for } j \neq i. \tag{2.72}$$

The above defines the "jump rate" from i to j. If we know the rates, we are able to construct the rest of the Markov process:

Let $\lambda_i = \sum_{j \neq i} q(i,j)$ be the rate at which X_t leaves i, and assume $0 < \lambda_i < \infty$. Let $r(i,j) = \frac{q(i,j)}{\lambda_i}$. This is the "probability that the chain goes to j when it leaves i," [8].

We are now able to obtain the transition probabilities from the jump rates via the Kolmogorov equations:

Using the backward equations, and that as $h \to 0,\ t+h \to t$ we can write:

$$\begin{aligned} p_{t+h}(i,j) &= \sum_k p_h(i,k)p_t(k,j) \\ p_{t+h}(i,j) - p_t(i,j) &= \left[\sum_k p_h(i,k)p_t(k,j)\right] - p_t(i,j) \\ &= \sum_{k\neq i} p_h(i,k)p_t(k,j) + p_h(i,i)p_t(i,j) - p_t(i,j) \\ &= \sum_{k\neq i} p_h(i,k)p_t(k,j) + [p_h(i,i)-1]p_t(i,j). \end{aligned} \tag{2.73}$$

We are about to divide by h and take the limit, as in the previous cases. But first note that, since by definition, $q(i,j) = \lim_{h\to 0} \frac{p_h(i,j)}{h},\ $ for $i \neq j$, and assuming we can interchange the limit and the sum, we have

$$\lim_{h\to 0} \frac{1}{h} \sum_{k\neq i} p_h(i,k)p_t(k,j) = \sum_{k\neq i} q(i,k)p_t(k,j). \tag{2.74}$$

Also note that $1 - p_h(i,i) = \sum_{k\neq i} p_h(i,k)$, so

$$\begin{aligned} &\lim_{h\to 0} \frac{p_h(i,i)-1}{h} = -\lim_{h\to 0} \sum_{k\neq i} \frac{p_h(i,k)}{h} = -\sum_{k\neq i} q(i,k) = -\lambda_i. \\ &\text{Thus } \lim_{h\to 0} \sum_{k\neq i} \frac{[p_h(i,i)-1]\, p_t(i,j)}{h} = -\lambda_i p_t(i,j). \end{aligned} \tag{2.75}$$

Therefore, we have:

$$p_t'(i,j) = \sum_{k\neq i} q(i,k)p_t(k,j) - \lambda_i p_t(i,j). \tag{2.76}$$

Noticing that the right hand side can be written in matrix notation, we can define

$$\mathbf{Q}(i,j) = \begin{cases} q(i,j), & \text{if } i \neq j, \\ -\lambda_i, & \text{if } i = j, \end{cases} \tag{2.77}$$

and thus the right hand side of equation (2.77) can be written $\mathbf{Q}p_t$. We therefore can write (2.77) as a matrix equation $p_t' = \mathbf{Q}p_t$.

For the forward equations:

$$\begin{aligned} &p_{t+h}(i,j) - p_t(i,j) = \sum_k p_t(i,k)p_h(k,j) - p_t(i,j) = \\ &\sum_{k\neq j} p_t(i,k)p_h(k,j) + [p_h(j,j) - 1]p_t(i,j), \\ &\text{and } p_t'(i,j) = \sum_{k\neq j} p_t(i,k)q(k,j) - p_t(i,j)\lambda_j, \\ &\text{or } p_t' = p_t\mathbf{Q}. \end{aligned} \tag{2.78}$$

We now use this result to define the Q-matrix:

Definition. Let I be a countable set. A **Q-matrix** on I is a matrix $\mathbf{Q} = \{q(i,j) : i,j, \in I\}$ satisfying the following:

1) $0 \leqslant -q(i,i) < \infty$ for all i;

2) $q(i,j) \geqslant 0$ for all $i \neq j$;

3) $\sum_{j\in I} q(i,j) = 0$ for all i.

As with usual differential equations, for *finite* Q-matrices, we can solve $p_t' = \mathbf{Q}p_t$ with $p_t = e^{\mathbf{Q}t}$, where the matrix exponential is defined as follows:

$$e^{\mathbf{Qt}} \equiv \sum_{n=0}^{\infty} \frac{(\mathbf{Q}t)^n}{n!} = \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n t^n}{n!} \tag{2.79}$$

Assuming we can interchange summation and differentiation, we can check the solution:

$$\frac{d}{dt}e^{\mathbf{Q}t} = \sum_{n=1}^{\infty} \frac{\mathbf{Q}^n t^{n-1}}{(n-1)!} = \sum_{n=1}^{\infty} \mathbf{Q}\frac{\mathbf{Q}^{n-1} t^{n-1}}{(n-1)!} = \mathbf{Q}e^{\mathbf{Q}t}. \tag{2.80}$$

We can also see that $p_t\mathbf{Q} = \mathbf{Q}p_t$, because

$$\mathbf{Q}e^{\mathbf{Q}t} = \sum_{n=0}^{\infty} \mathbf{Q}\frac{(\mathbf{Q}t)^n}{n!} = \sum_{n=0}^{\infty} \frac{(\mathbf{Q}t)^n}{n!}\mathbf{Q} = e^{\mathbf{Q}t}\mathbf{Q}, \tag{2.81}$$

(the only matrices being multiplied are powers of $\mathbf{Q}$ which commute).

The Q-matrix, for a *finite* state space, can be used to make some interesting connections, which suggest the way that more advanced methods such as semigroups of operators may become involved:

Theorem. Let $\mathbf{Q}$ be a matrix on a finite set I. Set $p_t = e^{\mathbf{Q}t}$. Then $p_t, t \geqslant 0$, has the following properties:

1) $p_{s+t} = p_s p_t$. This is the semigroup property.

2) $p_t, t \geqslant 0$, is the unique solution of the forward equation $p_t' = p_t\mathbf{Q}, p_0 = \mathbf{I}$, and the backward equation $p_t' = \mathbf{Q}p_t, p_0 = \mathbf{I}$.

3) For $k = 0, 1, 2, \ldots$, we have $\left(\frac{d}{dt}\right)^k\Big|_{t=0} p_t = \mathbf{Q}^k$

Proof:

1) The matrices $\mathbf{Q}s$ and $\mathbf{Q}t$ commute, so $e^{\mathbf{Q}s}e^{\mathbf{Q}t} = e^{\mathbf{Q}(s+t)}$.

3) The matrix-valued power series $p_t = \sum_{k=0}^{\infty} \frac{(\mathbf{Q}t)^k}{k!}$, for finite matrices, has an infinite radius of convergence, so each component is differentiable by term-by-term differentiation:

$$\begin{aligned}\left(\frac{d}{dt}\right)^k\Big|_{t=0} p_t = \left(\frac{d}{dt}\right)^k\Big|_{t=0} e^{\mathbf{Q}t} = \sum_{n=1}^{\infty} \frac{\mathbf{Q}^n t^{n-k}}{(n-k)!}\Big|_{t=0} = \\ \sum_{n=1}^{\infty} \mathbf{Q}^k \frac{\mathbf{Q}^{n-k}t^{n-k}}{(n-k)!}\Big|_{t=0} = \mathbf{Q}^k e^{\mathbf{Q}t}\big|_{t=0} = \mathbf{Q}^k.\end{aligned} \tag{2.82}$$

This also shows that p_t satisfies the forward and backward equations, proving the first part of (2).

2) uniqueness: Suppose $M(t)$ satisfies the forward equation. Then

$$\begin{aligned} \frac{d}{dt}\left(M(t)e^{-\mathbf{Q}t}\right) &= \left(\frac{d}{dt}M(t)\right)e^{-\mathbf{Q}t} + M(t)\left(\frac{d}{dt}e^{-\mathbf{Q}t}\right) \\ &= M(t)\mathbf{Q}e^{-\mathbf{Q}t} + M(t)(-\mathbf{Q})e^{-\mathbf{Q}t} = 0. \end{aligned} \tag{2.83}$$

This means that $M(t)e^{-\mathbf{Q}t}$ is a constant function. This is only possible if

$$M(t) = e^{\mathbf{Q}t} = p_t. \quad \square \tag{2.84}$$

Finally we have:

Theorem. A matrix $\mathbf{Q}$ on a finite set I is a Q-matrix if and only if $p_t = e^{\mathbf{Q}t}$ is a stochastic matrix for all $t \geqslant 0$. $\square$

For infinite, countable state space Markov processes, more advanced methods such as measure theory and functional analysis are required to obtain similar results. Some of these techniques and results are discussed in chapter 4.

We now consider limiting behavior for the general continuous-time Markov chain.

Limiting behavior

Because the exponential waiting times are random variables, we cannot have periodic behavior in continuous-time Markov chains. Therefore, the limiting distributions are easier to discover. Again we have some similar definitions:

Definition. Irreducible- A continuous time Markov chain is irreducible if, for any two states x and y, it is possible to go from x to y in a finite number of jumps.

Definition. Stationary distribution- Now we have $\pi p_t = \pi$ for all $t > 0$.

We can now state the following theorems:

Theorem. If a continuous-time Markov chain X_t is irreducible and has stationary distribution π, then

$$\lim_{t\to\infty} p_t(i,j) = \pi(j). \tag{2.85}$$

Furthermore, if r_j is a function of the state space, and $\sum_j |r_j| < \infty$, then

$$\text{as } t \to \infty,\ \frac{1}{t}\int_0^t r(X_s)ds \longrightarrow \sum_y \pi(y)r(y). \quad \square \tag{2.86}$$

Note, however, it is difficult to determine whether a particular distribution is stationary because the equation must be checked for all t. Fortunately, we have another means of checking using the Q-matrix, via the

Theorem. π is a stationary distribution if and only if $\pi\mathbf{Q} = 0$. $\square$

Existence and Uniqueness:

The most general requirements for the existence of the above solutions, and also for our ability to use the above definitions for the "jump rates" obtained via differentiation of the transition probabilities, as well as those instances where orders of operations such as summation and differentiation were interchanged, are obtained by more advanced techniques. The simplest proof of a requirement for differentiability will be given in the chapter 4.

Because the definition of the forward equations requires a uniformity condition, it has been argued that the forward equations are less connected to deeper properties of the system. According to Feller, "It will be recalled that [the uniformity assumption] is of a purely analytical character and was introduced only for convenience... Thus the backward equations express probabilistically meaningful conditions and lead to interesting processes, but the same cannot be said for the forward equations. This explains why the whole theory of Markov processes must be based on the backward equations (or abstractly, on semi-groups of transformations of functions rather than probability measures)," [13].

Feller was able to provide a proof of the existence of minimal solutions for the backward and forward systems. It was shown that there always exists a minimal solution,

$$p_{(\tau,t)}(i,j) \leqslant \overline{p}_{(\tau,t)}(i,j), \tag{2.87}$$

where $p_{(\tau,t)}(i,j)$ is the minimal solution, and $\overline{p}_{(\tau,t)}(i,j)$ is any other solution satisfying both systems of differential equations as well as the Chapman-Kolmogorov identity [13]. Notice the presence of two time variables- this result is general and applies to processes which are not time-homogeneous as well, in those cases the time position must be specified in addition to the time step.

When the minimal solution is not "defective," (when the total probabilities sum to one), then it is the only probabilistically meaningful solution, because all other solutions will sum to more than one, and cannot be used to define a probability function. In this case the Markov process is uniquely determined by a solution to either system. Otherwise, there may exist infinitely many Markov processes which satisfy the given backward equations [13].

2.5 Markov Processes in Continuous Time With Continuous State Space

When a stochastic process has a continuous state space, its possible realization may be continuous or discontinuous functions. A Markov process in continuous time, with a continuous state space but for which almost all sample functions are discontinuous may still be called a "jump process." On the other hand, a diffusion is a Markov process in continuous time but whose sample functions are continuous as well. The simplest type of diffusion process is called the Weiner process, or Brownian motion. Terminology differs between the physics community and the mathematical community, and sometimes within each of these as well. In this section I will present some introductory definitions of diffusion processes, simply be-

cause in the Kurtz theorem, as well as in an alternate expansion of the master equation in the thermodynamic limit given by physicist N.G. van Kampen [51], the Markov chains are found to converge in particular ways to diffusion processes, the means of which approach the deterministic equations. Furthermore, alternate ways of modeling chemical and physical processes involve the use of the Fokker-Planck equation, which is a partial differential equation for the transition probabilities of a Markov process that is usually assumed to be a diffusion [15].

The mathematical condition for the existence of continuous sample paths was derived by Gikhman and Skorohod [15]. It is sometimes referred to as the Lindeberg condition (for Markov processes):

Theorem. With probability one, the sample paths of a Markov process are continuous functions of t, if, for any $\varepsilon > 0$, we have

$$\lim_{\Delta t \to 0} \frac{1}{t} \int_{|x-z|>\varepsilon} dx \; p(x, t + \Delta t | z, t) = 0, \tag{2.88}$$

uniformly in z, t, and Δt. $\square$

This means that the probability for x to be different from z goes to zero faster than Δt, as Δt goes to zero.

One dimensional Brownian motion, the simplest example of a diffusion, can be defined by the following two assumptions [54]:

Definition. Brownian Motion:

1) For all $\delta > 0$,

$$\lim_{s \to 0} \frac{\mathbb{P}(|X_{t+s} - X_t| \geqslant \delta)}{s} = 0, \tag{2.89}$$

2) Each increment is independent not only of the previous values (the Markov property), but also of the current value of X_t, and the increments are time-homogeneous. This implies that the increments $X_{t+s} - X_t$ are normally distributed.

Thus we see that Brownian motion is a process with independent increments, such that for an interval of length t, the distribution μ_t of the increments is normal $(mt, \sigma^2 t)$, where m is an arbitrary constant. A *general* diffusion process is defined similarly, but the mean and variance are no longer constant. They may depend on the time, the position, or both [53].

The mathematical definition of Brownian motion, or the Weiner process, and the associated extensions to general diffusions, induce a rich and complex mathematical theory. This involves the definition of the Weiner measure, defining appropriate function spaces, and has led to the formulation of the basis for stochastic integration and ergodic theory.

Chapter 3

Modeling Natural Phenomena With Stochastic Processes

3.1 Markovian descriptions in physical sciences

A chemical system can usually be described by a finite-dimensional vector corresponding to quantities of essential constituents of the system [10]. Chemical kinetics describes the rates of changes in the quantities of these vector components. The first mathematical descriptions of chemical reactions were deterministic: systems of coupled differential equations. This is know as the macroscopic approach. Of course, such a model assumes that the process is continuous and completely predictable [7]. This description was based on an empirical agreement between measured experimental changes in concentrations of compounds of interest and formulas assumed to follow the law of mass action. The law of mass action is a postulate in the phenomenological theory of chemical reaction kinetics, and as such, it is not directly derivable from first principles of particle physics. However, it has been experimentally validated consistently and repeatedly over the course of several centuries, and is therefore accepted as an inherent property of chemical processes [10]. It states that the rate of a reaction is proportional to products of powers of the concentrations of the reactants

involved.

When the deterministic description becomes excessively complicated, such as in the description of the positions and momentums of a large number of gas particles, we have no choice but to use a simplified model. Assuming a system can be described by some probability distribution, and then connecting the averages of certain properties to mass phenomena, such as diffusion, was such a solution. The introduction of the use of probability theory in physics began for this exact reason– with the invention of statistical mechanics due to Gibbs and Boltzmann, and the description of Brownian motion given by Einstein and Smolochowski—in the beginning of the twentieth century. The mathematical description of Brownian motion is the "oldest and best-known example" of a successful description of a physical process by a Markov model [51]. The description can be created in two ways, using two time scales. On the shortest time scale, velocity is Markovian (the velocity autocorrelation decays quickly). On a much larger time scale, particle position can be assumed to be Markovian. Thus the "same physical system can be described by two different Markov processes," [51].

As an illustrative example, it is possible, given complete information about all of the interacting forces involved in the toss of a die from a hand, that we could actually predict the exact result of the throw. However, it is obvious that this complete description would be impossible, and even if it were not, the differential equations resulting from such a description may be impossible to solve. Thus we have become accustomed to the use of probability to describe the outcome of a roll of a die. We must be satisfied with a description of average aspects of the process. In studying physical processes that evolve in time, probabilistic methods require the choice of a time interval such that very rapid phenomena are averaged out, but overall change is still apparent. *The existence of such an increment of time is necessary for the use of the model.* This is the "stochastic mesoscopic description." It is "semi-phenomenological," in that the equations cannot in general be derived from exact microscopic descriptions. But the mesoscopic description is more fine-grained than the macroscopic description [51]. According to Keizer [26], chemical reaction systems separate

into four time scales, each of which can be accurately described by a separate model. The third time scale is of the order of the "chemical relaxation time." The time scale for completion of a chemical reaction in gas phase is of the same order as the time scale for the collision of two particles. This is the minimum time for the use of the discrete-space Markovian description [15]. Keizer showed, for a first-order reaction, that the stochastic description is not valid until this time scale has arrived. Furthermore, he demonstrated that the time-homogeneous Markovian stochastic model can only be used for near-equilibrium phenomena [26].

Another reason to use a probabilistic description is as an attempt to study inherent fluctuations in the system. Deterministic models may be satisfactory as long as fluctuations from macroscopic averages are negligible [10], but for very noisy systems, mass action descriptions may fail. Highly fluctuating signals can be found in systems with small numbers of particles, and in reactions that occur near instability points in the deterministic solution space. In certain situations, the fluctuations provide important information, and accurately describing their properties mathematically is highly desirable. A deep result know as the fluctuation-dissipation theorem in statistical physics was discovered during the twentieth century. In summary, it states that "dissipative processes leading to equilibrium are interconnected with the fluctuations around the equilibrium point," [10]. The microscopic fluctuations seem to be intimately connected with the transport processes that describe mass phenomena, such as diffusion: "fluctuations are the driving force of the [macroscopic] theory— constantly examining nearby states and testing them for stability. . . because of this intimate connection between fluctuations and the dynamics, fluctuations simply cannot be ignored," [26].

According to atomic and molecular theory, ultimately a change in a chemical population is a discrete process. A truly fine-grained description of a chemical reaction would thus necessarily be discrete in state space. Quantitatively, magnitudes of fluctuations of properties of a chemical system about their average values have been estimated to have order $\sqrt{N}$, where N is related to particle number. The ratio of fluctuations to the number of particles is then

$\frac{\sqrt{N}}{N}$ or $\frac{1}{\sqrt{N}}$. This is not much when N is large, but ratios like 25% are chemically significant, as seen when N is small [7].

With the desire to model physical processes that occur over time with probability theory, came the need for more robust mathematical descriptions of such random processes. The creation of a random function for which time is a parameter, by the mathematical community, was partially in response to this need in the physical sciences. As pointed out by B.V. Gnedenko, such need and its resulting stimulus was highly adventitious to mathematics because it resulted in a rich, stand-alone branch of mathematics. Work initiated, of course, by Kolmogorov, and additionally, Khintchine, Lévy, Lindeberg, and Feller, among others, formulated the descriptions in terms of measure theory and functional analysis [19]. Gnedenko elaborates:

> "This theory serves as a beautiful example of the organic synthesis of mathematical and scientific thought, in which the mathematician, in mastering the physical essence of the main problem of some science, finds a suitable mathematical language in which to express it," [19].

Chemical reactions can be modeled by discrete state space, continuous time Markov chains or "Markovian jump processes." The time evolution of the process is thus the time evolution of the transition probabilities, described by the Kolmogorov equations. In chemistry and physics, these equations have come to be known as "the master equation," [7, 10, 15, 17, 47, 51]. Important information to obtain from these models, in addition to quantification of the fluctuation in terms of the mean and variance, includes qualitative insights, such as can be obtained from studying the recurrence, stationarity, and ergodic properties of the model [10].

The most important property of the model, however, is the correspondence of the mean of the process, as a function of time, to the deterministic description, in the thermodynamic limit. This is because the deterministic description is not only validated by extensive empirical evidence, but because the truly microscopic physical description using the Liouville

equation has been found to converge, in the thermodynamic limit, to the deterministic descriptions, for certain systems, such as the ideal-gas at equilibrium, thus validating them theoretically and providing consistency between the two models. Any other models must therefore be consistent as well, to be considered valid [26, 27, 42].

3.1.1 Correspondence in mean

For certain specific stochastic models, the mean of the stochastic process is exactly the solution to the differential equations of the deterministic model. In these situations, the thermodynamic limit is not required. Initially, based on these types of results, it was thought that taking the mean of the process would be sufficient for models of all chemical reactions. Later, it was found that this was not the case.

Poisson Equations: a simple example

A zeroth-order reaction in chemistry, $\mathrm{A}_{\text{on surface}} \longrightarrow$ (products), which occurs in surface reactions, can be modeled by the deterministic differential equation

$$\frac{d[\mathrm{A}]}{dt} = -k, \tag{3.1}$$

whose solution is, of course, $[\mathrm{A}]_t = [\mathrm{A}]_0 - kt$, where [A] is the concentration of substance A, $[\mathrm{A}]_0$ is the initial concentration, and $[\mathrm{A}]_t$ is the concentration at time t. For the zeroth-order reaction, the amount of product does not affect the reaction rate, as it can for other types of reactions.

If we consider not the disappearance of some chemical species, but rather, its appearance on the product side of the equation, we have, for the deterministic model with initial condition $[\mathrm{B}]_0 = 0$, (where B is the product formed),

$$\frac{d[\mathrm{B}]}{dt} = k, \ [\mathrm{B}]_t = kt. \tag{3.2}$$

This type of system can also be modeled as a Poisson process, similar to the way particles being emitted from radioactive disintegration have been modeled. For the Poisson process, we know the mean is of the form λt, which corresponds in functional form to the deterministic solution.

Birth-Death Equations

The Malthusian model for geometric population growth, presented in 1798, was a deterministic model- a very simple differential equation [20]:

$$\frac{dN}{dt} = \lambda N,\ N(t) = N_0 e^{\lambda t}, \tag{3.3}$$

In 1924, Udny Yule created a stochastic model for populations of genetic variants that resulted in formations of new species, under very simple assumptions [13, 56]. This model, consisting of pure birth equations with linear rate functions, know as the Yule process, was presented in chapter 2. However, here the correspondence in mean to the solution of the Malthusian differential equation (3.3) is shown.

Correspondence in mean of the Yule process to the Malthusian model:

For an initial population of 1, our solution to the Yule process differential equation was found in chapter 2 to be:

$$P_n(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}. \tag{3.4}$$

This is a geometric distribution with parameter $\rho = e^{-\lambda t}$, and mean $\frac{1}{\rho}$, so

$$\mathrm{E}[P_n(t)] = \frac{1}{e^{-\lambda t}} = e^{\lambda t}, \tag{3.5}$$

which is the functional form of the solution of the Malthusian equation for an initial population of one. For initial population N_0 arbitrary, we found the generating function

$$g_{N_0}(s) = \left[\frac{se^{-\lambda t}}{1 - (1 - e^{-\lambda t})s}\right]^{N_0} = ([1 - (1 - e^{-\lambda t})s]^{-N_0} s^{N_0})e^{-\lambda t N_0}. \tag{3.6}$$

To find the mean, we differentiate $g_{N_0}(s)$ and set $s = 1$:

$$\frac{\partial g_{N_0} s}{\partial s} = e^{-\lambda t N_0}[-N_0[1-(1-e^{-\lambda t})s]^{-(N_0+1)}[-(1-e^{-\lambda t}]+[1-(1-e^{-\lambda t})s]^{-(N_0}(N_0)s^{(N_0-1)} \quad (3.7)$$

at $s = 1$:

$$\begin{aligned} \left.\frac{\partial g_{N_0} s}{\partial s}\right|_{s=1} &= e^{-\lambda t N_0}[N_0(e^{\lambda t(N_0+1)})(1-e^{-\lambda t}) + N_0 e^{\lambda t N_0}] = \\ & e^{-\lambda t N_0}[N_0 e^{\lambda t(N_0+1)} - N_0 e^{\lambda t(N_0+1)})e^{-\lambda t} + N_0 e^{\lambda t N_0}] = \\ & N_0[e^{\lambda t} - e^{\lambda t N_0} e^{-\lambda t N_0} + e^{\lambda t N_0} e^{-\lambda t N_0}] = N_0 e^{\lambda t}, \end{aligned} \quad (3.8)$$

which is the functional form for the deterministic solution for an initial population of N_0. Note that to obtain *exact* correspondence of solutions, **the same rate constants must be used in both the deterministic equations and the stochastic ones** (i.e. we would have to have $\lambda = k$).

Why use the stochastic models for these types of population studies? Although the means of the stochastic processes correspond to the solutions of the deterministic versions, the *fluctuations about the mean* offer models of the possible variations from this mean, which is not the case in the deterministic description. Because of this, it was thought that the stochastic models were superior to the deterministic ones, as they could predict the magnitude of the fluctuations, and this resulted in different estimates for potential population values. Population sizes predicted from stochastic models can attain values not predicted in the deterministic method, due to the fluctuations, which may be important from a practical standpoint. Furthermore, there is no need to assume that the population was a continuous variable: its discrete nature was more readily apparent in this model [20].

Use of birth-death equations in physics and chemistry

In 1937, Furry [14] used the same approach as Yule to model fluctuations in average behaviors of atomic particles, such as those found in cosmic rays, and their effects on and relationships to unusual phenomena such as rapidly multiplying particle "showers." In 1953,

Singer studied chemical reactions with a similar technique [45]. Some chemical reactions were found to be irreproducible due to occurrence of fast sequences of rare events that cascade to form an explosive but rare end-reaction. These events are attributed to fluctuations around the average (deterministic) behavior of the process, and possible reliance on intense sensitivity to initial conditions. Examples include the nucleation of crystal formation from solutions. Singer used birth-death type Markov models of changes in particle numbers over time, to obtain estimates of mean behavior as a function of time, thus attempting to quantify fluctuations and deviations from deterministic models. He used approximate methods as well as the method of generating functions to obtain some preliminary solutions for the probabilities, for specific examples.

In 1958, Anthony Bartholomay, a chemistry student at Harvard, presented the potential for using such models in chemistry in a more expounded, formal way, and demonstrated the use of the Q-matrix decomposition as well as generating function techniques in this context [2]. The rates for the Q matrix were assumed to be the same numbers as the rates obtained from the deterministic models, and he obtained correspondence to the deterministic models by taking the mean of the stochastic process after solving the Kolmogorov equations for the transition probabilities using generating functions. Because the linear birth-death stochastic model's solutions corresponded in mean to the deterministic solution, he proposed that this method could be used for studying fluctuations about average values of macroscopic models in areas such as chemical kinetics.

Pure death model of the first order reaction

In a second article by Bartholomay, also from 1958, the above method was applied to a simple unimolecular chemical reaction, and much like the classic example of Yule's model, correspondence between the mean and the deterministic solution was again obtained [3]. Bartholomay, like Singer, claimed that the fluctuations predicted and quantified by the stochastic model were a more accurate description both of the time course of the process and of the deterministic rate constant: he proposed a new method of calculating the rate

constant, not by fitting a curve corresponding to the deterministic model to the experimental data using least squares, but by fitting the stochastic model to calculations of experimental fluctuations over intervals of time. The method, using maximum likelihood analysis, was developed in the subsequent paper [4]. There, Bartholomay used maximum likelihood estimation and data from a time-series of concentration values for sucrose inversion, for two different temperatures, to obtain the rate values. In both instances, his values agreed with the deterministic, accepted values with high accuracy. He claimed that this method not only provided an accurate mean value for the time-course of the process, but also provided estimates of the fluctuation behavior via the variance about the mean as a function of time.

In 1963 a chemist named Donald McQuarrie expanded upon the Bartholomay technique [38]. He was able to solve the Kolmogorov equations for some specific models of unimolecular chemical reactions using generating functions and methods of cumulants. He considered three types of reactions, including parallel reactions, and also obtained correspondence of the stochastic means to the deterministic equations in each case. He, too, claimed that this method will provide a more accurate and informative description of chemical equations and could be used to describe small systems. McQuarrie had found an article in a Hungarian journal by A. Rényi, in which, he stated, it had been proved that the "law of mass-action (even for ideal systems) does not hold," [38].

Here I present a simple example of the model of first order reaction that has now become common in chemical kinetics textbooks:

For a first-order, unimolecular reaction such as A $\longrightarrow$ B, we have:

$$\frac{d[\mathrm{A}]}{dt} = -kt,\ [\mathrm{A}]_t = [\mathrm{A}]_0 e^{-kt}, \tag{3.9}$$

for the deterministic model. The stochastic model is a "pure death" model, similar to the pure birth model, but for which only "deaths" can occur:

$$\begin{aligned} \frac{dP_n}{dt} &= -knP_n(t), \text{ for } n = n_0, \\ \frac{dP_n}{dt} &= k(n+1)P_{n+1}(t) - knP_n(t), \text{ for } n = 1, 2, \ldots, n_0 - 1, \end{aligned} \tag{3.10}$$

with initial conditions:

$$P_n(0) = \begin{cases} 1 & n = n_0 \\ 0 & n \neq n_0. \end{cases} \tag{3.11}$$

We have $P_{n_0}(t) = e^{-kn_0t}$, and

$$\frac{dP_{n_0-1}}{dt} = k(n_0 - 1 + 1)e^{-kn_0t} - k(n_0 - 1)P_{n_0-1}(t) \tag{3.12}$$

This equation can be solved iteratively, or with the use of generating functions. The solution is:

$$P_n(t) = \binom{n_0}{n}(e^{-kt})^n(1 - e^{-kt})^{n_0-n}, \tag{3.13}$$

which is the binomial distribution with parameter $\rho = e^{-kt}$, whose mean is known to be $n_0\rho$, or n_0e^{-kt}. Again, the mean of the stochastic process has the same functional form as the deterministic solution.

3.1.2 Bimolecular reactions and the failure of correspondence in mean

In 1964, McQuarrie described various types of bimolecular reactions, A + B $\longrightarrow$ products, using the above methods. He obtained much more complicated differential equations for the transition probabilities [40]. He nonetheless solved some of them, using Gegenbauer polynomials, Legendre polynomials, and Jacobi polynomials, and approximated solutions to other equations using various methods. This time, the mean values of the stochastic

processes were *not* found to be the same functions as the deterministic equations used to model the same processes:

$$\begin{aligned}
&\textbf{deterministic: } \frac{d[\mathrm{A}]}{dt} = -k[\mathrm{A}]^2, \text{ solution: } [\mathrm{A}]_t = \frac{[\mathrm{A}]_0}{1 + kt[\mathrm{A}]_0} \\
&\textbf{stochastic: } \text{mean, } \langle x \rangle = \sum_{n=2}^{x_0} (-A_n) T_n(t), \text{where } x_0 \text{ is the number} \\
&\text{of molecules of } x, \text{ and} \\
&A_n = \left(\frac{1-2n}{2^n} \right) \left[\frac{\Gamma(x_0+1)\Gamma(\frac{x_0-n+1}{2})}{\Gamma(x_0-n+1)\Gamma(\frac{x_0+n+1}{2})} \right], \; T_n(t) = e^{-\frac{1}{2}kn(n-1)t}
\end{aligned} \tag{3.14}$$

However, it was found that for *increasing particle numbers*, the mean values of the stochastic solutions approached the values of the deterministic model in all cases studied [39, 40]. This result corresponded with the idea of the thermodynamic limit used in physics. McQuarrie stated that Rényi had in fact shown, that the unimolecular reaction models will *always* correspond in mean to the deterministic models, using arguments about the implicit assumptions of the independent random variables found in the models. Furthermore, he said that Rényi had shown, using a similar argument, that the unimolecular reactions are the *only* chemical reactions for which such correspondence can possibly hold [39].

Nevertheless, convergence in the thermodynamic limit was a promising result, even though it had only been demonstrated for a few equations, and had not been proved. Increasing interest and use of such models in the chemical community was occurring at that time. Perhaps it was hoped that other mathematical techniques could be found by which to solve or better approximate solutions to the more complicated Kolmogorov equations for chemical reactions beyond the simple ones already studied.

3.1.3 Convergence in mean in the thermodynamic limit

One of Kurtz's few citations for his short article in the *Journal of Chemical Physics*, 1972, was an article by Oppenheim, Shuler and Weiss, published in 1969 [42]. At that time, research

into the use of stochastic models to describe chemical kinetics was over a decade old. The Kolmogorov equations for the transition probabilities had become know as the "chemical master equations." The authors of this article discuss the required correspondence between the mean of the stochastic description and the deterministic model, in the thermodynamic limit.

They are able to show that, for two specific types of chemical reactions (described in a more general way than previously), the mean does in fact converge, in the thermodynamic limit, to the deterministic equation. The two reactions are the general bimolecular reaction, and the general bimolecular reaction coupled with a unimolecular reaction. To obtain convergence, authors use expansions of functions in power series and the introduction of arbitrary parameters that assist in making approximations and taking limits. Additionally, they make assumptions about the orders of certain terms, as is common in mathematical physics. Using a series of approximations, rearrangements, and solutions of iterative equations involving the moments of the probabilities, the authors are able to show that that in the thermodynamic limit, the smaller-order terms vanish, and convergence holds [42].

This paper paved the way for Kurtz's more mathematically rigorous version in which he showed that, using similar ideas, a system of equations of a much more general form but similar description can also converge to the deterministic solution in the thermodynamic limit.

3.2 The Kurtz Theorem

Using advanced mathematical methods, Kurtz was able to avoid the approximations and assumptions used by Oppenheim, Shuler, and Weiss, and achieve a more rigorous proof, and a more general result, for the convergence of certain types of discrete-state Markov processes in the thermodynamic limit (where the particle number and the volume of the system goes to infinity, but the ratio of the particle number to the volume— the density, or concentration—

is held constant).

The Kurtz theorem can be proved in two ways. In his article in *J. Chem. Phys.* [36], Kurtz gives three citations for where the proof can be found. The 1971 paper in the *Journal of Applied Probability* uses martingale techniques [35]. The other citations, for Kurtz's 1967 PhD thesis from Stanford University, entitled "Convergence of Operator Semigroups with Applications to Markov Processes," [31], and the 1970 paper from *Journal of Applied Probability*, [34], use operator semigroup methods. The use of operator semigroups in the study of Markov processes was an extension of the Hille, Yosida, and Phillips theory of semigroups from functional analysis to probability spaces, and was developed heavily by Dynkin and Feller [11, 53]. It is a fusion of ideas and methods relating to dynamical systems and potential theory, and operator theory. It was this approach that interested me the most, especially because of the use of similar techniques in other areas of mathematical physics such as quantum perturbation theory [51]. My presentation of his proof in the next chapter will follow the semigroup approach.

3.2.1 The Kurtz theorem for chemical reactions- summary form

This section follows the exposition presented by Thomas G. Kurtz in his paper published in the *Journal of Chemical Physics*, [36]. The system of chemical reactions modeled allows for movement in both directions, i.e., the reactions are reversible. In this most general model, any finite number of reactants can combine to form any finite number of products, and similarly for the reverse reactions:

$$\begin{cases} \mathrm{R}_1 = k_1\mathrm{A}_1 + k_2\mathrm{A}_2 + \cdots + k_k\mathrm{A}_k \rightleftarrows j_1\mathrm{B}_1 + j_2\mathrm{B}_2 + \cdots + j_j\mathrm{B}_j \\ \mathrm{R}_2 = d_1\mathrm{C}_1 + d_2\mathrm{C}_2 + \cdots + d_d\mathrm{C}_d \rightleftarrows g_1\mathrm{D}_1 + g_2\mathrm{D}_2 + \cdots + g_g\mathrm{D}_m \\ \vdots \\ \mathrm{R}_n = r_1\mathrm{G}_1 + r_2\mathrm{G}_2 + \cdots + r_r\mathrm{G}_r \rightleftarrows s_1\mathrm{H}_1 + s_2\mathrm{H}_2 + \cdots + s_s\mathrm{D}_s \end{cases} \tag{3.15}$$

Here the Rs represent each of the n reactions. On the left hand sides of the double arrows are the reactants represented by capital letters, and their required numbers represented by lowercase coefficients. On the right hand side of the arrows are the products. We can write this more concisely in matrix form, where the matrix $\mathbf{C} = [(c_{nm})]$ contains the number of molecules of the m^{th} reactant required for the n^{th} reaction, and $\mathbf{D} = [(d_{nm})]$ is the number of molecules of the m^{th} product formed in the n^{th} reaction. Note: because the reactions are reversible, the roles of "reactant" and "product" are reversed in the reverse reaction. Thus $[(d_{nm})]$ is also the number of molecules of the m^{th} reactant required for the n^{th} reverse reaction. We let M represent the total number of chemical species involved in either direction of the system, and N represent the total number of reactions. The Markov chain for this sytsem can then be expressed by the random vector $\mathbf{X^V}(t)$, representing the number of molecules of each reactant present at time t, for a given volume V:

$$\mathbf{X^V}(t) = [X_1^V(t), X_2^V(t), \ldots, X_M^V]. \tag{3.16}$$

Thus $\mathbf{X^V}(t)$ is a vector-valued Markov chain. We assume that the probability of x molecules reacting in a short amount of time is proportional to $(V^{x-1})^{-1}$ and the number of ways of selecting the x molecules (this is the "combinatorial model"). Then the probability of the n^{th} reaction occurring in the forward direction during a short time Δt is:

$$P_{nF}(t + \Delta t) = \alpha_n (V^{c_n - 1})^{-1} \left[\prod_{m=1}^{M} \binom{i_m}{c_{nm}} \right] \Delta t, \tag{3.17}$$

where i_m is the number of molecules of the m^{th} reactant and α_n is the rate constant for reaction n in the forward direction. Note that i_m and the m^{th} element of $\mathbf{X^V}(t)$ are equal, for a fixed t; i_m is introduced for ease of notation. The elements of $\mathbf{X^V}(t)$ and thus the i_m are nonnegative integers.

For the reverse reaction, we have

$$P_{nB}(t+\Delta t) = \beta_n (V^{d_n - 1})^{-1} \left[\prod_{m=1}^{M} \binom{i_m}{d_{nm}} \right] \Delta t, \tag{3.18}$$

where β_n is the reverse rate constant.

Using matrix notation, and letting $\mathbf{Y^V}(t)$ be the vector containing the number of times each reaction has occurred in the forward direction, minus the number of times it has occurred in the reverse direction by time t:

$$\mathbf{Y^V}(t) = [Y_1^V(t), Y_2^V(t), \ldots Y_N^V(t)], \tag{3.19}$$

we can also express the process by:

$$\mathbf{X^V}(t+s) = \mathbf{X^V}(s) + [\mathbf{Y^V}(t+s) - \mathbf{Y^V}(s)](\mathbf{D} - \mathbf{C}), \tag{3.20}$$

which may provide a clearer description of certain systems, such as a single bimolecular reversible reaction.

The expressions for the probabilities can be rewritten using the vector form as well:

$$\begin{aligned}
&P_{nF} = V f_n^V(\mathbf{X^V}(t))\Delta t \\
&P_{nB} = V g_n^V(\mathbf{X^V}(t))\Delta t, \text{ where} \\
&f_n^V(\mathbf{X^V}(t)) \equiv \alpha_n \left[\prod_{m=1}^{M} (V^{c_{nm}})^{-1} \binom{i_m}{c_{nm}} \right], \text{ and} \\
&g_n^V(\mathbf{X^V}(t)) \equiv \beta_n \left[\prod_{m=1}^{M} (V^{d_{nm}})^{-1} \binom{i_m}{d_{nm}} \right].
\end{aligned} \tag{3.21}$$

The system now is formulated like the general birth-death process introduced in chapter 2, where Vf_n^V and Vg_n^V are equivalent to the λ_n and μ_n rates which were allowed to be functions of species number. What Kurtz has added to the model, which was not used by

Bartholomay or McQuarrie, is explicit inclusion of the volume of the system, as well as its inverse proportionality to the reactions' occurrences, in the expression for the probabilities [36].

Two additional functions are now introduced. For some nonnegative-integer-valued vector $\mathbf{X} = [X_1, X_2, \ldots, X_m]$, define

$$\begin{aligned} f_n(\mathbf{X}) &\equiv \alpha_n \prod_{m=1}^{M} \frac{x_m^{c_{nm}}}{c_{nm}!}, and \\ g_n(\mathbf{X}) &\equiv \beta_n \prod_{m=1}^{M} \frac{x_m^{d_{nm}}}{d_{nm}!}. \end{aligned} \tag{3.22}$$

Then $f_n(V^{-1}\mathbf{X}^{\mathbf{V}}(t))$ is the usual deterministic model of chemical kinetics for the forward reaction n, using the law of mass action, as in equation (3.14), and $g_n(V^{-1}\mathbf{X}^{\mathbf{V}}(t))$ is the same for the reverse reaction.

We have that

$$\begin{aligned} f_n^V(\mathbf{X}^{\mathbf{V}}(t)) &= f_n(V^{-1}\mathbf{X}^{\mathbf{V}}(t)) + O(V^{-1}), \text{ and} \\ g_n^V(\mathbf{X}^{\mathbf{V}}(t)) &= g_n(V^{-1}\mathbf{X}^{\mathbf{V}}(t)) + O(V^{-1}), \end{aligned} \tag{3.23}$$

where if $g = O(f)$, then as $x \to \infty, g(x) \leqslant M|f(x)|$, for some positive constant M. In the above case, as $V \to \infty$, the function goes to zero, so as the volume becomes very large, the last terms in both equations of (3.23) go to zero and equality is established between f_n^V and f_n and also for g_n^V and g_n.

This is far from enough to prove convergence in mean in the thermodynamic limit of the two models, however. If we define

$$\begin{aligned} F_m^V(\mathbf{X}^{\mathbf{V}}(t)) &= \sum_{n=1}^{N} (d_{nm} - c_{nm})[f_n^V(\mathbf{X}^{\mathbf{V}}(t)) - g_n^V(\mathbf{X}^{\mathbf{V}}(t))], \text{ and} \\ \mathbf{F}^{\mathbf{V}}(\mathbf{X}^{\mathbf{V}}(t)) &= [F_1^V(\mathbf{X}^{\mathbf{V}}(t)), \ldots, F_m^V(\mathbf{X}^{\mathbf{V}}(t))], \end{aligned} \tag{3.24}$$

we now have the Kolmogorov equations for the probability functions governing $\mathbf{X}^{\mathbf{V}}(t)$, the

numbers of each of the m reactants at time t. Next, define

$$\begin{aligned} F_m(\mathbf{X}) &= \sum_{n=1}^{N}(d_{nm} - c_{nm})[f_n(\mathbf{X}) - g_n(\mathbf{X})], \text{ and} \\ \mathbf{F}(\mathbf{X}) &= [F_1(\mathbf{X}), \ldots, F_m(\mathbf{X})]. \end{aligned} \tag{3.25}$$

What Kurtz proved, using advanced methods, was the following [36]:

Theorem. Let $\mathbf{X}(t, \mathbf{x_0})$ be the solution of the initial value problem

$$\frac{\partial \mathbf{X}(t, \mathbf{x_0})}{\partial t} = \mathbf{F}(\mathbf{X}(t, \mathbf{x_0})), \quad \mathbf{X}(0, \mathbf{x_0}) = \mathbf{x_0}. \tag{3.26}$$

Then, if $\lim_{V\to\infty} V^{-1}\mathbf{X}^{\mathbf{V}}(0) = \mathbf{x_0}$,

$$\text{we have that } \lim_{V\to\infty} \mathbb{P}\left\{\sup_{s\leqslant t} |V^{-1}\mathbf{X}^{\mathbf{V}}(s) - \mathbf{X}(s, \mathbf{x_0})| > \epsilon\right\} = 0, \tag{3.27}$$

for every t and $\epsilon > 0$. $\square$

In order to prove this theorem, Kurtz uses the "convergence of generators" of Markov processes, to establish "weak convergence," or a type of convergence in distribution, of a *sequence of Markov processes* to the required deterministic description. This can be done in several ways: in Kurtz's thesis, it was done directly, but required a number of preliminary theorems and lemmas [31]. In later publications, Kurtz accomplishes the same result in a more concise, elegant way, first proving convergence of the model to a Brownian motion, [32]. In the infinite volume limit, it can be shown that *this* process then converges to the solution of the differential equations that is the deterministic model.

The operator semigroup technique uses the idea that, if X is a time-homogeneous Markov process, then the expectation of an appropriate function of X defines a semigroup of operators:

$$\mathrm{E}[f(X(t))|X(0) = x] = T(t)f(x), \tag{3.28}$$

Therefore, convergence theorems for semigroups can be applied to sequences of the Markov

processes. H.F. Trotter proved several important results on this type of convergence in the general semigroup setting, and Kurtz was able to extend some of these results so that they became applicable in the context of these types of stochastic processes [32]. In chapter 5, I provide an introduction to semigroup methods and their use in probability theory, in addition to other advanced topics. But first, it is important to discuss the impact of the above theorem and the applicability of this type of model to the physical sciences.

3.2.2 Importance and usefulness of the theorem

Kurtz's result contributed significantly to theoretical and computational chemistry and physics. Besides confirming the validity of an experimental technique which was already being applied, and generalizing the allowable description, it was the final missing link in the connection between several different methods of modeling molecular processes at the "mesoscopic" level. This link allowed for consistency and interchange between Fokker-Planck-type, Langevin-type, and master equation-type descriptions. Connections had been made by Keizer and others between Fokker-Planck equations and Langevin-type descriptions. Kurtz's addition of a connection between master equations and the Fokker-Planck method unified all of the modeling schemes in the thermodynamic limit and connected them all to the deterministic model's results [26].

Moreover, it validates the use of computationally easier techniques which otherwise would stand on a less rigorous platform. Two important such techniques are the van Kampen expansion of the master equation, and the Gillespie simulation algorithm. Finally, it gave the most rigorously derived mathematical description of the distribution of the fluctuations about the mean of the birth-death model in the thermodynamic limit, which were found to be described by a Gaussian, non-Markovian stochastic process.

For physical systems that consist of discrete particles, the use of jump processes as models is highly preferred. Here the fluctuations are not added onto the deterministic equations as an arbitrary, unrelated noise term, such as is done in the Langevin descriptions, but are

an essential part of the inherent description of the process. Thus the mean values of the stochastic process correspond to the deterministic part, and the fluctuations about the mean as the additional noise [15]. For this reason the use of master equations as a starting point is desired, however, convergence in mean to the deterministic model is essential.

Unfortunately, as was seen in McQuarrie's attempts to solve some of the more complicated master equations for bimolecular reactions, solutions to the Kolmogorov equations may be very difficult to obtain. In fact, "only in rare cases is it possible to solve the master equation explicitly," [51]. A more computationally promising technique is the van Kampen expansion of the master equation. This was developed in the seventies by physicist N.G. van Kampen. Expansion of the master equation starts with the *assumption* that the mean of the process corresponds to the deterministic equations. Then the master equation is expanded using the introduction of a parameter that generally corresponds to the system volume as done in Kurtz's method. One obtains a Fokker-Planck equation for which the mean and the variance can be calculated. The Fokker-Planck equation is the forward Kolmogorov equation for a type of diffusion process (continuous state, continuous time), but is simpler to solve for important properties like the mean and variance than the associated Kolmogorov equations [51]. This Fokker Planck equation describes a diffusion process equivalent to that found by Kurtz in his proof.

When the Fokker-Planck equation is used, the mathematical description is separable into a deterministic part and a fluctuating part, and when this Fokker-Planck equation has been derived via an expansion of the master equation, the fluctuating parts coefficients are provided by the initial master equation. Thus it is a more preferable description of discrete processes than a Langevin-based description. Furthermore, it is often possible to write down the appropriate Fokker Planck equation corresponding to an expanded master equation without ever expressing the master equation itself [15]. Thus the van Kampen method can allow for reduced computational difficulty, but maintains a connection with the original model, and allows for a description of fluctuations. Without Kurtz's theorem, the

initial assumption of correspondence used by van Kampen would be questionable.

Kurtz's mathematical methods, used in his article and his thesis, could also be used for specific cases, in a similar type of expansion leading to a Fokker-Planck equation. Kurtz's results are the "most general results on the asymptotic form of the solutions of master equations," and his theorem "applies to a broad class of sequences of jump processes," [26]. Furthermore, van Kampen's method is considered less rigorous [15, 26]. However, for physicists, the van Kampen method is less challenging, more familiar, and thus more immediate. For practical purposes, the van Kampen expansion method is equivalent to the Kurtz method: "In terms of quantities that can be calculated and measured, means, variances, etc., Kurtz's apparently stronger result is equivalent to van Kampen's system size expansion," [15].

The Gillespie algorithm was presented by physicist Daniel T. Gillespie in 1976 as a way to create a computational simulation of a Markovian birth-death process using the Q-matrix and randomly generated exponential holding times [16]. Using rates from the deterministic model, one can then create a simulated process whose mean and variance can be directly calculated, as opposed to analytically solving the Kolmogorov equations. Obviously, the use of these rates relies on the correspondence proved by Kurtz. The Gillespie algorithm has been used in many scientific applications from chemistry to biology [17, 50].

3.3 Accuracy of the Model

The Kurtz theorem showed that if a chemical system can be described by a certain type of birth-death stochastic process, under specific initial conditions, the description can correspond to the deterministic model. The next question for the physical scientist is whether the "combinatorial," mesoscopic model is appropriate and correct for a chemical reaction. Finally, under what conditions can this model be used, and what are the limitations?

Oppenheim, Shuler and Weiss in their 1969 article investigated the validity of the linear

birth-death model from a rigorous physical perspective. They concluded that such a model may be too over-simplified to be a truly accurate description of the interaction of particles in a chemical reactions. The most accurate description of the physical system would be a complete Liouville equation for the interacting particles, which includes correlations between all of the particles. The authors state that in the case of a dilute gas system, the Liouville equation description does converge, in the thermodynamic limit, to the "stochastic version of the Boltzmann equation in the limit $N \to \infty$," wherein all transition probabilities factorize (i.e. the variables become independent). This implies the deterministic description. Therefore, in this same limit, the Markovian stochastic description of the same gas-phase processes must converge to the deterministic description to be valid. However, the authors state that although this condition is necessary, "it is by no means a sufficient one," in that the Markovian description may still not accurately describe the fluctuations and the process itself outside of the thermodynamic limit.

The use of the Markovian description outside the thermodynamic limit was one of the reasons it was so actively developed by chemists: they wished to model systems with small particle numbers and predict the fluctuations that would occur in such systems. A complete Liouville description of a chemical system with an Avogadro's number of particles is intractable, thus it is not readily possible to prove the sufficiency condition- that a stochastic process that converges to the correct deterministic model is the correct mesoscopic model of a particular microscopic phenomenon. "The Markov character of the chemical process represented by the component vector has not been derived from microscopic models. Therefore Markovicity is not more (and of course not less) than a plausible assumption," [10].

The plausibility of the assumption has been addressed in some detail. Gillespie has shown [18] that for a "gas-phase system that is kept well stirred and in thermal equilibrium," the combinatorial stochastic model of a chemical reaction, and the resulting Kolmogorov equations that have become known as the chemical master equations are in fact derivable from a fairly rigorous physical description. Starting with the physical assumptions that

the chemical reaction occurs as a result of the collision of individual molecules of reactant species, and that the system is well stirred and in thermal equilibrium, Gillespie explains that it follows that the density of particles can be considered as spatially homogeneous, and therefore that the probability of finding a molecule in any subregion of the system is expressible using the uniform distribution. Furthermore, the velocities of the particles can be assumed to have the Maxwell-Boltzmann distribution, which is equivalent to a normal distribution. From this, and considering the molecules as spheres with a "reactive core," and using other physically plausible assumptions, Gillespie is able to justify the use of the time-homogeneous Markov model, the use of linear intensities that are scalar functions of only the number of particles and are independent of time, and the assumption that certain terms are $o(h)$, as required in the theory [18].

The above assumptions fail in chemical reactions such as enzyme reactions. In such reactions, a complex surface has to interact with another complex surface in intricate ways, therefore the reaction would not be accurately described by the model of reactant species as hard spheres colliding [10]. However, in recent years the use of the birth-death model for reactions and its subsequent chemical master equation has been applied successfully to noise in gene expression. The gene expression process involves many enzyme interactions, with other enzymes, and with other molecules having highly specific, complicated three-dimensional interaction surfaces [50]. It is therefore possible that to provide a course-grained approximation of a molecular process that is more detailed than the macroscopic model, a Markovian assumption and a combinatorial model is adequate. Stochastic theory, despite objections as to rigor, can often provide a "first order approximation" that can be satisfactory in many situations [7].

In fact, on the physical level, no process is truly Markovian. Under the appropriate description and conditions, however, it can be successfully modeled as such: "The art of the physicist is to find those variables that are needed to make the description (approximately) Markovian," [51]. For instance, in the Fokker-Planck and Langevin models, a process is

often assumed to have a continuous sample path, and is modeled as a diffusion, for ease of computation. For a physical particle, if one were to focus on the particle's motion on a fine enough scale to see continuous sample paths, the process would most likely not be Markovian. The forces affecting the particle in the past would have some effect on the particle's current state on such a time scale. Thus a particle with a Markovian distribution and a continuous sample path does not exist physically. On the other hand, such continuous state-space processes exist mathematically, and are often convenient to use, and when they provide an adequate approximate description of a process, they can be used effectively [15].

As for the thermodynamic limit, and the assumption that the number of particles as well as the volume are infinite, it seems that for usual experimental conditions with finite quantities of substances, this assumption is still appropriate. The difference between such important values as the free energy density, evaluated for a system with Avogadro's number of particles, and a system with an infinite number, is "smaller than experimental error," [48]. Practically, for many common experimental conditions the number of particles and volume can effectively be considered infinite.

Chapter 4

Advanced methods

4.1 General continuous-time Markov chains: differentiability and existence

We saw that even for simple processes such as the birth-death process, the assumptions about $o(h)$ and differentiability, required for the forward Kolmogorov equations, did not follow directly from the postulates defining the process, but required additional conditions. For the *general* continuous-time Markov chain with time-homogeneous transition probabilities, advanced techniques are required to prove differentiability properties. What has been shown via these methods is the following [25]:

Theorem. Let X_t be a continuous-time, discrete-state Markov process with transition probabilities $p_t(i,j) = \mathbb{P}\{X_{t+s} = j | X_s = i\}$. In addition to the usual assumptions on $p_t(i,j)$:

1) $p_t(i,j) \geqslant 0$,

2) $\sum_j p_t(i,j) = 1$,

3) $\sum_k p_t(i,k) p_h(k,j) = p_{t+h}(i,j),\ t,h > 0$,

we also assume that:

4) the $p_t(i,j)$ are continuous for $t > 0$, and

$$5)\lim_{t\to 0} p_t(i,j) = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

This type of transition probability function is called "standard," and with these conditions, it turns out that the $p_t(i,j)$ are differentiable for every $t \geqslant 0$. [25]. The proof of this is quite involved, even for the simplest case that the $p_t(i,j)$ are differentiable at $t = 0$.

Recall that we formally defined the derivatives of the $p_t(i,j)$ at zero as

$$q(i,j) = \lim_{h\to 0} \frac{p_h(i,j)}{h} \text{ for } j \neq i, \tag{4.1}$$

and

$$-\lambda_i = -\sum_{j\neq i} q(i,j) = \lim_{h\to 0} \frac{p_h(i,i) - 1}{h}, \tag{4.2}$$

in chapter 2, and also assumed that $\lambda_i < \infty$, for all i. These definitions are describing a continuous time Markov chain that is called "conservative." Generally, as long as the Markov process satisfies the above differentiability conditions, we have that $q(i,j) = \lim_{h\to 0} \frac{p_h(i,j)}{h}$ for $j \neq i$, and $-\lambda_i = \lim_{h\to 0} \frac{p_h(i,i)-1}{h}$, but not necessarily the remaining assumptions. With the differentiability conditions satisfied, we always have, for every i and j, $j \neq i$, $q(i,j) = \lim_{h\to 0} \frac{p_h(i,j)}{h}$ exists and is finite. However, we may have that $\lim_{h\to 0} \frac{p_h(i,i)-1}{h} = -\lambda_i$ exists but λ_i is infinite. Furthermore, we do not necessarily have $\lambda_i = \sum_{j\neq i} q(i,j)$. For a conservative process, we know that any Markov chain associated with the $q(i,j)$ must at least satisfy the backward equations. On the other hand, if $\lambda_i = \sum_{j\neq i} q(i,j)$ for all i, but λ_i is not necessarily finite, we know that we either have a unique Markov process, or an infinite number of them. If we only have that $\sum_{j\neq i} q(i,j) \leqslant \lambda_i$ for all i, any attempts at results classifying the process and its associated infinitesimal Q-matrix become very complicated.

For a conservative process, it can be rigorously proved that there is a unique minimal process that can be constructed using the Q-matrix and exponential holding times: the holding times are exponentially distributed with parameter λ_i. This result is crucial for the

use of the Gillespie simulation algorithm. This proof requires a separability assumption for the process. Finding the actual minimal transition probability function requires "rather deep measure theory methods," [25].

A state i for which $0 \leqslant \lambda_i < \infty$ is called **stable**. A state i for which $\lambda_i = \infty$ is called an **instantaneous state**. A continuous time Markov chain can be constructed with only instantaneous states. This leads to problems involving pathological processes whose sample paths may be difficult to determine.

Rigorous results about existence and types of solutions to the Kolmogorov equations almost always involve the use of the strong Markov property. For a discrete time Markov chain, the strong Markov property was uncomplicated. However, for the continuous time case, the strong Markov property "cannot properly be understood without measure theory," [41]. Furthermore, in the continuous time case, there exist Markov processes that are not strong Markov [25]. Recall that the strong Markov property requires that, for any stopping time σ, (σ is a random variable), the probability distribution of

$$X_{t_1+\sigma}, X_{t_2+\sigma}, \ldots X_{t_k+\sigma}, \quad (t_1 < t_2 < \cdots < t_k), \tag{4.3}$$

given $X_s, s \leqslant \sigma$ and $X_\sigma = x$, is identical with the probability distribution of

$$X_{t_1}, X_{t_2}, \ldots X_{t_k}, \quad (t_1 < t_2 < \cdots < t_k), \tag{4.4}$$

given $X_0 = x$. This is "not a direct consequence of the statement of the Markov property since the original formulation requires fixed times," [25].

Fortunately, using advanced methods, it has been proved that "any continuous time, conservative Markov chain with only stable states is strong Markov," [25]. This is very important for the use of renewal relations that can help to describe long-term behavior of the stochastic process. Furthermore, "almost all continuous time Markov chains arising in applications have only stable states," [25].

It is clear from the above exposition that without measure theoretic probability and possibly, techniques from other advanced fields such as functional analysis, further development of the theory of stochastic processes from a rigorous perspective is impossible. I now present an introduction to these advanced methods, which help to provide a clearer idea about the way that Kurtz proved his theorem.

4.2 Measure Theory and Probability

This section follows Rosenthal [43], Norris [41], Gray [22], Skorokhod [46], and Varadhan [52, 53].

4.2.1 Measure theoretic probability basics

Probability was placed in an axiomatic setting by work pioneered by Kolmogorov in the 1930s. Using the concepts of countably additive measures on sigma-algebras, it became possible to determine exactly what types of questions could be answered by probability theory and remain consistent with the axioms. For instance, the question about what types of sets of events could be assigned a probability was answered. Not all sets can be assigned a probability measure, as is seen from the Vitali construction of non-measurable sets; some subsets of an event space cannot be understood in terms of a probability measure. The sets to which a probability measure can be assigned are the sigma-algebra of subsets of a given event space.

Besides placing probability theory on a firm axiomatic footing, measure theoretic probability, or modern probability theory, unifies the notions of discrete and continuous probability distributions, and allows for mixed distributions containing both discrete and continuous parts, as integration theory does not make the artificial distinction between the two. Furthermore, modern probability theory allows for the computation of quantities such as the expectation of random variables which may not be Riemann integrable, but are Lebesgue

integrable. Thus measure theory extends the applicability of probability theory.

In the following sections, I assume familiarity with basic measure and integration theory, but not measure-theoretic probability theory. We begin by defining a probability space, similar to a measure space, but normalized to [0,1] as is required for probabilities:

Definition. A **probability space**, or **probability triple**, is a triple $(\Omega, \Sigma, \mathbb{P})$, where:

1) Ω is the sample space, or any non-empty set of points representing "outcomes,"

2) Σ is a σ-algebra of Ω, consisting of sets of points of Ω called "events,"

3) $\mathbb{P}$ is a "probability measure," which is a countably additive mapping from Σ to [0,1], with $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$.

Recall that a σ-algebra is a collection of all measurable sets of Ω with respect to the measure $\mathbb{P}$, contains $\emptyset$ and Ω, and is closed under formation of complements, countable unions, and countable intersections. It also follows that if $A \subseteq B$, then $\mathbb{P}(A) \leqslant \mathbb{P}(B)$ (the monotonicity property).

The Lebesgue measure on [0,1] is constructed using the Carethéodory Extension theorem as is done in general introductory measure theory. Recall that this is accomplished via the extension of a countably additive probability measure on an algebra of sets, such as that generated by the set of all open intervals on [0,1], using the definition of the outer measure $\mathbb{P}^*$, to all sets in the space, and then by defining the σ-algebra of measurable sets with respect to this measure, $\mathcal{M}$, as consisting of all sets E whose outer measure $\mathbb{P}^*(A) = \mathbb{P}^*(A \cap E) + \mathbb{P}^*(A \cap E^c)$ for *any* set $A \subset [0, 1]$. In terms of probability, this measure, also known as the Lebesgue measure on [0,1], is the same as the uniform distribution on [0,1].

In probability theory, we often do not wish to use the entire set $\mathcal{M}$ as our σ-algebra of measurable sets. This is because $\mathcal{M}$ is much bigger than $\mathcal{B}$, the *Borel sets*, or Borel σ-algebra, the σ-algebra generated by the intervals. However, although $\mathcal{B} \subseteq \mathcal{M}$, the Lebesgue measure restricted to $\mathcal{B}$ is not *complete*. By complete we mean that all subsets of sets having measure zero are measurable and also have measure zero. Because we often would like to

use $\mathcal{B}$ as our σ-algebra, we can define a measure that completes our space, or exclude those problematic subsets from being considered as "events." See Gray [22] for further details. Complete measures are important in defining uniqueness properties of random variables.

The distribution function:

Consider a class of subsets of the real numbers, $\mathcal{J} = \{I_{a,b} : -\infty \leqslant a < b \leqslant \infty\}$ where $I_{a,b} = \{x : a < x \leqslant b\}$ if $b < \infty$, and $I_{a,\infty} = \{x : a < x < \infty\}$, or the collection of intervals that are left-open and right-closed. The class of sets of finite, disjoint unions of members of $\mathcal{J}$ is an algebra, call it $\mathcal{F}$, if the empty set is added. The Borel σ-algebra on the real line is the σ-algebra generated by $\mathcal{F}$. If we define a nondecreasing function $F(x)$ on the real line that satisfies $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$, and define $\mathbb{P}(I_{a,b}) = F(b) - F(a)$ for intervals, then extending this measure to be finitely additive on $\mathcal{F}$ by taking sums, we can then attain the following relation:

Theorem. $\mathbb{P}$ is *countably* additive on $\mathcal{F}$ if and only if $F(x)$ is a right-continuous function of x. Therefore, for each right-continuous, nondecreasing function $F(x)$ with $F(-\infty) = 0$ and $F(\infty) = 1$, there is a unique probability measure $\mathbb{P}$ on the Borel subsets of the line such that $F(x) = \mathbb{P}(I_{-\infty,x})$. Conversely, every countably additive probability measure $\mathbb{P}$ on the Borel subsets of the real line comes from some F. The correspondence between $\mathbb{P}$ and F is one-to-one. $\square$

The function F is called the distribution function corresponding to the probability measure $\mathbb{P}$.

Definition. A **random variable** is a measurable function: given a probability triple $(\Omega, \Sigma, \mathbb{P})$, it is a map $f(\omega) : \Omega \to \mathbb{R},\ \omega \in \Omega$, such that for every Borel set $B \subset \mathbb{R},\ f^{-1}(B) = \{\omega : f(\omega) \in B\}$ is a measurable subset of $\Omega :\ f^{-1}(B) \in \Sigma$.

Integration is built up from definitions using simple functions, then nonnegative functions, up to arbitrary measurable functions, with respect to the probability measure $\mathbb{P}$, just as in

basic integration theory, to derive the Lebesgue integral. The expectation of a random variable, if it is integrable, is defined as its Lebesgue integral:

$$\mathrm{E}[X] = \int X(\omega) d\mathbb{P} \tag{4.5}$$

Important theorems such as the Bounded Convergence Theorem, Dominated Convergence Theorem, Monotone Convergence Theorem and Fatou's Lemma transfer exactly. Some terminology is often used in the case of convergence in the probability setting that replaces the usual terminology:

1) Almost sure convergence (a.s.): equivalent to convergence almost everywhere (a.e.),

2) Convergence in probability: equivalent to convergence in measure,

3) Infinitely often (i.o.) and almost always (a.a.): given a sequence of events $A_n \in \Sigma$, $\{A_n \ i.o.\}$ is equivalent to $\limsup_n A_n$, and $\{A_n \ a.a.\}$ to $\liminf_n A_n$.

When the underlying probability triple is complete, if X is a random variable, and $Y : \Omega \to \mathbb{R}$, such that $\mathbb{P}(X = Y) = 1$, then Y is also a random variable.

Given a random variable X on a probability triple $(\Omega, \Sigma, \mathbb{P})$, its **distribution** is the probability measure *induced* on $(\mathbb{R}, \mathcal{B})$, i.e. it is the function μ defined on $\mathcal{B}$, the Borel sets on $\mathbb{R}$, such that

$$\mu(B) = \mathbb{P}(X^{-1}(B)), \ B \in \mathcal{B}. \tag{4.6}$$

If μ is the distribution of some random variable, then $(\mathbb{R}, \mathcal{B}, \mu)$ is a valid probability triple. As in elementary probability theory, we have the result that distributions of a random variables completely specify their expected values, and those of functions of the random variables. In measure theoretic probability theory this is proved via the

Theorem. (Change of Variables): Given a probability triple $(\Omega, \Sigma, \mathbb{P})$, let X be a random variable having distribution μ. Then for any Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\int_\Omega f(X(\omega)) \mathbb{P}(d\omega) = \int_{-\infty}^{\infty} f(t) \mu(dt), \tag{4.7}$$

provided that each side is well-defined. In other words, the expectation of the random variable $f(X)$ with respect to the probability measure $\mathbb{P}$ on Ω is equal to the expectation of the function f with respect to the measure μ on $\mathbb{R}$.

Proof:

If $f = \mathbf{1}_B$ is the indicator function of a Borel set $B \subseteq \mathbb{R}$, the claim in equation (4.7) is just the definition of measurability and the induced measure. By linearity, the claim extends to simple functions, and by uniform limits to bounded measurable functions. Then using monotone limits, we extend the claim to nonnegative functions, and then considering positive and negative parts of general measurable functions separately, we are done. $\square$

The *distribution function induced on* $\mathcal{B}$ is the function

$$F(x) = \mu((-\infty, x]) = \mathbb{P}[\omega : X(\omega \leqslant x)]. \tag{4.8}$$

The above results can be extended to the more general case of vector-valued random variables.

The following section introduces an important idea involved in the convergence of sequences of probability distributions. In fact, it is this notion upon which the methods relating to the Kurtz theorem is based. Additionally, it is required for a precise statement and proof of the Central Limit Theorem.

4.2.2 Weak convergence

Weak convergence is a way to establish a type of "distance" between probability distributions, in the sense that if two probability measures are "close," they should assign similar probabilities to the same sets. There are several ways to define weak convergence, and a classic theorem connects the definitions. We must note, however, that any notion of weak convergence implicitly uses aspects of the underlying topology of the space being studied.

We will see later in this chapter that the most general setting for the more interesting aspects of probability theory, including analysis of stochastic processes, involves a metric space in addition to a measure.

Definition. Weak convergence 1: A sequence μ_n of probability distributions on $\mathbb{R}$ is said to *converge weakly* to a probability distribution μ if $\lim_{n\to\infty} \mu_n[I] = \mu[I]$ for any interval $I = [a, b]$ such that the single-point sets a and b have probability 0 under μ.

This is the most basic definition, yet the very idea of a closed interval, in the more general setting, involves the idea of some type of distance between elements of the space. The next definition also implicitly uses topological notions as it involves the continuous functions:

Definition. Weak convergence 2: A sequence μ_n of probability distributions on $\mathbb{R}$ is said to *converge weakly* to a probability distribution μ if $\int_{\mathbb{R}} f d\mu_n \to \int_{\mathbb{R}} f d\mu$, for all bounded, continuous functions $f : \mathbb{R} \to \mathbb{R}$.

The reason that the topological connections are pointed out in the above definitions is that the idea of weak convergence, and especially definition 2, corresponds to the weak* topology from functional analysis, with the function space being the set of all continuous functions on $\mathbb{R}$ vanishing at infinity, with norm defined by $||f|| = \sup_{x\in\mathbb{R}} |f(x)|$, and with dual space consisting of all finite signed Borel measures on $\mathbb{R}$. This topology is used often by Dynkin and by Kurtz in establishing some of their results using operator semigroups on probability function spaces. The next definition is stated with respect to the distribution functions:

Definition. Weak convergence 3: A sequence μ_n of probability distributions on $\mathbb{R}$ with distribution functions $F_n(x)$ is said to *converge weakly* to a probability distribution μ with distribution function $F(x)$ if $\lim_{n\to\infty} F_n(x) = F(x)$ for every x that is a continuity point of F.

If μ_n converges weakly to μ, we often write $\mu_n \Rightarrow \mu$, and sometimes $\mu_n \xrightarrow{w} \mu$. The above

definitions are equivalent, and additionally, there are several other statements that can be shown to be equivalent as well. I give two of these below:

Definition. Weak convergence 4: $\mu_n \Rightarrow \mu$ if there are random variables $Y, Y_1, Y_2, \ldots$ defined jointly on some probability triple, with the distribution of $Y = \mu$, and the distribution of $Y_n = \mu_n$ for each $n \in \mathbb{N}$, such that as $n \to \infty$, $Y_n \to Y$ with probability 1.

Definition. Weak convergence 5: (Recall that the boundary of a set $A \subseteq \mathbb{R}$ is $\partial A = \{x \in \mathbb{R} : \forall\ \epsilon > 0, A \cap (x - \epsilon, x + \epsilon) \neq \emptyset, A^c \cap (x - \epsilon, x + \epsilon) \neq \emptyset\}$). $\mu_n \Rightarrow \mu$ if $\mu_n(A) \to \mu(A)$ for all measurable sets A such that $\mu(\partial A) = 0$.

Connections to other types of convergence:

Proposition. If $\{X_n\} \to X$ in probability, then the distributions of the X_n converge weakly to the distribution of X. $\square$

Proposition. Suppose the distributions of X_n converge weakly to the distribution of X, with $X_n \geqslant 0$. Then $\mathrm{E}(X) \leqslant \liminf(\mathrm{E}(X_n))$.

Proof:

By definition 4 above, we can find random variable Y_n and Y with distribution of Y_n = distribution of X_n for each n, distribution of Y = the distribution of X, and $Y_n \to Y$ with probability 1. Then, using Fatou's Lemma,

$$\mathrm{E}(X) = \mathrm{E}(Y) = \mathrm{E}(\liminf Y_n) \leqslant \liminf(\mathrm{E}(Y_n)) = \liminf(\mathrm{E}(X_n)) \quad \square \tag{4.9}$$

4.2.3 Independence

Definition. Two events A and B are independent if $\mathbb{P}(A \cup B) = \mathbb{P}(A)\mathbb{P}(B)$.

Two random variables X, Y are independent if for any two Borel sets S_1, S_2 on $\mathbb{R}$, $\mathbb{P}(X^{-1}(S_1) \cap Y^{-1}(S_2)) = \mathbb{P}(X^{-1}(S_1))\mathbb{P}(X^{-1}(S_2))$.

A finite collection of random variables, $\{X_j : 1 \leqslant j \leqslant n\}$ is said to be independent if for any n Borel sets $A_1, A_2, \ldots A_n$ on $\mathbb{R}$,

$$\mathbb{P}\left[\bigcap_{1\leqslant j\leqslant n} [X_j \in A_j]\right] = \prod_{1\leqslant j\leqslant n} \mathbb{P}[X_j \in A_j]. \tag{4.10}$$

An infinite collection of random variables is independent if every finite subcollection is independent.

Given two sets Ω_1, Ω_2, the Cartesian product $\Omega = \Omega_1 \times \Omega_2$ is the set of pairs (ω_1, ω_2) with $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$. $\mathcal{F}$ is the field of finite disjoint measurable "rectangles" formed by the product of a set from each σ-algebra associated to Ω_1, Ω_2. The **product σ-algebra** is the σ-algebra generated by $\mathcal{F}$. The **product measure** is created by extension from the measure of a rectangle $\mathbb{P}(A_1 \times A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ to a countably additive measure on the σ-algebra generated by $\mathcal{F}$.

Proposition. Two random variables X, Y defined on $(\Omega, \Sigma, \mathbb{P})$ are independent if and only if the measure induced on $\mathbb{R}^2$ by (X, Y) is the product measure $\mu_1 \times \mu_2$, where μ_1 and μ_2 are the distributions on $\mathbb{R}$ induced by X and Y, respectively. $\square$

Existence of sequences of independent random variables:

The most basic types of stochastic processes can be built up from sums of outcomes of sequences of independent events, for instance, gambling games such as sequences of coin tosses. Furthermore, ideas such as a stochastic process with independent increments, and the strong Markov property, requires the existence of sequences of independent random variables. The problem with existence is such: how do we know that there exists a probability measure on the space of all sequences of random variables, each with a specified distribution, and that this space is consistent with all projections onto finite dimensional spaces of sequences?

The proof of this is provided by the Kolmogorov Consistency Theorem. It also provides the conditions for the existence of stochastic processes in general, and defines the type of

"questions" that can be answered about a process probabilistically. The theorem is presented in the section on stochastic processes later in this chapter, but a version is given here. First, we construct a measure $\mathbb{P}_n$ on $\mathbb{R}^n$, for every n, that is the joint distribution of the first n random variables, using the construction of product measures described above. These measures are *consistent*, in that if the spaces $\mathbb{R}^{n+1}$ are projected onto $\mathbb{R}^n$, the measures on the projections are $\mathbb{P}_n$. This is one way to define a *consistent family of finite-dimensional distributions*.

Let Ω be the space of all real sequences, $\mathbb{R}^\infty$, and the Σ be the σ-field generated by the field of *finite-dimensional cylinder sets*, $B = \{\omega : (x_2, x_2, \ldots, x_n) \in A\}$, for all $A \in \mathcal{B}(\mathbb{R}^n)$ and all $n \in \mathbb{N}$.

Theorem. The Kolmogorov Consistency Theorem, version 1: Given a consistent family of finite-dimensional distributions $\mathbb{P}_n$, there exists a unique $\mathbb{P}$ on (Ω, Σ) such that for every n, under the natural projection $\pi_n(\omega) = (x_1, x_2, \ldots, x_n)$, the induced measure $\mathbb{P}(\pi_n^{-1}) = \mathbb{P}_n$ on $\mathbb{R}^n$. $\square$

Finite dimensional distributions and the Kolmogorov Consistency Theorem can also be expressed using the idea of π-systems and λ-systems:

Definition. π-system: Let Ω be a set. A π-system $\mathcal{A}$ on Ω is a collection of subsets of Ω which is closed under finite intersections. Then $\sigma(\mathcal{A})$ is the σ-algebra generated by $\mathcal{A}$, and if $\sigma(\mathcal{A}) = \Sigma$, we say $\mathcal{A}$ generates Σ.

Definition. λ-system: Let $\mathcal{D}$ be a collection of sets. Then $\mathcal{D}$ is a λ-system if it has the following properties:

1) $\Omega \in \mathcal{D}$,

2) $(A, B \in \mathcal{D}$ and $A \subseteq B) \Rightarrow B \setminus A \in \mathcal{D}$,

3) $(A_n \in \mathcal{D},$ and $A_1 \subset A_2 \subset \ldots) \Rightarrow \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{D}$.

Lemma. Let $\mathcal{A}$ be a π-system and let $\mathcal{D}$ be a λ-system. Suppose $\mathcal{A} \subseteq \mathcal{D}$. Then $\sigma(\mathcal{A}) \subseteq \mathcal{D}$. $\square$

Using this lemma, we can prove the following version of the consistency theorem:

Theorem. Let (Ω, Σ) be a measurable space. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be probability measures on (Ω, Σ) which agree on a π-system generating Σ. Then $\mathbb{P}_1 = \mathbb{P}_2$.

Proof:

Let $\mathcal{D} = \{A \in \Sigma : \mathbb{P}_1(A) = \mathbb{P}_2(A)\}$. We have assumed $A \subseteq \mathcal{D}$. We can check that $\mathcal{D}$ is a λ-system. Since $\mathcal{A}$ generates $\Sigma, \Sigma \subseteq \mathcal{D}$, and $\mathbb{P}_1 = \mathbb{P}_2 \quad \square$.

Using results from this theorem, and additional considerations involving the measure-theoretic version of conditional probability, which is introduced in the next section, we also have the following theorem about independence:

Theorem. Let $\mathcal{A}_1$ be a π-system generating Σ_1 and $\mathcal{A}_2$ be a π-system generating Σ_2. Suppose that

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2) \text{ for all } A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2. \tag{4.11}$$

Then Σ_1 and Σ_2 are independent. $\square$

4.2.4 Conditioning

Conditional probability and expectation are essential concepts for defining Markov processes, as well as for many other aspects of probability theory. The measure theoretic version of conditioning is a concept relatively different from the intuitive version from elementary probability and more technically involved. This is because we would like for our definition,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \tag{4.12}$$

to be able to be defined for sets A whose probabilities equal zero. The above ratio, with $\mathbb{P}(A) = 0$, suggests a type of derivative may be involved. In fact, the use of the Radon-Nikodym derivative is needed to prove existence of conditional probabilities and expectations as they are defined with measure theory.

We define the conditional probability through the definition of the conditional expectation in this approach. The conditional expectation is now thought of as a type of random variable itself. It is defined as conditional on a sub-σ-algebra.

Definition. Given any sub-σ-algebra $\mathcal{F} \subseteq \Sigma$, and given an integrable function G on Σ, we define the **conditional expectation** of G given $\mathcal{F}$ as the function $g(\omega)$ such that

$$\int_A g(\omega) d\mathbb{P} = \int_A G(\omega) d\mathbb{P} \tag{4.13}$$

for all $A \in \mathcal{F}$.

If we take G to be the indicator function, $G = \chi_B(\omega)$ for $B \in \Sigma$, we have $\mathrm{E}(G|\mathcal{F})$ is the **conditional probability** of B given $\mathcal{F}$.

Given two random variables Y and X, if $\mathrm{E}|Y| < \infty$, $\mathrm{E}(Y|X)$ is a conditional expectation of Y given X if it is a $\sigma(X)$-measurable random variable and, for any Borel $S \subseteq \mathbb{R}$,

$$\mathrm{E}(\mathrm{E}(Y|X)\mathbf{1}_{X \in S}) = \mathrm{E}(Y\mathbf{1}_{X \in S}). \tag{4.14}$$

Likewise, given an event A, the conditional probability of A given X, $\mathbb{P}(A|X)$ is a $\sigma(X)$-measurable random variable and, for any Borel $S \subseteq \mathbb{R}$,

$$\mathrm{E}(\mathbb{P}(A|X)\mathbf{1}_{X \in S}) = \mathbb{P}(A \cap \{X \in S\}). \tag{4.15}$$

4.2.5 Stochastic Processes

Stochastic processes are here presented in a measure-theoretic setting, with additional concepts using metric spaces and other advanced constructs. The theory of stochastic processes generally studies infinite families of random variables.

Definition. A stochastic process, $\{x(\theta) : \theta \in \Theta\}$, is a collection $X_\theta(\omega)$ of random variables defined on some *shared* probability space $(\Omega, \Sigma, \mathbb{P})$, and indexed by $\theta \in \Theta$, where Θ is any

non-empty index set. θ is called the *parameter* of the process and often refers to time. A stochastic process can also be thought of as a random function $x(\theta) = x(\theta, \omega)$.

We can also define a stochastic process more generally, as a measurable mapping from one measure space to another. Either of the measure spaces may also be defined to be a metric space by assigning an appropriate metric. A general measurable mapping is a mapping $x(\theta, \omega) : \Theta \times \Omega \to X$, where $(X, \mathcal{B})$ is another measurable space. Sometimes such a generalization is called a *random element* [46]. The function $x(\theta, \omega) : \Theta \times \Omega \to X$, such that $x(\theta, \omega)$ is a random element in $(X, \mathcal{B})$ for every $\theta \in \Theta$ is a stochastic process.

Thus a random element may also be a function. A stochastic process may also be defined as a mapping onto the space of random functions:

Definition. A stochastic process is a measurable mapping $x(\theta, \omega) : \Theta \times \Omega \to X$, where

$$X = \prod_{\theta \in \Theta} \mathbb{R} \tag{4.16}$$

is the space of $\mathbb{R}$-valued *functions* on Θ.

For a general statement of this definition, where X is the set of E-valued functions on Θ, E must be assumed to be a metric space with additional properties that provide certain implicitly assumed requirements, such as separability.

The space $(X, \mathcal{B}, \mu)$, where X is defined as in (4.16) and μ is the induced distribution on X, can be the primary focus of study, as a probability space of its own. This space can be studied in a functional analysis setting, as a space of functions with a metric or norm, and without considering the underlying probability space $(\Omega, \Sigma, \mathbb{P})$. This is a "direct representation of the process," and was used by Kolmogorov [22]. In either case, we can create an appropriate metric on the function space X which provides the space with a topology that allows for easier analysis: for instance, creates a "Polish" space. Polish spaces are discussed below.

In physical applications, the stochastic process describes a state of a certain system that is varying with time, and generally, $\Theta = \mathbb{R}^{\geqslant 0}$. The probability space represents the types of questions which can be answered probabilistically about the system. The space $(X, \mathcal{B})$ is often referred to as the *phase space* in physical applications. Elements of the above space X of $\mathbb{R}$-valued functions are sometimes called "waveforms."

A third way of defining a stochastic process is as a single sequence or function on a probability space, together with a transformation or family of transformations on the space, which "moves the sequence along in time." This is the *dynamical systems* version of a stochastic process. An example of such a construct is given here:

Let $\{X_n; n \in \mathbb{N}\}$ be a direct representation of the stochastic process. For each N, X_n takes on some value from a set A, for instance, $A = \mathbb{R}$ in (4.16) above. Then the action of "time" can be described by defining a left-shift *transformation* $\mathbf{T} : A^\infty \to A^\infty$, where A^∞ is the space of infinite sequences of values from A, defined by:

$$\mathbf{T}(\dots, x_{n-1}, x_n, x_{n+1}, \dots) = (\dots, x_n, x_{n+1}, x_{n+2}, \dots). \tag{4.17}$$

$\mathbf{T}$ takes a sequence and moves every symbol in the sequence one slot to the left. Now the time, n, can be written as a function of the sequence:

$$X_n(x) = X_0(\mathbf{T}^n x), \tag{4.18}$$

and now the process is not viewed as an infinite collection of random variables on a shared probability space, but as a single random variable or random function X_0 together with a transformation on the *space*. The collection $(A^\infty, \mathcal{B}_A, \mu, \mathbf{T})$ is then called a dynamical system.

The finite-dimensional distributions introduced in section 4.2.3. define all the measurable sets of events described by a stochastic process and thus describe all that can be determined probabilistically about the process. The finite-dimensional distributions and another version

of the Kolmogorov consistency theorem can be used to establish the existence of stochastic processes.

Recall the definition of a family of *finite-dimensional distributions*: For a collection of random variables, $X_t, t \in T = \mathbb{R}^{\geqslant 0}$, defined on $(\Omega, \Sigma, \mathbb{P})$, and taking values in $\mathbb{R}$, we define the Borel probability measure $\mu_{t_1,t_2,\ldots t_k}, k \in \mathbb{N}$ on $\mathbb{R}^k$ by

$$\mu_{t_1,t_2,\ldots t_k}(H) = \mathbb{P}((X_{t_1}, \ldots, X_{t_k}) \in H), \quad H \subseteq \mathbb{R}^k \text{ Borel.} \tag{4.19}$$

The distributions $\{\mu_{t_1,t_2,\ldots,t_k}; k \in \mathbb{N}, t_k \in T \text{ distinct}\}$ are called the finite-dimension distributions of $\{X_t; t \in T\}$. The finite dimensional distributions satisfy two *consistency conditions*:

Definition. Consistency conditions for finite-dimensional distributions

1) If $(s(1), s(2), \ldots, s(k))$ is any permutation of $(1, 2, \ldots, k)$, then for any distinct $t_1, \ldots, t_k \in T$ and any Borel $H_1, \ldots, H_k \subseteq \mathbb{R}$, we have

$$\mu_{t_1,t_2,\ldots,t_k}(H_1 \times \cdots \times H_k) = \mu_{s(1),\ldots,s(k)}(H_{s(1)} \times \cdots \times H_{s(k)}). \tag{4.20}$$

2) For distinct $t_1, \ldots, t_k \in T$ and any Borel $H_1, \ldots, H_{k-1} \subseteq \mathbb{R}$, we have

$$\mu_{t_1,t_2,\ldots,t_k}(H_1 \times \cdots \times H_{k-1} \times \mathbb{R}) = \mu_{t_1,\ldots,t_{k-1}}(H_1 \times \cdots \times H_{k-1}). \tag{4.21}$$

These conditions are ordinary and expected of any probability distributions. Remarkably, they are all that is required to define a stochastic process associated with a given set of finite-dimensional distributions satisfying them:

Theorem. Kolmogorov Consistency Theorem: Existence of a Stochastic Process

A family of Borel probability measures $\{\mu_{t_1,t_2,\ldots,t_k}; k \in \mathbb{N}, t_k \in T \text{ distinct}\}$, with $\mu_{t_1,t_2,\ldots,t_k}$ a measure on $\mathbb{R}^k$, satisfies the consistency conditions above if and only if there exists a probability triple $(\Omega, \Sigma, \mathbb{P})$ and a random function $x(t, \omega) : (T \times \Omega) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that

for all $k \in \mathbb{N}$, distinct $t_1, \ldots, t_k \in T$, and Borel $H \subseteq \mathbb{R}^k$, we have

$$\mathbb{P}(\omega : (x(t_1, \omega), \ldots, x(t_k, \omega)) \in H) = \mu_{t_1, t_2, \ldots, t_k}(H). \quad \square \tag{4.22}$$

One way to prove this theorem involves letting $\Omega = \prod_{\theta \in \Theta} \mathbb{R}$, the space of all $\mathbb{R}$-valued functions on T, and $\Sigma = \sigma\{\{x(t) \in H\}; t \in T, H \subseteq \mathcal{B}(\mathbb{R})\}$. $\mathbb{P}$ is defined for finite-dimensional sets, and then extended to Σ in a similar way to the creation of the Lebesgue outer measure.

For the proof of the general case, the underlying metric space must be a Polish space— a complete, separable metric space. For discrete time processes, the finite-dimensional distributions and their associated measurable sets give a complete description of all events related to the process. This is also true for processes in continuous time but with a countable state space. However, for stochastic processes with uncountable time index and uncountable state space, certain events or "questions relating to the process" can not be addressed using the measurable sets created via the method described above. This is because notions such as boundedness, continuity, differentiability, etc., depend on knowledge of *all* values of the uncountable number of the $x(t, \omega)$. Thus the "probability that $x(t)$ is a continuous function," for instance, can not be determined.

Several methods are used to provide solutions to the above dilemma. One technique is to consider the equivalency classes of those random variables that have the same finite-dimensional distributions, which are called versions of each other, and prove theorems with respect to these versions. Constraints may be placed on the types of random variables that are considered, such as those whose moments satisfy certain conditions. These types of techniques are used in developing the theory of diffusions. Alternately, one can choose to consider only processes whose spaces consist of certain types of "nice" functions, such as the space of all almost surely continuous functions, or the space of functions that are almost surely continuous except for jumps. The standard Banach space $C[a, b]$ of continuous functions on the closed interval is a Polish space with its usual norm $||f|| = \sup_{t \in [a,b]} |f(t)|$. The

space $D[a,b]$ is the space of right-continuous functions with left limits, or *càdlàg* functions, using the classic french abbreviation "continue à droite, limite à gauche." The metric for these functions is unusual and several versions of it were created by Skorokhod; J_1 is used most frequently and is often referred to as the "Skorokhod metric," creating the "Skorokhod topology."

Because Kurtz considered the space of operators on the space of random functions for his proof, he was able to use the sup norm, and to modify some theorems on convergence of sequences of operator semigroups to be applicable to semigroups of jump Markov processes. However, he was forced to define "extended limits" with respect to which convergence was achieved. The sample functions of jump processes with countable state spaces can be studied as a topological space $D_E[0,\infty)$, and creation of compact sets allows convergence of probability measures using topological arguments [11].

We are now able to define Markov processes in the measure-theoretic setting. Discrete and continuous time Markov processes with a general state spaces, which may be countable or uncountable, are defined in the following. Given a general state space X, which is any non-empty set together with a σ-algebra $\mathcal{B}$ of measurable subsets, we have the following processes:

Definition. Discrete-time Markov process.

The time index is $n \in \mathbb{N}$. For each time step, the one-step transition probabilities are given by $\{P(x,A)\}_{x\in X, A\in\mathcal{B}}$, and intuitively, can be thought of as the probability that the process will be in some set A after one step, given it is currently in a state x. We make the following two assumptions:

1) For each fixed $x \in X$, $P(x,\cdot)$ is a probability measure on $(X,\mathcal{B})$.

2) For each fixed $A \in \mathcal{B}$, $P(x,A)$ is a measurable function of $x \in X$.

We now define the initial distribution ν, which can be any probability distribution on $(X,\mathcal{B})$. We then have a discrete-time, general state space, time-homogeneous Markov process

$X_0, X_1, X_2, \ldots$, where

$$\mathbb{P}(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \int_{x_0 \in A_0} \nu(dx_0) \int_{x_1 \in A_1} P(x_0, dx_1) \ldots \\ \ldots \int_{x_{n-1} \in A_{n-1}} P(x_{n-2}, dx_{n-1}) \int_{x_n \in A_n} P(x_{n-1}, dx_n). \tag{4.23}$$

These integrals are well-defined because of assumption 2 above.

Definition. Continuous-time Markov process.

The transition probability $P(s, t, x, A), 0 \leqslant s < t,\ x \in A,\ A \in \mathcal{B}$ defines the probability of the event that the system found in the state x at the moment s will belong to the set A at the moment t. As in the previous case, the transition probability has the properties that

1) it is a probability measure with respect to $A \in \mathcal{B}$,

2) it is a measurable function with respect to $x \in X$.

The Chapman-Kolmogorov equation is satisfied for all $0 \leqslant s < t < u$:

$$P(s, x, u, A) = \int P(t, y, u, A) P(s, x, t, dy) \tag{4.24}$$

For a time-homogeneous, continuous-time Markov process having initial distribution ν

$$\mathbb{P}(X_0 \in A_0, X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \\ \int_{A_0} \int_{A_1} \ldots \int_{A_n} \nu(dx_0) P_{t_1}(x_0, dx_{t_1}) \ldots P_{t_n - t_{n-1}}(x_{t_{n-1}}, dx_{t_n}) \tag{4.25}$$

for all times $0 \leqslant t_1 < \cdots < t_n$ and all subsets $A_1, \ldots, A_n \in \mathcal{B}$.

Define the distribution of a random variable to be a *point-mass* δ_c if $\delta_c(B) = \mathbf{1}_B(c)$ for any measurable set B; in other words, $\delta_c(B)$ equals 1 if $c \in B$ and 0 otherwise. Then, letting $P_0(x, \cdot)$ be a point-mass at x, it then follows that

$$P_{s+t}(x, A) = \int P_s(x, dy) P_t(y, A),\ s, t \geqslant 0,\ x \in X,\ A \in \mathcal{B}, \tag{4.26}$$

the Chapman-Kolmogorov equation in the time-homogeneous case.

4.2.6 Countable state spaces and finite-dimensional distributions

As we saw above, the probability of *any* event depending on a right-continuous process can be determined by its finite-dimensional distributions, that is, from the probabilities

$$\mathbb{P}(X_{t_0} = i_0, X_{t_1} = i_1, \ldots, X_{t_n} = i_n), \text{ for } n \geqslant 0,\ 0 \leqslant t_0 \leqslant t_1 \leqslant \cdots \leqslant t_n \text{ and } i_0, \ldots, i_n \in I, \tag{4.27}$$

where I is a countable set, and $X_t : \Omega \to I,\ 0 \leqslant t \leqslant \infty$ is a family of random variables.

We used this information implicitly in chapters 2 and 3 when we calculated probabilities of events that were not of the form $\{X_0 = i_0, \ldots, X_n = i_n\}$. For instance, the notion of a stopping time, and the definition of a jump process via its jump rates and exponential holding times, depend on the measurability of sets of the form

$$\{X_t = i \text{ for some } t > 0)\}. \tag{4.28}$$

which may depend on all $(X_t)_{t\geqslant 0}$. These sets may not be measurable with respect to $\sigma((X_t) : t \geqslant 0)$, because sigma algebras involve *countable* unions and intersections. But for right-continuous processes with countable state spaces, events of the form $\{X_0 = i_0, \ldots, X_n = i_n\}$ form a π-system which generates the σ-algebra $\sigma((X_t) : t \geqslant 0)$.

As an example, consider the notion of a stopping time. The concept of "depending only on" can only be made precise, in a probabilistic sense, as measurability with respect to some σ-algebra. Thus we let $(X_t)_{t\geqslant 0}$ be a right-continuous process with values in a countable set I. Let Σ_t be the σ-algebra generated by $\{X_s : s \leqslant t\}$, in otherwords, all sets $\{X_s = i\}$ for $s \leqslant t$ and $i \in I$.

Definition. A **stopping time** of $(X_t)_{t\geqslant 0}$ is a random variable T with values in $[0, \infty]$ if $\{T \leqslant t\} \in \Sigma_t$ for all $t \geqslant 0$.

For stopping times T, we define the collection of sets $\Sigma_T = \{A \in \Sigma : A \cap \{T \leqslant t\} \in \Sigma_t \text{ for all } t \geqslant 0\}$. This is a way to precisely define the idea of "those sets that depend only on $\{X_t : t \leqslant T\}$".

The following results can be proved using countable unions and intersections of sets and the assumption of right-continuity:

Lemma. Let S and T be stopping times of $(X_t)_{t\geqslant 0}$. Then both X_T and $\{S \leqslant T\}$ are Σ_T-measurable. $\square$

Theorem. Strong Markov property. Let $(X_t)_{t\geqslant 0}$ be a (minimal) right-continuous, discrete-state Markov process, and let T be a stopping time of $(X_t)_{t\geqslant 0}$. Then, conditional on $T < \zeta$ and $X_T = i$, $(X_{T+t})_{t\geqslant 0}$ is Markov and independent of Σ_T. $\square$

4.2.7 Martingales

An alternate way to describe stochastic processes is via the conditional expectations of a sequence of random variables X_n with respect to a sequence of sub-σ-algebras, instead of by using the conditional probability distributions.

Definition. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. A discrete-time, discrete-state stochastic process $X_0, X_1, \ldots$, together with its sequence of corresponding sub-σ-algebras Σ_1, Σ_2 is a **martingale** if:

1) For all n, X_n is measurable with respect to Σ_n, and $\mathrm{E}|X_n| < \infty$,

2) The sub-σ-algebras are increasing: $\Sigma_n \subset \Sigma_{n+1}$ for every n,

3) For every n, and with probability , $\mathrm{E}(X_{n+1}|\Sigma_n) = X_n$.

The sequence is a **submartingale** if $\mathrm{E}|X_n| < \infty$ for all n, and $\mathrm{E}(X_{n+1}|\Sigma_n) \geqslant X_n$.

A discrete-time Markov chain with transition probabilities $P_{i,j}$ and with $\mathrm{E}|X_n| < \infty$ is a martingale provided that

$$\sum_{j\in S} jP_{i,j} = i \quad i \in S. \tag{4.29}$$

An example of this type of Markov chain is the simple symmetric random walk, where $S = \mathbb{Z}, X_0 = 0$, and $P_{i,i-1} = P_{i,i+1} = \frac{1}{2}$.

In the general case, for a continuous-time process, we first define a **filtration**. Given a σ-algebra $\mathcal{F}$ on a set Ω, a filtration is a collection of sub-σ-algebras of $\mathcal{F}$, $\{\mathcal{F}_t, t \in [0, \infty)\}$, if $\mathcal{F}_s \subseteq \mathcal{F}_t$ for all $s < t$. A stochastic process $(X_t)_{t \geqslant 0}$ is said to be *adapted to a filtration* $\{\mathcal{F}_t, t \geqslant 0\}$ if X_t is $\mathcal{F}_t$-measurable for all t. We often think of the filtration as the information available to the observer by time t.

Definition. A real-valued process $(X_t)_{t \geqslant 0}$ with $\mathrm{E}|X_t| < \infty$ for all $t \geqslant 0$ and adapted to a filtration $\{\mathcal{F}_t\}$ is a **martingale** if:

$$\mathrm{E}[X_t | \mathcal{F}_s] = X_s \text{ for } 0 \leqslant s < t, \tag{4.30}$$

and a **submartingale** if

$$\mathrm{E}[X_t | \mathcal{F}_s] \geqslant X_s \text{ for } 0 \leqslant s < t. \tag{4.31}$$

Results from martingale theory include convergence theorems and descriptions of underlying stochastic processes, and can be used to characterize Markov processes and establish limiting approximations of sequences of Markov processes. Kurtz used martingale theory to establish bounds and approximations of the limiting probability with respect to which the stochastic models converged to the deterministic equation [35]. Furthermore, the original proof of Kurtz's theorem was revisited in [11] using martingales. An example of a basic martingale convergence theorem is given here:

Theorem. Let $\{X_n\}$ be a discrete-time submartingale. Suppose that $\sup_n \mathrm{E}|X_n| < \infty$. Then there is a finite random variable X such that $X_n \to X$ almost surely. $\square$

Chapter 5

Functional analysis, operator semigroups, and the Kurtz theorem

Probability theory has "exploded" in the last 50 years, and now it is impossible to contain the entirety of the theory in a single book or even collection. Specialization and integration of methods from numerous other branches of mathematics have turned probability theory into a vast field. From the basic foundations described in the previous chapters, many branching trajectories are possible, depending on the particular problem and the methods chosen. The use of functional analysis, operator theory, and specifically operator semigroup techniques to study stochastic processes is one such example. An additional several hundred pages would be required to provide a comprehensive exposition of these techniques, [23].

Treatments based solely on properties derived from the finite-dimensional distributions constituted initial approaches in stochastic processes, but work pioneered by Doob, among others, led to the realization that "most processes of interest have right-continuous versions with left-hand limits," [23]. Thus more general results can be attained by considering the spaces of such processes. The "modern setup for Markov processes," advanced by Dynkin in the 50s and 60s, is one where a process is defined on a path space, with a filtration, a family of shift operators, and a collection of probability measures [23].

Kurtz's thesis, his paper in the *Journal of Applied Probability*, and his subsequent explanation of the applicability of his theorem to chemistry in the *Journal of Chemical Physics* occurred during a time when a number of new methods in the theory of stochastic processes were being developed. In fact, Kurtz himself helped to create multiple advances in this area from 1969 through 1986 and later. Thus the original way that he proved the theorem [31, 34], using less-specialized theory centered around convergence of generators of semigroups, quickly became outdated as more specialized methods appeared, and he revisited the proofs several times in the years following, using new approaches that allowed for less cumbersome, more elegant presentations and sometimes led to more general results. [11].

Nevertheless, the treatment of stochastic processes in the context of functional analysis and operator theory originated in the early part of the twentieth century. Wiener, in his approach to Brownian motion, was first to describe the distribution of a stochastic process as a measure on a function space. [6, 23]. This insight even pre-dated the establishment of measure theory in a truly rigorous framework. Kolmogorov, in 1935, was able to unify the theory of Markov chains and diffusion processes, which were previously considered to be disjoint fields, by describing transition kernels as operators, and using ideas from spectral theory to describe local characteristics via associated infinitesimal generators. [23, 55].

This chapter follows expositions of Kallenberg [23], Kreyzig [30], Bobrowski [6], Ethier and Kurtz [11], Goldstein [21], Keller-Ressel [29], Engel and Nagel [9], and Karlin [25] .

5.1 Functional analysis and probability spaces

Functional analysis became the modern setting for limit theorems and approximation results for probability and stochastic processes because of the natural wide scope and generality offered by its approach. Functional analysis does not focus on distinctions between elements of a space, but rather on establishing general properties that apply to an entire class of these elements. Once these general results are attained, they apply to entire groups of processes or

measures. For instance, by creating a function-space setting for weak convergence theory, using the appropriate topology, convergence of sequences of functions, measures, and processes, which were previously "derived by cumbersome embedding and approximation techniques" now become "accessible by straightforward compactness arguments," [23]. Such limit theorems, like the Skorokhod embedding and its subsequent use in proving the functional version of the Central Limit Theorem, traditionally used purely probabilistic and classical analysis techniques based on characteristic functions and Taylor expansions for their derivations.

One topology created for the functional analysis method of proving convergence in distribution is founded on the Prohorov metric, introduced on the space of probability measures on a separable space S. Convergence with this metric in this space *is* weak convergence. Furthermore, it turns out that if S is a Polish space, then so is the space of probability measures on S with the Prohorov metric, [6]. A property called tightness is found to be connected to a type of compactness called relative distributional compactness. Using these concepts, classical compactness-related results can be transformed into results useful in the probabilistic setting. The Arzelà-Ascoli theorem, for example, can be applied in this way for the space of continuous processes.

The Hilbert space provides a setting for a number of fundamental examples of the contributions of functional analysis to probability and stochastic processes. In 1940, von Neumann initiated the Hilbert space approach to conditioning, which makes the connection between the conditional expectation and orthogonal projections [6, 23]. As a result, we find that "the simplest and most intuitive general approach to conditioning is via projection" [6] in the Hilbert space setting, and it avoids the difficulties encountered in the traditional approach when having to use the Radon-Nikodym theorem [23]. The ergodic theorem, essential for mathematical physics, was first rigorously proved by von Neumann in the Hilbert space setting, after it was noted that there was a "connection between measure-preserving transformations and unitary operators on a Hilbert space," [23]. Modern treatments of Brownian motion, diffusions, and stochastic integration all rely on the Hilbert space approach which

provides the best-equipped framework for the most direct, general and insightful results in these fields.

The modern setting for investigating stochastic processes using functional analysis usually uses as the function space the path space of sample functions of the processes. Thus for diffusions, the space is $C(K, S)$ of continuous functions from the space (K, d) to (S, ρ), where K is compact and S is separable and complete. K is the index or parameter set, and the topology is that induced by the Prohorov metric. For jump processes, we have the space $D(\mathbb{R}_+, S)$, where D is the space of right continuous functions with left limits $f : \mathbb{R}_+ \to S$, and the jump processes have paths in D. We use one of the Skorokhod topologies on this space to achieve similar results as the Prohorov metric created for continuous paths: convergence in this space becomes convergence in distribution.

The approach taken by Kurtz to prove his theorem in its first two iterations did not use this path space approach, but rather considered the space of operators mapping random functions onto their integrals by means of a *probability kernel*:

Definition. Given two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a mapping $\mu : S \times \mathcal{T} \to \hat{\mathbb{R}}_+$, (where $\hat{\mathbb{R}}_+$ is the compactification of $\mathbb{R}_+$, or the extended real line) is called a **probability kernel** from S to T if the function $\mu(s, B)$ is $\mathcal{S}$-measurable in $s \in S$ for fixed $B \in \mathcal{T}$ and a probability measure in $B \in \mathcal{T}$ for fixed $s \in S$.

A probability kernel determines an associated operator that maps suitable functions $f : T \to \mathbb{R}$ into their integrals $\mu f(s) = \int \mu(s, dt) f(t)$. A Markov transition function is such a kernel, and an associated operator is, for example, the expectation with respect to this transition function.

As we see in the next section, the space of certain kinds of operators from one function space to another forms its own space that can be a Banach space, and can thus enjoy some important results that rely on completeness. For this type of approach, we do not have to define a new type of metric, such as the Prohorov metric or the Skorokhod metric. However, as we will see in the following sections, we do have to create extended definitions to allow for

the proper convergence results when using operator spaces in the stochastic process setting.

5.2 Operator Theory

For this section, it is assumed the reader is familiar with basic notions about vector spaces and function spaces, such as the definition of a norm, a normed linear space, a Banach space, etc. Definitions and results relating to operator theory are included here for completeness, however.

Operator theory is a branch of functional analysis that focuses on spaces of operators (mappings) from one function space to another. Formally, an **operator** is a mapping from a normed space X into a normed space Y. Operator theory is often concerned primarily with bounded linear operators, since they "fully take advantage of the vector space structure," [30]. One reason for this is the fundamental result that bounded linear operators are necessarily continuous. With the appropriate norm, the space of bounded linear operators forms a Banach space.

The notion of a Banach space incorporates two important components: an algebraic component, and a topological component [6]. The topological component comes from the definition of the norm which creates the open sets. Continuous functions on normed linear spaces take advantage of the topological component because they preserve the opens sets. The algebraic component of a Banach space, that the elements can be added together and multiplied by scalars, is important in the creation of topological groups and semigroups, addressed in the next section.

The study of Banach spaces forms a central part of functional analysis because some important properties can be proved for them which cannot be proved for incomplete normed spaces, for instance, the Hahn Banach theorem, the Uniform Boundedness theorem, the Open mapping theorem, and the Closed Graph theorem [30]. The last two are especially important in the operator semigroup methods discussed in the next section, as certain oper-

ators considered are not bounded but are closed, or are closable. This provides an alternate criterion for proving continuity.

Necessary definitions and results for use in the next section will now be introduced.

Definition. Linear Operator: A linear operator T is an operator such that

1) The domain $\mathcal{D}(T)$ of T is a vector space and the range $\mathcal{R}(T)$ lies in a vector space over the same field.

2) For all $x, y \in \mathcal{D}(T)$ and scalars α, $T(x+y) = Tx + Ty$, and $T(\alpha x) = \alpha Tx$.

Definition. Bounded linear operator: Let X and Y be normed spaces and $T : \mathcal{D}(T) \to Y$ be a linear operator, where $\mathcal{D}(T) \subset X$. The operator T is said to be bounded if there is a real number c such that for all $x \in \mathcal{D}(T)$, $||Tx|| \leqslant c||x||$, where the norm on the left is that on Y, and the norm on the right is that on X.

From the above definition we can see that a bounded linear operator maps bounded sets in $\mathcal{D}(T)$ onto bounded sets in Y. This is one reason that the bounded linear operators prove to be so useful.

If c in the above definition is set to equal 1, then the operator T is called a **contraction**. Contraction operators from an essential part of the semigroup theory used in the next section. The above definition also serves to create the definition of a norm for a bounded linear operator:

Definition. For $x = 0$, $Tx = 0$. For $x \neq 0$, we define the **operator norm** of T, also written $||T||$, as

$$||T|| = \sup_{x \in \mathcal{D}(T), x \neq 0} \frac{||Tx||}{||x||} \tag{5.1}$$

An equivalent formula for this expression, also called the operator norm of T, can be shown to be

$$||T|| = \sup_{x \in \mathcal{D}(T), ||x|| = 1} ||Tx|| \tag{5.2}$$

Definition. Continuous operator An operator T is continuous at a point $x_0 \in \mathcal{D}(T)$ if for every $\epsilon > 0$, there is a $\delta > 0$ such that $||Tx - Tx_0|| < \epsilon$ for all $x \in \mathcal{D}(T)$ satisfying $||x - x_0|| < \delta$. T is continuous if T is continuous at every $x \in \mathcal{D}(T)$.

Now we can state the fundamental

Theorem. (Continuity and Boundedness): Let $T : \mathcal{D}(T) \to Y$ be a linear operator, where $\mathcal{D}(T) \subset X$ and X, Y are normed spaces. Then

1) T is continuous if and only if T is bounded, and

2) If T is continuous at a single point, T is continuous. $\square$

Definition. The **restriction** of an operator $T : \mathcal{D}(T) \to Y$ to a subset $B \subset \mathcal{D}(T)$ is denoted by $T|_B$, and is the operator defined by $T|_B : B \to Y,\ T|_B x = Tx$ for all $x \in B$.

Definition. An **extension** of an operator T to a set $M \supset \mathcal{D}(T)$ is an operator $\widetilde{T} : M \to Y$, such that $\widetilde{T}|_{\mathcal{D}(T)} = T$, that is, $\widetilde{T}x = Tx$ for all $x \in \mathcal{D}(T)$.

Many possible extensions exist for T when $\mathcal{D}(T)$ is a proper subset of M, for instance, an extension from a dense set in X to all of X, or an extension from a normed space to its completion. Some of the most useful extensions are those that preserve an important property, such as boundedness or linearity. That way, for instance, a result about convergence in the original space can be extended to a similar result in the extended space, under certain criteria. This type of operator extension is used several times by Kurtz in the each version of the proof of his theorem. The following theorem gives a basic but important example of such an extension theorem:

Theorem. Let $T : \mathcal{D}(T) \to Y$ be a bounded linear operator, where $\mathcal{D}(T)$ lies in a normed space X, and Y is a Banach space, and let $\overline{\mathcal{D}(T)}$ be the closure of $\mathcal{D}(T)$. Then T has an extension $\widetilde{T} : \overline{\mathcal{D}(T)} \to Y$, where $\widetilde{T}$ is a bounded, linear operator of norm $||\widetilde{T}|| = ||T||$. $\square$

The following results show that the space of bounded linear operators from one normed space to another is also a normed space, and can be a Banach space.

Theorem. The vector space $\mathcal{B}(X, Y)$ of all bounded linear operators from one normed linear space X into a normed linear space Y is itself a normed space, with norm defined by

$$||T|| = \sup_{x \in \mathcal{D}(T), x \neq 0} \frac{||Tx||}{||x||} = \sup_{x \in \mathcal{D}(T), ||x||=1} ||Tx||. \quad \square \tag{5.3}$$

Theorem. If Y is a Banach space, then $\mathcal{B}(X, Y)$ is a Banach space. $\square$

The Banach space of bounded linear operators creates a setting for several types of convergence of sequences of these operators. Three types of convergence form the most often used and studied. These are convergence in the norm, strong convergence, and weak convergence. Recall that weak convergence was defined on measure spaces in the previous chapter. This idea can be generalized for normed spaces. Also, a type of convergence familiar from elementary analysis, which is convergence in the norm, is called **strong convergence**. The two versions for normed spaces are listed below. But first, we need some more terminology.

Definition. Bounded linear functional: A bounded linear functional f is a bounded linear operator with range in the scalar field over which the normed space X in which the domain $\mathcal{D}(f)$ of f lies. Thus, there exists a real number c such that for all $x \in \mathcal{D}(f)$, $|f(x)| \leqslant c||x||$.

Definition. Dual Space X': Let X be a normed space. Then the set of all bounded linear functionals on X is also a normed space with norm defined by

$$||f|| = \sup_{x \in X, x \neq 0} \frac{|f(x)|}{||x||} = \sup_{x \in X, ||x||=1} |f(x)|. \tag{5.4}$$

This space is called the dual space of X and is denoted by X'.

Theorem. The dual space X' of a normed space X is a Banach space, regardless of whether X is. $\square$.

Definition. A sequence (x_n) in a normed space X is said to be strongly convergent if there

is an $x \in X$ such that $\lim_{n\to\infty} ||x_n - x|| = 0$. We can also write $\lim_{n\to\infty} x_n = x$, or simply $x_n \longrightarrow x$.

Definition. A sequence (x_n) in a normed space X is said to be weakly convergent if there is an $x \in X$ such that for every $f \in X'$, $\lim_{n\to\infty} f(x_n) = f(x)$. We can also write $x_n \Rightarrow x$.

In finite-dimensional spaces there is no difference between strong and weak convergence. However, in infinite-dimensional spaces, strong convergence implies weak convergence, but the converse is not true.

Definition. Convergence in the operator norm is also called **uniform convergence**, and a sequence (T_n) of operators $T_n \in \mathcal{B}(X, Y)$ is said to be **uniformly operator convergence**, or simply uniformly convergent, if (T_n) converges in the norm on $\mathcal{B}(X, Y)$, or or there is an operator $T : X \to Y$ such that $||T_n - T|| \longrightarrow 0, \text{ as } n \to \infty$

Definition. A sequence (T_n) of operators $T_n \in \mathcal{B}(X, Y)$ is said to be **strongly operator convergent**, or simply strongly convergent, if $(T_n x)$ converges strongly in Y for every $x \in X$, or there is an operator $T : X \to Y$ such that $||T_n x - Tx|| \longrightarrow 0$ for all $x \in X$.

Definition. A sequence (T_n) of operators $T_n \in \mathcal{B}(X, Y)$ is said to be **weakly operator convergent**, or simply weakly convergent, if $(T_n x)$ converges weakly in Y for every $x \in X$, or there is an operator $T : X \to Y$ such that $|f(T_n x) - f(Tx)| \longrightarrow 0$ for all $x \in X$ and all $f \in Y'$.

Uniform convergence implies strong convergence, which implies weak convergence. Convergence in the operator norm is called uniform convergence because it is strong convergence that is uniform in any ball. Uniform convergence turns out not to be a useful requirement for proving convergence in many instances, especially for stochastic processes, because it can be much too restrictive. More often, in such settings, we see the use of strong convergence, weak convergence, or some other type of convergence related to these, and defined specifically for the proof at hand. The definition of additional types of operator limits is central to the way that the Kurtz theorem is proved.

A few additional definitions in operator theory are required for the presentation of operator semigroup methods in stochastic processes. Not all operators which are encountered in applications are bounded. For instance, the differential operator is not bounded, yet it has an important role in many areas of mathematics. However, a different property that is useful which the differential possess is that of being closed. For the operator semigroup approach, this is another important property to consider.

Definition. Let X and Y be normed spaces and $T : \mathcal{D}(T) \to Y$ be a linear operator with domain $\mathcal{D}(T) \subset X$. Then T is called a **closed linear operator** if its graph $\mathcal{G}(T) = \{(x, y)|x \in \mathcal{D}(T), y = Tx\}$ is closed in the normed space $X \times Y$. The two algebraic operations in $X \times Y$ are defined in the usual way, as

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1, y_2), \text{ and } \alpha(x, y) = (\alpha x, \alpha y) \tag{5.5}$$

(with α a scalar). The norm on $X \times Y$ can be given as $||(x, y)|| = ||x|| + ||y||$, or a number of other versions.

The following theorems illustrate some important results about closed linear operators.

Theorem. Closed Graph Theorem. Let X and Y be Banach spaces and $T : \mathcal{D}(T) \to Y$ a closed linear operator, where $\mathcal{D}(T) \subset X$. Then if $\mathcal{D}(T)$ is closed in X, the operator T is bounded. $\square$

Theorem. Let $T : \mathcal{D}(T) \to Y$ be a linear operator, where $\mathcal{D}(T) \subset X$, and X and Y are normed spaces. Then if T is closed if and only if it has the following property. If $x_n \to x$, where $x_n \in \mathcal{D}(T)$, and $Tx_n \to y$, then $x \in \mathcal{D}(T)$ and $Tx = y$. $\square$

Subsequently, we have the closable operator:

Definition. Closable Operator: If a linear operator T has an extension $\tilde{T}$ which is a closed linear operator, then T is said to be **closable**, and $\tilde{T}$ is called a closed linear extension.

Definition. Closure of an operator: A closed linear extension $\overline{T}$ of a closable linear operator T is said to be **minimal** if every closed linear extension T_L of T is a closed linear extension of $\overline{T}$. This minimal extension of T, if it exists, is said called the **closure** of T.

Definition. An operator T is called **densely defined** in a space $\mathcal{B}$, if $\mathcal{D}(T)$ is dense in $\mathcal{B}$.

5.2.1 Spectral Theory and Resolvents

Spectral theory developed in part to be used in the solution of problems in differential and integral equations, such as in the Sturm-Liouville theory and Fredholm theory. It studies inverse operators associated with particular originally-defined operators, and describes how their properties are related to the original. Elements of spectral theory of operators helps to characterize properties of the operators considered, and in this way, can be useful for convergence and approximation analyses as we shall see in the next section.

The most important aspect of spectral theory that applies to the operator semigroup method is the idea of the resolvent. For the finite-dimension spaces that are studied in elementary mathematics courses, we have the characteristic equation from which we are able to find eigenvectors and eigenvalues of a matrix that relates to a set of differential equations. In infinite spaces, an analogous approach is taken, but the theory becomes more complicated. However, as in the finite case, we begin with the operator $T_\lambda = T - \lambda I$, associated with $T : \mathcal{D}(T) \to X$, for $X \neq \{0\}$, a normed space, and $\mathcal{D}(T) \subset X$. λ is a scalar, and I is the identity operator on $\mathcal{D}(T)$.

If T_λ has an inverse, we denote it by $R_\lambda(T) = T_\lambda^{-1} = (T - \lambda I)^{-1}$, and call it the **resolvent operator** of T. The notation is often shortened to R_λ, and called simply the resolvent of T, if the meaning is clear. The name "resolvent" comes from the fact that the resolvent helps to solve the equation $T_\lambda x = y$. In many texts, the resolvent is alternately defined by $(\lambda I - T)^{-1}$.

The resolvent has a connection to the Laplace transform method of the one-dimensional, nonrandom case. If $g(t) = e^{\alpha t}$ for some scalar α, we can take the Laplace transform of $g(t)$

to recover α:

$$\mathcal{L}(g)(\lambda) = \int_0^\infty e^{-\lambda t} g(t) dt = \frac{1}{\lambda - \alpha}, \ \lambda \geqslant 0. \tag{5.6}$$

If α was not a scalar, but an operator A, in certain situations, we can write a representation of e^{At}, and a representation of the Laplace transform of e^{At}, and thus obtain $\frac{1}{\lambda - A} = (\lambda I - A)^{-1}$. Of course, the operator $\lambda I - A$ has to be invertible. A scalar λ is said to be in the resolvent set of a bounded linear operator A if $\lambda I - A$ has a bounded linear inverse. As we shall see in the next section, these types of generalizations of finite-dimensional results of real analysis to function-valued operators on function spaces is what constitutes the theory of operator semigroups. The types of operators considered must be carefully specified, as an exponential representation does not make sense for all operators.

5.3 Operator semigroups

At the end of section 2.4.6 of chapter 2, we saw that for finite state spaces, we have a finite $\mathbf{Q}$-matrix that allows for an exponential representation $p_t = e^{\mathbf{Q}t}$ for the solution to the Kolmogorov equations, and that the rate matrix $\mathbf{Q}$ was recoverable by taking the derivative of p_t at zero, provided this exists. We would like to be able to extend this notion to the infinite-state space case, and to the general setting using infinite-dimensional operators on probability spaces. This type of theory was created as a functional analysis approach to the Abstract Cauchy Problem:

$$\begin{aligned} &\frac{d}{dt} u(t) = A[u(t)], \ (t \geqslant 0), \\ &u(0) = f, \end{aligned} \tag{5.7}$$

where we would like A to be an operator on a Banach space $\mathcal{B}(X, X)$. From this desire grew the theory of operator semigroups. Consider again the one-dimensional, nonrandom case, and recall that if $U(t)$ is a nonconstant, real-valued function satisfying $U(t+s) = U(t)U(s)$, with $t, s \geqslant 0$, and $|U(t)| \leqslant 1$ for $t \geqslant 0$, then $U(t)$ must necessarily be of the form e^{at} for some constant a, and furthermore, when $U(t)$ is continuous at $t = 0$, so that $\lim_{t \to 0} U(t) = 1$,

then a is finite, $U(t)$ is differentiable, and $a = \frac{d}{dt}U(t)|_{t=0}$. Alternately, using the Laplace transform, as above, we can also recover a, showing the connection between the derivative at zero of $U(t)$, and the resolvent.

The property described by the Chapman-Kolmogorov equation is precisely the relation $T(s+t) = T(s)T(t)$, where $T(t)f(x) = \mathrm{E}[f(X_t)|X_0 = x]$. Thus, this expectation operator, for Markov processes, can be treated in a similar way to the method proposed above, provided the needed requirements are met, including those that allow for a definition of an operator exponentiation representation, a resolvent, and a "derivative" at zero.

Note that the exponential function and the expectation operator above depend on a continuous parameter, t. Continuous binary operations on algebraic groups of operators form the theory of topological groups, and for those sets of operators that do not necessarily have inverses or an identity element, continuous mapping $\cdot$ of a set $G \times G \to G$ creates a topological semigroup. Because it is continuous, this map is also measurable with respect to the appropriate Borel σ-algebras.

At the heart of this theory, we have the idea of the exponent of an operator. For bounded linear operators, we have the following

Theorem. Let A be a bounded linear operator in a Banach space X and let $A^n = A \cdot A^{n-1},\ n \geqslant 2$, be its n^{th} power. Then the series

$$\sum_{n=0}^{\infty} \frac{t^n A^n}{n!} \tag{5.8}$$

converges for all $t \in \mathbb{R}$ in the operator topology. $\square$

Operator semigroup theory was initiated by Hille and Yosida in the late 1940s to study problems related to generalized and abstract initial value problems for differential equations, for instance, such as the Abstract Cauchy Problem above. It was later applied to the context of Markov processes.

In order to have a generalization of the finite, one-dimensional results above, we must

have appropriate definitions and conditions for equation 5.7 above. First, assuming the function $u(t)$ takes values in some set X, we must make sense of the expression

$$\frac{du(t)}{dt} = \lim_{h\to 0} h^{-1}[u(t+h) - u(t)]. \tag{5.9}$$

To meet these requirements, X must be a vector space, and one in which limits make sense. Thus it is most common for X to be a Banach space. Furthermore, for the equation $\frac{du}{dt} = Au$ to make sense, $u(t)$ must belong to the domain of A, and also, we must have that

$$\lim_{h\to 0} ||h^{-1}[u(t+h) - u(t)] - A[u(t)]|| = 0, \tag{5.10}$$

where $||\cdot||$ denotes the norm in X. As is usual in the theory of differential equations, we would like to have a well-posed problem, Therefore, we would like to prove existence, uniqueness, and stability of our solutions, where stability refers to a proof showing that the solution depends continuously on the initial condition and the operator A. For these criteria to be met for the Abstract Cauchy Problem, we find that assumptions of time-homogeneity, the semigroup property, continuity of operators, and density of the domain, among other things, are important.

Let the solution $u(t+s)$ at time $t+s$ be computed as $T(t+s)f$. Alternately, we have that $u(t+s) = T(t)(T(s)f)$. Uniqueness of the solution implies the semigroup property $T(t+s) = T(t)T(s),\ t,s > 0$. The initial condition, f, must be required to belong to the domain of A, which must be dense in X. Additionally, each $T(t)$ must be linear if A is linear.

To create such a setting, we introduce the concept of the "strongly continuous one-parameter semigroup of bounded linear operators on a Banach space $\mathcal{H}$." This type of semigroup is called a (C_0)-semigroup. The space C_0 is the space of all continuous functions vanishing at infinity. Formally, we have the following definition:

Definition. ($\mathbf{C_0}$) semigroup: A family $T = \{T(t) : 0 \leqslant t < \infty\}$ of linear operators from a Banach space $\mathcal{H}$ to a Banach space $\mathcal{H}$ is called a (C_0) semigroup if

1) $||T(t)|| < \infty$ where $||\cdot||$ is the supremum (operator) norm,

2) $T(t+s)f = T(t)T(s)f$ for all $f \in \mathcal{H}$ and all $t, s \geqslant 0$,

3) $T(0)f = f$ for all $f \in \mathcal{H}$,

4) $t \to T(t)f$ is continuous for $t \geqslant 0$ for each $f \in \mathcal{H}$

in addition, if

5) $||T(t)f|| \leqslant ||f||$ for all $t \geqslant 0$ and all $f \in \mathcal{H}$, (i.e. $||T(t)|| \leqslant 1$ for each $t \geqslant 0$),

T is called a $(\mathbf{C_0})$ **contraction semigroup**.

Contraction semigroups are important in this theory, and are used in its application to stochastic processes. The importance of contraction mappings in proving existence and uniqueness in differential equations comes from their relation to the existence and uniqueness of fixed points, and to the methods of successive approximations. Every contraction mapping defined in a complete metric space has one and only one fixed point, thus the equation $f(x) = x$ has one and only one solution. Furthermore, a contraction on a metric space is a continuous mapping.

An operator semigroup of class C_0 is also called of class "continuous at the origin," because we have from the definition above, that $\lim_{t\to 0} T_t f = f$, for all $f \in \mathcal{H}$. Thus we can define the "derivative at zero" of $T(t)$ in analogy to the finite-dimensional case above. Formally, this represents the suggestion that $T(t) = e^{At}$, where $A = \frac{d}{dt}T(t)|_{t=0}$, and that the solution of equation 5.7 is given by $u(t) = T(t)f$, where T is the "semigroup generated by A". Therefore, A is called the generator of T. In earlier literature, this was also called the "infinitesimal generator" or "infinitesimal operator."

Definition. The **generator** $A : \mathcal{D}(A) \subseteq X \to X$ of a strongly continuous semigroup $(T(t))_{t\geqslant 0}$ on a Banach space X is the operator

$$Af = \lim_{h\to 0} \frac{(T(h) - I)f}{h} = \lim_{h\to 0} \frac{T(h)f - f}{h}, \tag{5.11}$$

where f is in the domain of A, $\mathcal{D}(A)$ if and only if this limit exists. The domain $\mathcal{D}(A)$ is an

essential part of the definition of the generator A.

In order to use the generator to describe a semigroup, the domain of the generator must be carefully characterized, and multiple theorems address this type of characterization. The generator is not a bounded operator, thus much of the theory of bounded linear operators may not be applicable. Instead, we must use results about closed operators, and properties relating to closed and compact sets. Therefore, the domain of the generator is preferentially dense, and we find that a densely defined generator is the type that generates a strongly continuous semigroup. We would like to know when equivalence of the generators of two semigroups implies equivalence of the semigroups themselves, and additionally, when convergence of sequences of generators or resolvents imply convergence of the semigroup. This is because in the application of operator semigroup theory to Markov processes, strong convergence of sequences of semigroups can imply weak convergence of the processes. This exact method was how Kurtz was able to prove his result– that of convergence in distribution of a sequence of Markov processes to a limiting process.

In order to do this, however, in the setting of jump processes converging to a deterministic process, multiple technicalities had to be addressed, and multiple extensions of both operators and existing theorems had to be undertaken. The elegance of the approach comes from the recognition that both the Markov processes, and the deterministic process described by the system of differential equations, could be described by operator semigroup theory. Thus if the convergence of the semigroups could be described by means of convergence of the respective generators, the result would follow. One of the difficulties is that the spaces that the Markov processes are defined in, may not be the same spaces that the limiting process is defined in, and furthermore, that each process in the sequence may be defined in a different space. Thus we also must consider the convergence of sequences of spaces that contain the processes.

A brief overview of basic theorems about characterization of operator semigroups by their generators follows, including two versions of the famous Hille-Yosida generation theorems.

Theorem. The generator of a strongly continuous contraction semigroup is a closed and densely defined linear operator that determines the semigroup uniquely. $\square$

Corollary. For a strongly continuous semigroup $(T(t))_{t\geqslant 0}$ on a Banach space X with generator $(A, \mathcal{D}(A))$, the following are equivalent:

1) The generator is bounded.

2) The domain $\mathcal{D}(A)$ is all of X.

3) The domain $\mathcal{D}(A)$ is closed in X.

4) The semigroup $(T(t))_{t\geqslant 0}$ is uniformly continuous.

5) The semigroup is given by:

$$T(t) = e^{At} := \sum_{n=0}^{\infty} \frac{t^n A^n}{n!},\ t \geqslant 0. \quad \square \tag{5.12}$$

Of course, we can not expect the generator to be bounded, so we must consider alternative properties. Let A be a closed linear operator on a real Banach space L, for some real λ, $\lambda - A$ ($\equiv \lambda I - A$) is one-to-one, the range $\mathcal{R}(\lambda - A) = L$, and $(\lambda - A)^{-1}$ is a bounded linear operator on L. Then denote the resolvent set of A as $\rho(A)$, and the resolvent (at λ) of A as $R_\lambda(A)$. We then have the following

Proposition. Let $(T(t))_{t\geqslant 0}$ be a strongly continuous contraction semigroup on L with generator A. Then $(0, \infty) \subset \rho(A)$, and

$$R_\lambda g = \int_0^\infty e^{-\lambda t} T(t) g dt \tag{5.13}$$

for all $g \in L$ and $\lambda > 0$.

Theorem. Hille-Yosida Theorem, contraction case, version 1: For a linear operator A on a real Banach space L, the following two properties are equivalent:

1) A generates a strongly continuous contraction semigroup.

2) A is closed, densely defined, and for every $\lambda > 0$, one has $\lambda \in \rho(A)$, and

$||\lambda R_\lambda(A)|| \leqslant 1.$ $\square$

In order to use the Hille-Yosida theorem to decide if a given operator is the generator of a strongly continuous contraction semigroup, it is necessary to estimate the operator norm of the resolvent. This may not always be possible. An alternate version of the Hille-Yosida theorem allows for the characterization of the generator without explicit knowledge of the resolvent. In order to state this version, we need another operator definition:

Definition. A linear operator A on L is said to be **dissipative** if $||\lambda f - Af|| \geqslant \lambda||f||$ for every $f \in \mathcal{D}(A)$ and $\lambda > 0$.

First we state the following useful

Lemma. Let A be a dissipative operator on L and let $\lambda > 0$. Then A is closed if and only if the range of $(\lambda - A)$ is closed. $\square$

Theorem. Hille-Yosida Theorem, contraction case, version 2: A linear operator A on a real Banach space L is the generator of a strongly continuous contraction semigroup on L if and only if:

1) $\mathcal{D}(A)$ is dense in L,

2) A is dissipative, and

3) The range, $\mathcal{R}(\lambda - A) = L$, for some $\lambda > 0$. $\square$

We can now state one of the first results relating relationships between generators to relationships between their corresponding semigroups:

Proposition. Let $(T(t))_{t\geqslant 0}$ and $(S(t))_{t\geqslant 0}$ be strongly contiuous contraction semigroups on L with generators A and B, respectively. If $A = B$, then $T(t) = S(t)$ for all $t \geqslant 0$. $\square$

An extension of the Hille-Yosida theorem is often used, which considers a closable A. Recall that if A is closable, then the closure of A is the minimal closed linear extension of A. First we have a lemma:

Lemma. Let A be a dissipative linear operator on L with $\mathcal{D}(A)$ dense in L. Then A is closable and $\overline{\mathcal{R}(\lambda - A)} = \mathcal{R}(\lambda - \overline{A})$ for every $\lambda > 0$. $\square$

Theorem. A linear operator A on L is closable and its closure $\overline{A}$ is the generator of a strongly continuous contraction semigroup on L if and only if:

1) $\mathcal{D}(A)$ is dense in L,

2) A is dissipative, and

3) $\mathcal{R}(\lambda - A)$ is dense in L, for some $\lambda > 0$. $\square$

Often the domain of the generator is difficult to determine. Instead, we may find that it is sufficient to characterize a subset of the space that serves as the domain of another operator, whose extension is the operator we are concerned with. For situations like this, we find that the definition of a **core** for a generator is useful.

Definition. Let A be a closed linear operator on L. A subspace D of $\mathcal{D}(A)$ is said to be a **core** for A if the closure of the restriction of A to D is equal to A.

Proposition. Let A be the generator of a strongly continuous contraction semigroup on L. Then a subspace D of $\mathcal{D}(A)$ is a core for A if and only if D is dense in L and $\mathcal{R}(\lambda - A|_D)$ is dense in L for some $\lambda > 0$. $\square$

For example, given a dissipative linear operator A with $\mathcal{D}(A)$ dense in L, we may want to show that $\overline{A}$ generates a strongly continuous contraction semigroup on L. The problem at hand is then to verify the range conditions. This type of approach is used by Kurtz in his theorems, and additionally, other extensions and "extended limits" are introduced. For instance, the dissipative operator initially considered may not be single valued. This situation arises in Kurtz's proof. For example, for $n = 1, 2, \ldots$, let L_n and L be Banach spaces, and let $\pi_n : L \to L_n$ be a bounded linear transformation. Assume that $\sup_n ||\pi_n|| < \infty$. If $A_n \subset L_n \times L_n$ is linear for each $n \geqslant 1$, the **extended limit** of the sequence $\{A_n\}$ is defined

by

$$\begin{aligned}&\text{ex}-\lim_{n\to\infty} A_n = \\ &\{(f,g)\in L\times L:\ \text{there exists} \\ &(f_n,g_n)\in A_n \text{ for each } n\geqslant 1 \text{ such that } ||f_n-\pi_n f||\to 0 \text{ and } ||g_n-\pi_n g||\to 0\}.\end{aligned} \tag{5.14}$$

Then $\text{ex}-\lim_{n\to\infty} A_n$ is necessarily closed in $L\times L$, but need not be single valued, even if each A_n is. This extended limit is essential for the precursor proofs to the Kurtz theorem from 1970, namely, his extensions of the Trotter approximation theorems discussed in the next section. Additional extensions are required to show measurability, because for jump processes, integrals must be computed in the Lebesgue sense, as functions are discontinuous. These are the types of considerations required for the use of semigroup methods in the context of stochastic processes.

5.3.1 Operator semigroups in stochastic processes

To characterize a generators of semigroup that correspond to stochastic processes, suitable regularity conditions must be established in order to use the Hille-Yosida theorems. The establishment of these regularity conditions can be an exceedingly complicated process, and here only the most basic case is described. For this case, we consider a probability kernel μ on a measurable space $(S,\mathcal{S})$, and introduce the associated transition operator given by

$$Tf(x) = (Tf)(x) = \int \mu(x,dy)f(y),\ x\in S, \tag{5.15}$$

where we assume $f: S\to\mathbb{R}$ to be measurable and either bounded or nonnegative. By approximation by simple functions, we find that Tf is then also measurable on S. Furthermore, T is a positive contraction operator. The identity operator, I, corresponds to the kernel $\mu(x,\cdot)\equiv\delta_x$, the distribution of the point mass at x. We see that if the probability kernels satisfy Chapman-Kolmogorov relation, the transition operators $(T(t))_{t\geqslant 0}$ have the

semigroup property. Now assuming that S is a locally compact, separable metric space, we introduce the class $C_0 = C_0(S)$ of continuous functions $f : S \to \mathbb{R}$ with $f(x) \to 0$ as $x \to \infty$. We make C_0 into a Banach space by using the sup norm, $||f|| = \sup_x |f(x)|$.

We now introduce the notion of a Feller semigroup, named after William Feller for his extensive development of this theory. The term was apparently not commonly used when Kurtz was writing his thesis, as he did not make use of it, nor did he in the 1970 version of the proof in *Journal of Applied Probability*. However, in his 1986 textbook with Ethier, the proof is explained using this terminology. The idea of a Feller semigroup and the associated results helps to organize and generalize convergence results concerning semigroup methods in stochastic processes.

Definition. A semigroup of positive contraction operators $(T(t))_{t\geqslant 0}$ on C_0 is called a **Feller semigroup** if it has the additional regularity properties

1) $T(t)C_0 \subset C_0,\ t \geqslant 0$,

2) $T(t)f(x) \to f(x)$ as $t \to 0$, for all $f \in C_0,\ x \in S$.

Property 1 and 2 above, together with the semigroup property, imply the required strong continuity property for use of the Hille-Yosida theory:

3) $T(t)f \to f$ as $t \to 0$, for all $f \in C_0$.

Furthemore we can prove the following results:

Theorem. Let $T(t)f(x) \to f(x)$ be a Feller semigroup on C_0 with resolvents $R_\lambda, \lambda > 0$. Then the operators λR_λ are injective contractions on C_0 such that $\lambda R_\lambda \to I$ strongly as $\lambda \to \infty$. Furthermore, the range $\mathcal{D} = R_\lambda C_0$ is independent of λ and dense in C_0, and there exists an operator A on C_0 with domain $\mathcal{D}$ such that $R_\lambda^{-1} = \lambda - A$ on $\mathcal{D}$ for every $\lambda > 0$. Finally, A commutes on $\mathcal{D}$ with every $T(t)$. $\square$

Lemma. A Feller semigroup is uniquely determined by its generator.

Proof:

The operator A determines $R_\lambda = (\lambda - A)^{-1}$ for all $\lambda > 0$. By the uniqueness of Laplace transforms, it then uniquely determines the measure $\mu(dt) = T(t)f(x)dt$ on $\mathbb{R}_+$ for any $f \in C_0$ and $x \in S$. Since the density $T(t)f(x)$ is right-continuous in t for fixed x, the assertion follows. $\square$

Theorem. Let $(T(t))_{t\geqslant 0}$ be a Feller semigroup with generator $(A, \mathcal{D}(A))$. Then $(T(t))_{t\geqslant 0}$ is strongly continuous and satisfies

$$T(t)f - f = \int_0^t T(s)Af ds,\ f \in \mathcal{D}(A),\ t \geqslant 0. \tag{5.16}$$

Furthermore, $T(t)f$ is differentiable at 0 if and only if $f \in \mathcal{D}(A)$, in which case

$$\frac{d}{dt}(T(t)f) = T(t)Af = AT(t)f,\ t \geqslant 0 \quad \square \tag{5.17}$$

To give an idea of the probabilistic applications of these conditions, we make a further simplification and assume that S is compact, and also that $(T(t))_{t\geqslant 0}$ is conservative, in other words, $T(t)\mathbf{1}_S = \mathbf{1}_S$ for all t. For every initial state x, let $X_x(t), t \geqslant 0$ be an associated Markov process with transition operators $T(t)$. The following properties hold, where the numbers refer to the regularity conditions 1-3 from the definition of a Feller semigroup above:

Proposition. Let $(T(t))_{t\geqslant 0}$ be a conservative transition semigroup on a compact metric space (S, ρ). Then

1) Feller property 1 holds if and only if $X_x(t) \Rightarrow X_y(t)$ as $x \to y$ for fixed $t \geqslant 0$.

2) Feller property 2 holds if and only if $X_x(t) \to x$ in probability as $t \to 0$, for fixed x.

3) Feller property 3 holds if and only if $\sup_x \mathrm{E}_x[\rho(X(s), X(t)) \wedge 1] \to 0$ as $s - t \to 0$. $\square$

As a simple example of a generator for a Markov jump process, consider the compound Poisson process. The compound Poisson process is created as follows:

Let $X_n,\ n \in \mathbb{N}$ be a sequence of i.i.d. random variables with distribution function $F(x)$ and let $N(t)$ be a Poisson process with parameter λ. Let $S_n = X_1 + \cdots + X_n$ be the n^{th} partial sum over the X_i. Then the compound Poisson process is defined by

$$Y(t) := \sum_{n \geqslant 1} S_n \mathbf{1}_{\{N(t)=n\}} \tag{5.18}$$

We assume f belongs to the space $C_b(\mathbb{R})$ of bounded continuous functions on $\mathbb{R}$, and define the transition operator $T(t)$ of Y by $\mathrm{E}_x[Y(t)]$. Also define an operator L by

$$Lf(x) = \mathrm{E}[f(x + X_1)] = \int_{\mathbb{R}} f(x + y)F(dy), \tag{5.19}$$

in order to make calculations easier. Note that $L^n f(x) = \mathrm{E}[f(x + X_1 + \cdots + X_n)] = \mathrm{E}[f(x + S_n)]$. Now we have that

$$\begin{aligned}(T(t)f)(x) = \mathrm{E}_x[f(Y(t)] = \sum_{n \geqslant 0} \mathrm{E}[f(x + S_n)]\mathbb{P}(N(t) = n) = \sum_{n \geqslant 0} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\mathrm{E}[f(x + S_n)] \\ \sum_{n \geqslant 0} e^{-\lambda t}\frac{(\lambda t)^n}{n!}L^n f(x) = (e^{\lambda t(L-I)}f)(x),\end{aligned} \tag{5.20}$$

thus we can explicitly see that the transition semigroup can be expressed as e^{At}, where A is given by

$$Af(x) = \lambda(L - I)f(x) = \lambda \int_{\mathbb{R}} (f(x + y) - f(x))F(dy). \tag{5.21}$$

5.4 The Kurtz theorem

The proof of the convergence in distribution of a sequence of Markov jump processes to the solution of a deterministic differential equation modeling the same physical process, in the thermodynamic limit, as was done by Kurtz in 1967 and 1970, requires the use of several

precursory results on the convergence of semigroups. These convergence results were based on the establishment of sufficient conditions under which convergence of the generators of the semigroups implied convergence of the semigroups themselves. The two main results were proved by Trotter in 1958 [49], and Kurtz in 1969 [33].

5.4.1 Precursory Theorems

Theorems proved by H.F. Trotter and published in 1958 [49] established conditions for convergence of sequences of semigroups in terms of both the generators and the resolvents, in the context of sequences of Banach spaces each containing an element of a sequence of semigroups. The idea of the convergence of a sequence of Banach spaces X_n to a limiting space X is introduced, in the following sense: given a sequence of Banach spaces X_n, together with a sequence of linear maps P_n with $||P_n|| \leqslant 1$, mapping the limiting space onto each element of the sequence of spaces $(P_n : X \to X_n)$, we have that $\lim_{n\to\infty} ||P_n f|| = ||f||$ for every $f \in X$. The maps P_n become isomorphisms, in the limit, in a certain way that is made precise in Trotter's paper [49]. Trotter introduces new definitions which extend ideas about uniformity and convergence and allow him to prove convergence of random walks to one-dimensional diffusion processes using these methods.

Kurtz further extended Trotter's theorems to be applicable to general Markov processes, by using the extended limit described above, and furthermore, replacing the need for a uniform type of convergence with a bounded, pointwise convergence, among other additions [33]. This required the use of associated extensions of the generator. The theorem was further extended and revised by Sova and Mackevičius, and can be presented as a cumulated combination of results called the "Trotter, Sova, Kurtz, Mackevičius" theorem [23]. The Kurtz version as presented in the Ethier and Kurtz 1986 text is given here [11]:

Theorem 1. Let L and $L_n, n = 1, 2, \ldots$ be Banach spaces with norms all denoted by $||\cdot||$, and $\pi_n : L \to L$ be a bounded linear transformation. Assume also that $\sup_n ||\pi_n|| < \infty$. Let the notation $f_n \to f$ indicate that $f_n \in L_n$ for each $n \geqslant 1$, $f \in L$, and $\lim_{n\to\infty} ||f_n - \pi_n f|| = 0$.

For $n = 1, 2, \ldots$, let $(T_n(t))_{t\geqslant 0}$ and $(T(t))_{t\geqslant 0}$ be strongly continuous contraction semigroups on L_n and L, respectively, with generators A_n and A. Let D be a core for A. Then the following are equivalent:

1) For each $f \in L,\ T_n(t)\pi_n f \to T(t)f$ for all $t \geqslant 0$, uniformly on bounded intervals.

2) For each $f \in L,\ T_n(t)\pi_n f \to T(t)f$ for all $t \geqslant 0$.

3) For each $f \in D$, there exists $f_n \in \mathcal{D}(A_n)$ for each $n \geqslant 1$, such that $f_n \to f$ and $A_n f_n \to Af$, i.e., $\{(f, Af) : f \in D\} \subset \text{ex} - \lim_{n\to\infty} A_n$. $\square$

5.4.2 Original presentations of the theorem and later version

In the original versions of the Kurtz theorem, various terms and results that were created, proved and solidified later did not exist. The thesis version of the proof required that each step of the multiple extensions and approximations required to achieve the results were fleshed out individually, which made for a more cumbersome and lengthy exposition. In the 1986 textbook version the use of the Feller semigroup terminology and associated results is found [11].

The steps of the theorem are summarized here, based on the later version [11], together with a simple example illustrating the method. The basis of the proof requires the definition of a **density dependent family** of processes:

Definition. Let β_l be a collection of nonnegative functions with $l \in \mathbb{Z}^d$, defined on a subset $E \subset \mathbb{R}^d$. Set $E_n = E \cap \{n^{-1}k : k \in \mathbb{Z}^d\}$, and assume that if $x \in E_n$ and $\beta_l(x) > 0$ then $x + n^{-1}l \in E_n$. A **density dependent family** of processes corresponding to the β_l is a sequence $\{X_n\}$ of jump Markov processes such that X_n has state space E_n and transition rates

$$q^{(n)}_{x,y} = n\beta_{n(x-y)}(x),\ x, y \in E_n. \tag{5.22}$$

For the simple example, we begin by defining a sequence of processes as follows:

For $n \geqslant 1$, define

$$\begin{aligned} \lambda_n(x) &= 1 + 3x\left(x - \frac{1}{n}\right), \\ \mu_n(x) &= 3x + x\left(x - \frac{1}{n}\right)\left(x - \frac{2}{n}\right), \end{aligned} \tag{5.23}$$

and let $\{Y_n\}$ be a sequence of birth-death processes in $\mathbb{Z}_+$, with transition probabilities satisfying

$$\begin{aligned} \mathbb{P}\{Y_n(t+h) = j+1 | Y_n(t) = j\} &= n\lambda_n\left(\frac{j}{n}\right)h + o(h) \\ \mathbb{P}\{Y_n(t+h) = j-1 | Y_n(t) = j\} &= n\mu_n\left(\frac{j}{n}\right)h + o(h) \end{aligned} \tag{5.24}$$

This process represents the number of molecules of a chemical species R in a volume n undergoing the following chemical reactions:

$$A \underset{3}{\overset{1}{\rightleftharpoons}} R, \quad B + 2R \underset{1}{\overset{3}{\rightleftharpoons}} 3R \tag{5.25}$$

The deterministic model corresponding to this process, based on the law of mass-action, would be $\frac{d[R]}{dt} = [R] + 3[R]^2 - [R]^3 - 3[R]$. We rescale and renormalize for ease of calculation, setting

$$X_n(t) = n^{\frac{1}{4}}(n^{-1}Y_n(n^{\frac{1}{2}}t) - 1), \ t \geqslant 0. \tag{5.26}$$

Let $E_n = \{n^{\frac{1}{4}}(n^{-1}y - 1) : y \in \mathbb{Z}_+\}$, and note that

$$T_n(t)f(x) \equiv \mathrm{E}[f(X_n(t))|X_n(0) = x] \tag{5.27}$$

defines a semigroup $(T_n(t))_{t\geqslant 0}$ on $\mathcal{B}(E_n)$ (the Borel sets of E_n) with generator

$$G_n f(x) = n^{\frac{3}{2}}\lambda_n(1 + n^{-\frac{1}{4}}x)[f(x + n^{-\frac{3}{4}}) - f(x)] + n^{\frac{3}{2}}\mu_n(1 + n^{-\frac{1}{4}}x)[f(x - n^{-\frac{3}{4}}) - f(x)] \tag{5.28}$$

Now let X be another process with rates $\lambda(x) = 1 + 3x^2$, and $\mu(x) = 3x + x^3$, and generator $Gf(x) = 4f''(x) - x^3 f'(x)$. We can show using a Taylor expansion of $G_n f(x)$ that for

appropriate f,

$$\lim_{n\to\infty} \sup_{x\in E_n} |G_n f(x) - Gf(x)| = 0. \tag{5.29}$$

Using theorems such as those described above, it can be shown that

$$A \equiv \{(f, Gf) : f \in C[-\infty, \infty] \cap C^2(\mathbb{R}),\ Gf \in C[-\infty, \infty]\} \tag{5.30}$$

is the generator of a Feller semigroup $(T(t))_{t\geqslant 0}$ on $C[-\infty, \infty]$. By use of some additional theorems, it can be shown that there is a diffusion process X corresponding to $(T(t))_{t\geqslant 0}$. To prove that $X_n \Rightarrow X$, it suffices to show that equation 5.29 holds for all f in a core D for the generator A [11]. This diffusion process then can be shown to converge in the same way to the deterministic process, as long as the initial conditions also converge. Note the similarity between the expression for the rates of the diffusion and the differential equation of the deterministic version above. For the proof of the general version, as summarized in chapter 3, a number of proofs relating to the specific forms of the functions β_l are required as well.

Chapter 6

Modern Applications and Conclusion

We have seen that, from more basic mathematical beginnings such as those used to show convergence in the mean of the simple unimolecular reaction to the deterministic solution without the use of limits, the theory can explode in complexity. For a convergence proof such as the one given by Kurtz to show convergence in the thermodynamic limit, quite advanced methods were used. Of course, Kurtz's aim was not to prove applicability of the use of stochastic models in chemistry *per se*, as his theorems have much broader consequences and apply to much more general aspects of the pure theory than for this singular purpose. However, even the van Kampen expansion is quite advanced, and we see that when there is a limit involved such that the state space becomes infinite, extreme difficulties can arise. For the semigroup approach, this is of course because a finite transition matrix becomes an infinite operator, and therefore representations such as the exponent and the derivative at zero become difficult to justify without extreme care.

On the other hand, we saw that even for some bimolecular reactions, solving for the transition function itself can be daunting, if not intractable. Therefore, the alternate description of a process via its semigroup or the generator of the semigroup can often offer another avenue. The connections found between jump Markov processes, diffusions, and deterministic processes as viewed through the lens of their associated semigroups offers a deep and

cohesive way to understand dynamical systems of all types from a similar standpoint.

A revival in the birth-death type description in the sciences has occurred over the past two decades. It has been found that gene expression in cells seems to contain a random component [50]. Although, according to Gillespie's rigorous derivation of the master equation, this type of simple birth-death model is only expected to be applicable in an ideal-gas, well-stirred system at equilibrium, it seems that these same basic, combinatorial stochastic models offer remarkably appropriate predictions of gene expression. This is especially surprising due to the fact that the reactions between enzymes in the cell occur between complex three-dimensional surfaces and involve numerous co-factors. Additionally, the crowded solution-state environment is very far from gas phase. Models are computationally created using the Gillespie algorithm or one of its subsequent variants, and often the rates that are used are from kinetic experiments and based on deterministic models. An interesting question is whether the success of these types of models, for such systems that are theoretically completely inappropriate, results from a cancelation of errors, or if there is another way to justify the approximation in these situations. For instance, the use of these types of models to describe population dynamics (for humans or animals, etc.) has not been rigorously validated, based on first principles and atomic details, the way that the physics community required of the chemical master equation. However, they have nevertheless proved to be useful models and can make predictions that can, in some situations, be more accurate than the deterministic models provide. We may have that, for processes in biochemistry such as complex enzymatic pathways in the cell, we are already at such a macroscopic scale that we are no longer modeling a chemical reaction with the birth-death model, but a biological population much like the canonical foxes and rabbits models do. Thus the atom-based, rigorous derivation is not required or appropriate.

It should be noted that the Kurtz theorem (in its 1970 version) was originally intended to show correspondence between the stochastic population models and the deterministic ones, as well as the application in chemistry. For the deterministic population model, there is

no validation from first principles as there is in chemical physics. However, we still have convergence of the two. It is possible that something in our approach to creating a model of a dynamical system is captured by the operator semigroup description, and that the same underlying ideas expressed by both models are also expressed mathematically by the transition semigroups and the associated generators for both the stochastic processes and the deterministic version.

Bibliography

[1] D.F. Anderson, *Introduction to stochastic processes with applications in the Biosciences*, freely available at: `www.math.wisc.edu/~anderson/605F13/Notes/StochBio.pdf`, 2013.

[2] Anthony Bartholomay, *On the linear birth and death processes of biology as Markoff chains*, Bulletin of Mathematical Biophysics **20** (1958), 97–118.

[3] ———, *Stochastic models for chemical reactions I: Theory of the unimolecular reaction process*, Bulletin of Mathematical Biophysics **20** (1958), 175–190.

[4] ———, *Stochastic models for chemical reactions II: The unimolecular rate constant*, Bulletin of Mathematical Biophysics **21** (1959), 363–373.

[5] A.T. Barucha-Reid, *Elements of the theory of Markov processes and their applications*, McGraw-Hill Book Company, Inc., New York, 1960.

[6] Adam Bobrowski, *Functional analysis for probability and stochastic processes*, Cambridge University Press, New York, 2005.

[7] Laxmangudi Krishnamurthy Doraiswamy and Balkrishna Dattatraya Kulkarni, *The analysis of chemically reacting systems: a stochastic approach*, Vol. 4, Taylor & Francis, New York, 1987.

[8] Rick Durrett, *Essentials of stochastic processes*, Springer-Verlag, New York, 1999.

[9] Klaus-Jochen Engel and Rainer Nagel, *One-parameter semigroups for linear evolution equations*, Springer-Verlag, New York, 2000.

[10] Péter Érdi and János Tóth, *Mathematical models of chemical reactions: Theory and applications of deterministic and stochastic models*, Princeton University Press, Princeton, New Jersey, 1989.

[11] Stewart N. Ethier and Thomas G. Kurtz, *Markov processes: Characterization and convergence*, Wiley, Hoboken, New Jersey, 1986.

[12] W. Feller, *On the theory of stochastic processes, with particular reference to applications* (Jerzey Neyman, ed.), Proceedings of the first Berkeley symposium on mathematical statistics and probability, August 13-18, 1945 and January 27-29, 1946, University of California Press, Berkeley, Calif., 1949.

[13] William Feller, *An introduction to probability theory and its applications, vol. 1*, John Wiley and Sons, New York, 1957.

[14] W.H. Furry, *On fluctuation phenomena in the passage of high energy electrons through lead*, Physical Review **52** (1937), 569–581.

[15] Crispin Gardiner, *Stochastic methods*, Springer, Berlin, 2009.

[16] Daniel T. Gillespie, *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*, Journal of Computational Physics **22** (1976), 403–434.

[17] ———, *Markov processes: An introduction for scientists and engineers*, Academic Press, Boston, 1992.

[18] ———, *A rigorous derivation of the chemical master equation*, Physica A **188** (1992), 404–425.

[19] B.V. Gnedenko, *The theory of probability. B.D. Seckler, trans.*, Chelsea Publishing, New York, 1967.

[20] Narendra Goel and Nira Richter-Dyn, *Stochastic models in biology*, Academic Press, New York, 1974.

[21] Jerome A. Goldstein, *Semigroups of linear operators*, Oxford University Press, New York, 1985.

[22] Robert M. Gray, *Probability, random processes, and ergodic properties*, Springer, New York, 2009.

[23] Olav Kallenburg, *Foundations of moddern probability*, Springer, New York, 2002.

[24] Samuel Karlin and Howard M. Taylor, *A first course in stochastic processes*, Academic Press, San Diego, 1975.

[25] ———, *A second course in stochastic processes*, Academic Press, Inc., New York, 1981.

[26] Joel Keizer, *Examination of the stochastic process underlying a simple isomerization reaction*, Journal of Chemical Physics **56** (1972), no. 12, 5775–5783.

[27] ———, *Concentration fluctuations in chemical reactions*, Journal of Chemical Physics **63** (1975), no. 11, 5037–5043.

[28] ———, *A simple proof of Kurtz's theorem, In: Probability, statistical mechanics, and number theory: a volume dedicated to Mark Kac* (Gian-Carlo Rota, ed.), Academic Press, New York, 1986, 1–23.

[29] Martin Keller-Ressel, *An intuitive introduction to operator semigroups*, freely available at: www.math.tu-dresden.de, 2006.

[30] Erwin Kreyzig, *Introduction to functional analysis with applications*, John Wiley and Sons, Inc., New York, 1978.

[31] T.G. Kurtz, *Convergence of operator semigroups with applications to Markov processes*, PhD Dissertation, Stanford University, 1967.

[32] ———, *Approximation of stochastic processes. In L. Arnold and R. Lefever, eds., Stochastic nonlinear systems*, Springer-Verlag, Berlin (1981), 22–35.

[33] Thomas G. Kurtz, *Extensions of trotter's operator semigroup approximation theorems*, Journal of Functional Analysis **3** (1969), 354–375.

[34] ———, *Solutions of ordinary differential equations as limits of pure jump Markov processes*, Journal of Applied Probability **7** (1970), no. 1, 49–58.

[35] ———, *Limit theorems for sequences of pure jump Markov processes approximating ordinary differential processes*, Journal of Applied Probability **8** (1971), no. 2, 344–356.

[36] ———, *The relationship between stochastic and deterministic models for chemical reactions*, The Journal of Chemical Physics **57** (1972), no. 7, 2976–2978.

[37] Gregory F. Lawler, *Introduction to stochastic processes*, Taylor & Francis, Boca Raton, FL, 2006.

[38] Donald McQuarrie, *Kinetics of small systems I.*, The Journal of Chemical Physics **38** (1963), no. 2, 433–436.

[39] ———, *Stochastic approach to chemical kinetics*, Journal of Applied Probability **4** (1967), no. 3, 413–478.

[40] Donald McQuarrie, C.J. Jachimowski, and E. Russell, *Kinetics of small systems II.*, The Journal of Chemical Physics **40** (1964), no. 10, 2914–2921.

[41] J.R. Norris, *Markov chains*, Cambridge University Press, New York, 1997.

[42] I. Oppenheim, K.E. Shuler, and G.H. Weiss, *Stochastic and deterministic formulation of chemical rate equations*, The Journal of Chemical Physics **50** (1969), no. 1, 460–466.

[43] Jeffrey S. Rosenthal, *A first look at rigorous probability theory*, World Scientific, Hackensack, New Jersey, 2008.

[44] Sheldon Ross, *Introduction to probability models*, Academic Press, Boston, 2003.

[45] K. Singer, *Application of the theory of stochastic processes to the study of irreproducible chemical reactions and nucleation processes*, Journal of the Royal Statistical Society. Series B (Methodological) **15** (1953), no. 1, 92–106.

[46] A.V. Skorokhod, *Lectures on the theory of stochastic processes*, VSP Science Press, Utrecht, The Netherlands, 1996.

[47] Jeffrey I. Steinfeld, Joseph S. Francisco, and L. Hase William, *Chemical kinetics and dynamics*, Prentice Hall, Upper Saddle River, New Jersey, 1989.

[48] Daniel F. Styer, *What good is the thermodynamic limit?*, American Journal of Physics **72** (2004), no. 1, 25–29.

[49] H.F. Trotter, *Approximation of semigroups of operators*, Pacific Journal of Mathematics **8** (1958), no. 4, 887–919.

[50] Lev S. Tsimring, *Noise in biology*, Reports on Progress in Physics **77** (2014), 026601.

[51] N.G. van Kampen, *Stochastic processes in physics and chemistry*, Elsevier, Boston, 2007.

[52] S.R.S. Varadhan, *Probability theory: Courant lecture notes 7*, American Mathematical Society, Providence, Rhode Island, 2001.

[53] ———, *Stochastic processes: Courant lecture notes 16*, American Mathematical Society, Providence, Rhode Island, 2007.

[54] Jan Vrbik and Paul Vrbik, *Informal introduction to stochastic processes with Maple*, Springer, New York, 2013.

[55] Kosaku Yosida, *Functional analysis*, Springer-Verlag, New York, 1980.

[56] G.U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis*, Philos. Trans. Roy. Soc. London Ser. B **213** (1924), 21–87.