Stochastic Processes

Jacob Denson

October 2, 2017

Table Of Contents

Chapter 1

Stochastic Processes

The theory of dynamical systems allows us to determine the motions of objects under deterministic actions. In Newton's mechanics, past and future events can be predicted exactly from the position and velocity of all objects at a particular point in time. In reality, one can never measure the data required to determine the state of a system in precision. Inexactness shrouds the determinism of a system, which invalidates the application of Newton's model. Stochastic processes are the probabilistic variant of dynamical systems. Rather than a deterministic rule determining the evolution of a state over time, a stochastic rule is employed leading to a randomized state over time. Formally, a **stochastic process** is a collection $\{X_t\}$ of random variables defined over the same probability space Ω, with range in the same **state space** S, indexed over some linearly ordered set T.

Example. *To model the uncertainty of weather, we may take a stochastic process with state space $S = \{\textbf{\textit{sunny}}, \textbf{\textit{rainy}}\}$. For $i \in \mathbf{Z}$, we may model the weather by a random variable $X_i : \Omega \to \mathcal{S}$, modelling the weather on a certain day i. Then $\{X_i : i \in \mathbf{Z}\}$ is a stochastic process.*

Example. *To model how the value of stocks change over time, we take S to be the real numbers, and let X_i be the value of a certain stock at time i, for $i \in \mathbf{R}$. This is a continuous time random process, because the values are indexed over time, and the states are also continuous. We will study a generalization of this process, Brownian motion, in the sequel.*

Example. *To estimate the cumulative density function of an independant and*

identically distributed sample $X_1, \ldots, X_n \sim F$. we can take the estimate

$$\hat{F}(t) = \frac{\sum \mathbf{I}[X_i \leqslant t]}{n}$$

For a fixed $t \in \mathbf{R}$, $\hat{F}(t)$ is a random variable, and considering t as the time variable lets us view $\hat{F}$ as a stochastic process.

Every problem in probability theory involving collections of random variables can be formulated as a statement about stochastic processes. The right application of the theory of stochastic processes may shed a different light to a problem, giving an intuitive perspective to the problem. On the other hand, we can't say much about stochastic processses in general, because of how widely they can be applied. The fun of stochastic processes results when we add additional relationships between the random variables, and study the resultant properties.

In order to study the relations between a stochastic process $\{X_t\}$ at different time points, it makes sense to consider the **marginal distributions** of the random variables. For infinitely many random variables, the corresponding marginal distribution is very difficult to study, so we often focus on the marginal distribution given by a finite subset of time points $t_1 \ldots, t_n$ of the process. On the other hand, we often want to generate a stochastic process from the finite dimensional marginal distributions. In the discrete setting, this is often easy to explicitly construct, but in the continuous setting the construction does not seem so easy. The Kolmogorov theorem tells us that this is a valid method of constructing a process.

To introduce this theorem, we introduce some temporary notation. Consider a state space S, which forms a subset of the real numbers, and some index set T. Suppose that for each finite subset $R \subset T$ we have determined a probability distribution $\mathbf{P}_R$ over the borel σ-algebra of S^R. If $K \subset R \subset T$, then we have a projection map $\pi_{R \to K} : S^R \to S^K$, and we say that the family of probability distributions chosen over the index sets are **consistant** if the projection maps are all measure preserving, in the sense that $\mathbf{P}_K(A) = \mathbf{P}_R(\pi^{-1}(A))$ for all Borel measurable A. If we want to construct a stochastic process whose finite dimensional marginal distributions are given by the $\mathbf{P}_K$, consistancy is a necessary requirement, but Kolmogorov's theorem shows that this condition is also sufficient.

Theorem 1.1 (Kolmogorov's extension theorem)**.** *For any consistant family of distributions, there exists a stochastic process whose finite dimensional marginal distributions agree with the distribution family.*

Proof. The proof uses the Hahn-Kolmogorov / Carathëodory extension theorem to construct a probability measure on $\mathcal{S}^T$, which can then be taken as the sample space of our random variables $X_i = \pi_i$. We leave the technical details to the reader. The proof should extend to any Polish (separable and completely metrizable) space, but this is not needed here. The random variables specified are not unique. We call any other solution a **version** of the same stochastic process. □

The Kolmogorov theorem is used to construct measures, most importantly when T is uncountable. To gain intition, we will begin studying discrete time processes, for which most paradoxes is unavoidable. When $T = \mathbf{N}$, we need only specify consistant distributions on initial segments $\{0, 1, \ldots, K\}$.

Chapter 2

Finite Markov Chains

By the beginning of the 20th century, the work of the Poisson, Chebyshev, and the Bernoulli brothers had cemented the law of large numbers in mathematical culture. Given a number of independent and identically distributed random variables, well behaved asymptotic behaviour of the mean is guaranteed. It took the genius of Markov to realize that one can derive similar results for random variables which are not independent, nor distributed identically, but follow well behaved rules that exhibit asymptotic behaviour in the long run.

Markov had a stong and abrasive relationship with his colleagues. This extended beyond his professional life to the revolutionary atmosphere of 20th century Russia. When Leo Tolstoy was excommunicated from the Orthodox church, Markov requested that he too be excommunicated in solidarity. Markov's acrimony was most strongly directed towards his mathematical rival, Pavel Nekrasov, who had attempted to apply probability theory (rather loosely) to philosophical arguments. Nekrasov compared acts of free will to independent events. Since crime statistics obey the law of large numbers, this data should imply that human decisions are independent events – ergo, human free-will exists. What Nekrasov had assumed was that the law of large numbers only applies to independant events. Nekrasov had not commited an isolated mistake in applying this principle – mathematicians back to the Bernoullis had made the mistake. Markov's vitriol towards Nekrasov gave him the motivation to disprove this principle. He introduced Markov chains, families of dependant random events which still have a well defined law of large numbers.

Let $X_1, X_2, \ldots$ be a discrete time stochastic process, with a discrete, at

most countable state space. This process satisfies the **discrete Markov property** if, for any n, and for any states $x_1, \ldots, x_n, x_{n+1}$,

$$\mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \mathbf{P}(X_{n+1} = x_{n+1} | X_n = x_n)$$

A **Markov chain** is a stochastic process satisying the Markov property. In the theory of Newtonian mechanics, if we know the position and velocity of a particle at any single point in time, we can predict all past and future motion. The Markov property is a stochastic equivalent to this. We might not predict the future from the present, but we can gain as much information as possible from the present about the future, and we don't need to worry about the past.

Example. *All independent families of random variables $\{X_t\}$ satisfy the Markov property, since we cannot learn anything from previous results,*

$$\begin{aligned}\mathbf{P}(X_{t_{n+1}} = y | X_{t_n} = x_n, \ldots, X_{t_1} = x_1) &= \mathbf{P}(X_{t_{n+1}} = y) \\ &= \mathbf{P}(X_{t_{n+1}} = y | X_{t_n} = x_n)\end{aligned}$$

Independant processes are the least interesting example of a markov process.

Example. *If $\{X_i\}_{i \in \mathbf{Z}}$ is any stochastic process, we can create a Markov chain by 'memorizing' previous states of the system. We define $Y_k = (X_0, \ldots, X_k)$. Then one may verify that*

$$\begin{aligned}&\mathbf{P}(Y_{n+1} = (x_{n+1}, \ldots, x_0) | Y_n = (x_0, \ldots, x_n), Y_{n-1} = (x_0, \ldots, x_{n-1}), \ldots, Y_0 = x_0) \\ &\quad = \mathbf{P}(Y_{n+1} = (x_{n+1}, \ldots, x_0) | Y_n = (x_0, \ldots, x_n))\end{aligned}$$

This shows that $\{Y_k\}$ satisfies the Markov property, so one can always keep a copy of the past in the present so that we don't need to 'look back' to remember what happened.

For any three random variables X, Y, Z mapping into a discrete state space, we find

$$\mathbf{P}(X = x | Z = z) = \sum_y \mathbf{P}(X = x | Y = y, Z = z) \mathbf{P}(Y = y | Z = z)$$

If $i < j < k$, then in a Markov chain we may write

$$\mathbf{P}(X_k = x | X_i = z) = \sum_y \mathbf{P}(X_k = x | X_j = y) \mathbf{P}(X_j = y | X_i = z)$$

This is the **Chapman-Kolmogorov equation**, relating various transition probabilities of a markov chain. If we know $\mu_0(x) = \mathbf{P}(X_0 = x)$ and transition probability functions $p_k(x,y) = \mathbf{P}(X_{k+1} = y | X_k = x)$, then it is possible to calculate the probability distribution of X_n for every n. Conversely, given some μ_0 and p_k, we can always find a Markov chain $X_0, X_1, \ldots$ with these functions at the initial distribution and transition function (We can just consider a sample space $S^{\mathbf{N}}$ where $X_i(x) = x_i$ and such that

$$\mathbf{P}(\varnothing) = 0 \qquad \mathbf{P}(x_0 \times S^{\mathbf{N}-\{0\}}) = \mu_0(x_0)$$

$$\mathbf{P}(x_0, \ldots, x_n \times S^{\mathbf{N}-[n]}) = \mathbf{P}(x_0, \ldots, x_n) P_{n-1}(x_{n-1}, x_n)$$

Then $\mathbf{P}$ is a probability measure on $2^{[n]} \times S^{\mathbf{N}-[n]}$ for each integer n, assuming that μ_0 is a probability measure, and $p_n(x, \cdot)$ is a probability measure for each state x. Then $\mathbf{P}$ is defined on a ring of sets, since the family is certainly closed under a pairwise intersection, and

$$A \times S^{\mathbf{N}-[n]} - B \times S^{\mathbf{N}-[n]} = (A - B) \times S^{\mathbf{N}-[n]}$$

and $\mathbf{P}$ certainly satisfies countable additivity, so the Caratheódory extension theorem guarantees that $\mathbf{P}$ extends uniquely to a measure on the σ algebra generated by the subsets in question. The random variables are obviously measurable, and it is easy to verify the Markov property.

The nicest theory of Markov chains occurs when we assume the chain is 'time homogenous'. A Markov chain is **time homogenous** if we can specify the transition probabilities such that $p(x,y) = p_n(x,y)$ does not depend on n. We shall find that the best way to understand time homogenous chains is to vary the initial probability distribution μ_0 and studying how the chain varies. The main mechanism to this analysis is to view the transition probabilities as an operator on the space of all initial distributions (a convex subset of the Banach space $l_1(S)$ of summable functions on S). Studying the distributions of time-homogenous chains on a finite state space reduces to operator theory, and in the finite dimensional case, matrix algebra.

Let us define the transition operator P by the formula

$$(\mu P)(y) = \sum \mu(x) p(x,y)$$

Thus P takes a probability distribution over states to the probabilities of states one step into the future. In general, this means that μP^n gives the

probability distribution n steps into the future (this is formally verified by the Chapman-Kolmogorov equations). If the state space is finite, then μ can be viewed as a row vector, and then P as a finite dimensional matrix with $P_{xy} = p(x,y)$. Then μP can be literally interpreted as matrix multiplication. P is an example of a **stochastic matrix**, a matrix whose rows sum to one. Any such matrix with these rows specifies the transition probabilities of a time-homogenous Markov chain.

The space of probability distributions can be viewed in some way as functionals on the vector space $\mathbf{R}^S$ of real functions on S. Given a distribution μ and function f, we can define $\mathbf{E}_\mu(f) = \sum \mu(x) f(x)$, which is the expected value of f one step into the future given that we start at the initial distribution μ. In particular, we let $\mathbf{E}_x$ denote the expectation with respect to the initial distribution concentrated at x with probability one. Since P acts on the right in the family of probability distributions, we should have a natural operator on the family of functions on S, with

$$(Pf)(x) = \sum_y P(x,y) f(y) = \mathbf{E}[f(X_{n+1}) | X_n = x]$$

Given a function f, the formal calculation

$$\mathbf{E}_{\mu P}(f) = \sum (\mu P)(x) f(x) = \sum \mu(x) P(x,y) f(y) = \mathbf{E}_\mu(Pf)$$

verifies that P really does act like a dual operator.

Example. *It is a useful simplification to assume that the transition between states of weather from one day to the next is time-homogenous. After collecting data in a particular region, we might choose a transition matrix like the one below*

$$\begin{array}{c} \\ \textbf{sunny} \\ \textbf{rainy} \end{array} \begin{array}{c} \begin{array}{cc} \textbf{sunny} & \textbf{rainy} \end{array} \\ \begin{bmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{bmatrix} \end{array}$$

Thus there is a 60% chance of it being rainy the day after it is sunny, and an 80% change of it being sunny the day after it is rainy. We will find that, in the long run, the days will be sunny about 57% of the time, and rainy 43% of the time.

Example. *Consider a queueing system (for a phone-hold system, etc.) which can only hold 2 people at once. Every time epoch, there is a certain chance p*

that a new caller will attempt to access the system, and a chance q that we will finish with a person in the queue. Assuming these events are independent, we can model this as a time homogenous markov process with transition matrix

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{bmatrix} 1-p & p & 0 \\ (1-p)q & (1-q)(1-p)+pq & p(1-q) \\ 0 & q(1-p) & (1-q)+pq \end{bmatrix} \end{array}$$

Given a large amount of time, it is of interest to the maker of the queing system to know the average number of people in the queue at a certain time. This leads to the study of asymptotics of Markov chains, of which we will soon find a complete characterization.

Example. *Consider a random walk on a graph. This means that at each vertex, we have an equal chance of moving from one vertex to any other vertex connected by an edge. The simplest example of such a process is the random walk on the vertices $\{0,1,\ldots,n\}$, where each integer is connected to adjacent integers. The transition probabilities are given by*

$$P(i,i+1) = P(i,i-1) = \frac{1}{2} \qquad i \in \{1,\ldots,n-1\}$$

$$P(0,1) = P(n,n-1) = 1$$

If one connects the end vertices to themselves, then one obtains another form of the random walk. The former is known as the reflecting random walk, and the latter the partially reflecting.

2.1 Asymptotics of Markov chains

As was Markov's goal, we want to determine the asymptotic behaviour of a Markov chain $\{X_i\}$ after large lengths of time. In most cases, we will show the chains X_i converge in distribution, or at least that the averages $n^{-1}(X_1 + \cdots + X_n)$ convege in distribution.

Example. *Consider a homogenous process with the transition matrix*

$$P = \begin{pmatrix} 3/4 & 1/4 \\ 1/6 & 5/6 \end{pmatrix}$$

We may write $P = QDQ^{-1}$, where

$$Q = \frac{1}{2}\begin{pmatrix} 2 & -3 \\ 2 & 2 \end{pmatrix} \quad D = \begin{pmatrix} 1 & 0 \\ 0 & 7/12 \end{pmatrix} \quad Q^{-1} = \frac{1}{5}\begin{pmatrix} 2 & 3 \\ -2 & 2 \end{pmatrix}$$

Hence,

$$\begin{aligned} \lim_{n\to\infty} P^n &= \lim_{n\to\infty} (QDQ^{-1})^n = Q(\lim_{n\to\infty} D^n)Q^{-1} \\ &= \frac{1}{10}\begin{pmatrix} 2 & -3 \\ 2 & 2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 2 & 3 \\ -2 & 2 \end{pmatrix} = \begin{pmatrix} 2/5 & 3/5 \\ 2/5 & 3/5 \end{pmatrix} \end{aligned}$$

Regardless of the initial distribution of the markov chain, $\mu_0 P^n \to (2/5, 3/5)$, so the asymptotics are well defined.

Some initial distributions work very nicely when taking limits of the stochastic matrix: Suppose μ is a left eigenvector of P ($\mu P = \mu$). Then $\mu P^n = \mu$, and so, taking $n \to \infty$, we find μ is the limiting distribution of the Markov chain it generates. One can check that $(2/5, 3/5)$ is a left eigenvector for the probability matrix in the last example. If all initial distribution converge to the same value, then they must converge to this distribution. Identifying these vectors therefore seems important in order to identify the limiting distribution of the matrix. An **invariant**, or **stationary probability distribution** for P is a probability distribution μ such that $\mu P = \mu$. We will show that a large class of stochastic processes have a unique invariant probability density, which represents the 'average' time spent in each state, and assuming a slightly stronger condition, the distribution on the states converges to the distribution.

2.2 Irreducibility

Let x and y be two states. We say x **communicates** with y if there is some n with $P^n_{xy} > 0$. If we divide the states of a process into equivalence classes of states, all of which communicate between one another, then we obtain a family of **communication classes** for the stochastic process. A Markov chain with one communication class is **irreducible**. We can further classify the communication classes of a reducible markov chain by looking at one-sided communication. A state x may communicate with a state y without the converse being true. A communication class which only communicates with itself is know as **recurrent** whereas if a communication class

communicates with other classes, it is known as **transient**. By reordering the entries of P, we may assume the states in the same communication class occur continguously, and that all the recurrent communication classes occur before the transient classes. We can then write

$$P = \begin{pmatrix} P_1 & & & \\ & \ddots & & 0 \\ & & P_n & \\ & S_1 & & Q \end{pmatrix}$$

where each P_i is a stochastic matrix over a particular recurrent class. For any m,

$$P^m = \begin{pmatrix} P_1^m & & & \\ & \ddots & & 0 \\ & & P_n^m & \\ & S_m & & Q^m \end{pmatrix}$$

Each P_i acts as it's own 'sub Markov process', which we can analyze on their own, and then put them together to understand the full Markov process.

We claim that $Q^m \to 0$ as Q tends to ∞. This means exactly that transient states almost surely enter recurrent states over time. if U is the set of transient states on a Markov process, then $\mathbf{P}(X_k \in U) \to 0$ (this is the limit of the probability of a decreasing family of sets, so the limit certainly exists). Since our state space is infinite, there is $\varepsilon > 0$ and n such that for any state $x \in U$, there is $0 \leqslant n \leqslant m$ and some recurrent state y such that $P^n(x,y) > \varepsilon$. Then

$$\begin{aligned} \mathbf{P}(X_{(n+1)m} \in U) &= \mathbf{P}(X_{nm} \in U) - \mathbf{P}(X \text{ leaves } U \text{ on } (nm,(n+1)m]) \\ &\leqslant (1-\varepsilon)\mathbf{P}(X_{nm} \in U) \end{aligned}$$

So $\mathbf{P}(X_{nm} \in U) \leqslant (1-\varepsilon)^n$, which converges to zero as $n \to \infty$.

2.3 Irreducibility and Potential Functions

There is a one-to-one correspondence between left eigenvectors of P and right eigenvectors of P. We shall determine the uniqueness of invariant probabilities by analyzing the right eigenvectors. Strangely, the proof

mimics the analysis of harmonic functions on Euclidean space. We say a function f is **harmonic** if $Pf = f$. This can be interpreted as saying the average value of f beginning from a particular state is equal to the value at the state itself.

Lemma 2.1. *A harmonic function on an irreducible markov chain is constant.*

Proof. Let s^* be a state maximizing a harmonic function f. If $P(s^*, s) > 0$, then it cannot be true that $f(s) < f(s^*)$, for then

$$\begin{aligned} f(s^*) = Pf(s^*) = \sum_x P(s^*, x) f(x) = \sum_{x \neq s} P(s^*, x) f(x) + P(s^*, s) f(s) \\ \leqslant (1 - P(s^*, s)) f(s^*) + P(s^*, s) f(s) < f(s^*) \end{aligned}$$

This implies $f(s) = f(s^*)$. Furthermore, it implies that the function must be constant on the communication class of s^*. In particular, since an irreducible markov chain consists of one connected component, f must be constant. □

Corollary 2.2. *Invariant probability vector for irreducible processes are unique if they exist.*

Proof. The space of harmonic functions on an irreducible process is one dimensional, which implies that the space of left eigenvectors for the transition matrix is also one dimensional. This means that there is at most one eigenvector of eigenvalue one with non-negative entries whose entries sum to one. □

The theorem above is an analogy of the maximum modulus principle for harmonic functions – which states that, if a function attains its maximum value on an open set, the function must be constant on the connected component upon which it is defined. Classically, electromagnetics modelled the electrical potential in space by such a harmonic function. In the continuous case, the charge distributes itself across the entire space. In the discrete finite case, the electric potential must occur at one of the points where the electricity flows, so the flow must be constant throughout.

2.4 Existence of a Stationary Distribution

Given a state x on a Markov process $X_0, X_1, \ldots$, define a random variable $\tau_x = \min\{t \geqslant 0 : X_t = x\}$, and $\tau_x^+ = \min\{t > 0 : X_t = x\}$. τ is known as the **hitting time** of the state x. If $X_0 = x$, then we call τ_x^+ the **first return time**.

Lemma 2.3. *For any two states x and y on an irreducible chain, $\mathbf{E}_x(\tau_y^+) < \infty$.*

Proof. Because we are working on a finite state space, there is an integer n and $\varepsilon > 0$ such that for any two states x and y, there is $m \leqslant n$ with $P_n(x,y) > \varepsilon$. Thus

$$\begin{aligned}\mathbf{P}_x(\tau_y^+ > kn) &= \mathbf{P}_x(\tau_y^+ > (k-1)n) - \mathbf{P}_x((k-1)n \leqslant \tau_y^+ < kn) \\ &\leqslant (1-\varepsilon)\mathbf{P}_x(\tau_y^+ > (k-1)n)\end{aligned}$$

so we conclude $\mathbf{P}_x(\tau_y^+ > kn) \leqslant (1-\varepsilon)^n$, so

$$\mathbf{E}_x(\tau_y^+) = \sum_{k=0}^{\infty} \mathbf{P}_x(\tau_y^+ > k) \leqslant n \sum_{k=0}^{\infty} \mathbf{P}_x(\tau_y^+ > kn) \leqslant n \sum_{k=0}^{\infty} (1-\varepsilon)^k < \infty$$

and thus the expected value is finite. □

We will soon see that on irreducible Markov chains, there is a unique invariant probability distribution μ_*, and for any initial distribution μ,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \mathbf{P}_\mu(X_k = x) = \mu_*(x)$$

which is the long term chance of going to x. The intuition is that if we start at x, and let $\tau_x^n = \min\{k > \tau_x^{n-1} : X_k = x\}$ denote the k'th return time to x, with $\tau_x^0 = 0$. Then the $\tau_x^{n+1} - \tau_x^n$ are intuitively i.i.d random variables with mean $\mathbf{E}[\tau_x^+]$, so the strong law of large numbers guarantees that almost surely,

$$\lim_{n\to\infty} \frac{\tau_x^n}{n} = \mathbf{E}[\tau_x^+]$$

So pointwise, we find $\tau_x^n \approx n\mathbf{E}[\tau_x^+]$, implying that we visit x n times in time roughly proportional to $n\mathbf{E}[\tau_x^+]$. But the theorem we desire says that in m steps we visit x $m\mu_*(x)$ times. Setting $m = n\mathbf{E}[\tau_x^+]$ gives $n = n\mathbf{E}[\tau_x^+]\mu^*(x)$, so we conclude that $\mu^*(x) = \mathbf{E}_x[\tau_x^+]^{-1}$. Though this is a heuristic argument, we wll show that the measure $\mu^*(x) = \mathbf{E}_x[\tau_x^+]^{-1}$ is actually an invariant measure, which we will soon show is unique.

Theorem 2.4. *Every irreducible chain has an invariant probability measure.*

Proof. Let x denote an arbitrary state of the chain. Define

$$\begin{aligned}\tilde{\pi}(y) &= \mathbf{E}_x(\text{number of visits to } y \text{ before returning to } x) \\ &= \sum_{k=0}^{\infty} \mathbf{P}_x(X_k = y, \tau_x^+ > k)\end{aligned}$$

Then $\tilde{\pi}(y) \leqslant \mathbf{E}(\tau_x^+) < \infty$. We claim $\tilde{\pi}$ is stationary. For a fixed y,

$$\sum_z \tilde{\pi}(z)P(z,y) = \sum_z \sum_{k=0}^{\infty} \mathbf{P}(X_k = z, \tau_x^+ > k)P(z,y)$$

Now by the Markov property, because the event $\{\tau_x^+ > k\}$ is determined by $X_0, \ldots, X_k$, one can use conditional probabilities to show

$$\mathbf{P}_x(X_k = z, X_{k+1} = y, \tau_x^+ > k) = \mathbf{P}_x(X_k = z, \tau_x^+ > k)P(z,y)$$

so interchanging the summation, we find

$$\begin{aligned}\sum_z \sum_{k=0}^{\infty} \mathbf{P}(X_k = z, \tau_x^+ > k)P(z,y) &= \sum_{k=0}^{\infty} \mathbf{P}(X_{k+1} = y, \tau_x^+ > k) \\ &= \tilde{\pi}(y) - \mathbf{P}_x(X_0 = y, \tau_x^+ > 0) + \sum_{k=1}^{\infty} \mathbf{P}_x(X_k = y, \tau_x^+ = k) \\ &= \tilde{\pi}(y) - \mathbf{P}_x(X_0 = y, \tau_x^+ > 0) + \mathbf{P}(X_{\tau_x^+} = y) = \tilde{\pi}(y)\end{aligned}$$

which is easily seen regardless of whether $x = y$ or $x \neq y$. Normalizing $\tilde{\pi}$ by

$$\sum \tilde{\pi}(y) = \sum_{k=0}^{\infty} \mathbf{P}_x(\tau_x^+ > k) = \mathbf{E}[\tau_x^+]$$

Since $\tilde{\pi}(x) = 1$, we conclude $\pi(x) = \mathbf{E}[\tau_x^+]^{-1}$. Since π is unique, we may repeat this proof for all states to conclude that the equation $\pi(y) = \mathbf{E}[\tau_y^+]^{-1}$ holds for all states y. □

A **stopping time** is a $\mathbf{N} \cup \{\infty\}$ valued random variable τ such that the event $\{\tau = k\}$ is determined by $X_0, \ldots, X_k$. In the proof above, we can substitute an arbitrary stopping time provided $\mathbf{P}_x(\tau < \infty) = \mathbf{P}_x(X_\tau = x) = 1$,

and we still obtain that $\tilde{\pi}$ is stationary. If τ is any stopping time and m is an integer, then

$$\mathbf{P}_{x_0}(X_{m+1} = x_1, \ldots, X_{m+n} = x_n | \tau = m, X_1, \ldots, X_m) = \mathbf{P}_{X_m}(X_1 = x_1, \ldots, X_n = x_n)$$

which is an immediate consequence of the Markov property. This is known as the **strong Markov property**, which is less obvious in the continuous setting.

2.5 Perron-Frobenius

There is an incredibly useful theorem of analytical linear algebra to help prove the existence of invariant distributions on finite markov chains.

Theorem 2.5 (The Perron-Frobenius Theorem)**.** *Let M be a positive square matrix. Then there is a positive eigenvalue λ of maximal modulus, called the* **Perron root** *of M, with one dimensional eigenspace which contains a positive vector.*

Proof. Let $v \leqslant w$ represent that $v_i \leqslant w_i$ for all i. For the purposes of this proof, we let $|v|$ denote the vector v with $|v|_i = |v_i|$. We proceed in a series of steps:

> (Claim 1) If $v \geqslant 0$, but $v \neq 0$, then $Mv > 0$: This follows because if $v_i > 0$, then for any j,
>
> $$(Mv)_j = \sum M_{jk} v_k \geqslant M_{ji} v_i > 0$$
>
> Because of this, if $v \geqslant 0$, we may define $g(v) = \sup\{\lambda : Mv \geqslant \lambda v\}$.
>
> (Claim 2) The function $g(v)$ is continuous for $v \neq 0$: We can write $g = \min(g_1, \ldots, g_d)$, where $g_i(v) = \sup\{\lambda : (Mv)_i \geqslant \lambda v_i\}$, and it suffices to prove the functions g_i are continuous as maps into $(0, \infty]$. If $v_i \neq 0$, then $g_i(v) < \infty$, because
>
> $$(Mv)_i = \sum M_{ik} v_i \leqslant v_i \left(\frac{(Mv)_i}{v_i} \right)$$

so $g_i(v) \leqslant (Mv)_i v_i^{-1}$. If $v_i, w_i \neq 0$, and $(Mv)_i \geqslant \lambda v_i$, then

$$\begin{aligned}(Mw)_i &= (Mv)_i - (M(v-w))_i \\ &\geqslant \lambda v_i - \sum M_{ij}(v_j - w_j) \geqslant \lambda v_i - \|M\|_\infty \|v-w\|_\infty \\ &= v_i \left(\lambda - \frac{\|M\|_\infty \|v-w\|_\infty}{v_i}\right) \geqslant v_i \left(\lambda - \frac{\|M\|_\infty \|v-w\|_\infty}{\min(v_i, w_i)}\right)\end{aligned}$$

It follows that $|g_i(v) - g_i(w)| \leqslant \|M\|_\infty \|v-w\|_\infty \min(v_i, w_i)^{-1}$, which gives continuity at v_i if $v_i \neq 0$. On the other hand, for any w with $w_i \neq 0$, we conclude

$$(Mw)_i = \sum M_{ik} w_k \geqslant w_i \left(M_{ik} \frac{w_j}{w_i}\right)$$

so $g_i(w) \geqslant M_{ik} w_j w_i^{-1}$, so if $w \to v$, where $v_i = 0$ and $v_j \neq 0$, then w_j remains bounded while $w_i \to 0$, so $g_i(w) \to \infty$. This concludes the proof of continuity.

Since g is continuous, and $g(\alpha v) = g(v)$ for all $\alpha, v \neq 0$, we conclude that g attains it's maximum α, because the problem reduces to finding the maximum over the non-negative elements of the unit sphere, which forms a compact set.

1. (Claim 3) If $g(v) = \alpha$, then $Mv = \alpha v$, and all its components are strictly positive: We know that $Mv \geqslant \alpha v$. We know $Mv \geqslant \alpha v$, so if $v \neq \alpha v$, we conclude $Mv > \alpha Mv$, so $g(Mv) > \alpha$, contradicting the maximality of α at v. But since $v \geqslant 0$, $Mv = \alpha v > 0$, so we conclude all elements of v are positive.

2. (Claim 4) If λ is any other eigenvalue of M, then $|\lambda| < \alpha$: If v is an eigenvector for λ, and we define $w = (|v_1|, \ldots, |v_n|)$, then

$$|\lambda v_i| = \left|\sum M_{ik} v_k\right| \leqslant \sum M_{ik} |v_k|$$

hence $\alpha \geqslant g(|v|) \geqslant |\lambda|$. If $|\lambda| = \alpha$, then we conclude that $g(|v|) = \alpha$ and thus $|v|$ is an eigenvector with eigenvalue λ, so

$$\left|\sum M_{ik} v_k\right| = \sum M_{ik} |v_k|$$

This equation holds only when there is a complex number z of norm one such that $v = z|v|$ for some $t \geqslant 0$. But then

$$\lambda v = Mv = zM|v| = z|\lambda||v| = |\lambda|v$$

so $\lambda = |\lambda|$.

3. (Claim 4) Any two positive eigenvectors of eigenvalue α are linearly independant: Let v and w be non-negative eigenvectors of eigenvalue α. Choose ε small enough that $v - \varepsilon w \geqslant 0$, and $v_i - \varepsilon w_i = 0$. If $v \neq \varepsilon w$, then $v - \varepsilon w \neq 0$, and so $M(\alpha^{-1}(v - \varepsilon w)) = v - \varepsilon w > 0$, a contradicition proving $v = \varepsilon w$.

4. (Claim 5) The eigenvalues of any $n-1 \times n-1$ submatrix of M are strictly less than α: Let B be any such submatrix obtained from deleting the ith row and jth column. Then $B_{kl} = A_{f(k)g(l)}$, where

$$f(k) = \begin{cases} k & : k < i \\ k+1 & : k \geqslant i \end{cases} \quad g(l) = \begin{cases} l & : l < j \\ l+1 & : l \geqslant j \end{cases}$$

B satisfies the hypothesis of the Frobenius theorem, so if β maximizes the g function on B, there is a non-negative vector w with $Bw = \beta w$, and if we consider the vector v with

$$v_k = \begin{cases} w_k & : k < j \\ \varepsilon : k = j & \\ w_{k-1} & : k > j \end{cases}$$

Then $(Mv)_k = \lambda v_k + \varepsilon M_{kj} > \lambda v_k$ for $k \neq j$, and we can choose ε small enough that $(Mv)_j > \lambda\varepsilon = \lambda v_j$, because w is a positive vector and so $(Mv)_j = \sum M_{jk} v_k \geqslant \sum_{k \neq j} M_{jk} v_k$, which does not depend on ε. We conclude that $\beta < \alpha$.

5. (Claim 6) Consider the characteristic polynomial $f(\lambda) = \det(M - \lambda)$. Then $f'(\lambda) = -\sum_{i=1}^{n} \det(M_i - \lambda)$, where M_i is obtained from M by deleting the ith row and ith column: We consider the expansion

$$f(\lambda) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} (M - \lambda)_{i\sigma(i)}$$

Then, by the product rule,

$$\begin{aligned} f'(\lambda) &= -\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \sum_{i=\sigma(i)} \prod_{j \neq i} (M-\lambda)_{j\sigma(j)} \\ &= -\sum_{i=1}^{n} \sum_{\sigma \in S_{n-1}} \operatorname{sgn}(\sigma) \prod_{j=1}^{n} (M_i - \lambda)_{j\sigma(j)} \\ &= -\sum_{i=1}^{n} \det(M_i - \lambda) \end{aligned}$$

Since α exceeds the modulus of any eigenvalue of M_i, and $\det(M_i - \lambda) \to \pm\infty$ as $\lambda \to \infty$ (with the sign determined by the dimension of M_i, and thus constant over all M_i, we conclude that $f'(\lambda) \neq 0$, so α has a one dimensional eigenspace, since it is a simple root of the characteristic polynomial.

Looking back over the claims, we have proven all we set out to do. □

Now suppose P is a stochastic, positive matrix. Then we may apply Perron-Frobenius to P, obtaining a Perron root λ. We must have $|\lambda| \leqslant 1$, since all entries of the matrix are less than one, and so for any vector v, $|(Av)_i| \leqslant |v_i|$. Because $(1,\ldots,1)^t$ is a right eigenvector for P of eigenvalue 1, $\lambda = 1$. Thus P can be modified, under some change of basis matrix Q, such that

$$D = QPQ^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & M \end{pmatrix}$$

Where M is a square matrix such that $\lim_{n\to\infty} M^n = 0$ (Use the Jordan Canonical Form, and the fact that all eigenvalues of M are less than one). But then

$$\lim_{n\to\infty} P^n = Q^{-1}(\lim_{n\to\infty} D^n)Q = Q^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} Q = \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix}$$

where μ is a row vector which sums to one. μ is the unique invariant distribution to the process, because $\mu P = (\lim_{n\to\infty} \mu P^n)P = \lim_{n\to\infty} \mu P^{n+1} = \mu$.

This argument can be considerably strengthened. Let P be a stochastic matrix such that P^n is positive, for some n. The eigenvalues of P^n are

simply the eigenvalues of P taken to the power of n. Perron and Frobenius tell us that 1 is the Perron root of P^n (since P^n is stochastic), so that P has a maximal eigenvalue which is an n'th root of unity. Since P^{n+1} also has all positive entries, the maximal eigenvalue of P must also be an $n+1$'th root of unity, and this is only true if the eigenvalue is 1. If v is an eigenvector of eigenvalue 1, it must also be an eigenvector of P^n, so the eigenvectors of P are the same as the eigenvectors of P^n, and we may choose an eigenvector which is also a distribution - an invariant distribution to which the matrix converges. Note, however, that we cannot expect all homogenous matrices to satisfy this theorem.

Example. *Consider a process with transition matrix*

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

Then $P^n = I$ for even n, and $P^n = P$ for odd n. Thus P^n cannot converge. This is because the matrix is periodic – it oscillates between values. Note that P^n never has all positive entries. The only time μP^n converges is when $\mu = (1/2, 1/2)$.

Example. *Consider a process whose transition matrix is the identity matrix I. Then $P^n \to I$, so $\mu P^n \to \mu$ for all distributions μ. Thus we can always take limits of probability distributions, but different initial distributions give rise to different asymptotics. This is because the process is reducible – there is not enough 'mixing' among all possible states to generate a homogenous distribution.*

2.6 Aperiodicity and Irreducibility

Our problem thus reduces to classifying those stochastic matrices P for which P^n is positive, for some n. This property will reduce to identifying two concepts on which the positivity fails: periodicity and irreducibility.

There is one other way an irreducible Markov chain can fail to converge like we would like. For any state x, let $J(x) = \{n \in \mathbf{N} : P^n_{xx} > 0\}$. Then $J(x)$ is closed under addition. The greatest common divisor of $J(x)$ is known as the **period** of s. A Markov chain for which every state has period one is known as **aperiodic**. Note that two states in the same communication class share a common period. Thus we may talk about the periodicity of a irreducible markov chain.

Theorem 2.6. *Let P be a stochastic matrix, which determines an aperiodic, irreducible Markov chain. Then there is a unique vector μ for which $\mu P = P$, and for any other probability distribution π, $\lim_{n\to\infty} \pi P^n = \mu$.*

Proof. We just need to verify that P^n is a positive matrix for a large enough n. Since P is aperiodic, for large enough m, $P^m_{ii} > 0$ for all i. If $j \neq i$, there is some k for which $P^k_{ij} > 0$. Then, for large enough m, $P^m_{ij} > 0$, since

$$P^m_{ij} \geqslant P^k_{ij} P^{m-k}_{ii} > 0$$

Taking m large enough so that the argument above works for all i and j, we find $P^m_{ij} > 0$ for all i, j. It follows that we may apply Perron-Frobenius to P^m, and we find our invariant distribution. □

Corollary 2.7. *On every aperiodic, irreducible Markov chain on a finite state space there exists a unique stationary distribution.*

We call an irreducible, aperiodic Markov chain **ergodic**, which is why the theorem is known as the ergodic theorem for Markov chains. An ergodic chain is a chain with enough 'mixing' to generate an invariant distribution for the process. In terms of Ergodic theory, the pushforward map T on $S^{\mathbf{N}}$ given by mapping $x_0, x_1, \ldots$ to $x_1, \ldots$ is measure preserving under measure induced by the random variables $X_0, X_1, \ldots$. In terms of general ergodic theory, this map is ergodic if and only if the chain is irreducible, and mixing if and only if the chain is aperiodic.

Example. *Let us consider the asymptotics of a two state time homogenous markov chain on two states x and y. There are parameters $0 \leqslant p, q \leqslant 1$ such that the transition matrix has the form*

$$P = \begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \end{array}$$

If $p = 0$ or $q = 0$, the chain is reducible. If $p = 1$ and $q = 1$, then the chain is periodic, swinging back and forth deterministically between the two states. In any other case, the Markov chain is ergodic, and since

$$\left(\frac{q}{p+q}, \frac{p}{p+q}\right) \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \left(\frac{q}{p+q}, \frac{p}{p+q}\right)$$

the unique invariant probability distribution is $\mu^* = (p+q)^{-1}(q,p)$*, and this is the limiting distribution. Given an arbitrary initial distribution* μ_0*, if we define* $\Delta_n = \mu_n - \mu^*$*, then*

$$\begin{aligned}\Delta_{n+1}(x) &= (1-p)\mu_n(x) + q\mu_n(y) - \frac{q}{p+q} \\ &= (1-p-q)\mu_n(x) + q - \frac{q}{p+q} \\ &= (1-p-q)\left(\mu_n(x) - \frac{q}{p+q}\right) = (1-p-q)\Delta_n(x)\end{aligned}$$

And since $\Delta_n(y) = -\Delta_n(x)$*, we conclude* $\Delta_n = (1-p-q)^n\Delta_0$*, so the distribution converges linearly at a rate* $1-p-q$*.*

Example. *Consider a random walk on a connected graph with n vertices and m edges. Then the process is irreducible, and since*

$$\sum_{vw\in E} \deg(v)P(v,w) = \sum_{vw\in E} \frac{\deg(v)}{\deg(v)} = \deg(w)$$

so the distribution $\mu(v) = \deg(v)/2m$ *is invariant. We say a graph is regular if every vertex has the same degree, in which case* μ *is the uniform distribution.*

2.7 Periodicity and Average State Distributions

If a chain has period greater than one, say of period n, then the limiting properties of the process are not so simple. We may divide the states into a partition $K_1, K_2, \ldots, K_n$, for which states in K_i can only transition to states in K_{i+1}, or from K_n to K_1. If we only look at the time epochs $t_1, t_2, \ldots$ where the chain is guaranteed to be in a certain partition, then we obtain an aperiodic markov chain, which in the irreducible case reduces to invariant distributions on the states. If our chain has period m, our chain converges to m distributions $\mu_{t_1}, \ldots, \mu_{t_m}$. The limit $\lim_{n\to\infty} \mu P^n$ may not exist, but the chebyshev limit

$$\lim_{n\to\infty} \frac{\sum_{k=0}^n \mu P^n}{n} = \mu \lim_{n\to\infty} \frac{\sum_{k=0}^n P^n}{n} = \frac{\mu_{t_1} + \cdots + \mu_{t_m}}{m}$$

will always exists. It represents the overall, accumulated average of which states we visit over the whole time period the chain is ran for.

Example. *Take a markov chain of period 2, with transition matrix*

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

We may diagonalize this matrix, letting $P = QDQ^{-1}$, where

$$Q = \begin{pmatrix} 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1 & 1 & 0 & 0 & -1 \\ -1 & 1 & -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad D = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 1/\sqrt{2} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Taking matrix limits, we see that only the first two rows of D become relavant far into the future, so that for large n, for any μ,

$$P^n \approx \begin{pmatrix} 1/8 & 1/4 & 1/4 & 1/4 & 1/8 \\ 1/8 & 1/4 & 1/4 & 1/4 & 1/8 \\ 1/8 & 1/4 & 1/4 & 1/4 & 1/8 \\ 1/8 & 1/4 & 1/4 & 1/4 & 1/8 \\ 1/8 & 1/4 & 1/4 & 1/4 & 1/8 \end{pmatrix} + (-1)^n \begin{pmatrix} 1/8 & -1/4 & 1/4 & -1/4 & 1/8 \\ -1/8 & 1/4 & -1/4 & 1/4 & -1/8 \\ 1/8 & -1/4 & 1/4 & -1/4 & 1/8 \\ -1/8 & 1/4 & -1/4 & 1/4 & -1/8 \\ 1/8 & -1/4 & 1/4 & -1/4 & 1/8 \end{pmatrix}$$

On even states, P^n converges to a different matrix than on odd states. Nonetheless, the Chebyshev limit exists for any distribution μ, and is given by

$$\frac{1}{2}[(1/4, 0, 1/2, 0, 1/4) + (0, 1/2, 0, 1/2, 0)] = (1/8, 1/4, 1/4, 1/4, 1/8)$$

This is not the distribution at a certain time point, but the distribution of averages over a long time period.

For instance, if $\{X_i\}$ is a irreducible markov chain, we would like to know the proportional number of times a certain state x is visited. We would like to determine the expected value of

$$S_x = \lim_{n \to \infty} \sum_{k=0}^{n} \frac{\mathbf{I}(X_k = x)}{n}$$

$$\mathbf{E}(S_x) = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{1}{n}\mathbf{E}(\mathbf{I}(X_k = x)) = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{\mathbf{P}(X_k = x)}{n}$$

And this is just the invariant probability of the process – the Chebyshev limit.

2.8 Stopping Times

We would like to finish our discussion of finite state space Markov chains by analyzing a certain class of random variables – representing the time at which a certain event happens.

Definition. A **Stopping Time** for a process $\{X_0, X_1, \dots\}$ is a $\mathbf{Z} \cup \{\infty\}$ valued random variable τ, such that, if we know the values of $X_0, \dots, X_n$, we can tell if $\tau = n$. Rigorously, $\mathbf{I}(\tau = n)$ is a function of the $X_0, \dots, X_n$.

A stopping time basically encapsulates a decision process. After observing the $X_0, \dots, X_n$, we decide whether we want to finish observing the Markov process. We can't look into our future, and must decide at that time point to leave.

Example. *Let $\{X_0, X_1, \dots\}$ be a stochastic process on a state space $\mathcal{S}$. Fix a state s, and define the* **hitting time** *τ_s to be*

$$\tau_s = \min\{n : X_n = s\}$$

Since $\mathbf{I}(\tau_s = n) = \mathbf{I}(X_0 \neq s, \dots, X_{n-1} \neq s, X_n = s)$, this is a stopping time.

Example. *Let $\{X_0, X_1, \dots\}$ be a stochastic process on a state space $\mathcal{S}$. Fix a state s, and suppose that $\mathbf{P}(X_0 = s) = 1$. The* **return time** *ρ_s of the process is defined*

$$\rho_s = \min\{n \geqslant 1 : X_n = s\}$$

And is a stopping time.

Since stopping times are valued on the time epochs upon which a process is defined, we can do interesting things to combine the time with the

process. For instance, we may consider a random variable X_τ. In the case that τ is the hitting or return time for a state s, then $X_\tau = s$. One wonders whether the Markov property behaves nicely with respect to a stopping time. This is the strong Markov property.

Definition. let $\{X_t\}$ be a markov process, and τ a stopping time. X_t satisfies the **strong Markov property** with respect to τ, if, for $t_1 < \cdots < t_n < \tau$,

$$\mathbf{P}(X_\tau = y | X_{t_n} = x_n, \ldots, X_1 = t_1) = \mathbf{P}(X_\tau = y | X_{t_n} = x_n)$$

In other words, X_t forgets history with respect to the stopping time. By letting $\tau = n$ be a fixed integer, we obtain the normal markov property.

Theorem 2.8. *All discrete markov processes satisfies the strong markov property with respect to any stopping time.*

Proof. Let $\{X_0, X_1, \ldots\}$ be a markov process, and τ a stopping time. Then, assuming $t_1 < \cdots < t_n < \tau$

$$\begin{aligned}\mathbf{P}(X_\tau = y | X_{t_n} = x_n, \ldots, X_{t_1} = x_1) &= \sum_{k=t_n+1}^{\infty} \mathbf{P}(\tau = k)\mathbf{P}(X_k = y | X_{t_n} = x_n) \\ &= \mathbf{P}(X_\tau = y | X_{t_n} = x_n)\end{aligned}$$

So the process is strongly Markov. □

Let us use our tools to derive the expected return time $\mathbf{E}(\rho_s)$. First, let $\mathcal{J}_0 = 0$, $\mathcal{J}_1 = \rho_s$ and, more generally, define $\mathcal{J}_k$ to be the k'th time we return to s, $\mathcal{J}_k = \min\{n > J_{k-1} : X_n = s\}$. Then the strong markov property shows $\mathcal{J}_{k+1} - \mathcal{J}_k$ are independant and identically distributed, by the law of large numbers, as $n \to \infty$,

$$\sum_{k=1}^{n} \frac{\mathcal{J}_k - \mathcal{J}_{k-1}}{n} = \frac{\mathcal{J}_n}{n} \to \mathbf{E}(\rho_s)$$

After a large enough n, each state will be approximately visited $n\mu_s$ times. Thus $\mathcal{J}_n \approx n/\mu_s$, and $\mathbf{E}(\rho_s) = 1/\mu_s$.

Now we can analyze Markov chains with transient states. Recall that we can write the transition matrix of such a process as

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 \\ 0 & P_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & P_n & 0 \\ \dots & S_1 & \dots & Q \end{pmatrix}$$

We have $Q^n \to 0$ as $n \to \infty$, since we are guarenteed to leave a transient state and never return.

All eigenvalues of Q are less than one in absolute value, so $I - Q$ is invertible. A small computation shows that

$$\sum_{k=0}^{\infty} Q^k = (I - Q)^{-1}$$

provided the sum on the right converges, which it must, since the series converges absolutely (and the space is Banach). Q^k_{ij} is the probability that $X_k = x_j$ given $X_0 = x_i$, so $(\sum_{k=0}^{n} Q^k)_{ij}$ is the expected number of visits to x_j from time epoch n starting from x_i. Taking $n \to \infty$, we find the expected number of visits to the state before hitting a recurrent state is $(I - Q)^{-1}_{ij}$. If we sum up row i, we get the expected number of states before hitting a recurrent state starting from i.

We can also use this method in an irreducible chain to find the expected time to reach a state x_j starting at x_i, for $i \neq j$. We modify the Markov process by making it impossible to leave x_j once it has been entered. This makes all other states transient. Then the expected number of visits before enterring a recurrent state is the expected number of states until we hit x_j.

How about determining the probability of entering a specific recurrent class starting from a transient state. To simplify our discussion, let each recurrent class consist of a single vertex, whose probability of return to itself equals 1. First, to simplify the situation, assume each recurrent class consists of a single vertex (we may 'shrink' any Markov process so that each class consists of a single vertex for our situation). For each transient x and recurrent y, let $\alpha(x, y)$ be the probability of ending up at y starting

at x. We have

$$\alpha(x,y) = \sum_{z \text{ transient}} P(x,z)\alpha(z,y) + P(x,y)$$

Let $\{x_1,\dots,x_n\}$ be the recurrent states of the process, and $\{y_1,\dots,y_m\}$ the transient states. If we define a matrix $A_{ij} = \alpha(x_i,y_j)$, then the equation above tells us that $A = S + QA$, where we write

$$P = \begin{pmatrix} 1 & \dots & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \dots & S & \dots & Q \end{pmatrix}$$

Hence $(I-Q)A = S$, and so $A = (I-Q)^{-1}S$. This is the limiting values of P^n on S as $n \to \infty$.

Example. *Consider a gambler who's going 'all in'. He won't leave without obtaining a certain amount of money N, unless he runs out of money and goes bust. We want to find out the probability that he will go home happy rather than broke. The situation of the gambler can be modelled by a random walk on $\{0,1,\dots,N\}$. We assume each integer represents how much money the gambler has at a certain time, and that each bet either costs or wins the gambler a single unit of money. If $p > 0$ is the probability of winning the bet, then the transition probabilities of the random walk are*

$$P(i,i+1) = p \quad P(i,i-1) = (1-p) \quad P(0,0) = P(N,N) = 1$$

This is a reducible markov chain with transient states. We are trying to determine the probability of entering the different recurrent classes, starting from a certain transient state M. Using our newly introduced technique, we write $\alpha(x,0)$ and $\alpha(x,N)$ to be the probabilities of going home rich or poor. The matrix notation is ugly for our purposes, so we just use the linear equations considered,

$$\alpha(1,1) = p\alpha(2,0) \quad \alpha(n-1,1) = p + (1-p)\alpha(n-1,1)$$

$$\alpha(k,1) = p\alpha(k+1,1) + (1-p)\alpha(k-1,1)$$

These are a series of linear difference equations. If we assume $\alpha(k,0) = \beta^k$, *then* $\beta^k = p\beta^{k+1} + (1-p)\beta^{k-1}$. *This equation has the solution* $\beta = \left\{1, \frac{1-p}{p}\right\}$, *and thus a general solution is of the form*

$$\alpha(k,1) = c_0 + c_1 \left(\frac{1-p}{p}\right)^k$$

The boundary conditions $\alpha(0,1) = 0$, $\alpha(N,1) = 1$ *tells us that*

$$c_0 + c_1 = 0 \qquad c_0 + c_1 \left(\frac{1-p}{p}\right)^N = 1$$

So

$$c_1 = \frac{1}{\left(\frac{1-p}{p}\right)^N - 1} \qquad c_0 = \frac{1}{\left(1 - \frac{1-p}{p}\right)^N}$$

And the general form is

$$\alpha(k,1) = \frac{1 - \left(\frac{1-p}{p}\right)^k}{1 - \left(\frac{1-p}{p}\right)^N}$$

provided, of course, that $p \neq 1/2$. *In this case,* 1 *is a double roots of the characteristic equation, so*

$$\alpha(k,1) = c_0 + c_1 k$$

and $c_0 + c_1 = 0$, $c_0 + c_1 N = 1$, *so* $c_1 = \frac{1}{N-1}$,

$$\alpha(k,1) = \frac{k-1}{N-1}$$

Our discussion of the classical theory of ergodic finite state space markov chain has been effectively completed.

Chapter 3

Countable-State Markov Chains

3.1 General Properties

Let us now consider time homogenous Markov chains on a countable state space. For instance, we may consider random walks on $\mathbf{N}, \mathbf{Z}$, and $\mathbf{Z}^2$. Most finite space techniques extend to the countable situation, but not all. We may continue to talk of irreducibility, periodicity, the Chapman Kolmogorov equation, communication, and the like. Recurrence and transcience is a little more cimplicated, since in a single 'recurrence class' of infinite size, it may still be very rare for a state to return to itself.

3.2 Recurrence and Transience

We call a state **recurrent** if the markov chain is almost certain to return to itself infinitely many times. If a state in a class is recurrent, all states in a class is recurrent, then all states in the same class are recurrent. A state is **transient** if it is not recurrent. In the finite case, these new definitions agree with previous terminology.

How do we reliably determine if a process is transient? Let S_x be the total number of visits to x, assuming we start at x

$$S_x = \sum \mathbf{I}(X_n = x)$$

Calculating recurrence reduces to calculating $\mathbf{P}(S_x = \infty)$. Since S_x is a

random variable, we can take expectations

$$\mathbf{E}(S_x) = \sum_n \mathbf{P}(X_n = x | X_0 = x) = \sum_n P^n(x,x)$$

If $\mathbf{E}(S_x) < \infty$, then $\mathbf{P}(S_x < \infty) = 0$, so x is transient. Consider the hitting time τ_x. If $\mathbf{P}(\tau_x < \infty) = 1$, then by time homogeneity we conclude that x is hit infinitely many times. Suppose instead that $\mathbf{P}(\tau_x < \infty) = q < 1$. We have $\mathbf{P}(S_x = m) = q^{m-1}(1-q)$. Thus

$$\mathbf{E}(S_x) = \sum_{m=1}^{\infty} m\mathbf{P}(S_x = m) = \sum_{m=1}^{\infty} mq^{m-1}(1-q) = \frac{1}{1-q} < \infty$$

Hence a state is transient if and only if the expected number of returns is finite, that is,

$$\sum_{n=0}^{\infty} P^n(x,x) < \infty$$

Example. *Let us find whether symmetric random walk on* $\mathbf{Z}$ *is recurrent or symmetric. The chain is irreducible, so we only need determine the transience of a single point, say, 0. The number of paths from 0 to itself of length* $2n$ *is the number of choices of* n *down movements given* $2n$ *ups and downs, so*

$$\mathbf{P}(X_{2n} = 0 | X_0 = 0) = \frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(n!)^2 4^n}$$

For large n*, Stirling's formula tells us that* $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$*, so*

$$\mathbf{P}(X_{2n} = 0 | X_0 = 0) \approx \sqrt{\frac{1}{\pi n}} \left(\frac{2n}{e}\right)^{2n} \left(\frac{e}{n}\right)^{2n} 4^{-n} = \sqrt{\frac{1}{\pi n}}$$

Since $\sum(\pi n)^{-1/2} \to \infty$*, so must our sum, so the process is recurrent.*

Now take a random walk on $\mathbf{Z}^d$*. The number of paths from 0 to itself of length* $2n$ *is*

$$\sum_{2k_1 + \cdots + 2k_n = 2n} \binom{2n}{2k_1, \ldots, 2k_d} = \sum_{k_1 + \cdots + k_d = n} \frac{(2n)!}{(2k_1)! \ldots (2k_d)!}$$

FINISH HERE and the walk is recurrent for $d \leqslant 2$*, and transient for* $d > 2$*.*

Here's yet another method for determining recurrence. Fix a state y on an irreducible markov chain, and define $\alpha(x) = \mathbf{P}(X_n = y \text{ for some } n \geqslant 0 | X_0 = x)$. Then $\alpha(y) = 1$, and $\alpha(x) = \sum P(x,z)\alpha(z)$ for $z \neq y$. If the chain is recurrent, then $\alpha(z) = 1$ for all z. Less obviously, if y is a transient state, $\inf\{\alpha(z)\} = 0$. We shall prove later that if y is recurrent, there is no solution α with these properties, and if y is transient, α exists, and is unique.

Even if a chain is recurrent, an invariant distribution may not exist, due to the fact that we have an infinite number of states to work around. Let's specialize again. A chain is **null recurrent** if it is recurrent, but $\lim_{n\to\infty} P^n(x,y) = 0$, and is **positive recurrent** otherwise. An invariant probability is a function μ for which $\mu P = \mu$. We won't show it, but every irreducible, aperiodic, positive recurrent Markov chain has a distribuition μ. Moreover, such a chain is positive recurrent if and only if its has an invariant distribution μ. The return time τ_x has $\mathbf{E}(\tau_x | X_0 = x) = 1/\mu(x)$. For null recurrent chains, $\mathbf{E}(\tau_x | X_0 = x) = \infty$.

Example. *Let us derive the equations for a random walk on* $\mathbf{Z}$. *We have* $P(x,x-1) = q$, *and* $P(x,x+1) = 1-q$, *for some fixed* $0 \leqslant q \leqslant 1$. *We attempt to solve the equations to determine that the chain is recurrent.*

$$\alpha(x) = q\alpha(x+1) + (1-q)\alpha(x-1)$$

Using the rules of linear difference equations, if α *exists, it satisfies* $\alpha(x) = \beta^x$. *We have*

$$\beta^x = q\beta^{x+1} + (1-q)\beta^{x-1}$$

$$q\beta^2 - \beta + (1-q) = 0$$

$$\beta = \frac{1 \pm \sqrt{1-4q(1-q)}}{2q} = \frac{1 \pm (2q-1)}{2q} = \left\{1, \frac{1-q}{q}\right\}$$

Thus $\alpha(x) = c_0 + c_1\left(\frac{1-q}{q}\right)^x$ *If* $q < 1/2$, $c_1 = 0$ *because* α *must be bounded. But then* $\alpha(0) = 1$, *so* $c_0 = 1$, *and this contradicts that* $\inf \alpha(x) = 0$. *Hence the process is recurrent. For* $q > 1/2$, *we may pick* $c_1 = 1$, *so the process is recurrent. For* $q = 1/2$, *we have* $\alpha = c_0 + c_1 t$, *which cannot be bounded, so the process is recurrent.*

Let us try and determine if the random walk is positive or null recurrent for $q \leqslant 1/2$. *We need* μ *with* $\sum \mu(x) = 1$, *and* $\sum \mu(x)P(x,y) = \mu(y)$. *In this*

example we therefore need

$$\mu(x-1)q + \mu(x+1)(1-q) = \mu(x)$$
$$q\lambda^{x-1} + (1-q)\lambda^{x+1} = \lambda^x$$
$$\mu(x) = c_0 + c_1\left(\frac{q}{1-q}\right)^x$$

We must have $c_0 = 0$, and $c_1 > 0$. If $q = 1/2$, we cannot solve for μ, so the chain must be null recurrent. For $q < 1/2$ we find that

$$\sum_{x=-\infty}^{\infty}\left(\frac{q}{1-q}\right)^x = \sum_{x=0}^{\infty}\left(\frac{q}{1-q}\right)^x + \sum_{x=0}^{\infty}\left(\frac{1-q}{q}\right)^x - 1$$

This is infinite, so the process is null recurrent.

3.3 Branching Processes

Victorian upper-class culture strongly valued history and heritage. It was therefore a concern to these people when it was noticed that venerable surnames were dying out. If a male dies without producing a male heir, then a branch disappears from the family tree. If no males produce an heir in a generation, then the name completely dies out. Some believed that the exceeding comfort of upper-class life encouraged sterility, and that soon lower-classes would dominate England. Worried about this problem, the polymath Francis Galton put up a bulletin in "The Educational Times", challenging mathematicians to determine the cause of the problem. The reverend Henry William Watson took him up on this offer, and together they attempted a probabilistic analysis of the problem.

Galton and Watson represented the spread of families by a succeeding discrete number of generations $X_0, X_1, \ldots$, where the initial generation X_0 produces the offspring X_1, which produces the offspring X_2, and so on, through the ages. Each time epoch will represent a generation of a species, so that at each time interval, offspring are generated, and the current population dies off. Though it may seem a simplification to assume that generations do not overlap, assuming that each offspring reproduces independently, one can just consider the process as a family tree, independent of time. X_0 just represents the initial roots of the tree, X_1 represents

the offspring on the first layer of the tree, and so on and so forth, regardless of which order they came into being.

We now make the assumption that each member of the species, regardless of which generation the species is in, has an equal chance of producing offspring, and that the population produces asexually and independantly – considering only men as heirs to the family results in asexual reproduction. The first assumption is obviously not true over a long time period, but given that the probabilites do not seem to change too rapidly over direct successions, our results should not alter too much. These assumptions are equivalent to saying that X_t is a Markov chain with a certain probability transition function, which we shall now define.

Definition. Fix some distribution p over $\mathbf{N}$, and initial population distribution X_0, also over $\mathbf{N}$. We define a stochastic process $\{X_i\}$ by defining the transition probabilities

$$\mathbf{P}(X_{t+1} = m | X_t = n) = (p * p * \cdots * p)(m) \tag{3.1}$$

Where $(p * p * \cdots * p)$ is the n-fold convolution of p. More vicerally, one can define n independant random variables $Y_1, \ldots, Y_n \sim p$, and define

$$\mathbf{P}(X_{t+1} = m | X_t = n) = \mathbf{P}\left(\sum_{i=1}^{n} Y_i = m\right) \tag{3.2}$$

The Markov chain created is known as a **Branching Process**.

We shall start by understanding the evolution of the mean size of the population. Let μ denote the mean offspring a single individual will possess. Then we conclude by (3.2) that

$$\mathbf{E}(X_{t+1} | X_t = k) = \mathbf{E}\left(\sum_{i=1}^{k} Y_i\right) = \sum_{i=1}^{k} \mathbf{E}(Y_i) = k\mu \tag{3.3}$$

$$\mathbf{E}(X_{t+1}) = \mathbf{E}[\mathbf{E}(X_{t+1} | X_t)] = \sum_{k=0}^{\infty} \mathbf{P}(X_t = k)(k\mu) = \mu \mathbf{E}(X_t) \tag{3.4}$$

And therefore $\mathbf{E}(X_k) = \mu^k \mathbf{E}(X_0)$. We can already conclude from these calculations the intuitive fact that

1. If $\mu < 1$, then the average population tends to extinction.
2. If $\mu = 1$, the average population is maintained.
3. If $\mu > 1$, the average population becomes unbounded.

It shall turn out that extinction is guarenteed even in the case that $\mu = 1$.

Regardless of your average population growth, provided $p_0 > 0$ there is a chance that the population will eventually become extinct.The problem in this section will be in deriving this probability in terms of the reproduction probabilities. Let $a_n(k)$ denote the probability of extinction after n generations given that we start with k individuals. Then, as we have derived above, the possibility of general extinction from k individuals is $a(k) = \lim_{n\to\infty} a_n(k)$. Since all k branches of the population act independantly, we have $a_n(k) = a_n(1)^k$, and it suffices to determine $a(1)$, which we shall denote by a. If we look one generation ahead, then

$$a = \mathbf{P}(\text{extinction}|X_0 = 1) \tag{3.5}$$

$$= \sum_{k=0}^{\infty} \mathbf{P}(X_1 = k|X_0 = 1)\mathbf{P}(\text{extinction}|X_1 = k) = \sum_{k=0}^{\infty} p_k a^k \tag{3.6}$$

Thus $a = \varphi(a)$, where φ is the generating function of X_1, assuming $X_0 = 1$. Since $z > 0$ implies

$$\varphi'(z) = \sum_{k=0}^{\infty} k p_k z^{k-1} > 0$$

We know that φ is monotonically increasing on the positive numbers. If we let φ_n be the generating function of X_n, then

$$\varphi_n(z) = \sum_{k=0}^{\infty} \mathbf{P}(X_n = k) z^k = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \mathbf{P}(X_1 = j)\mathbf{P}(X_n = k|X_1 = j) z^k$$

$$= \sum_{j=0}^{\infty} p_j \sum_{k=0}^{\infty} \mathbf{P}(X_n = k|X_1 = j) z^k = \sum_{j=0}^{\infty} p_j \sum_{k=0}^{\infty} \mathbf{P}(X_{n-1} = k|X_0 = j) z^k$$

Now $\mathbf{P}(X_{n-1}|X_0 = j)$ is the distribution of the sum of j independent random variables $Y_1, \ldots, Y_j \sim \mathbf{P}(X_{n-1}|X_0 = 1)$, so that each has a generating function φ_{n-1}, and so, continuing the calculation

$$\varphi_n(z) = \sum_{j=0}^{\infty} p_j \mathbf{E}(z^{X_{n-1}}|X_0 = j) = \sum_{j=0}^{\infty} p_j \mathbf{E}(z^{X_{n-1}}|X_0 = 1)^j \tag{3.7}$$

$$= \sum_{j=0}^{\infty} p_j \varphi_{n-1}(z)^j = \varphi(\varphi_{n-1}(j)) \tag{3.8}$$

Using (3.8), we can recursively determine $a_n(1) = \mathbf{P}(X_n = 0|X_0 = 1) = \varphi_n(0)$. We are now ready to show that a is the smallest positive root of the equation $x = \varphi(x)$, assuming $a \neq 0$. Using (3.6), we know a is a root of this equation. Let $\hat{a}$ denote the least such positive root of the equation. We will verify that $a_n(1) \leqslant \hat{a}$. Certainly $a_0(1) = 0 \leqslant \hat{a}$. If $a_{n-1}(1) \leqslant \hat{a}$, then

$$a_n(1) = \varphi(\varphi_{n-1}(0)) = \varphi(a_{n-1}(1)) \leqslant \varphi(\hat{a}) = \hat{a} \tag{3.9}$$

Taking limits on (3.9), we find $a \leqslant \hat{a}$. Thus, assuming $a > 0$, equality is obtained. We have deduced that

Theorem 3.1. *If $X_1, X_2, \ldots$ is a branching process with reproduction probabilities p, and if $p_0 > 0$, then the extinction probability is the smallest positive root of $\varphi(z) = z$, where $\varphi(z) = \sum p_k z^k$.*

Let's do some examples. Let $p_0 = 1/4$, and $p_2 = 3/4$. The extinction probability is the smallest positive root of $z = 1/4 + 3/4z^2$, which can be calculated by the quadratic formula to be $a = 1/3$. Let $p_0 = 1/2$, $p_1 = 1/4$, and $p_2 = 1/4$. The extinction probability is the smallest root of $z = 1/2 + 1/4z + 1/4z^2$. The roots of this equation are 1 and 2. Hence $a = 1$. We could have seen this from noticing the mean growth rate μ is less than one.

When exactly is the population almost surely going to extinction ($a = 1$)? Suppose $\mu = 1$. Then, for any $0 < z < 1$,

$$1 - \varphi(z) = \int_z^1 \varphi'(x)dx < \int_z^1 \varphi'(1) = \mu(1 - z) = 1 - z \tag{3.10}$$

Here we have used the monoticity of φ. From (3.10), we conclude $z < \varphi(z)$ on $[0, 1)$. Assuming extinction is possible, we obtain the surprising result

that extinction is guaranteed if the average number of descendents is less than or equal to one! Since the average population remains to be 1 though, we know that, if we are not extinct after a long time, then we will probably have a large population.

Chinese surnames are ancient. Applying our model, we see that the names that have survived over the generations should be very prominant. There are approximately 3,000 Chinese last names in use nowadays, as compared to 12,000 in the past, even though there are far more Chinese people in the world than in the past. This is the reason Gatson and Walton concluded upper class surnames were going extinct in Victorian Britain. The elite few who had these names were in populations that were likely to die out very soon, whereas the common names are names which will last much longer.

3.4 Reversibility

Some Markov chains have a certain symmetry which enables us to easily understand them. If we watch the markov chain as it proceeds from state to state, it forms a kind of 'movie'. A Markov chain is reversible if the markov chain has the same probability distribution when we watch the movie backwards. That is, if $X_0, X_1, \ldots, X_n$ are the first few frames of the movie, then $(X_0, \ldots, X_n)$ is distributed identically to $(X_n, \ldots, X_0)$. We have

$$\begin{aligned}\mu_0(x_0)P(x_0,x_1)\ldots P(x_{n-1},x_n) &= \mathbf{P}(X_0 = x_0, \ldots, X_n = x_n) \\ &= \mathbf{P}(X_0 = x_n, \ldots, X_n = x_n) = \mu_0(x_n)P(x_n,x_{n-1})\ldots P(x_1,x_0)\end{aligned}$$

Normally, being pairwise identically distributed is not enough to determine the independence of a larger family of variables. Nonetheless, in a homogenous markov chain, we need only verify the chain for pairs.

Definition. A Markov chain is **reversible** if there is a measure μ (which need not be a probability distribution) for which, for any two states $x, y \in \mathcal{S}$,

$$\mu(x)P(x,y) = \mu(y)P(y,x)$$

It follows that, if μ is a probability distribution, then $(X_0, X_1, \ldots, X_n)$ is identically distributed to $(X_n, \ldots, X_0)$, given that μ is the initial distribution of the chain.

Example. *Any symmetric markov chain (with $P(x,y) = P(y,x)$) is reversible, with $\mu(x) = 1$ for all x.*

Example. *Consider a random walk on a graph $G = (V, E)$. Let $\mu(x) = deg(x)$. Then*

$$\mu(x)P(x,y) = 1 = \mu(y)P(y,x)$$

So the walk is reversible with respect to μ.

Now let μ_0 be a reversible measure generating a reversible markov chain $\{X_t\}$. Suppose we watch a markov chain $(X_0, \ldots, X_N)$ for a really large N. Then, if a limiting distribution exists, it mustn't be too different from μ_N. If we watch the markov chain backwards $(X_N, \ldots, X_0)$, then it is equal in distribution by the properties of a markov chain. In particular, this means that the distribution of μ_0 is also the result of watching a Markov chain for a really long time – so we should expect μ_0 to be really close to the limiting distribution of the markov chain. In fact, since $\mu(x)P(x,y) = \mu(y)P(y,x)$, we have

$$\mu(x) = \sum_y \mu(x)P(x,y) = \sum_y \mu(y)P(y,x) = (\mu P)(x)$$

So μ is an invariant distribution, and is the convergent probability distribution on an ergodic markov chain.

In the past few chapters, we have thoroughly addressed the problem of finding the limiting distribution of a stochastic process. We now address the converse problem. We are given an invariant measure, and we must construct a markov process which has this invariant measure for an invariant distribution. This is useful for approximating the invariant distribution when it is computationally too difficult to calculate.

For instance, consider the set of all $N \times N$ matrices with entries in $\{0,1\}$. We may assign the uniform distribution to these matrices. There are 2^{N^2} different matrices of this form, so the probability of any matrix being picked is $1/2^{N^2}$. What about if we consider the set $\mathcal{T}$ of all matrices such that no two entries of the matrix are one at the same time. At face-value, there is no immediate formula we may use to count these matrices. Nonetheless, if we construct a markov chain whose limiting distribution is the uniform distribution, we can approximate the number of matrices by simulation – we just count the average number of times a matrix is visited out of a certain number of trials.

Consider a markov chain with the following transition. We start with an initial matrix X_0 in $\mathcal{T}$, and we pick a random entry (i,j). Let Y be the matrix resultant from flipping the X_{ij} on or off. If $Y \in \mathcal{T}$, let $X_1 = Y$. Otherwise, let $X_1 = X$. Continue this process indefinitely. This is an irreducible, symmetric markov process in $\mathcal{T}$, with transitions

$$P(A,B) = \begin{cases} \frac{1}{N^2} & : A \text{ and } B \text{ differ by one entry} \\ 1 - \sum_{C \neq A} P(A,C) & : B = A \\ 0 & : \text{elsewhere} \end{cases}$$

Since the Markov chain is symmetric, the distribution converges to the uniform distribution on all of $\mathcal{T}$ – and we can use this to attempt to determine the distribution on the set.

How do we simulate a Markov chain? We shall accept that a computer is able to generate psuedorandom numbers distributed uniformly on any finite state space and on an interval $[0,1]$. A **random mapping representation** of a markov chain $\{X_i\}$ is a function $f : \mathcal{S} \times \Lambda \to \mathcal{S}$ together with a Λ-valued random variable Z for which

$$\mathbf{P}(X_{n+1} = x_{n+1} | X_n = x) = \mathbf{P}(f(x,Z) \in x_{n+1})$$

If we generate a sequence $Z_1, \ldots, \infty$ of random variables independant and identically distributed to Z, Then $X_{n+1} = f(X_n, Z_{n+1})$. Conversely, we can use a random mapping representation to generate a markov chain.

There is a general method by which we can construct a markov chain to converge to a distribution. Suppose we have a distribution β defined on a state space $\mathcal{S}$, with $\sum \beta(x) = B < \infty$. In addition, assume that we already have a symmetric state transition set P. We shall use this state to generate

a new process. Define a Markov chain with probabilities

$$P'(x,y) = P(x,y)\,\min(1, \frac{\beta(y)}{\beta(x)}) \qquad x \neq y$$

$$P'(x,x) = 1 - \sum_{y \neq x} P'(x,y)$$

We 'slow' down the chain at certain points to make it reversible with respect to β, and hence converges to $\mu = \beta/B$. This is the Metropolis-Hastings algorithm for computing a distribution β up to a multiplicative constant. It is important that the algorithm only depends on the ratios of β. Frequently, β is of the form $h(x)/Z$ for some very large normalizing constant Z. Because the algorithm only depends on the ratios, we do not needs to calculate β at all.

Chapter 4

Martingales

4.1 Modern Conditional Expectation

Most of the theory of random processes is connected with understanding the relationship of the values of the process at different times in the theory. When studying Markov chains, we tried to understand the relationship between random variables by considering directly the processes' transition coefficients. We now understand random processes by looking at how the evolution of a stochastic process changes as we fix the state at certain time periods. The primary tools in our analysis will be **conditional probabilities** and **conditional expectations**, which allow us to quantify how the distribution of a random variable changes when we fix the value of another random variable.

For discrete random variables X and Y, we can calculate the conditional probabilities and expectations by

$$\mathbf{P}(Y = y|X = x) = \frac{\mathbf{P}(Y = y, X = x)}{\mathbf{P}(X = x)} \qquad \mathbf{E}(Y|X = x) = \sum_y y\mathbf{P}(Y = y|X = x)$$

whenever $\mathbf{P}(X = x) \neq 0$, and for continuous random variables,

$$\mathbf{P}(Y \in A|X = x) = \int_A \frac{f_{Y,X}(y,x)}{f_X(x)}dy \qquad \mathbf{E}(Y|X = x) = \int y\frac{f_{Y,X}(y,x)}{f_X(x)}dy$$

where $f_X(x) \neq 0$ is assumed. However, these two classical formulations are insufficient to cover condition expectations for the general random variables we will investigate in the study of stochastic processes. Kolmogorov,

one of the founders of measure theoretic probability theory, found the most elegant way to define $\mathbf{P}(Y \in A|X = x)$ and $\mathbf{E}(Y|X = x)$ for the random variables we're interested in. We note that we will focus only on defining the conditional *expectations* $\mathbf{E}(Y|X = x)$, because we can then define $\mathbf{P}(Y \in A|X = x) = \mathbf{E}(\mathbf{I}(Y \in A)|X = x)$.

The trick to Kolmogorov's method is to think of conditional values as 'best guesses' of the values given some information about the system is known. Our first revelation is to think of $\mathbf{E}(Y|X)$ as a random variable on the same sample space as X and Y, taking value $\mathbf{E}(Y|X = X(\omega))$ on input $\omega \in \Omega$. In both classical definitions, the conditional expectation possesses two important properties:

- $\mathbf{E}(Y|X)$ is a function of the random variable X. That is, we only need to know $X(\omega)$ to predict the value $\mathbf{E}(Y|X)(\omega)$.
- For any subset of the sample space of the form $B = X^{-1}(A)$, where $A \subset \mathbf{R}$ is a set of real values, we have a measure theoretic integral

$$\int_B \mathbf{E}(Y|X)d\mathbf{P} = \int_B Y d\mathbf{P}$$

For discrete random variables, this equation takes the form

$$\sum_{a \in A} \mathbf{P}(X = a)\mathbf{E}(Y|X = a) = \sum_{a \in A} \sum y\mathbf{P}(X = a, Y = y)$$

and for continuous random variables,

$$\int_A f_X(x)\mathbf{E}(Y|X = x)\ dx = \int_A \int y f_{X,Y}(x, y)\ dxdy$$

Note, in particular, that the equation for discrete random variables uniquely defines the conditional expectation whenever $\mathbf{P}(X = a) \neq 0$ by taking $A = \{a\}$. In the case of continuous random variables, we can only conclude that $f_Y(y)\mathbf{E}(X|Y = y) = \int x f_{X,Y}(x, y)\ dx$ holds almost everywhere, and this defines $\mathbf{E}(X|Y = y)$ up to a set of measure zero if we ignore the values y' where $f_Y(y') = 0$. If we are able to choose $\mathbf{E}(X|Y = y)$ to be continuous as a function of y, then it is the unique continuous function satisfying the integral.

In general, we say a general random variable $\mathbf{E}(X|Y)$ is a *version* (it is not unique) of a **conditional expectation** of X with respect to Y if the two conditions above are satisfied. However, we note that the definition can certainly be simplified by noting that once we consider $\mathbf{E}(X|Y)$ as a function on the sample space rather than on values of Y, the actual values of Y are not actually important to the definition of conditional expectation, but rather the ways the values spread out over the sample space. The integration condition only depends on the subsets of $\sigma(X)$ that we use. It is a theorem of Doob that if these random variables are real valued, this is equivalent to being able to write $Y = f(X)$, for some function f (the theorem is true for more general spaces, just not for arbitrary spaces).

Lemma 4.1. *A real-valued random variable Y is $\sigma(X)$ measurable if and only if $Y = f(X)$ for some Borel-measurable $f : \mathbf{R} \to \mathbf{R}$.*

Proof. Suppose that Y is simple, with $Y(\omega) = \sum a_i \mathbf{I}(\omega \in A_i)$. Then Y is $\sigma(X)$ measurable if and only if $A_i = X^{-1}(B_i)$ for some Borel measurable sets B_i. In this case, the B_i are disjoint, and we can define a simple measurable function $f(t) = \sum a_i \mathbf{I}(t \in B_i)$ with $f(X) = Y$. Now if Y is a general random variable, let Y be the pointwise limit of simple $\sigma(X)$ random variables $Y_n = f_n(X)$, where the f_n are Borel measurable, and can be chosen to be equal to zero outside the range of X. Then, the f_n obviously converge pointwise outside the range of X, and for any sample point ω,

$$\lim_{n\to\infty} f_n(X(\omega)) = \lim_{n\to\infty} Y_n(\omega) = Y(\omega)$$

so the f_n converge pointwise on the range of X, to some measurable function f with $Y = f(X)$. □

The intuitive way we should think about the conditional values is as a 'best guess' of the probability values given that some information is known about the system at some time. If we think of the information about the random variable X being given by the σ-algebra $\sigma(X)$, then it isn't too much to consider arbitrary 'sets of information' about a probability space to be a sub σ-algebra of the σ-algebra defining the space. We say a random variable X is **adapted** to a σ-algebra Σ if X is measurable with respect to Σ. A random variable Y is adapted with respect to X if Y is adapted to the σ algebra $\sigma(X)$ generated by X. With all this terminology, we can now make the general definition. For a given Σ algebra and random variable X, a Σ measurable random variable $\mathbf{E}(X|\Sigma)$ is a version of a

conditional expectation if $\int_S \mathbf{E}(X|\Sigma) = \int_S X$ holds for any subset $S \subset \Sigma$. Despite the technical definition, the existence and uniqueness of conditional expectations is relatively easy to prove for any finitely integrable random variable.

Theorem 4.2. *If $\|X\|_1 < \infty$, then $\mathbf{E}(X|\Sigma)$ exists, $\|\mathbf{E}(X|\Sigma)\|_1 < \infty$, and is unique up to a set of measure zero.*

Proof. First, assume $X \geqslant 0$. Then the map $S \mapsto \int_S X d\mathbf{P}$ is a *finite measure* over Σ which is absolutely continuous with respect to $\mathbf{P}$, hence by the Radon-Nikodym theorem there exists a Σ measurable function Y such that for any set S, $\int_S Y d\mathbf{P} = \int_S X d\mathbf{P}$. Now it is easy to see that since $X \geqslant 0$, $Y \geqslant 0$ almost surely, and so

$$\|Y\|_1 = \int |Y| d\mathbf{P} = \int Y d\mathbf{P} = \int X d\mathbf{P} = \|X\|_1 < \infty$$

To see uniqueness, note that if Y_0 and Y_1 are both conditional expectations for X, then for any set $S \in \Sigma$, $\int_S (Y_0 - Y_1) = \int_S (X - X) = 0$, and this implies $Y_0 = Y_1$ almost surely. □

Note that since $\mathbf{E}(X|\Sigma)$ is unique up to a set of measure zero, and if X and Y agree almost everywhere, then $\mathbf{E}(X|\Sigma) = \mathbf{E}(Y|\Sigma)$, and so $\mathbf{E}(\cdot|\Sigma)$ can be viewed as an operator on $L^1(\Omega)$. By elementary linearity properties of the integral

- $\mathbf{E}(aX + bY|\Sigma) = a\mathbf{E}(X|\Sigma) + b\mathbf{E}(Y|\Sigma)$
- $\|\mathbf{E}(X|\Sigma)\|_1 = \|X\|_1$.
- $\mathbf{E}(\mathbf{E}(X|\Sigma)) = \mathbf{E}(X)$, and if $\Gamma \subset \Sigma$, $\mathbf{E}(\mathbf{E}(X|\Sigma)|\Gamma) = \mathbf{E}(X|\Gamma)$.

Since conditional expectations are defined with respect to the Lebesgue integral, we also get convergence results based on the main results of integration theory.

- (Monotone Convergence) If $0 \leqslant X_1 \leqslant X_2 \cdots \to X$, then $\mathbf{E}(X_i|\Sigma)$ converge pointwise monotonely almost everywhere to $\mathbf{E}(X|\Sigma)$.
- (Fatou) If $X_n \geqslant 0$ then $\mathbf{E}((\liminf X_n)|\Sigma) \leqslant \liminf \mathbf{E}(X_n|\Sigma)$ a.s.
- (Dominated Convergence) If $|X_n| \leqslant Y$, $\int Y < \infty$, and $X_n \to X$ pointwise a.s, then $\mathbf{E}(X_n|\Sigma) \to \mathbf{E}(X|\Sigma)$ pointwise almost surely.

- (Jensen) If f is a convex L^1 function then $\mathbf{E}(f(X)|\Sigma) \geqslant f(\mathbf{E}(X|\Sigma))$ almost surely. An important corollary of this is that $\|\mathbf{E}(X|\Sigma)\|_p \leqslant \|X\|_p$ for $p \geqslant 1$.

On another interesting thread, we can also prove the existence of conditional expectations for L^2 measurable functions, and we also note that $\mathbf{E}(X|\Sigma)$ is the best approximation of X is the square mean, which explains why conditional expectations occur so often in applications to statistics.

Theorem 4.3. *If $\|X\|_2 < \infty$, then $\|\mathbf{E}(X|\Sigma)\|_2 < \infty$, and is the orthogonal projection of X onto the subspace of L^2 functions which are Σ measurable.*

Proof. If we let $\mathbf{E}(X|\Sigma)$ be the orthogonal projection of X onto the subspace of L^2 functions which are Σ measurable, then orthogonality implies that for any Σ measurable function Y,

$$\int Y[\mathbf{E}(X|\Sigma) - X] = 0$$

which can be rewritten as

$$\int Y\mathbf{E}(X|\Sigma) = \int YX$$

and letting Y be an indicator function over some element of Σ, we obtain easily that $\mathbf{E}(X|\Sigma)$ satisfies the properties of a conditional expectation, hence we have proven that the conditional expectation is square integrable. □

If X is already Σ measurable, then $\mathbf{E}(X|\Sigma) = X$. In particular, $\mathbf{E}(X|X) = X$. More generally, if X is Σ measurable, then $\mathbf{E}(XY|\Sigma) = X\mathbf{E}(Y|\Sigma)$, which follows from the next lemma.

Lemma 4.4. *For any Σ-measurable function Y, and $A \subset \Sigma$ for which the equation*

$$\int_A Y\mathbf{E}(X|\Sigma) = \int_A YX$$

makes sense, the equation holds.

Proof. By linearity, and splitting X into its positive and negative parts, we may assume $X \geqslant 0$, and $\mathbf{E}(X|\Sigma) \geqslant 0$ a.e. The lemma is true if Y is an

indicator function by definition, and so by linearity, the lemma is also true if Y is a simple function. Then we can prove the theorem if $Y \geqslant 0$ by the monotone convergence theorem, and then for all Y by splitting Y into its positive and negative parts. □

We say two sigma algebras Σ and Δ are independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, for any $A \in \Sigma$, $B \in \Delta$. This generalizes to an arbitrary family of σ algebras (where we take arbitrary intersections), and a family of random variables $\{X_\alpha\}$ are independent if and only if $\{\sigma(X_\alpha)\}$ is independent. X and Y are independent with respect to Σ if $\sigma(X) \cap \Sigma$ is independent of $\sigma(Y) \cap \Sigma$. If X and Y are independant with respect to Σ, then $\mathbf{E}(XY|\Sigma) = \mathbf{E}(X|\Sigma)\mathbf{E}(Y|\Sigma)$. Most importantly, if X is independent with respect to Σ, then $\mathbf{E}(X|\Sigma) = \mathbf{E}(X)$.

As a stochastic process evolves, we gain more and more information about the future of the process. This also needs to be modelled in a measure theoretic manner. An evolving system of information is known as a **filtration**. Let T be a linearly ordered set. A **filtration** is an increasing family of σ algebras. Precisely, a family $\{\mathcal{F}_t\}$ of σ algebras with respect to T is a **filtration** if $\mathcal{F}_t \subset \mathcal{F}_u$ for $t < u$. It is the measure theoretic equivalent of a topological filter.

4.2 Martingales

We wish to discuss a markov chain which represents a 'fair bet'. Each point of time in the process represents the amount of money in a gambler's pocket. if we have a certain amount of money at a time, and we watch the process evolve. We should expect us to find the same amount of money with us on average. This is where our new definition of expectation comes into play. A **martingale** with respect to a filtration $\{\mathcal{F}_t\}$ is a process $\{X_t\}$, with each $X_t \in L_1(\mathbf{P})$ such that, for $s < t$, $\mathbf{E}(X_t|\mathcal{F}_s) = X_s$. Normally, we will assume the filtration is $\mathcal{F}_t = \sigma(X_s : s \leqslant t)$.

Example. *Let $\{X_t\}$ be a sequence of independent, identically distributed random variables in $L_1(\mathbf{P})$ with average 0. Define $S_n = \sum X_k$ to be the average*

value of X_t. Then

$$\begin{aligned}\mathbf{E}(S_n|S_k,\ldots,S_0) &= \sum_{i\leqslant k}\mathbf{E}(X_i|S_0,\ldots,S_k)+\sum_{k>i}\mathbf{E}(X_i|S_k,\ldots,S_0)\\ &= \sum_{i\leqslant k}\mathbf{E}(S_i-S_{i-1}|S_0,\ldots,S_k)+\sum_{k>i}\mathbf{E}(X_i|S_k,\ldots,S_0)\\ &= S_k+\sum_{k>i}\mathbf{E}(X_i)=S_k\end{aligned}$$

Each S_k is in L_1, so the sequence S_k is a martingale. In general, if X_t has mean μ, the $S_n-n\mu$ is a martingale – the fairness of the bet is tipped on one side, so we need the other factor to rebalance it.

Example. *Let us consider an actual betting example. Let us flip a fair coin, determining a sequence of independant and identically distributed Bernoulli random variables $\{X_n\}$, taking values in $\{\pm 1\}$. At each point in time, we make a bet W_n, in L_1 and adapted to $X_0,\ldots,X_{n-1}$. The stochastic process we now observe is the accumlation of our winnings*

$$M_n=\sum_{k=0}^{n}W_kX_k$$

Since

$$\mathbf{E}(M_n|M_s,\ldots,M_0)=\sum_{k=0}^{n}\mathbf{E}(W_kX_k|M_s,\ldots,M_0)=\sum_{k=0}^{s}W_kX_k+\sum_{k=s+1}^{n}W_k\mathbf{E}(X_k)=M_s$$

Thus M_n is a martingale.

For discrete martingales, we need only verify that $\mathbf{E}(M_n|M_{n-1})=M_{n-1}$, since then $\mathbf{E}(M_n|M_{n-2})=\mathbf{E}(\mathbf{E}(M_n|M_{n-1},M_{n-2}),M_{n-2})=\mathbf{E}(M_{n-1}|M_{n-2})=M_{n-2}$, and so on.

Example. *Consider the Polya urn model of the process. We start with one white ball and one black ball in an urn. At each time epoch, we draw a random ball from the urn, and put the ball back in addition to another ball of the same colour. Let X_k be the number of white balls after drawing k balls, and let $M_n=X_n/(n+2)$ be the relative proportion of white balls in the urn at a certain time. Then M_n is bounded, hence in L_1, and*

$$\mathbf{E}(M_n|M_{n-1})=\frac{\mathbf{E}(X_n|X_{n-1})}{n+2}=\frac{1}{n+2}[X_{n-1}+\frac{X_{n-1}}{n+1}]=\frac{X_{n-1}}{n-1}=M_{n-1}$$

We are therefore observing a Martingale.

4.3 The Optional Sampling Theorem

Probability was created to analyze gambling games, to calculate the manner in which one may succeed. In the 18th century, a strategy was discovered which could be applied to 'beat' certain gambling games. It became known as the martingale. The idea is simple – consider the martingale $M_n = \sum_{k=0}^{n} W_k X_k$ described above. Let $W_k = 2^k$. Let $\tau = \min\{k : X_k = 1\}$. Then

$$M_\tau = 2^\tau - \sum_{k=1}^{\tau-1} 2^k = 2^\tau - (2^\tau - 1) = 1$$

So we always come out of the bet with a profit on 1 unit of money, if we follow this strategy. The Martingale became all the rage in the 18th century. Casanova was one of many figures known to apply the strategy. The key problem with the strategy is that is that it assumes one is able to bet inevitably over and over again – we assume we have an infinite amount of money to gamble with. When we have a finite amount of money, we're running a gambler's ruin – we either bet until we run out of money, or gain a single unit of money. Thus the strategy is somewhat risky.

The martingale is one of the reasons casinos now put limits on the table. You can't bet an unbounded amount of money anymore. Effectively, the casino restricts the martingales you play on to a restricted family, upon which the martingale doesn't work anymore. The optional sampling theorem shows that no stopping time will enable us to gain an average profit. In fact, we will always end up with the same amount of money we started off with on average. First, a rigorous definition of a stopping time.

Definition. A stopping time with respect to a process $\{X_t\}$ and filtration $\mathcal{F}_t$ is a random variable τ mapping into time, such that for any time t, $\{\omega : \tau(\omega) \leqslant t\} \in \mathcal{F}_t$.

Lemma 4.5. *If $\{M_0, M_1, \dots\}$ is a discrete martingale with respect to $\mathcal{F}_n$, and τ is a bounded stopping time, then $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

Proof. Let $\tau \leqslant K$. Then we may write $M_\tau = \sum_{k=0}^{K} \mathbf{I}(\tau = k) M_k$. Then $\tau = K$ is measurable with respect ot $\mathcal{F}_{K-1}$, since the event $\tau = K$ is the same as

$\tau > K-1$

$$\mathbf{E}(M_\tau|\mathcal{F}_{K-1}) = \sum_{k=0}^{K} \mathbf{E}(\mathbf{I}(\tau=k)M_k|\mathcal{F}_{K-1}) = \sum_{k=0}^{K-2} \mathbf{I}(\tau=k)M_k + \mathbf{I}(\tau \geqslant K-1)M_{n-1}$$

Repeating this proof on $\mathcal{F}_{K-2}$, we obtain that $\mathbf{E}(M_\tau|\mathcal{F}_{K-2}) = \sum_{k=0}^{K-3} M_k + \mathbf{I}(\tau \geqslant K-2)M_{n-2}$. By induction, we determine that

$$\mathbf{E}(M_\tau|\mathcal{F}_0) = M_0$$

and therefore, by iterated expectation,

$$\mathbf{E}(M_\tau) = \mathbf{E}\mathbf{E}(M_\tau|\mathcal{F}_0) = \mathbf{E}(M_0)$$

□

We would like to conclude the same theorem for unbounded stopping times. Our idea is to approximate a proper stopping time by bounded stopping times. Let τ be a stopping time, and define $\tau_n = \min\{\tau, n\}$. We may write

$$M_\tau = M_{\tau_n} + \mathbf{I}(\tau > n)M_\tau - \mathbf{I}(\tau > n)M_n$$

We would hope that the middle two factors do not contribute much to the process for large n. If $M_{\tau_n} \to M_\tau$ as $n \to \infty$, then $\mathbf{E}(M_0) = \mathbf{E}(M_{\tau_n}) \to \mathbf{E}(M_\tau)$. To obtain this, we require that the extraneous factors on the right vanish in expectation. The first extraneous factor is easy to remove, if $\mathbf{E}(|M_\tau|) < \infty$, then $\mathbf{E}(\mathbf{I}(\tau > n)M_\tau) \to 0$. The second factor dissapears if

$$\lim_{n\to\infty} \mathbf{E}(|M_n|\mathbf{I}(\tau > n)) = 0$$

and this provides the conditions for our theorem to hold.

Theorem 4.6 (Optional Stopping)**.** *If $\{M_n\}$ is a martingale, and τ a stopping time for which $M_\tau \in L_1(\mathbf{P})$, $\lim_{n\to\infty} \mathbf{E}(|M_n|\mathbf{I}(\tau > n)) = 0$, and $\mathbf{P}(\tau < \infty) = 1$. The $\mathbf{E}(M_\tau) = \mathbf{E}(M_0)$.*

Example. *Consider the fair Gambler's ruin problem, where we start with M units of money, and play a fair game until we go bust, or we end up with N units of money. Let $\tau = \min\{X_k : X_k = 0 \text{ or } X_k = N\}$. Since X_n is bounded, and $\mathbf{P}(\tau < \infty) = 1$, this process satisfies the optional stopping theorem, so*

$$N\mathbf{P}(X_\tau = N) = \mathbf{E}(M_\tau) = \mathbf{E}(X_0) = M$$

And therefore $\mathbf{P}(X_\tau = N) = M/N$.

There is a much easier condition, which allows us to conclude the consequences of the optional stopping theorem. Call a family $\mathcal{C}$ of random variables **uniformly integrable** if, for any ε, there is K such that for all $X \in \mathcal{C}$, $\mathbf{E}(|X|\mathbf{I}(X > K)) < \varepsilon$.

Lemma 4.7. *If $\{X_n\}$ is a family of random variables for which these is C where $X_n^2 < C$, then the X_n are uniformly integrable.*

Now let $\{M_n\}$ be a uniformly integrable martingale, and τ a stopping time for which $\mathbf{P}(\tau < \infty) = 1$. Then $\lim_{n\to\infty} \mathbf{E}(|M_n|\mathbf{I}(\tau > n)) = 0$, since $\mathbf{P}(\tau > n) \to 0$. Thus the optional stopping theorem holds for the M_n and τ, provided $\mathbf{E}(|M_\tau|) < \infty$.

4.4 Martingale Convergence

Definition. A **submartingale** $\{M_0, M_1, \dots\}$ with respect to a filtration $\{\mathcal{F}_k\}$ is a process such that for $n < m$, $\mathbf{E}(M_m|\mathcal{F}_n) \geqslant M_n$. A **supermartingale** satisfies $\mathbf{E}(M_m|\mathcal{F}_n) \leqslant M_n$.

Theorem 4.8. *If M_n is a submartingale, for which there is C such that $\mathbf{E}(max(M_n, 0)) \leqslant C < \infty$, then there is $M_\infty \in L_1(\mathbf{P})$ suc that $M_n \to M_\infty$ almost surely,*

Chapter 5

Continuous Markov Processes

In some mathematical circumstances, we may approximate a continuous system by a simpler system, which enables us to derive approximate results more simply. For instance, we often replace a Newtonian system by its linear approximation, which enables us to use the fleshed-out theory of linear differential equations to obtain an analytic formula for how the system develops. Nonetheless, in some mathematical systems it is worthwhile keeping a continuous system, which leads to more precise and concise results.

In the last chapter, we considered a discrete-time queue, with individuals arriving and exiting at each separate time epoch. In this chapter, we will extend this model to a real-time queue, with individuals arriving and exiting at separate moments occuring at any real time-epoch.

5.1 Poisson Processes

Our first trick to modelling a real-time queueing system $\{Y_t\}$ is to split the queue into two parts, $Y_t = X_t - Z_t$. The first split, X_t, is a counter, which tells us how many people in total have ever entered the queue. The second part, Z_t, tells us how many people in total have left the queue. By understanding these processes separately, we can understand Y_t.

What assumptions do we make about the 'counting process' $\{Y_t\}_{t\in[0,\infty)}$. Firstly, the counter shuold be increasing: the total number of people who have entered the store should not decrease over time. Secondly, to simplify things, we shall assume that the average number of customers arriving is

constant, and that the number of customers arriving at disjoint intervals are independant of one another. This is a Poisson process.

Definition. A stochastic process $\{X_t\}$ valued in $\mathbf{N}$ is Poisson with arrival length $\lambda > 0$ if:

1. $X_0 = 0$, and $i \leqslant j$ implies $X_i \leqslant X_j$.
2. Disjoint intervals (i_k, j_k) have independant differences $X_{j_k} - X_{i_k}$, and if $i \leqslant j$, then $X_j - X_i$ is equal in distribution to X_{j-i}.
3. The Process satisfies the equations

$$\mathbf{P}(X_t = 1) = \lambda \Delta t + o(t) \tag{5.1}$$
$$\mathbf{P}(X_t = 0) = 1 - \lambda \Delta t + o(t) \tag{5.2}$$
$$\mathbf{P}(X_t > 1) = o(t) \tag{5.3}$$

These axioms determine a unique probability distribution. Define $P_k(t) = \mathbf{P}(X_t = k)$. We have $P_0(0) = 1$, and $P_k(0) = 0$ for $k > 0$. Then

$$\begin{aligned}
P_k(t + \Delta t) &= \mathbf{P}(X_{t+\Delta t} = k, X_t = k) \\
&\quad + \mathbf{P}(X_{t+\Delta t} = k, X_t = k - 1) \\
&\quad + \mathbf{P}(X_{t+\Delta t} = k, X_t < k - 1) \\
&= \mathbf{P}(X_{t+\Delta t} - X_t = 0, X_t - X_0 = k) \\
&\quad + \mathbf{P}(X_{t+\Delta t} - X_t = 1, X_t - X_0 = k - 1) \\
&\quad + \mathbf{P}(X_{t+\Delta t} = k, X_t - X_0 < k - 1) \\
&= P_0(\Delta t) P_k(t) + P_1(\Delta t) P_{k-1}(t) + o(\Delta t) \\
&= P_k(t) - \lambda \Delta t P_k(t) + \lambda \Delta t P_{k-1}(t) + o(\Delta t)
\end{aligned}$$

It therefore follows that $P_k' = \lambda(P_{k-1} - P_k)$. This is just an ordinary differential equation. Altering the derivation above, noting we only need the first term for $k = 0$, we have

$$P_0' = -\lambda P_0 \qquad P_0(0) = 1$$

So $P_0(t) = e^{-\lambda t}$. We shall now show that $P_k(t) = t^k/k!e^{-\lambda t}$. Define $f_k(t) = P_k(t)e^{\lambda t}$ (so that, if our theorem is true $f_k(t) = t^k/k!$). We have

$$f_k'(t) = \lambda P_k(t)e^{\lambda t} + \lambda(P_{k-1}(t) - P_k(t))e^{\lambda t} = P_{k-1}e^{\lambda t} = f_{k-1}'(t)$$

And it follows that $f_k(t) = t^k/k!$, since $f_k(0) = P_k(0) = 0$.The Poisson distribution Poisson(k, λ) is just the distribution of P_k.

Another natural way to understand Poisson processes is by directly studying the discrete timepoints at which the counter of the process increments. Fix a Poisson process $\{X_t\}$, and define a stopping time $\tau_k = \inf\{t : X_t \geqslant k\}$. Since X_t is monotonic, this variable is well-defined. The variables $\tau_{k+1} - \tau_k$ should be independent and identically distributed, and the τ_k should satisfy the 'memory loss' property

$$\mathbf{P}(\tau_{k+1} - \tau_k \geqslant s + t | \tau_{k+1} - \tau_k \geqslant t) = \mathbf{P}(T_k \geqslant s)$$

The only left-continuous non-zero real-valued functions f which satisfies $f(s + t) = f(s)f(t)$ are the family of exponential functions $f(t) = e^{-\lambda t}$. Hence any variables $\{\tau_k\}$ satisfying the properties above have $\mathbf{P}(\tau_{k+1} - \tau_k \geqslant t) = \mathbf{P}(\tau_1 \geqslant t) = e^{-\lambda t}$ for some λ.

Given any variables τ_k satisfying the assumptions above, define $X_t = \inf\{k : \tau_k \geqslant t\}$. Then $X_0 = 0$, $\{X_t\}$ is increasing, and

$$\mathbf{P}(X_t = 1) = \mathbf{P}(\inf\{k : \tau_k \geqslant t\} = 1) = \mathbf{P}(\tau_1 \leqslant t) = 1 - e^{-\lambda t} = \lambda t + o(t)$$

$$\mathbf{P}(X_t = 0) = \mathbf{P}(\tau_1 \geqslant t) = e^{-\lambda t} = 1 - \lambda t + o(t)$$

If (i_k, j_k) are disjoint, then $X_{j_k} - X_{i_k} = \inf\{k : \tau_k - \tau_{k-1} \geqslant j_k\} - \inf\{k : \tau_k \geqslant i_k\}$. Hence $\{X_t\}$ is a Poisson process.

Consider the following calculation

$$\mathbf{E}(\tau_1) = \int_0^\infty \frac{\lambda t}{e^{\lambda t}} dt = \left.\frac{t + \lambda^{-1}}{e^{\lambda t}}\right|_{t=\infty}^0 = \lambda^{-1} - \lim_{t\to\infty} \frac{t + \lambda^{-1}}{e^{\lambda t}} = \lambda^{-1} - \lim_{t\to\infty} \frac{1}{\lambda e^{\lambda t}} = \lambda^{-1}$$

So that in a Poisson process, we should expect to wait on average λ^{-1} for each event.

5.2 Continuous Time Markov Process

Let's now consider an arbitrary Markov process $\{X_t\}$ in continuous time on a denumerable state space. For each time point t and u, we have the

transition probabilities $P_{u,t}(x,y) = \mathbf{P}(X_t = y | X_u = x)$. We still have the Kolmogorov equation

$$P_{u,v}(x,z) = \sum_t P_{u,t}(x,y) P_(t,v)(y,z) \tag{5.4}$$

We shall also assume a continuity requirement that

$$\lim_{j \to i^+} \mathbf{P}(X_j = x | X_i = y) = \delta_{x,y} \tag{5.5}$$

A process is **time-homogenous** if

$$P_{u,t}(x,y) = P_{t-u,0}(x,y) \tag{5.6}$$

If we define a transformation $P_t(x,y) = \mathbf{P}(X_t = y | X_0 = x)$, as well as a multiplication rule $(P_t P_s)(x,y) = \sum_z P_t(x,z) P_s(z,y)$, then we obtain from (5.4) and (5.6) that $P_{t+s} = P_t P_s$, so that $\{P_t\}$ forms a commutative monoid.

To obtain genuine derivations of probability distributions on homogenous Markov processes, we shall restrict ourselves to probability distributions which are differentiable. Apparently (I haven't seen the proof), any time-homogenous Markov process can be written

$$P_t(x,y) = t\alpha(x,y) + o(t)$$

for some value $\alpha(x,y)$, where $x \neq y$. We call $\alpha(x,y)$ the infinitismal generator of the system – we think of it as the rate at which a state x changes to a state y. We then obtain

$$P_t(x,x) = 1 - \sum_{x \neq y} P_t(x,y) = 1 - \sum_{x \neq y} [t\alpha(x,y) + o(t)]$$

In the finite case, we may conclude $P_t(x,x) = 1 - \sum_{x \neq y} t\alpha(x,y) + o(t)$. It thus makes sense to define $\alpha(x) = \sum_{y \neq x} \alpha(x,y)$ (even if our state space is denumerable) – it is the rate at which the process leaves x. This constitutes the definition of a process.

Definition. The **rates** of a time-homogenous Markov process $\{X_t\}$ are values α for which

$$\mathbf{P}(X_t = x | X_0 = x) = 1 - \alpha(x)t + o(t)$$

$$\mathbf{P}(X_t = y | X_0 = x) = \alpha(x,y)t + o(t)$$

The average amount of time for a state to transition out of a state x is $1/\alpha(x)$. The probability that the next state we will end up at is y from x is $\alpha(x,y)/\alpha(x)$. The waiting time is an exponential distribution, with $\mathbf{P}(\tau_x \leqslant t | X_0 = x) = 1 - e^{-\alpha(x)t}$.

Assume our state space is finite, and enumerate the states $x_1, \ldots, x_n$. Define a matrix P by $P_{i,j} = \alpha(x_i, x_j)$ for $i \neq j$, and $A_{i,i} = -\alpha(x_i)$. We call A the infinitisimal generator of the chain. If μ_t denotes the probability mass function at a certain time (seen as a row vector), then via an analogous proof to when we analyzed Poisson processes, we can verify that

$$\mu'(t) = \mu_t P$$

By the theory of linear differential equations, this means

$$\mu_t = \mu_0 e^{tA} = P(0) \sum_{k=0}^{\infty} \frac{(tA)^k}{k!}$$

In general, we consider the action $\mu P(y) = \sum \mu(x) P(x,y)$. Then $(\mu P)' = \mu P$ holds for countable state-spaces.

Example. *Consider a Markov chain with infinitisimal generator*

$$\begin{pmatrix} -1 & 1 \\ 2 & -2 \end{pmatrix}$$

We may diagonalize this matrix as $Q^{-1}AQ$, where

$$Q^{-1} = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{pmatrix} \quad A = \begin{pmatrix} 0 & 0 \\ 0 & -3 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix}$$

Hence

$$\mu_t = \mu_0 \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & -1/3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & e^{-3t} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -2 \end{pmatrix} = \frac{P(0)}{3} \begin{pmatrix} 2+e^{-3t} & 2-2e^{-3t} \\ 1-e^{-3t} & 1+2e^{-3t} \end{pmatrix}$$

As $t \to \infty$, $\mu_t \to (2/3, 1/3)$.

In general, we shall find that for an irreducible markov chain, there is a single eigenvector with eigenvalue zero, and all other eigenvectors have negative eigenvalue (we need not worry about periodicity for continous chains). The μ_t will converge to the single eigenvector, invariant of the initial distribution, and this is the unique μ for which $\mu P = 0$.

Suppose we want to compute the mean passage times $\mathbf{E}(\rho_y)$, where $\rho_y = \min\{t : X_t = y | X_0 = x\}$. Define $\beta(x)$ be the average time it takes to get to y given we start in x. Then

$$\beta(y) = 0 \qquad \beta(x) = 1/\alpha(x) + \sum_{z \neq y} \frac{\alpha(x,z)}{\alpha(x)} \beta(z)$$

Then $\alpha(x)\beta(x) = 1 + \sum \alpha(x,z)\beta(z)$. We can write this as $0 = 1 + \tilde{A}\beta$, where $\tilde{A}$ is obtained from A by deleting the row and column representing y, which has the solution $\beta = -\tilde{A}^{-1}1$.

5.3 Birth and Death Processes

Definition. A Birth and Death process is a continuous markov-process taking states in $\mathbf{N}$, with rates $\alpha(n, n+1) = \lambda_n$, and $\alpha(n, n-1) = \mu_n$, with $\mu_0 = 0$ (no-one can die if no-one is alive). Thus

$$\mathbf{P}(X_{t+\Delta t} = n | X_t = n) = 1 - (\mu_n + \lambda_n)\Delta t + o(\Delta t)$$

$$\mathbf{P}(X_{t+\Delta t} = n+1 | X_t = n) = \lambda_n \Delta t + o(\Delta t)$$

$$\mathbf{P}(X_{t+\Delta t} = n-1 | X_t = n) = \mu_n \Delta t + o(\Delta t)$$

We have already considered a special case of birth and death processes. We can convert these equations into a system of differential equations,

defining $P_n(t) = \mathbf{P}(X_t = n)$.

$$P_n'(t) = \mu_{n+1}\mathbf{P}_{n+1}(t) + \lambda_{n-1}P_{n-1}(t) - (\mu_n + \lambda_n)P_n(t)$$

This has a unique solution given a starting point n, so $P_n(0) = 1$, and $P_m(0) = 0$ for $m \neq n$.

Example. *A Poisson process with rate λ is a birth and death process with $\lambda_n = \lambda$ and $\mu_n = 0$, for all n. Our differential equation was*

$$P_n'(t) = \lambda P_{n-1}(t) - \lambda P_n(t)$$

Which we solved recursively.

Here we shall address queueing theory, the main application of continuous markov chains. There are many different types of queues, and in the literature there is a standard code for describing a specific type of queue. The basic code uses 3 characters, and is written $A/S/c$, where A S and C are substituted for common letters. Here we will be considering $M/M/c$ queues. A is the type describing the distribution of customers arriving at a queue and M means arrivals are be memoryless, or Markov. S describes the distribution time to serve a customer. Here, S will be M, since the distribution will also be markov. Finally, c stands for the number of servers, which can range from $1, 2, \ldots, \infty$.

An $M/M/1$ queue has only one person being served at each time. Thus, modelling the queue as a birth and death process, $\lambda_i = \lambda$ for some fixed λ, and $\mu_i = \mu$ for a fixed μ. In an $M/M/c$ queue, for $1 < k < \infty$, up to c people may be served at any time. Thus if n people have arrived in the queue, with $n \leqslant c$, then the queue 'kills' n times faster than if one server was working, so $\lambda_k = \lambda$, and $\mu_k = \min(c,k)\mu$, for some μ. This formula also works if $c = \infty$.

We can understand a birth and death process via our understanding of discrete time markov chains. Let X_n be the discrete process which 'follows the chain when it moves'. The transition probabilities are $P(n, n+1) = \frac{\lambda_n}{\mu_n+\lambda_n}$, and $P(n, n-1) = \frac{\mu_n}{\mu_n+\lambda_n}$. The discrete process is recurrent if and only if the continuous process is recurrent. Thus we define $\alpha(x)$ to be the probability of returning to 0 starting at x. We have

$$(\mu_n + \lambda_n)\alpha(x) = \mu_n\alpha(n-1) + \lambda_n\alpha(n+1)$$

This can be rewritten

$$\alpha(n) - \alpha(n+1) = \frac{\mu_n}{\lambda_n}[\alpha(n-1) - \alpha(n)]$$

By induction,

$$\alpha(n) - \alpha(n+1) = \frac{\mu_n \cdots \mu_1}{\lambda_n \ldots \lambda_1}[\alpha(0) - \alpha(1)]$$

Hence

$$\alpha(n+1) = \alpha(n+1) - \alpha(0) + \alpha(0) = 1 + [\alpha(1) - 1] \sum_{j=0}^{n} \frac{\mu_j \cdots \mu_1}{\lambda_j \ldots \lambda_1}$$

And thus the chain is transient if and only if

$$\sum_{j=0}^{\infty} \frac{\mu_1 \cdots \mu_j}{\lambda_1 \ldots \lambda_j} < \infty$$

Chapter 6

Brownian Motion

Brownian motion is one of fundamental continuous stochastic processes, modeling random continuous motion. It has a rich and beautiful theory. We say a real-valued $[0,\infty)$ time stochastic process $\{B_t\}$ is a **Brownian motion** if $B_0 = 0$, if the map $t \mapsto B_t(\omega)$ is continuous for almost all points ω in the sample space, and if $B_{t+h} - B_t$ is independant of $\{B_u : 0 \leqslant u \leqslant t\}$ for all $t, h \geqslant 0$, and is Gaussian distributed with mean zero and variance h. Another reason to study Brownian motion is it is an example of almost every interesting class of stochastic processes:

6.1 Brownian Motion is a Martingale

Since each $X_t \in L^1$ because it is $N(0,t)$ distributed. If $\Sigma_s = \sigma(X_t : t \leqslant s)$, then for $t \geqslant s$,

$$\mathbf{E}[X_t|\Sigma_s] = \mathbf{E}[X_t - X_s|\Sigma_s] + \mathbf{E}[X_s|\Sigma_s] = \mathbf{E}[X_t - X_s] + X_s = X_s$$

Furthermore, we find that $\mathbf{E}[(B_t - B_s)^2|\Sigma_s] = t - s$, and also

$$\mathbf{E}[(B_t - B_s)^2|\Sigma_s] = \mathbf{E}[B_t^2|\Sigma_s] - 2\mathbf{E}[B_t B_s|\Sigma_s] + \mathbf{E}[B_s^2|\Sigma_s] = \mathbf{E}[B_t^2|\Sigma_s] - B_s^2$$

so $B_t^2 - t$ is also a martingale. Once the theory is suitably developed, we will be able to prove that Brownian motion is the *only* continuous time martingale with continuous sample paths such that $B_t^2 - t$ is a martingale.

6.2 Brownian Motion is a Gaussian Process

A continuous time process X is Gaussian if for every finite set of indices $t_1, \ldots, t_n$, $(X_{t_1}, \ldots, X_{t_n})$ is normally distributed). The law of the process is then specified by the functions $\mu(t) = \mathbf{E}[X_t]$ and $\rho(s,t) = \mathrm{Cov}(X_s, X_t)$. Brownian motion is a Gaussian process. Given a set of time indices $t_1 < \cdots < t_n$, and $\lambda_1, \ldots, \lambda_n \in \mathbf{R}$, and if we let $B_0 = 0$, then

$$\sum_{k=1}^{n} \lambda_k B_{t_k} = \sum_{k=1}^{n} \mu_k (B_{t_k} - B_{t_{k-1}})$$

where $\mu_n = \lambda_n$ and $\mu_k = \lambda_k + \mu_{k+1}$ for all $1 \leqslant k \leqslant n$. Then we have represented the random variable as a linear combination of independent Gaussian random variables, and thus the random variable is Gaussian distributed. We find that $\mu = 0$, and $\rho(s,t) = \min(s,t)$, because if $s \leqslant t$ and

$$\mathbf{E}[X_t X_s] = \mathbf{E}[(X_t - X_s)X_s + X_s^2] = \mathbf{E}[X_t - X_s]\mathbf{E}[X_s] + \mathbf{E}[X_s^2] = 0 + s$$

If any Gaussian process has continuous sample paths, and has $\mu(t) = 0$, $\rho(s,t) = \min(s,t)$, then the process is a Brownian motion, since $X_0 = 0$, because it is Gaussian with mean zero and variance 0, and $X_{t+h} - X_t$ is independant of any finite family of X_s, because if $s \leqslant t$, then

$$\mathrm{Cov}(X_{t+h} - X_t, X_s) = \mathrm{Cov}(X_{t+h}, X_s) - \mathrm{Cov}(X_t, X_s) = s - s = 0$$

and thus $X_{t+h} - X_t$ is independant of $\{X_s : s \leqslant t\}$.

The Gaussian process condition makes it very easy to verify a process is a Brownian motion. In particular, if $\{B_t\}$ is a Brownian motion,

- $\{-B_t\}$ is a Brownian motion.
- If a is fixed then $\{B_{t+a} - B_t\}$ is a Brownian motion.
- If $c \neq 0$, then $\{cB_{t/c^2}\}$ is a Brownian motion.
- If we define $\tilde{B}_0 = B_0$, and $\tilde{B}_t = tB_{1/t}$, then $\tilde{B}$ is a Brownian motion. The only tricky part is verifying continuity, and this follows because

by continuity on $t \neq 0$,

$$\begin{aligned}\mathbf{P}\left(\lim_{t\downarrow 0}\tilde{B}_t = 0\right) &= \mathbf{P}\left(\lim_{n\to\infty}\lim_{m\to\infty}\bigcap\left\{|\tilde{B}_q| \leqslant n^{-1} : q \in \mathbf{Q}\cap(0,1/m]\right\}\right)\\ &= \mathbf{P}\left(\lim_{n\to\infty}\lim_{m\to\infty}\bigcap\left\{|B_q| \leqslant n^{-1} : q \in \mathbf{Q}\cap(0,1/m]\right\}\right)\\ &= \mathbf{P}\left(\lim_{t\downarrow 0} B_t = 0\right) = 1\end{aligned}$$

where we used the fact that $\tilde{B}$ is identically distributed to B. This means that $B = o(t)$ almost surely.

Brownian scaling is one of the most important of these points, for it implies that Brownian motion has a certain 'fractal' quality about it – the behaviour of Brownian motion on $[0,a]\times[-b,b]$ is the same as the behaviour of Brownian motion on $[0,t^2a]\times[-tb,tb]$.

Lemma 6.1. *We have* $\mathbf{P}(\sup B_t = \infty, \inf B_t = -\infty) = 1$.

Proof. Let $Z = \sup B_t$. By Brownian scaling, for any c, cZ is identically distributed to Z. This means that Z is concentrated on $\{0,\infty\}$, because

$$\mathbf{P}(0 < Z < N) = \mathbf{P}\left(0 < Z(\varepsilon N)^{-1} < \varepsilon^{-1}\right) = \mathbf{P}(0 < Z < \varepsilon^{-1})$$

Letting $\varepsilon \to 0$ gives $\mathbf{P}(0 < Z < N) = 0$, and we can the let $N \to \infty$ to conclude $\mathbf{P}(0 < Z < \infty) = 0$. Now

$$\begin{aligned}\mathbf{P}(\sup B_t = 0) &\leqslant \mathbf{P}(B_1 \leqslant 0 \text{ and } B_u \leqslant 0 \text{ for all } u \geqslant 1)\\ &= \mathbf{P}(B_1 \leqslant 0 \text{ and } \sup(B_{1+t} - B_t) = 0)\\ &= \frac{\mathbf{P}(\sup(B_{1+t} - B_t) = 0)}{2} = \frac{\mathbf{P}(\sup B_t = 0)}{2}\end{aligned}$$

hence $\mathbf{P}(\sup B_t = 0) = 0$, and so $\sup B_t = \infty$ almost surely. Since $-B_t$ is a Brownian motion, this gives $\inf B_t = -\infty$ almost surely. □

This lemma also implies that for each a, $\{t : B_t = a\}$ is almost surely not bounded above. Thus every a is a recurrent state of the process.

6.3 Brownian Motion is a Markov Process

Brownian motion is also a continuous time time-homogenous Markov process, because for any bounded Borel measurable f, $\mathbf{E}[f(B_{t+s})|\Sigma_t] = P_s(f)(B_t)$, where P_t is the transition semigroup operator $P_t f = p_t * f$, and $p_s(x) = (2\pi s)^{-1/2}\exp(-x^2/2s)$ is the transition density of the Brownian motion, and $p_0 = \delta$ is the Dirac delta. This is easily verified because

$$\begin{aligned}
\mathbf{P}(a \leqslant B_{t+s} \leqslant b|\Sigma_t) &= \mathbf{P}(a - B_t \leqslant B_{t+s} - B_t \leqslant b - B_t|\Sigma_t) \\
&= \mathbf{E}[\mathbf{P}(a - B_t \leqslant B_{t+s} - B_t \leqslant b - B_t|B_t)|\Sigma_t] \\
&= \mathbf{E}\left[\int_{a-B_t}^{b-B_t} p_s(y)\, dy \middle| \Sigma_t\right] \\
&= \mathbf{E}\left[\int_a^b p_s(B_t + y)\, dy \middle| \Sigma_t\right] \\
&= \int p_s(B_t + y)\chi_{[a,b]}(y)\, dy \\
&= (p_s * \chi_{[a,b]})(B_t)
\end{aligned}$$

and the general result follows by taking limits of simple functions. The time homogeneity follows because $p_t * p_s = p_{t+s}$ (easily verified by taking the Fourier transform), so $P_{t+s} = P_t \circ P_s$. We can define an infinitisimal generator

$$Af = \lim_{t\downarrow 0} \frac{P_t f - f}{t}$$

and provided $f \in C_b^2(\mathbf{R})$,

$$\begin{aligned}
\lim_{t\downarrow 0} \frac{(P_t f)(x) - f(x)}{t} &= \lim_{t\downarrow 0} \int_{-\infty}^{\infty} \frac{f(x+y) - f(x)}{t} \frac{e^{-y^2/2t}}{\sqrt{2\pi t}}\, dy \\
&= \lim_{t\downarrow 0} \int_{-\infty}^{\infty} \frac{f(x+\sqrt{t}y) - f(x)}{t} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy \\
&= \lim_{t\downarrow 0} \int_{-\infty}^{\infty} \frac{1}{t}\left(y\sqrt{t}f'(x) + (y^2 t/2)f''(x + \theta y\sqrt{t})\right) \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy \\
&= \lim_{t\downarrow 0} \int_{-\infty}^{\infty} (y^2/2)f''(x + \theta y\sqrt{t}) \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy = f''(x)/2
\end{aligned}$$

Thus, on $C_b^2(\mathbf{R})$, the infinitisimal generator of the Brownian motion is

$$\frac{1}{2}\frac{d^2}{dx^2}$$

This implies that for any $f \in C_b^2(\mathbf{R})$, and $s > 0$,

$$\frac{\partial P_t f}{\partial t} = \lim_{t \to 0} \frac{P_{t+s} f - P_t f}{t} = \frac{1}{2}\frac{\partial^2 P_t f}{\partial x^2}$$

Thus $P_t f$ is a solution to the *heat equation* for any sufficiently regular function f. Letting f converge to the Dirac delta function hints at the fact that

$$\frac{\partial p_t}{\partial t} = \frac{1}{2}\frac{\partial^2 p_t}{\partial x^2}$$

We can interpret this as saying the heat equation models the averages of particle behaviour undergoing brownian motion over a time period. This connects the classical study of diffusion in physics with the study of diffusion in probability theory. However, whereas the study of diffusion in physics gives results about the average behaviour of particles over a long period of time, whereas probability theory gives much stronger results of the behaviour of *individual* particles.