

# Stochastic Processes

Jacob Denson

September 8, 2017

# Table Of Contents

<b>1</b>	<b>Stochastic Processes</b>	<b>2</b>
<b>2</b>	<b>Finite Markov Chains</b>	<b>5</b>
2.1	Finite Markov Processes . . . . .	5
2.2	Asymptotics of Markov chains . . . . .	9
2.3	Aperiodicity and Irreducibility . . . . .	12
2.4	Potentials: Unique Invariant Probabilities . . . . .	15
2.5	Periodicity and Average State Distributions . . . . .	16
2.6	Stopping Times . . . . .	18
<b>3</b>	<b>Countable-State Markov Chains</b>	<b>23</b>
3.1	General Properties . . . . .	23
3.2	Recurrence and Transience . . . . .	23
3.3	Branching Processes . . . . .	26
3.4	Reversibility . . . . .	30
<b>4</b>	<b>Martingales</b>	<b>34</b>
4.1	Modern Conditional Expectation . . . . .	34
4.2	Martingales . . . . .	39
4.3	The Optional Sampling Theorem . . . . .	41
4.4	Martingale Convergence . . . . .	43
<b>5</b>	<b>Continuous Markov Processes</b>	<b>44</b>
5.1	Poisson Processes . . . . .	44
5.2	Continuous Time Markov Process . . . . .	46
5.3	Birth and Death Processes . . . . .	49

# 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  $\Omega$ , 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  $S = \{\text{sunny}, \text{rainy}\}$ . For  $i \in \mathbf{Z}$ , we may model the weather by a random variable  $X_i : \Omega \rightarrow S$ , 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  $X_1, \dots, X_n \sim F$ . we can take the estimate

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

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 processes 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  $t_1, \dots, t_n$  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  $\sigma$ -algebra of  $S^R$ . If  $K \subset R \subset T$ , then we have a projection map  $\pi_{R \rightarrow K} : S^R \rightarrow S^K$ , and we say that the family of probability distributions chosen over the index sets are **consistent** 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  $\mathbf{P}_K$ , consistency is a necessary requirement, but Kolmogorov's theorem shows that this condition is also sufficient.

**Theorem 1.1** (Kolmogorov's extension theorem). *For any consistent 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 / Carathéodory 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.  $\square$

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

# Chapter 2

## Finite Markov Chains

### 2.1 Finite Markov Processes

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 strong 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 independent events. Nekrasov had not committed 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, and introduce the Markov chain.

**Definition.** Let  $\{X_t\}_{t \in T}$  be a discrete stochastic process (defined on a discrete state space and discrete time interval). This process satisfies the **discrete Markov property** if, for any  $t_1 < t_2 < \dots < t_n < t_{n+1} \in T$  and any  $x_1, x_2, \dots, x_n, y \in S$ , we have

$$\mathbf{P}(X_{t_{n+1}} = y | X_{t_n} = x_n, \dots, X_{t_1} = x_1) = \mathbf{P}(X_{t_{n+1}} = y | X_{t_n} = x_n)$$

A **Markov chain** is a discrete stochastic process  $\{X_n\}_{n \in \mathbf{N}}$  satisfying 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 do not 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,

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

*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  $Y_k = (X_0, \dots, X_k)$ . Then one may verify that

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

*This shows that  $\{Y_k\}$  satisfies the Markov property, excusing of course the obvious problems of what exactly the state space of the process is.*

For any three random variables  $X, Y, Z$  mapping into a discrete state space  $\mathcal{S}$ , we have the equation

$$\mathbf{P}(X = x|Z = z) = \sum_{s, s' \in \mathcal{S}} \mathbf{P}(X = x|Y = s, Z = s') \mathbf{P}(Y = s|Z = s')$$

If  $i < j < k$ , we may simplify the equation in the case of a Markov chain where  $X = X_k, Y = X_j, Z = X_i$ ,

$$\mathbf{P}(X_k = x|X_i = z) = \sum_{y \in \mathcal{S}} \mathbf{P}(X_k = x|X_j = y) \mathbf{P}(X_j = y|X_i = z)$$

This is the **Chapman-Kolmogorov equation**, relating various transition probabilities of a markov chain. By the Kolmogorov extension theorem, a Markov chain on  $\mathbf{N}$  can be defined by the initial probability distribution  $\mathbf{P}(X_0 = x)$ , and the transition probabilities  $\mathbf{P}(X_k = x|X_{k-1} = y)$ . We will begin by focusing on ‘time homogenous’ Markov chains, which have the nicest theory.

**Definition.** A Markov chain can be defined in the following manner. First, we specify two functions  $P(\cdot, \cdot) : \mathcal{S}^2 \rightarrow [0, 1]$  and  $\mu_0 : \mathcal{S} \rightarrow [0, 1]$  such that, for a fixed  $s \in \mathcal{S}$ ,  $P(s, \cdot)$  is a probability distribution, and  $\mu_0$  specifies an ‘initial’ probability distribution. Then we may define

$$\mathbf{P}_{\{0,1,2,\dots,n\}}(x_0, x_1, \dots, x_n) = I(x_0) \prod_{k=0}^{n-1} P(x_k, x_{k+1})$$

and extend this to a probability density on  $\mathcal{S}^n$ . These distributions are sufficient for Kolmogorov’s theorem to apply, and the stochastic processes formed by the construction are called **time-homogenous markov chains**.

Our first insight to time-homogenous chains is that  $\mu_0$  does not really factor much into the time homogenous process. The main mechanism occurs in the transition probabilities  $P$ . To analyze time-homogenous chains, it will be convenient to fix  $P$ , and vary  $\mu_0$ . We may view  $P$  as an operator on



the space of all distributions (contained in the Banach space  $l_1$ ). Studying the distribution of time-homogenous chains on a finite state space reduces to operator theory – in particular for the finite dimensional case, matrix algebra.

What are the specifics of how we view  $P$  as an operator? First, we may consider  $P$  to take a distribution one step into the future. That is, if  $\mu$  is a distribution, then  $\mu P$  is the distribution of the Markov chain with initial probabilities  $\mu$  and transitions  $P$  one step into the future. In general,  $\mu P^n$  is the distribution  $n$  steps into the future. If we suppose  $\mathcal{S}$  is finite,  $\mathcal{S} = \{x_1, \dots, x_n\}$ , then we may consider  $P$  as a **stochastic matrix**, with  $P_{ij} = P(x_i, x_j)$ . The rows of  $P$  sum to one, and any such matrix with these rows specifies the transition probabilities of a time-homogenous markov chain. The probability mass functions  $\mu$  can be seen as row vectors; the Kolmogorov equation, in matrix form, tells us exactly that  $\mu P$  corresponds to matrix multiplication.

We can also see  $\mu$  as operating by  $P$  on the right. Let  $f$  be any column vector. Then

$$(P^n f)(x) = \sum_y P^n(x, y) f(y)$$

We can take  $f$  to be any real-valued function on the state space of the process, and then  $P^n f(x) = \mathbf{E}(f(X_n) | X_0 = x)$ , which we may define for any process, even if the state space is not finite.

**Example.** *It is a useful simplification to assume that the probability of weather is time-homogenous. We might choose a transition matrix like the one below*

$$\begin{pmatrix} & \text{sunny} & \text{rainy} \\ \text{sunny} & 0.6 & 0.4 \\ \text{rainy} & 0.8 & 0.2 \end{pmatrix}$$

*One may also specify this process with a transition matrix 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 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

$$\begin{pmatrix} & 0 & 1 & 2 \\ 0 & 1-p & p & 0 \\ 1 & (1-p)q & (1-q)(1-p)+pq & p(1-q) \\ 2 & 0 & q(1-p) & (1-q)+pq \end{pmatrix}$$

Given a large amount of time, it is of interest to the maker of the queuing system to know the average number of people in the queue at a certain time.

**Example.** Consider a random walk on a certain graph. At any 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  $\{0, 1, \dots, n\}$ , where each integer is connected to adjacent integers. The transition probabilities are given by

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

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

These are the transitions for the canonical reflecting random walk. If one connects the end vertices to themselves, then one obtains another form of the random walk.

## 2.2 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 that probabilities converge in the limit. It is most easy to perform the analysis on finite state space time-homogenous Markov chains. We shall show that in most cases there exists a unique **invariant distribution**  $\mu$  such that, for any initial distribution  $\mu_0$ ,  $\mu_0 P^n \rightarrow \mu$  as  $n \rightarrow \infty$ . Our analysis of these markov chains follows from Eigenvalue decomposition, via the use of the Perron-Frobenius theorem.

**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,

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

Regardless of the initial distribution of the markov chain,  $\mu_0 P^n \rightarrow (2/5, 3/5)$ .

Some initial distributions work very nicely when taking limits of the stochastic matrix: Suppose  $\mu$  is a left eigenvector of  $P$  ( $\mu P = \mu$ ). Then  $\mu P^n = \mu$ , and so, taking  $n \rightarrow \infty$ , we find  $\mu$  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.

**Definition.** Fix a stochastic matrix  $P$ . An **invariant/stationary probability distribution** for  $P$  is a distribution  $\mu$  such that  $\mu P = \mu$ . Note that this is simply a left eigenvector of eigenvalue 1.

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

**Theorem 2.1** (The Perron-Frobenius Theorem). *Let  $M$  be a positive square  $n \times n$  matrix. There is a positive eigenvalue  $\lambda$  of maximal modulus, called the **Perron root** of  $M$ , with one dimensional eigenspace which contains a positive vector.*

Now suppose  $P$  is a stochastic, positive matrix. Then we may apply Perron-Frobenius to  $P$ , obtaining a Perron root  $\lambda$ . We must have  $|\lambda| \leq 1$ , since all entries of the matrix are less than one. Because  $(1, 1, \dots, 1)^t$  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

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

Where  $M$  is a square matrix such that  $\lim_{n \rightarrow \infty} M^n = 0$  (Use the Jordan Canonical Form, and the fact that 1 is the maximal eigenvalue). But then

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

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

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.

**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  $\mu P^n$  converges is when  $\mu = (1/2, 1/2)$ .

**Example.** Consider a process whose transition matrix is the identity matrix  $I$ . Then  $P^n \rightarrow I$ , so  $\mu P^n \rightarrow \mu$  for all distributions  $\mu$ . This is an asymptotic result, but different initial distributions give rise to different asymptotics. This is because the process is reducible - there is not enough 'mixing' to occur to generate a homogenous distribution.

## 2.3 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 recognizing two concepts, that of periodicity and irreducibility.

**Definition.** Let  $s, m \in \mathcal{S}$  be two states in an arbitrary Markov chain. We say  $s$  communicates with  $m$  if there is some  $n$  with  $\mathbf{P}(X_n = s | X_0 = m) > 0$ . If we divide  $\mathcal{S}$  into classes of states, all of which communicate between one another, then we obtain a set of **communication classes**. 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

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

And for any  $m$ ,

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

We can work with each  $P_i$  as if it were its own stochastic process, and then put them together to form a stochastic process in full. We leave the analysis of the transient classes till later, and from now on assume an irreducible chain.

**Definition.** For any  $s \in \mathcal{S}$ , let  $\mathcal{J}(s) = \{n \in \mathbf{N} : \mathbf{P}(X_n = s \mid X_0 = s) > 0\}$ . Then  $\mathcal{J}(s)$  is closed under addition. The greatest common divisor of  $\mathcal{J}(s)$  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 an irreducible Markov chain.

**Theorem 2.2.** *If  $p$  is the period of a state  $s$ , then all but finitely many of  $\{0, p, 2p, \dots\}$  are contained in  $\mathcal{J}(s)$ .*

*Proof.* We note that the set above has the property that if  $x, y \in \mathcal{J}(s)$ , then  $x + y \in \mathcal{J}(s)$ , so that  $\mathcal{J}(s)$  forms a commutative semigroup. It follows that, if  $a, b > 0$ , then  $ax + by \in \mathcal{J}(s)$ . Let  $p = \sum a_i d_i$ , for  $a_i \in \mathbf{Z}$ ,  $d_i \in \mathcal{J}(s)$ . Let  $M = \sum |a_i| d_i^2$ . Then  $M \in \mathcal{J}(s)$ , and

$$M + kp = \sum_i (|a_i| d_i + k a_i) d_i$$

Provided  $k < \min(d_i)$ , then  $M + kp \in \mathcal{J}(s)$ . Let  $d_j = \min(d_i)$ . If  $k = d_j$ , then

$$M + d_j p = \left( \sum_{i \neq j} |a_i| d_i^2 \right) + (|a_j| d_j + p) d_j$$

All the coefficients of the  $d_i$  are positive, so  $M + d_j p \in \mathcal{J}(s)$ . In general, every integer  $m$  can be written as  $nd_j + m$ , where  $m < d_j$ . Repeating the argument above, adding  $n$  factors to  $d_j$ , we conclude  $M + mp \in \mathcal{J}(s)$ .  $\square$

**Theorem 2.3.** *Let  $P$  be a stochastic matrix, which determines an aperiodic, irreducible Markov chain. Then there is a unique vector  $\mu$  for which  $\mu P = \mu$ , and for any other probability distribution  $\pi$ ,  $\lim_{n \rightarrow \infty} \pi P^n = \mu$ .*

*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  $m$ ,  $P_{ii}^m > 0$  for all  $i$ . If  $j \neq i$ , there is some  $k$  for which  $P_{ij}^k > 0$ . Then, for large enough  $m$ ,  $P_{ij}^m > 0$ , since

$$P_{ij}^m \geq P_{ij}^k P_{ii}^{m-k} > 0$$

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

**Corollary 2.4.** *On every aperiodic, irreducible Markov chain 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.

The advantage of the above result is we may proceed brashly in derivations. First, we show that a chain is aperiodic and irreducible. Then we find a stationary distribution, and conclude that this is the limiting distribution.

**Example.** *Let us consider the asymptotics of a two state markov chain. Consider a transition matrix*

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

*If  $p = 0$  and  $q = 0$ , the chain is reducible (the two states never change). 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 we can determine the invariant distribution. If  $\mu = (\mu_1, \mu_2)$  is an initial distribution, then*

$$\mu P = ((1-p)\mu_1 + q\mu_2, p\mu_1 + (1-q)\mu_2)$$

*To find a stationary distribution, we try and equate  $\mu P = \mu$ . So*

$$(1-p)\mu_1 + q\mu_2 = \mu_1 \quad p\mu_1 + (1-q)\mu_2 = \mu_2$$

*Since  $\mu_2 = 1 - \mu_1$ , we may manipulate the equations to show*

$$\mu_1 = \frac{q}{p+q} \quad \mu_2 = \frac{p}{p+q}$$

*And this is the limiting distribution of the chain.*

## 2.4 Potentials: Unique Invariant Probabilities

There is another way to derive the uniqueness of invariant distributions that is of interest. Recall that an invariant probability is a function  $\mu$  for which  $\mu P = \mu$ , so that  $\mu$  is a left eigenvector for  $P$ . 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 complex differentiable functions.

**Definition.** A function  $\mu$  on states is harmonic with respect to a transition matrix  $P$  if  $P\mu = \mu$ . In other words, the average value of  $\mu$  beginning from a certain state is not affected by a transition period.

**Lemma 2.5.** *A harmonic function on an irreducible markov chain is constant.*

*Proof.* Let  $s_0$  be the state upon which  $\mu(s_0)$  attains its maximal value. If  $P(s_0, s) > 0$ , then it cannot be true that  $\mu(s) < \mu(s_0)$ , for

$$P\mu(s_0) = \sum_x P(s_0, x)\mu(x) = \mu(s_0)$$

Yet if  $\mu(s) < \mu(s_0)$ ,

$$\begin{aligned} \sum_x P(s_0, x)\mu(x) &= \sum_{x \neq s} P(s_0, x)\mu(x) + P(s_0, s)\mu(s) \\ &\leq (1 - P(s_0, s))\mu(s_0) + P(s_0, s)\mu(s) < \mu(s) \end{aligned}$$

So on the connected component of  $s_0$ , the function must be constant. Since an irreducible markov chain consists of one connected component,  $\mu$  must be constant.  $\square$

**Corollary 2.6.** *The invariant probability vector for a irreducible process is unique.*

*Proof.* The dimension of the eigenspace of  $P$  with eigenvalue 1 is one-dimensional – and thus the invariant probability is unique if it exists, which we have already verified.  $\square$



The theorem above is an analogy of the maximum modulus principle of complex analysis – 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 case, the electric potential must occur at one of the points where the electrical flows, so the flow must be constant throughout – provided we are working in a connected space; all the wires in the circuit are connected, or the markov chain is irreducible.

## 2.5 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  $K_1, K_2, \dots, K_n$ , 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  $t_1, t_2, \dots$  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  $\mu_{t_1}, \dots, \mu_{t_m}$ . The limit  $\lim_{n \rightarrow \infty} \mu P^n$  may not exist, but the chebyshev limit

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

will always exist. 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 relevant far into the future, so that for large  $n$ , for any  $\mu$ ,

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

On even states,  $P^n$  converges to a different matrix than on odd states. Nonetheless, the Chebyshev limit exists for any distribution  $\mu$ , 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

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

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

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

## 2.6 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  $\tau$ , 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**  $\tau_s$  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**  $\rho_s$  of the process is defined

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

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  $X_\tau$ . In the case that  $\tau$  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  $\tau$  a stopping time.  $X_t$  satisfies the **strong Markov property** with respect to  $\tau$ , if, for  $t_1 < \dots < t_n < \tau$ ,

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

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

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

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

So the process is strongly Markov.  $\square$

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  $s$ ,  $\mathcal{J}_k = \min\{n > \mathcal{J}_{k-1} : X_n = s\}$ . 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  $n \rightarrow \infty$ ,

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

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  $Q^n \rightarrow 0$  as  $n \rightarrow \infty$ , since we are guaranteed 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).  $Q_{ij}^k$  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  $n \rightarrow \infty$ , we find the expected number of visits to the state before hitting a recurrent state is  $(I - Q)_{ij}^{-1}$ . 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 entering 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  $n \rightarrow \infty$ .

**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

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

So

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

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 transience is a little more complicated, 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  $\tau_x$ . 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

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

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  $\sum (\pi n)^{-1/2} \rightarrow \infty$ , 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

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

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

Here's yet another method for determining recurrence. Fix a state  $y$  on an irreducible markov chain, and define  $\alpha(x) = \mathbf{P}(X_n = y \text{ for some } n \geq 0 | X_0 = x)$ . 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  $\alpha$  with these properties, and if  $y$  is transient,  $\alpha$  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  $\lim_{n \rightarrow \infty} P^n(x, y) = 0$ , and is **positive recurrent** otherwise. An invariant probability is a function  $\mu$  for which  $\mu P = \mu$ . We won't show it, but every irreducible, aperiodic, positive recurrent Markov chain has a distribution  $\mu$ . Moreover, such a chain is positive recurrent if and only if it has an invariant distribution  $\mu$ . The return time  $\tau_x$  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  $0 \leq q \leq 1$ . 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  $\alpha$  exists, it satisfies  $\alpha(x) = \beta^x$ . We have

$$\begin{aligned}\beta^x &= q\beta^{x+1} + (1-q)\beta^{x-1} \\ q\beta^2 - \beta + (1-q) &= 0 \\ \beta &= \frac{1 \pm \sqrt{1-4q(1-q)}}{2q} = \frac{1 \pm (2q-1)}{2q} = \left\{1, \frac{1-q}{q}\right\}\end{aligned}$$

Thus  $\alpha(x) = c_0 + c_1 \left(\frac{1-q}{q}\right)^x$ . If  $q < 1/2$ ,  $c_1 = 0$  because  $\alpha$  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  $q \leq 1/2$ . We need  $\mu$  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  $\mu$ , 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  $X_0, X_1, \dots$ , 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 probabilities 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

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

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

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

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  $\mu$  denote the mean offspring a single individual will possess. Then we conclude by (3.2) that

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

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

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 guaranteed 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  $a(k) = \lim_{n \rightarrow \infty} a_n(k)$ . Since all  $k$  branches of the population act independently, 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

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

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

Thus  $a = \varphi(a)$ , where  $\varphi$  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  $\varphi$  is monotonically increasing on the positive numbers. If we let  $\varphi_n$  be the generating function of  $X_n$ , then

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

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

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

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

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  $a_n(1) \leq \hat{a}$ . Certainly  $a_0(1) = 0 \leq \hat{a}$ . If  $a_{n-1}(1) \leq \hat{a}$ , then

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

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

**Theorem 3.1.** *If  $X_1, X_2, \dots$  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  $\mu$  is less than one.

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

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

Here we have used the monotonicity of  $\varphi$ . 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 prominent. 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  $X_0, X_1, \dots, X_n$  are the first few frames of the movie, then  $(X_0, \dots, X_n)$  is distributed identically to  $(X_n, \dots, X_0)$ . We have

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

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  $\mu$  (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  $\mu$  is a probability distribution, then  $(X_0, X_1, \dots, X_n)$  is identically distributed to  $(X_n, \dots, X_0)$ , given that  $\mu$  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  $\mu$ .

Now let  $\mu_0$  be a reversible measure generating a reversible markov chain  $\{X_t\}$ . Suppose we watch a markov chain  $(X_0, \dots, X_N)$  for a really large  $N$ . Then, if a limiting distribution exists, it mustn't be too different from  $\mu_N$ . If we watch the markov chain backwards  $(X_N, \dots, X_0)$ , then it is equal in distribution by the properties of a markov chain. In particular, this means that the distribution of  $\mu_0$  is also the result of watching a Markov chain for a really long time – so we should expect  $\mu_0$  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

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

So  $\mu$  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 pseudorandom 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  $f : \mathcal{S} \times \Lambda \rightarrow \mathcal{S}$  together with a  $\Lambda$ -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  $Z_1, \dots, \infty$  of random variables independent 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  $\beta$  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

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

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

We ‘slow’ down the chain at certain points to make it reversible with respect to  $\beta$ , and hence converges to  $\mu = \beta/B$ . This is the Metropolis-Hastings algorithm for computing a distribution  $\beta$  up to a multiplicative constant. It is important that the algorithm only depends on the ratios of  $\beta$ . Frequently,  $\beta$  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 need to calculate  $\beta$  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

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

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

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

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,

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

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  $f : \mathbf{R} \rightarrow \mathbf{R}$ .*

*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  $\omega$ ,

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

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

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  $\sigma$ -algebra  $\sigma(X)$ , then it isn’t too much to consider arbitrary ‘sets of information’ about a probability space to be a sub  $\sigma$ -algebra of the  $\sigma$ -algebra defining the space. We say a random variable  $X$  is **adapted** to a  $\sigma$ -algebra  $\Sigma$  if  $X$  is measurable with respect to  $\Sigma$ . A random variable  $Y$  is adapted with respect to  $X$  if  $Y$  is adapted to the  $\sigma$  algebra  $\sigma(X)$  generated by  $X$ . With all this terminology, we can now make the general definition. For a given  $\Sigma$  algebra and random variable  $X$ , a  $\Sigma$  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  $X \geq 0$ . Then the map  $S \mapsto \int_S X d\mathbf{P}$  is a *finite measure* over  $\Sigma$  which is absolutely continuous with respect to  $\mathbf{P}$ , hence by the Radon-Nikodym theorem there exists a  $\Sigma$  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  $X \geq 0$ ,  $Y \geq 0$  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.  $\square$

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

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

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  $X$  in 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  $\mathbf{E}(X|\Sigma)$  is the orthogonal projection of  $X$  onto the subspace of  $L^2$  functions which are  $\Sigma$  measurable.*

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

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

**Lemma 4.4.** *For any  $\Sigma$ -measurable function  $Y$ , and  $A \in \Sigma$  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  $X \geq 0$ , and  $\mathbf{E}(X|\Sigma) \geq 0$  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  $Y \geq 0$  by the monotone convergence theorem, and then for all  $Y$  by splitting  $Y$  into its positive and negative parts.  $\square$

We say two sigma algebras  $\Sigma$  and  $\Delta$  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  $\sigma$  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  $\Sigma$  if  $\sigma(X) \cap \Sigma$  is independent of  $\sigma(Y) \cap \Sigma$ . If  $X$  and  $Y$  are independent with respect to  $\Sigma$ , then  $\mathbf{E}(XY|\Sigma) = \mathbf{E}(X|\Sigma)\mathbf{E}(Y|\Sigma)$ . Most importantly, if  $X$  is independent with respect to  $\Sigma$ , 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  $\sigma$  algebras. Precisely, a family  $\{\mathcal{F}_t\}$  of  $\sigma$  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  $\mathcal{F}_t = \sigma(X_s : s \leq t)$ .

**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

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

Each  $S_k$  is in  $L_1$ , so the sequence  $S_k$  is a martingale. In general, if  $X_t$  has mean  $\mu$ , 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 independent 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  $X_0, \dots, X_{n-1}$ . The stochastic process we now observe is the accumulation of our winnings

$$M_n = \sum_{k=0}^n W_k X_k$$

Since

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

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

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

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 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  $\tau$  mapping into time, such that for any time  $t$ ,  $\{\omega : \tau(\omega) \leq t\} \in \mathcal{F}_t$ .

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

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

$$\tau > K - 1$$

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

Repeating this proof on  $\mathcal{F}_{K-2}$ , we obtain that  $\mathbf{E}(M_\tau | \mathcal{F}_{K-2}) = \sum_{k=0}^{K-3} M_k + \mathbf{I}(\tau \geq K-2) M_{n-2}$ . 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  $\tau$  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  $M_{\tau_n} \rightarrow M_\tau$  as  $n \rightarrow \infty$ , then  $\mathbf{E}(M_0) = \mathbf{E}(M_{\tau_n}) \rightarrow \mathbf{E}(M_\tau)$ . 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  $\mathbf{E}(\mathbf{I}(\tau > n) M_\tau) \rightarrow 0$ . The second factor disappears if

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

and this provides the conditions for our theorem to hold.

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

**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  $\varepsilon$ , 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 there is  $C$  where  $X_n^2 < C$ , then the  $X_n$  are uniformly integrable.*

Now let  $\{M_n\}$  be a uniformly integrable martingale, and  $\tau$  a stopping time for which  $\mathbf{P}(\tau < \infty) = 1$ . Then  $\lim_{n \rightarrow \infty} \mathbf{E}(|M_n|\mathbf{I}(\tau > n)) = 0$ , since  $\mathbf{P}(\tau > n) \rightarrow 0$ . Thus the optional stopping theorem holds for the  $M_n$  and  $\tau$ , 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  $n < m$ ,  $\mathbf{E}(M_m|\mathcal{F}_n) \geq M_n$ . A **supermartingale** satisfies  $\mathbf{E}(M_m|\mathcal{F}_n) \leq M_n$ .

**Theorem 4.8.** *If  $M_n$  is a submartingale, for which there is  $C$  such that  $\mathbf{E}(\max(M_n, 0)) \leq C < \infty$ , then there is  $M_\infty \in L_1(\mathbf{P})$  such that  $M_n \rightarrow M_\infty$  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 occurring 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 should 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 independent 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  $i \leq j$  implies  $X_i \leq X_j$ .
2. Disjoint intervals  $(i_k, j_k)$  have independent differences  $X_{j_k} - X_{i_k}$ , and if  $i \leq j$ , then  $X_j - X_i$  is equal in distribution to  $X_{j-i}$ .
3. The Process satisfies the equations

$$\mathbf{P}(X_t = 1) = \lambda \Delta t + o(t) \quad (5.1)$$

$$\mathbf{P}(X_t = 0) = 1 - \lambda \Delta t + o(t) \quad (5.2)$$

$$\mathbf{P}(X_t > 1) = o(t) \quad (5.3)$$

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

$$P'_0 = -\lambda P_0 \quad P_0(0) = 1$$

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  $\text{Poisson}(k, \lambda)$  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  $\tau_k = \inf\{t : X_t \geq k\}$ . 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  $\tau_k$  should satisfy the ‘memory loss’ property

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

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  $\mathbf{P}(\tau_{k+1} - \tau_k \geq t) = \mathbf{P}(\tau_1 \geq t) = e^{-\lambda t}$  for some  $\lambda$ .

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

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

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

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

Consider the following calculation

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

So that in a Poisson process, we should expect to wait on average  $\lambda^{-1}$  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

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

We shall also assume a continuity requirement that

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

A process is **time-homogenous** if

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

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 infinitesimal 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  $\alpha$  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  $\mathbf{P}(\tau_x \leq t | X_0 = x) = 1 - e^{-\alpha(x)t}$ .

Assume our state space is finite, and enumerate the states  $x_1, \dots, x_n$ . 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 infinitesimal generator of the chain. If  $\mu_t$  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 infinitesimal 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  $t \rightarrow \infty$ ,  $\mu_t \rightarrow (2/3, 1/3)$ .

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 continuous chains). The  $\mu_t$  will converge to the single eigenvector, invariant of the initial distribution, and this is the unique  $\mu$  for which  $\mu P = 0$ .

Suppose we want to compute the mean passage times  $E(\rho_y)$ , 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

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

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

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

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  $\lambda$  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 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  $1, 2, \dots, \infty$ .

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  $\lambda$ , and  $\mu_i = \mu$  for a fixed  $\mu$ . 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  $n \leq c$ , 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  $\mu$ . 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,

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

Hence

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

And thus the chain is transient if and only if

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