Stochastic Processes

Jacob Denson

October 21, 2017

Table Of Contents

Chapter 1

Stochastic Processes

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 2

Finite Markov Chains

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Let us define the transition operator P by the formula

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

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

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

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

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

Given a function f, the formal calculation

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

verifies that P really does act like a dual operator.

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

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

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

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

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

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

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

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

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

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

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

2.1 Asymptotics of Markov chains

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

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

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

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

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

Hence,

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

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

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

2.2 Irreducibility

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

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

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

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

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

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

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

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

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

2.3 Irreducibility and Potential Functions

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

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

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

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

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

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

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

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

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

2.4 Existence of a Stationary Distribution

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

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

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

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

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

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

and thus the expected value is finite. □

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

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

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

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

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

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

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

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

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

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

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

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

so interchanging the summation, we find

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

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

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

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

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

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

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

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

2.5 Perron-Frobenius

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

so $\lambda = |\lambda|$.

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

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

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

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

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

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

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

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

Then, by the product rule,

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

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

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

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

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

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

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

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

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

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

Example. *Consider a process with transition matrix*

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

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

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

2.6 Aperiodicity and Irreducibility

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.7 Periodicity and Average State Distributions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.8 Stopping Times

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

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

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

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

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

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

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

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

And is a stopping time.

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

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

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

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

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

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

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

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

So the process is strongly Markov. □

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

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

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

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

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

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

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

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

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

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

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

at x. We have

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

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

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

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

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

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

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

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

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

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

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

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

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

So

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

And the general form is

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

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

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

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

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

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

Chapter 3

Countable-State Markov Chains

3.1 General Properties

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

3.2 Recurrence and Transience

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

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

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

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

random variable, we can take expectations

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

example we therefore need

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

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

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

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

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

This is infinite, so the process is null recurrent.

Chapter 4

Branching Processes

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

Galton and Watson represented the spread of families by a succeeding discrete number of generations $X_0, X_1, \ldots$, where the initial generation X_0 produces the offspring X_1, which produces the offspring X_2, and so on, through the ages. Each time epoch represents 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 a family tree results in such an asexual process. These assumptions are equivalent to saying that X_t is a Markov chain with a certain probability transition function, which we now define. Fix some distribution ρ over $\mathbf{N}$, which represents the distribution of a particular person's children, and an initial probability distribution X_0, also over $\mathbf{N}$. We define a stochastic process $\{X_i\}$ by considering the transition probabilities

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

Where $(\rho * \rho * \cdots * \rho)$ is the n-fold convolution of ρ. More vicerally, we can construct X by considering an infinite grid of independent and identically distributed random variables $Y_{ij} \sim \rho$, and defining

$$X_{n+1} = \sum_{i=1}^{X_n} Y_{in}$$

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

4.1 The Distribution of the n'th Generation

One defining property of the random variables X_n is that they are defined in terms of sums of *independent* random variables. This means that the random variable will probably behave well under certain Fourier transform methods, which utilize the exponential function to transform sums in an easy to control way. One probabilistic Fourier transform method is to calculate probability generating functions. Given X_n, we consider the analytic function

$$G_n(t) = \mathbf{E}\left[t^{X_n}\right] = \sum_{k=0}^{\infty} \mathbf{P}(X_n = k) t^k$$

which is well defined and analytic for $0 \leqslant t \leqslant 1$. We can calculate

$$\begin{aligned}\mathbf{E}\left[t^{X_n} \middle| X_{n-1} = i\right] &= \sum_{j=0}^{\infty} t^j \mathbf{P}(X_n = j | X_{n-1} = i) \\ &= \sum_{j=0}^{\infty} t^j (\rho * \cdots * \rho)(j)\end{aligned}$$

We note that the k fold convolution $\rho * \cdots * \rho$ is the distribution of a sum of k independent random variables $Y_1, \ldots, Y_k$ distributed according to ρ, and so by independence

$$\sum_{j=0}^{\infty} t^j (\rho * \cdots * \rho)(j) = \mathbf{E}\left[t^{\sum Y_i}\right] = \prod \mathbf{E}\left[t^{Y_i}\right] = \mathbf{E}\left[t^Y\right]^k = G(t)^k$$

where $G(t)$ is the probability generating function corresponding to Y (this is a general consequence of the fact that Fourier methods turn convolution into multiplication). This implies that $\mathbf{E}[t^{X_n}|X_{n-1}] = G(t)^{X_{n-1}}$, and so if G is the probability generating function corresponding to Y. Applying the tower formula, we conclude that

$$G_n(t) = \mathbf{E}\left[G(t)^{X_{n-1}}\right] = G_{n-1}(G(t))$$

and so $G_n(t) = (G_0 \circ G^n)(t)$.

4.2 Mean Population Size

We shall start by understanding how the mean size of the evolution varies over time. Note that the power series representation of G_n guarantees that

$$G_n'(t) = \sum_{k=0}^{\infty} k\mathbf{P}(X_n = k)t^{k-1} = \mathbf{E}[X_n e^{tX_n}]$$

which implies $G_n'(0) = \mathbf{E}[X_n]$, so G_n can tell us the expectations of the functions X_n. The relation $G_{n+1}(t) = (G_n \circ G)(t)$ tells us that

$$G_{n+1}'(t) = G'(t)(G_n' \circ G)(t)$$

and in particular, this means

$$\mathbf{E}[X_{n+1}] = G_{n+1}'(0) = G'(0)G_n'(0) = \mu\mathbf{E}[X_n]$$

because $G'(0) = \mathbf{E}[Y]$, where $Y \sim \rho$. This means $\mathbf{E}[X_n] = \mu^n \mathbf{E}[X_0]$. 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 $\mu = 1$. The intuitive reason why is that even if the average population is maintained, eventually you will get unlucky and end up with a generation producing no offspring, which will end you entire family line.

4.3 Probability of Extinction

Regardless of your average population growth, provided $\rho(0) > 0$ there is a chance that the population will eventually become extinct. Indeed, $\mathbf{P}(X_{n+1} = 0 | X_n = k) = \rho(0)^k > 0$. We now discuss the probability $\pi \in [0,1]$ that that extinction occurs. We calculate that

$$\pi = \mathbf{P}\left(\liminf_{n\to\infty} \{X_n = 0\}\right) = \mathbf{P}\left(\lim_{n\to\infty} \{X_n = 0\}\right) = \lim_{n\to\infty} \mathbf{P}(X_n = 0) = \lim_{n\to\infty} \pi_n$$

where π_n is the probability of extinction in n steps. If $\mu < 1$, we know that processes almost surely become extinct, because we can apply Markov's inequality to conclude that

$$\pi_n = \mathbf{P}(X_n = 0) = 1 - \mathbf{P}(X_n \geqslant 1) \geqslant 1 - \mathbf{E}(X_n) = 1 - \mu^n \mathbf{E}(X_0)$$

For $\mu < 1$, this value converges to 1 as $n \to \infty$. It is more difficult to calculate the extinction probability for $\mu \geqslant 1$, but the probability generating function provides a powerful tool to calculate this probability.

For now, we assume that $X_0 = 1$. It then follows that $G_0(t) = t$, so $G_n(t) = G^n(t)$. The generating function's construction implies

$$\pi_n = G_n(0) = G(G_{n-1}(t)) = G(\pi_{n-1})$$

allowing us to calculate π_n recursively. Letting $n \to \infty$ on both sides of this equation, using the continuity of G on $[0,1]$, gives $\pi = G(\pi)$. We calculate that

$$G(0) = \rho(0) \geqslant 0$$

$$G(1) = \sum_{k=0}^{\infty} \mathbf{P}(Y = k) = 1$$

Since all the coefficients in the expansion of G are positive, we know that $G'(x), G''(x) \geqslant 0$ for $x > 0$, so G is convex and increasing in $(0,1)$. If $G'(x), G''(x)$ is *strictly* greater than zero on $(0,1)$, then we can conclude G is *strictly convex*. This occurs except in the special case that $\rho(0) + \rho(1) = 1$, and in these cases we find $G(x) = \rho(0) + \rho(1)x$, so either

- $\rho(1) = 1$: population size stays constant at every generation, and so if $X_0 = 1$, $\pi = 0$.
- $\rho(0) > 0$: We calculate $\pi_{n+1} = G(\pi_n) = \rho(0) + \rho(1)\pi_n$, which gives

$$\pi_n = \sum_{k=0}^{n-1} \rho(1)^k \rho(0) = \frac{1-\rho(1)^n}{1-\rho(1)}\rho(0) = 1 - \rho(1)^n$$

  and so we easily see that since $\rho(1) \neq 1$, $\pi_n \to 1$.

The fact that G is increasing and strictly convex implies $G(x) = x$ has *at most one* solution x_0 in $(0,1)$. If x_0 exists, then it follows that if $\pi_0 \leqslant x_0$, then $\pi_n \to x_0$, and if $\pi_0 > x_0$, then $\pi_0 \to 1$. In most cases, we assume that $X_0 = 1$, so that the extinction probability is always x_0.

- If $\mu \leqslant 1$, then because $G'' > 0$, we conclude $G'(x) < 1$ for all $x \in [0,1)$, which forces $x < G(x)$ for all $x \in [0,1)$. We conclude that $\pi_n \to 1$, so populations become extinct almost surely.
- If $\mu > 1$, then the fact that $G'(x)$ decreases continuously, we can conclude that $y < G(y)$ in a suitably small neighbourhood of 1. Since $G(0) = \rho(0) > 0$, we conclude by the intermediate value theorem that there is a point $x_0 \in (0,1)$ with $G(x_0) = x_0$, and this gives the convergence result considered above.

The case where $\mu = 1$ has one of the most interesting features of our model. For $X_0 = 1$, we conclude that $\mathbf{E}[X_n] = 1$ for all n, but $X_n \to 0$ almost surely. We can infer from this that $\mathbf{E}[X_n | X_n \neq 0] = \mathbf{E}[X_n]/\pi_n \to \infty$, so that if a population has survived for a long time, and is not extinct, we can guarantee that it has a huge population. This has applications to the theory of surnames that Gatson and Walton were reasoning about. Chinese surnames are ancient. Applying our model, we see that the names that have survived over the generations should be very prominant. There are approximately 3,000 Chinese last names in use nowadays, as compared to 12,000 in the

past, even though there are far more Chinese people in the world than in the past. This is the reason Gatson and Walton concluded upper class surnames were going extinct in Victorian Britain. The elite few who had these names were in populations that were likely to die out very soon, whereas the common names are names which will last much longer.

Example. *Suppose $\rho(0) = \rho(1) = 1/4$, $\rho(2) = 1/2$. Then the probability generating function is*

$$G(x) = \frac{2x^2 + x + 1}{4}$$

Solving the equation $G(x) = x$ gives $x = 1/2$, and this is the extinction probability of a branching process corresponding to ρ with $X_0 = 1$.

If $X_0 = k$ for some $k > 0$, one can prove that X is identically distributed to the sum of k independent branching processes $Y^1, \ldots, Y^k$, with $Y_0^i = 1$. It then follows that if we denote the probability that a population becomes extinct in n steps beginning with k people by $\pi_n(k)$, then

$$\pi_n(k) = \mathbf{P}(X_n = 0) = \mathbf{P}(Y_n^1 = 0, \ldots, Y_n^k = 0) = \prod \mathbf{P}(Y_n^i = 0) = \pi_n^k$$

Letting $n \to \infty$ gives $\pi(k) = \pi^k$. More generally, for any random variable X_0 the chance of dying is equal to $\mathbf{E}[\pi^{X_0}]$.

Example. *A simple variant of the branching process problem is to add the condition that some members of the population live on until the next generation to have more offspring. Thus we have a offspring distribution ρ, as well as some probability $q \in [0,1]$ of a particular individual dying at the end of each generation. If we assume the offspring production probabilities are independent of the probability that a member of the population dies off, then we can see this as just a case of the branching process with offspring distribution ν, where*

$$\nu(k) = q\rho(k) + (1-q)\rho(k-1)$$

If μ is the mean number of offspring given by ρ, then we find that if $Y \sim \nu$, then we find that the mean number of offspring given by ν is

$$\begin{aligned}
\sum_{k=0}^{\infty} k\nu(k) &= q \sum_{k=0}^{\infty} k\rho(k) + (1-q) \sum_{k=1}^{\infty} k\rho(k-1) \\
&= q\mu + (1-q) \sum_{k=1}^{\infty} (k-1)\rho(k-1) + (1-q) \\
&= \mu + (1-q)
\end{aligned}$$

Thus if $\mu \leqslant q$, the population is guaranteed to become extinct, and if $\mu > q$, then the population can sustain itself indefinitely.

4.4 Martingales and Branching Asymptotics

Using the Markov property of the branching process, we know that

$$\mathbf{E}[X_{n+1}|X_1,\ldots,X_n] = \mu X_n$$

If we define the process $M_n = X_n/\mu^n$ which in some sense measures the exponential growth of the process relative to μ, then we find

$$\mathbf{E}[M_{n+1}|M_1,\ldots,M_n] = M_n$$

This means that M_n is a *martingale* with respect to its natural filtration.

Now we can calculate that

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_n]}{\mu^n} = \frac{\mu^n \mathbf{E}[X_0]}{\mu^n} = \mathbf{E}[X_0]$$

so provided that $\mathbf{E}[X_0] < \infty$, we can conclude that M_n converges almost everywhere to an integrable random variable M_∞. This means that for almost all ω, we have $X_n(\omega) = [M_\infty(\omega) + o(1)]\mu^n$, so that the process essentially has exponential growth relative to μ^n. We might be tempted to think that M is now extended to be a martingale on $\{1,2,\ldots,\infty\}$, but we have to be a bit more careful. For instance, if $\mu \leqslant 1$, then extinction is almost sure to happen in a finite amount of time, and we can conclude that $M_\infty = 0$ almost everywhere (this implies $X_n(\omega) = o(1)\mu^n$ almost surely), whereas in all but the most trivial cases $\mathbf{E}[M_0] \neq 0$, so $\mathbf{E}[M_\infty] \neq \mathbf{E}[M_0]$. However, we can calculate that if $Y_1, Y_2, \ldots$ are independent random variables distributed according to ρ, then provided ρ has finite variance σ^2, we can conclude that

$$\mathbf{E}[X_{n+1}^2|X_n] = \mathbf{E}\left(\sum_{i=1}^{X_n} Y_i\right)^2 = \sum_{ij=1}^{X_n} \mathbf{E}[Y_i Y_j] = X_n^2\mu^2 + X_n\sigma^2$$

so

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

and in particular, this means

$$\mathbf{E}[M_{n+1}^2] = \mathbf{E}[M_n^2] + \mathbf{E}[M_n]\frac{\sigma^2}{\mu^{n+2}}$$

Reexpressing the reccurence gives

$$\mathbf{E}[M_n^2] = \mathbf{E}[M_0^2] + \sum_{k=0}^{n-1} \mathbf{E}[M_k]\frac{\sigma^2}{\mu^{k+2}} = \sum_{k=0}^{n-1} \mathbf{E}[M_0]\frac{\sigma^2}{\mu^{k+2}}$$

This means that the martingale M_n is bounded in $L^2(\Omega)$ if and only if $\mu > 1$ (except if $X_0 = 0$ of course). We may now apply Doob's L^2 convergence theorem to conclude that M_n converges in the L^2 norm to M_∞, so this gives that M_n converges to M_∞ in the L^1 norm. We can therefore conclude that

$$\mathbf{E}[M_\infty] = \mathbf{E}[M_0]$$

which tells us that X_n grows on average on the order of μ^n. What's more, we can conclude the additional fact that

$$\begin{aligned}\mathbf{V}[M_\infty] &= \lim_{n\to\infty} \mathbf{V}[M_n] = \lim_{n\to\infty} \mathbf{E}[M_0^2] + \sum_{k=0}^{n-1} \mathbf{E}[M_0]\frac{\sigma^2}{\mu^{k+2}} - \mathbf{E}[M_0]^2 \\ &= \mathbf{V}[M_0] + \frac{\sigma^2}{\mu(\mu-1)}\mathbf{E}[X_0]\end{aligned}$$

So M_∞ has low variance for large values of μ, implying that the growth of X_n is more steadily close to μ^n.

We can actually determine the distribution of M_∞ *exactly*, by using Fourier transform techniques. Unfortunately, M_n isn't defined over a discrete set, so the probability generating functions are no longer well defined, but we can consider the moment generating functions

$$H_n(t) = \mathbf{E}[e^{tM_n}] = G_n(\exp(t\mu^{-n}))$$

$$H_\infty(t) = \mathbf{E}[e^{tM_\infty}]$$

If $t \leqslant 0$, then $e^{tM_n} \leqslant 1$, because M_n is non-negative. We can therefore apply the dominated convergence theorem to conclude

$$H_\infty(t) = \lim_{n\to\infty} H_n(t)$$

This allows us to compute H_∞ on an interval, which by the analytic properties of the function will enable us, in theory, to calculate the distribution of M_∞. In practice, however, this is only computable in the most basic of examples.

We can derive a functional equation which will enable us to calculate H_∞. Note that if $X_0 = 1$, then

$$\begin{aligned} H_{n+1}(\mu t) &= G_{n+1}(\exp(t\mu^{-n})) \\ &= G^{n+1}(\exp(t\mu^{-n})) \\ &= G(G^n(\exp(t\mu^{-n})) = G(H_n(t)) \end{aligned}$$

Letting $n \to \infty$ on both sides tells us that $H_\infty(\mu t) = G(H_\infty(t))$.

Example. *In the example $\rho(0) = \rho(1) = 1/4$, $\rho(2) = 1/2$, we conclude that for $t \leqslant 0$*

$$H_\infty(5t/4) = \frac{1 + H_\infty(t) + 2H_\infty(t)^2}{4}$$

If we assume $H_\infty(-1) = \alpha$, then

$$\alpha = \frac{1 + H_\infty(-4/5) + 2H_\infty(-4/5)^2}{4}$$

$$0 = \frac{1 - 4\alpha + H_\infty(-4/5) + 2H_\infty(-4/5)^2}{4}$$

$$H_\infty(-4/5) = \frac{-1 + \sqrt{1 - 8(1 - 4\alpha)}}{4} = \frac{-1 + \sqrt{32\alpha - 7}}{4}$$

so $32\alpha \geqslant 8$, hence $\alpha \geqslant 1/4$. But it seems impossible to calculate the actual iterates of this map, so we can't calculate the actual distribution – however, we can use a computer to approximate these limits, and then perform an inverse transform to calculate the distribution of M_∞ approximately.

Example. *About the only mathematically feasible example where we can compute the distribution is when the number of children have a geometric distribution $\rho(k) = pq^k$, for some $0 < p < 1$, $q = 1 - p$. One way to think about the geometric distribution is as the distribution of waiting times until we see a first success in a series of independent Bernoulli trials, each with a success probability p. Thus we can see this example as where people keep having children as many times as possible, until the first failure (a miscarriage?) which causes*

the generation to die off. Now we can check that the probability generating function of this distribution is

$$G(t) = \sum_{k=0}^{\infty} p(qt)^k = \frac{p}{1-qt}$$

hence

$$G'(t) = \frac{pq}{(1-qt)^2}$$

and so

$$\mu = G'(1) = \frac{pq}{(1-q)^2} = \frac{q}{p}$$

we immediately see that if $q \leqslant p$ ($p \geqslant 1/2$), then the population is guaranteed to become extinct, and if $q > p$, then since

$$G(p/q) = \frac{p}{1-p} = p/q$$

we conclude the extinction probability is p/q. The nice fact about G is that it is a rational function of t, represented by the Möbius transformation

$$G(t) = \begin{pmatrix} 0 & p \\ -q & 1 \end{pmatrix}(t)$$

Thus

$$G^n(t) = \begin{pmatrix} 0 & p \\ -q & 1 \end{pmatrix}^n (t)$$

and by diagonalization, we can calculate

$$\begin{aligned}
\begin{pmatrix} 0 & p \\ -q & 1 \end{pmatrix}^n &= \frac{1}{q-p} \begin{pmatrix} 1 & p \\ 1 & q \end{pmatrix} \begin{pmatrix} p^n & 0 \\ 0 & q^n \end{pmatrix} \begin{pmatrix} q & -p \\ -1 & 1 \end{pmatrix} \\
&= \frac{1}{q-p} \begin{pmatrix} p^n & pq^n \\ p^n & q^{n+1} \end{pmatrix} \begin{pmatrix} q & -p \\ -1 & 1 \end{pmatrix} \\
&= \frac{1}{q-p} \begin{pmatrix} p^n q - pq^n & pq^n - p^{n+1} \\ p^n q - q^{n+1} & q^{n+1} - p^{n+1} \end{pmatrix} \\
&= \frac{p^{n+1}}{q-p} \begin{pmatrix} \mu - \mu^n & \mu^n - 1 \\ \mu - \mu^{n+1} & \mu^{n+1} - 1 \end{pmatrix}
\end{aligned}$$

so that

$$G^n(t) = \frac{p\mu^n(1-t) + qt - p}{q\mu^n(1-t) + qt - p}$$

If $\mu < 1$, $G^n(t) \to 1$, reflecting the fact that the process eventually dies out (the moment generating function of the dirac delta distribution is the constant 1 distribution). If $\mu = 1$, then this calculation doesn't quite work, but a modification shows $G^n(t) \to 1$ also. If $\mu > 1$, then we can calculate that

$$\begin{aligned} H_\infty(t) &= \lim_{n\to\infty} H_n(t) = \lim_{n\to\infty} G_n(\exp(-t/\mu^n)) \\ &= \lim_{n\to\infty} \frac{p\mu^n(1-e^{-t/\mu^n}) + qe^{-t/\mu^n} - p}{q\mu^n(1-e^{-t/\mu^n}) + qe^{-t/\mu^n} - p} \\ &= \lim_{n\to\infty} \frac{pt + q - p + O(t/\mu^n)}{qt + q - p + O(t/\mu^n)} = \frac{pt+q-p}{qt+q-p} \\ &= \frac{\pi t + (1-\pi)}{t + (1-\pi)} = \pi + \frac{(1-\pi)^2}{t+(1-\pi)} \\ &= \pi + \int_0^\infty (1-\pi)^2 e^{-tx} e^{-(1-\pi)x} dx \end{aligned}$$

where $\pi = p/q$ is the extinction probability. It follows that $\mathbf{P}(M_\infty = 0) = \pi$, and on $(0,\infty)$, M_∞ is a continous random variable with distribution function

$$f_{M_\infty} = (1-\pi)^2 e^{-(1-\pi)x}$$

which is certainly an interesting result.

It turns out that, though $M_\infty = 0$ almost surely when $\mu \leqslant 1$, we can still determine interesting asymptotic results when $\mu < 1$. Indeed, we ask what the distribution of M_n is, conditional on the fact that $M_n \neq 0$. Then

$$\mathbf{E}[t^{X_n} | X_n \neq 0] = \frac{G_n(t) - G_n(0)}{1 - G_n(0)} = \frac{\alpha_n t}{1 - \beta_n t}$$

where

$$\alpha_n = \frac{p-q}{p-q\mu^n} \qquad \beta_n = \frac{q(1-\mu^n)}{p-q\mu^n}$$

so $0 < \alpha_n < 1$ and $\alpha_n + \beta_n = 1$. As $n \to \infty$, $\alpha_n \to 1-\mu$, $\beta_n \to \mu$, so

$$\lim_{n\to\infty} \mathbf{P}(X_n = k | X_n \neq 0) = (1-\mu)\mu^{k-1}$$

so we see that the distribution grows asymptotically exponentially. If $\mu = 1$*, then one can see by induction that*

$$G_n(t) = \frac{n-(n-1)t}{(n+1)-nt}$$

and that

$$\mathbf{E}(e^{-tX_n/n}|X_n \neq 0) \to \frac{1}{1+t}$$

which corresponds to

$$\mathbf{P}(X_n/n > x|X_n \neq 0) \to e^{-x}$$

so we get a form of logarithmic growht.

Chapter 5

Reversibility

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

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

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

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

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

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

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

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

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

So the walk is reversible with respect to μ.

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

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

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

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

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

Consider a markov chain with the following transition. We start with an initial matrix X_0 in $\mathcal{T}$, and we pick a random entry (i, j). Let Y be

the matrix resultant from flipping the X_{ij} on or off. If $Y \in \mathcal{T}$, let $X_1 = Y$. Otherwise, let $X_1 = X$. Continue this process indefinitely. This is an irreducible, symmetric markov process in $\mathcal{T}$, with transitions

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

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

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

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

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

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

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

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

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

Chapter 6

Conditional Expectations

Most of the theory of random processes is connected with understanding how certain values of a process influence the process later on in time. When studying Markov chains, we tried to understand the relationship by directly considering the processes' transition coefficients. In the theory of martingales, we instead study random processes by looking at how the evolution of a stochastic process changes if we fix states to certain time points. 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. We find that the elementary definition of conditional expectations introduced in elementary probability theory breaks down when we begin to look at more general classes of random variables, and we introduce Kolmogorov's general definition of a conditional expectation with respect to a σ algebra to compensate.

Example. *Consider the Polya urn 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. We can then calculate*

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

This is the equation which we will see defines a martingale. Now we can calculate inductively that $\mathbf{P}(X_n = k) = (n+1)^{-1}$ for all $1 \leqslant k \leqslant n+1$, as

$\mathbf{P}(X_0 = 1) = 1$, *and*

$$\begin{aligned}\mathbf{P}(X_n = k) &= \sum_{m=1}^{n} \mathbf{P}(X_{n-1} = m)\mathbf{P}(X_n = k | X_{n-1} = m) \\ &= \mathbf{P}(X_{n-1} = k-1)\mathbf{P}(X_n = k | X_{n-1} = k-1) \\ &\quad + \mathbf{P}(X_{n-1} = k)\mathbf{P}(X_n = k | X_{n-1} = k) \\ &= \frac{1}{n}\frac{k-1}{n+1} + \frac{1}{n}\frac{n+1-k}{n+1} = \frac{1}{n+1}\end{aligned}$$

This means that $M_n = (n+2)^{-1}X_n$ *converges in distribution to a uniform distribution over* $[0,1]$. *On the other hand, suppose we start off with two white balls in the urn, and one black ball. Then we find* $M_n = X_n/(n+3)$ *still satisfies the martingale equation, but we find that for* $2 \leqslant k \leqslant n+2$,

$$\mathbf{P}(X_n = k) = \frac{2(k-1)}{n(n+1)}$$

which in a sense says that X_n *is much more likely to be bigger than smaller. As* $n \to \infty$, M_n *becomes much more concentrated at large values of* $[0,1]$. *In fact, one can calculate that* M_n *converges in distribution to a* β *distribution with parameters* $\alpha = 2$ *and* $\beta = 1$. *Thus changing the initial values of the process slightly caused an entirely different evolution of the process.*

6.1 Classical Conditioning

Recall that for discrete random variables X and Y, we can calculate conditional probabilities and expectations by

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

$$\mathbf{E}(Y | X = x) = \sum_{y} y\mathbf{P}(Y = y | X = x)$$

defined whenever $\mathbf{P}(X = x) \neq 0$. For continuous random variables, we can consider the joint densities $f_{X,Y}$, along with the individual densities f_X and f_Y, and then define

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

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

defined where $f_X(x) \neq 0$. However, these two classical formulations are insufficient to cover conditional expectations for the general random variables we encounter 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 $\mathbf{E}(Y|X=x)$ which works for almost every random variable we encounter in practice, and also leads to the most elegant definitions in the theory of martingales.

6.2 Kolmogorov's Realization

Kolmogorov realized we can think of conditional values as 'best guesses' of the values of a random variable given some known information about the system. 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 $B = X^{-1}(A)$, we have the equations

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

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

  where $A \subset \mathbf{R}$ is a set of real values, which have a common measure-theoretic equation

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

  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 $f_Y(y)\mathbf{E}(Y|X=x) = \int y f_{X,Y}(x,y)\, dy$ holds almost everywhere, and this defines $\mathbf{E}(Y|X=x)$ up to a set of measure zero, if we ignore the values y where $f_Y(y) = 0$. In particular, if we can choose $\mathbf{E}(Y|X=x)$ to be a continuous function of x, then it is the unique continuous function satisfying the integral equation.

In general, for *any* two random variables X and Y, we say a random variable $Z = f(X)$ is a *version* of $\mathbf{E}(Y|X)$ if the two conditions above are satisfied for Z. That is, if Z can be expressed as a function of X (the function f in the definition), and if for any set $B = X^{-1}(A)$,

$$\int_B Z d\mathbf{P} = \int_B X d\mathbf{P}$$

This can also be expressed as saying

$$\int_A f(x) d\mathbf{P}_X(x) = \int_A f(x) d\mathbf{P}_X(x)$$

However, the definition can certainly be simplified by noting that once we consider $\mathbf{E}(X|Y)$ as a function on Ω, rather than as a function of the 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. If we consider the σ algebra $\sigma(X)$, then if we know the value of χ_E for each $E = X^{-1}(F)$, this should be sufficient knowledge to calculate the expectation $\mathbf{E}(Y|X)$, rather than knowing the actual values of X. The Doob-Dynkin lemma guarantees that if X and Y are real-valued, then Y is a function of X if and only if Y is $\sigma(X)$ measurable. This means the first property of conditional expectation can be reduced to a statement about σ algebras.

Lemma 6.1 (Doob-Dynkin)**.** *A real-valued random variable Y is $\sigma(X)$ measurable if and only if $Y = f(X)$ for some Borel-measurable $f : \mathbf{R} \to \mathbf{R}$.*

Proof. If $Y = \sum a_i \chi_{A_i}$ is a simple function, 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 Borel measurable function $f = \sum a_i \chi_{B_i}$, and we find $Y = f(X)$. If Y is a general non-negative $\sigma(X)$ random variable, we can consider an increasing sequence $Y_1, Y_2, \ldots$ of non-negative simple $\sigma(X)$ random variables converging monotonely to

Y. Applying the previous result, we can write $Y_i = f_i(X)$ for some Borel measurable functions f_i. If we consider any sample point ω, then

$$Y(\omega) = \lim Y_i(\omega) = \lim f_i(X(\omega))$$

Thus if we set E to be the subset of points x in $\mathbf{R}$ where $f_i(x)$ converges, then $X(\Omega) \subset E$. Since the set E is Borel measurable, the functions $\chi_E f_i$ are Borel measurable, and converge everywhere to a Borel measurable function f, and it is easy to verify that $Y = f(X)$. □

The intuitive way we should think about the conditional values is as a 'best guess' of the probability values given that some information is known about the system at some time. If we think of the information about the random variable X being given by the σ-algebra $\sigma(X)$, then it isn't too much to model arbitrary 'sets of information' about a probability space by a sub σ-algebra Σ of the σ-algebra defining the space. In this case, we should interpret $\mathbf{E}(X|\Sigma)$ as giving a best guess of the value of X, given that we know the value of all Σ measurable functions ahead of time. We say a random variable X is **adapted** to a σ-algebra Σ if X is measurable with respect to Σ. This means, essentially, that we 'know' the value of X if we know the information contained in Σ.

6.3 General Conditional Expectations

With all this terminology set in stone, we can now formulate Kolmogorov's theory of conditional expectations. For a given Σ algebra, and a random variable X, a random variable $\mathbf{E}(X|\Sigma)$ is a *version* of a **conditional expectation** with respect to Σ if it is adapted to Σ, and if

$$\int_S \mathbf{E}(X|\Sigma) = \int_S X$$

for any $S \in \Sigma$. In a sense, $\mathbf{E}(X|\Sigma)$ is the best approximation to X, given that we know the information in Σ.

Example. *The σ algebra $\Sigma = \{\varnothing, \Omega\}$ is the smallest σ algebra over Ω, and represents a set of 'no information at all'. If X is any integrable random variable, then the constant function $\mathbf{E}(X)$ is a version of a conditional expectation*

for X, because

$$\int_{\varnothing} \mathbf{E}(X) = 0 = \int_{\varnothing} X \qquad \int_{\Omega} \mathbf{E}(X) = \mathbf{E}(X) = \int_{\Omega} X$$

Thus, given the presense of no information at all, the best constant approximation we can have of X is $\mathbf{E}(X)$.

Despite the technical definition, the existence and almost-sure uniqueness of conditional expectations is relatively easy to prove in $\mathcal{L}^1(\Omega)$, thanks to the Radon-Nikodym theorem in measure theory.

Theorem 6.2. *If $X \in \mathcal{L}^1(\Omega)$, then $\mathbf{E}(X|\Sigma)$ exists in $\mathcal{L}^1(\Sigma)r$, and is unique up to a set of measure zero.*

Proof. First, assume $X \geqslant 0$. Then the map $\mathbf{P}_\Sigma : S \mapsto \int_S X d\mathbf{P}$ is a *finite measure* over Σ which is absolutely continuous with respect to $\mathbf{P}$. The Radon-Nikodym theorem asserts that there is a Σ-adapted random variable Y such that for any set S,

$$\int_S Y d\mathbf{P}_\Sigma = \int_S X d\mathbf{P}$$

This shows exactly that Y is a conditional expectation for X. It is easy to see that since $X \geqslant 0$, $Y \geqslant 0$ almost surely, and so

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

To verify 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. If X is not necessarily positive, then we can write $X = X^+ - X^-$, and it is simple to verify that $\mathbf{E}(X^+|\Sigma) - \mathbf{E}(X^-|\Sigma)$ is a conditional expectation for X. □

It is also easy to show that $\mathbf{E}(X|\Sigma)$ exists if $X \geqslant 0$, because we can write X as the monotone limit of simple functions X_n, which are in $L^1(\Omega)$, and then we can apply the monotone convergence theorem to verify that the pointwise limit of the $\mathbf{E}(X_n|\Sigma)$ satisfy the required integral formulas. To verify uniqueness, we note that if Y and Z are versions of the conditional expectation of X, then we can apply the subtraction trick to conclude that $\mathbf{P}(Y \neq Z, Y < \infty) = 0$, $\mathbf{P}(Y \neq Z, Z < \infty) = 0$. But now the only set remaining to analyze is where $Y = Z = \infty$, and of course $\mathbf{P}(Y \neq Z, Y = Z = \infty) = 0$, so Y and Z are equal almost everywhere.

6.4 Properties of the Conditioning Operator

Since $\mathbf{E}(X|\Sigma)$ is unique up to a set of measure zero, and X and Y agree almost everywhere, then $\mathbf{E}(X|\Sigma) = \mathbf{E}(Y|\Sigma)$, and we can consider conditional expectation as an *operator* on $L^1(\Omega)$. In particular, it is easy to verify from properties of the Lebesgue integral that

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

We also get variants of the standard convergence results of Lebesgue theory.

- (Monotone Convergence) If $0 \leqslant X_1 \leqslant X_2 \cdots \to X$, then $\mathbf{E}(X_i|\Sigma)$ converges almost surely to $\mathbf{E}(X|\Sigma)$.
- (Fatou) If $X_n \geqslant 0$ then $\mathbf{E}((\liminf X_n)|\Sigma) \leqslant \liminf \mathbf{E}(X_n|\Sigma)$ almost surely.
- (Dominated Convergence) If $|X_n| \leqslant Y$, $\int Y < \infty$, and $X_n \to X$ pointwise almost surely, then $\mathbf{E}(X_n|\Sigma) \to \mathbf{E}(X|\Sigma)$ pointwise almost surely.
- (Jensen) If f is a convex function with $\|f(X)\|_1 < \infty$, then we can consider the function $\mathbf{E}(f(X)|\Sigma)$, and $f(\mathbf{E}(X|\Sigma)) \leqslant \mathbf{E}(f(X)|\Sigma)$ almost surely.

The general idea is that the integral equations defining conditional expectation can be manipulated using the standard theorems of Lebesgue integrals.

Proof. To prove the monotone convergence theorem, note that it is obvious that $\mathbf{E}(X_i|\Sigma)$ are increasing and non-negative almost everywhere, and therefore we can apply monotone convergence to conclude that for each set S,

$$\int_S \mathbf{E}(X|\Sigma) = \int_S X = \lim_{n\to\infty} \int_S X_n = \lim_{n\to\infty} \int_S \mathbf{E}(X_n|\Sigma)$$

If we let $T = \{\omega : \limsup \mathbf{E}(X_n|\Sigma)(\omega) \leqslant \mathbf{E}(X|\Sigma)(\omega) - \varepsilon\}$, then the reverse Fatou lemma gives

$$\lim_{n\to\infty} \int_T \mathbf{E}(X_n|\Sigma) \leqslant \int_T \limsup \mathbf{E}(X_n|\Sigma) \leqslant \int_T \mathbf{E}(X|\Sigma) - \varepsilon = \int_T \mathbf{E}(X|\Sigma) - \mathbf{P}(T)\varepsilon$$

It follows that $\mathbf{P}(T) = 0$, and letting $\varepsilon \to 0$ shows that $\limsup \mathbf{E}(X_n|\Sigma) \geqslant \mathbf{E}(X|\Sigma)$ almost surely. Similar results show that $\liminf \mathbf{E}(X_n|\Sigma) \leqslant \mathbf{E}(X|\Sigma)$ almost surely, so that $\mathbf{E}(X_n|\Sigma) \to \mathbf{E}(X|\Sigma)$ almost surely. Now let's prove Fatou's theorem. If we set $Y_n = \inf_{k \geqslant n} X_k$, then Y_n tends monotically to $\liminf X_n$, so

$$\mathbf{E}(\liminf X_n|\Sigma) = \lim \mathbf{E}(Y_n|\Sigma)$$

and it suffices to show that $\lim \mathbf{E}(Y_n|\Sigma) \leqslant \liminf \mathbf{E}(X_n|\Sigma)$ almost surely. But this follows because $Y_n \leqslant X_n$, so $\mathbf{E}(Y_n|\Sigma) \leqslant \mathbf{E}(X_n|\Sigma)$, and we may then take limits, and liminfs. Verifying dominated convergence is easy. If $S \in \Sigma$ is given, then using the dominated convergence theorem gives

$$\int_S \mathbf{E}(X_n|\Sigma) = \int_S X_n \to \int_S X = \int_S \mathbf{E}(X|\Sigma)$$

using the same techniques as in the theorems above, we can conclude that $\mathbf{E}(X_n|\Sigma) \to \mathbf{E}(X|\Sigma)$ almost surely. To verify Jensen's inequality, we can apply the standard Jensen's inequality to conclude

$$\int_S \mathbf{E}(f(X)|\Sigma) = \int_S f(X) \geqslant f\left(\int_S X\right) = f\left(\int_S \mathbf{E}(X|\Sigma)\right)$$

and this implies the almost sure inequality. □

Here is a notable use of the conditional Jensen's inequality.

Proposition 6.3. *If $X \in L^p(\Omega)$, then $\mathbf{E}(X|\Sigma) \in L^p(\Omega)$.*

Proof. Applying Jensen's inequality, using the fact that $x \mapsto |x|^p$ is convex, and $\mathbf{E}(|X|^p) < \infty$, we conclude that $|\mathbf{E}(X|\Sigma)|^p \leqslant \mathbf{E}(|X|^p|\Sigma)$, and so

$$\|\mathbf{E}(X|\Sigma)\|_p^p = \int |\mathbf{E}(X|\Sigma)|^p \leqslant \int \mathbf{E}(|X|^p|\Sigma) = \int |X|^p = \|X\|_p^p < \infty$$

Thus conditional expectation is a contraction on each L^p space. □

The conditional expectation of L^2 measurable functions has important orthogonality properties, which show $\mathbf{E}(X|\Sigma)$ is the best Σ adapted approximation of X in the square mean error, which explains why conditional expectations occur so often in statistical applications.

Theorem 6.4. *If $X \in L^2(\Omega)$, then $\mathbf{E}(X|\Sigma)$ is the orthogonal projection of X onto the subspace of Σ adapted L^2 functions.*

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

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

which can be rewritten as

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

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

If X is already Σ measurable, then it is obviously true that $\mathbf{E}(X|\Sigma) = X$. In particular, $\mathbf{E}(X|X) = X$. More generally, we find that if X is Σ measurable, and in $L^p(\Omega)$, for $1 \leqslant p \leqslant \infty$, and if Y is Σ measurable, and in $L^q(\Omega)$, then XY is in L^1, and $\mathbf{E}(XY|\Sigma) = X\mathbf{E}(Y|\Sigma)$. This follows from the next lemma.

Lemma 6.5. *If X is in $L^p(\Omega)$, and Y is in $L^q(\Omega)$, then for any set $S \in \Sigma$,*

$$\int_S \mathbf{E}(X|\Sigma)Y = \int_S XY$$

Similarily, if $X \geqslant 0$ and $Y \geqslant 0$ then the formula holds.

Proof. Assume first that $X \geqslant 0$, from which the general theorem will follow by taking $X = X^+ - X^-$. As we noted, if Y is the indicator function of some element of Σ, the theorem is obvious by definition of conditional expectation. Applying linearity of the equation, this means that the equation holds if Y is any simple function. If $Y \geqslant 0$ is the limit of simple functions $Y_1 \leqslant Y_2 \leqslant \ldots$, monotone convergence implies

$$\int_S \mathbf{E}(X|\Sigma)Y = \lim \int_S \mathbf{E}(X|\Sigma)Y_i = \lim \int_S XY_i = \int_S XY$$

and this proves the theorem. The proof for general positive random variables follows from monotone convergence. □

6.5 Conditional Probabilities

We have only been discussing conditional expectation so far, but generalizing the formula $\mathbf{P}(E) = \mathbf{E}(\chi_E)$ tells us we should be able to define

$$\mathbf{P}(E|\Sigma) = \mathbf{E}(\chi_E|\Sigma)$$

This means that $\mathbf{P}(E|\Sigma)$ is no longer a number, it is a random variable, like a number that can look 'ahead of time' into the information contained in Σ to randomly improve upon our approximation of the probability of an event happening. If $E_1, E_2, \ldots$ are a countable sequence of disjoint events with union E, then $\chi_E = \sum \chi_{E_i}$, and applying monotone convergence we conclude that $\mathbf{P}(E|\Sigma) = \sum \mathbf{P}(E_i|\Sigma)$ almost everywhere. Thus in some sense, conditional probabilities follow the same rules as regular probabilities. However, if we consider the class of all measurable E, then we obtain a family of uncountable sets, and it doesn't seem quite as likely that the conditional expectations of indicators functions will always behave like probability distributions. We define a **regular conditional probability** for a distribution $\mathbf{P}$ on an algebra Σ, with respect to a σ algebra Δ as a map $\mathbf{P}(\cdot|\Delta) : \Sigma \times \Omega \to [0,1]$ such that for ever $E \in \Sigma$, the map $\omega \mapsto \mathbf{P}(E|\Delta)(\omega)$ is a version of $\mathbf{P}(E|\Delta)$, and for each ω, the map $E \mapsto \mathbf{P}(E|\Delta)(\omega)$ is a probability measure on Σ.

Example. *If X and Y are continuous random variables, then the density function $f_{X|Y} = f_{X,Y}/f_Y$ is the density for a regular conditional probability, because for any Borel set $B \subset \mathbf{R}$, the function*

$$\omega \mapsto \int_B f_{X|Z}(x|Z = Z(\omega))dx$$

is a version of $\mathbf{P}(X \in B|Z)$, and it is easy to see that this defines a probability distribution if ω is fixed.

Example. *Let X and Y be independent continuous random variables with a common distribution function F. Let's calculate $\mathbf{P}(X \leqslant t|Z)$, where $Z = \max(X, Y)$. If $Z \leqslant t$, then $X \leqslant t$ is guaranteed. On the other hand, if $Z > t$, then it is first necessary that $Y = Z$, which happens independently of Z with probability $1/2$, and then we try to determine the chance that $X \leqslant t$, given that*

$X \leqslant Z$. This heuristically justifies that

$$\begin{aligned}\mathbf{P}(X \leqslant t|Z) &= \mathbf{I}(Z \leqslant t) + \mathbf{I}(Z > t)\frac{\mathbf{P}(X \leqslant t|X \leqslant Z)}{2} \\ &= \mathbf{I}(Z \leqslant t) + \mathbf{I}(Z > t)\frac{\mathbf{P}(X \leqslant t)}{2\mathbf{P}(X \leqslant Z)} \\ &= \mathbf{I}(Z \leqslant t) + \mathbf{I}(Z > t)\frac{F(t)}{2F(Z)}\end{aligned}$$

Lets verify this is formally a conditional expectation. It suffices to integrate this function over $Z \leqslant u$, where $u \leqslant \infty$, since this is a π system, and in this case we need to verify that

$$\mathbf{P}(Z \leqslant \min(t,u)) + \frac{F(t)}{2}\int_{t<Z\leqslant u}\frac{d\mathbf{P}}{F(Z)} = \mathbf{P}(X \leqslant t, Z \leqslant u)$$

If $u \leqslant t$, we reasoned above that $\mathbf{P}(X \leqslant t, Z \leqslant u) = \mathbf{P}(Z \leqslant u)$. On the other hand, since the distribution of (X,Y) is a product measure, since X and Y are independant, we can apply Fubini's theorem, calculating

$$\int_{t<Z\leqslant u}\frac{d\mathbf{P}}{F(Z)} = \int_{\substack{x\leqslant y \\ t<y\leqslant u}}\frac{dF(x)dF(y)}{F(y)} + \int_{\substack{x>y \\ t<x\leqslant u}}\frac{dF(y)dF(x)}{F(x)} = 2[F(u) - F(t)]$$

$$\mathbf{P}(X \leqslant t, Z \leqslant u) = \mathbf{P}(X \leqslant t, Y \leqslant u) = F(t)F(u)$$
$$\mathbf{P}(Z \leqslant t) = \mathbf{P}(X \leqslant t, Y \leqslant t) = F(t)^2$$

and the verification is complete. Note that the specification we have given induces a regular conditional probability distribution, because if ω is fixed, then $Z(\omega) = z$ is fixed, then provided $F(z) \neq 0$, the values

$$\mathbf{P}(X \leqslant t|Z = z) = \mathbf{I}(z \leqslant t) + \mathbf{I}(z > t)\frac{F(t)}{2F(z)}$$

define a right countinuous function, non-decreasing of t whose value at $-\infty$ is $F(-\infty)/2F(z) = 0$, and whose value at ∞ is 1. Since $\mathbf{P}(F(Z) = 0)$ occurs with probability zero, we can edit the conditional probability function over this set so that we get a probability distribution everywhere.

Regular conditional probabilities exist on almost every space encountered in practice (for instance, they exist if Σ is the Borel algebra on a *Lusin space* Ω, that is, a space homeomorphic to a Borel subset of a compact metric space). (TODO: PROVE THIS).

Example (Halmos, Dieudonné, Andersen, Jessen)**.** *Consider the probability space* $[0,1]$ *with the standard Borel* σ *algebra and Lebesgue measure* μ*. Using the axiom of choice, construct a set* A *with inner Lebesgue measure 0 and outer Lebesgue measure 1 (so* A^c *has outer Lebesgue measure 1 as well). Let* Σ *be the* σ *algebra generated from Borel sets and* A*. Then a typical element of* Σ *can be written in the form*

$$B = (A \cap E) \cup (A^c \cap F)$$

where E *and* F *are Borel sets. It follows then that* $\mu^*(B \cap A) = \mu(E)$*, and* $\mu^*(B \cap A^c) = \mu(F)$*. We can therefore define a probability measure on* Σ *by setting*

$$\mathbf{P}(B) = \frac{\mu^*(B \cap A) + \mu^*(B^c \cap A)}{2} = \frac{\mu(E) + \mu(F)}{2}$$

We have essentially hid 'two copies' of $[0,1]$ *in itself. Assume that we have a conditional probability measure for* $\mathbf{P}$ *over Borel sets. If* $B \in \Sigma$*, and* E *is Borel measurable, then*

$$\begin{aligned}\int_E \mathbf{P}(A \cap B|B[0,1])d\mathbf{P} &= \mathbf{P}(A \cap E \cap B) \\ &= \frac{\mu^*(A \cap E \cap B)}{2} = \frac{\mu^*(E \cap B)}{2} = \int_E \frac{\chi_B}{2} d\mathbf{P}\end{aligned}$$

Thus $\mathbf{P}(A \cap B|B[0,1]) = \chi_B/2$ *almost surely for each set* B*. Since* $B[0,1]$ *is generated by a countable* π *system, and the maps* $B \mapsto \mathbf{P}(A \cap B|\Sigma)(\omega)$*,* $B \mapsto \chi_B(\omega)$ *are both measures for every* ω*, we have that* $\mathbf{P}(A \cap B|B[0,1])(\omega) = \chi_B(\omega)/2$ *for every Borel set* B *if and only if* $\mathbf{P}(A \cap B|B[0,1])(\omega) = \chi_B(\omega)/2$ *for every element in the* π *system, and so we conclude*

$$J = \left\{\omega : \mathbf{P}(A \cap B|B[0,1])(\omega) = \frac{\chi_B(\omega)}{2} \text{ for all Borel } B\right\}$$

is also Borel, and $\mathbf{P}(J) = 1$*, since it is the countable intersection of probability one sets. This means that we may 'plug* J *into itself', ala Russell's paradox, so we conclude if* $\omega \in J$*, then* $J - \{\omega\}$ *is also Borel, and so*

$$\mathbf{P}(A \cap J|B[0,1])(\omega) = \frac{\chi_J(\omega)}{2} \neq \frac{\chi_{J-\{\omega\}}(\omega)}{2} = \mathbf{P}(A \cap [J - \{\omega\}]|B[0,1])(\omega)$$

so that $A \cap J \neq A \cap [J - \{\omega\}]$*. This means* $\omega \in A$*, so* $J \subset A$*. But* A *has inner Lebesgue measure zero, whereas* J *has measure 1, which is impossible. Thus the conditional probability could never exist in the first place.*

6.6 Independence and Conditional Expectation

We know that a series of random variables $X_1, \ldots, X_n$ is independent if

$$\mathbf{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbf{P}(X_1 \in A_1) \ldots \mathbf{P}(X_n \in A_n)$$

for any Borel set $A_1, \ldots, A_n$. This essentially means that the information in each random variables X_k does not give any information about the other random variables. In this form, there seems there should be an obvious extension to σ algebras. We say a family of sigma algebras $\Sigma_1, \ldots, \Sigma_n$ is independent if

$$\mathbf{P}(A_1 \cap \cdots \cap A_n) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_n)$$

if $A_1 \in \Sigma_1, \ldots, A_n \in \Sigma_n$, and an infinite family is independent if every finite family is independent. If Σ contains no relevant information to X, then $\mathbf{E}(X|\Sigma)$ should essentially have no extra information to predict X, so we would have to conclude $\mathbf{E}(X|\Sigma) = \mathbf{E}(X)$. We prove this in a more general form below.

Theorem 6.6. *If Σ is independent of X, then $\mathbf{E}[X|\Sigma] = \mathbf{E}[X]$.*

Proof. If $X = \chi_A$, then A is independent of any set in Σ, and so for any $B \in \Sigma$,

$$\int_B \mathbf{P}(A) = \mathbf{P}(A)\mathbf{P}(B) = \mathbf{P}(A \cap B) = \int_B \chi_A = \int_B \mathbf{P}(A|\Sigma)$$

This means $\mathbf{P}(A) = \mathbf{P}(A|\Sigma)$. By linearity, the equation holds if X is any simple function. If X is positive, we can take limits, and then we can just let $X = X^+ - X^-$ to get the general result. □

Corollary 6.7. *If Σ is independent of $\sigma(\sigma(X), \Delta)$, then*

$$\mathbf{E}[X|\sigma(\Sigma, \Delta)] = \mathbf{E}[X|\Delta]$$

Proof. If $A \in \Delta$, and $B \in \Sigma$, then $X\chi_A$ is independent of Σ, and so by the last theorem,

$$\int_{A \cap B} X = \int_B X\chi_A = \int_B \mathbf{E}[X\chi_A] = \mathbf{P}(B)\mathbf{E}[X\chi_A]$$

Now $\mathbf{E}[X|\Delta]\chi_A$ is also independent of Σ, so

$$\begin{aligned}\int_{A \cap B} \mathbf{E}[X|\Delta] &= \int_B \mathbf{E}[X|\Delta]\chi_A = \mathbf{P}(B)\mathbf{E}(\mathbf{E}[X|\Delta]\chi_A) \\ &= \mathbf{P}(B)\mathbf{E}(\mathbf{E}[X\chi_A|\Delta]) = \mathbf{P}(B)\mathbf{E}[X\chi_A]\end{aligned}$$

It follows by equating these two equations that $\mathbf{E}[X|\sigma(\Sigma, \Delta)] = \mathbf{E}[X|\Delta]$. □

Example. *Consider a sequence of independent and identically distributed random variables $X_1, X_2, \ldots$, and consider the partial sums S_n. Given that we know S_n, the best guess of each X_i should be $n^{-1}S_n$. Define*

$$\Sigma_n = \sigma(S_n, S_{n+1}, \ldots) = \sigma(S_n, X_{n+1}, \ldots)$$

Note that $X_{n+1}, X_{n+2}, \ldots$ is independent of X_i, S_n, so $\mathbf{E}[X_i|\Sigma_n] = \mathbf{E}[X_i|S_n]$. But if F is the common distribution of each X_i, then for any set $A = \{S_n \in B\}$, where B is Borel, then $X(A)$ is symmetric in each variable (because the sum is symmetric), and so

$$\int_A X_i d\mathbf{P} = \int_{X(A)} x_i dF(x_1) \ldots dF(x_n) = \int_{X(A)} x_j dF(x_1) \ldots dF(x_n) = \int_A X_j d\mathbf{P}$$

so $\mathbf{E}(X_i|S_n) = \mathbf{E}(X_j|S_n)$. Now, using the fact that $S_n = \sum \mathbf{E}(X_i|S_n) = n\mathbf{E}(X_i|S_n)$, and this completes the calculation. This calculation leads to a very nice proof of the strong law of large numbers, as we will soon see.

Chapter 7

Discrete Time Martingales

7.1 Filtrations

As a stochastic process evolves, the values we observe give us more and more information into the future of the process. We can model this 'increase' in information in a measure theoretic manner, by the object known as a filtration, which is just an increasing family of σ algebras $\Sigma_0 \subset \Sigma_1 \subset \ldots$, representing information at is available at a given time, which should increase over time as we learn more and more states of the process. It is the measure theoretic equivalent of a topological filter.

Example. *Let $X_0, X_1, \ldots$ be a stochastic process. The* **natural filtration** *corresponding to X is the filtration $\Sigma_0 = X_0$, $\Sigma_1 = X_1, \ldots$, which represents the increase in information over time if we only look at the values X_i.*

Example. *Suppose we have a random walk, where at each step we flip a coin $X \in \{-1, +1\}$ to decide which direction to travel, and then travel along this direction by the length of a Poisson distribution $Y \in \mathbf{Z}^+$. Then the natural information given to use after n steps is the value of the first n coin flips $X_1, \ldots, X_n$ and the values of the first n Poisson distributions $Y_1, \ldots, Y_n$. The natural filtration to use in this example is $\Sigma_n = \sigma(X_1, Y_1, \ldots, X_n, Y_n)$.*

It is often interesting to define $\Sigma_\infty = \lim \Sigma_n$, which represents all the information we could ever obtain in a process.

7.2 Martingales

We wish to discuss stochastic processes $X_1, X_2, \ldots$ representing 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, so that the bets we made were fair. This is where our new definition of expectation comes into play. A **martingale** with respect to a filtration $\Sigma_0, \Sigma_1, \ldots$ is a process $M_0, M_1, \ldots$ consisting of integrable martingales such that, for $m < n$, $\mathbf{E}(M_n|\Sigma_m) = M_m$. More generally, we say M is a **submartingale** if $\mathbf{E}(M_n|\Sigma_m) \geqslant M_m$, and M is a **supermartingale** if $\mathbf{E}(M_n|\Sigma_m) \leqslant M_m$. A submartingale is a game that's in the player's favour, and a supermartingale is a game against the player (A su*b*martingale increases, and a su*p*ermartingale decreases). We need only verify the defining property for $\mathbf{E}(M_n|\Sigma_{n-1})$, because we can apply the tower rule to obtain the general property recursively. Though it is helpful to think of these processes as modelling gambling, they seem to crop up in almost every aspect of probability theory.

Example. *Let $X_1, X_2, \ldots$ be a sequence of independent, integrable random variables with mean zero. Define $S_n = X_1 + \cdots + X_n$ to be the 'random walk' with respect to these random variables, with $S_0 = 0$. Then*

$$\begin{aligned}\mathbf{E}(S_n|X_1, \ldots, X_{n-1}) &= X_1 + \cdots + X_{n-1} + \mathbf{E}(X_n|X_1, \ldots, X_{n-1}) \\ &= S_{n-1} + \mathbf{E}(X_n) = S_{n-1}\end{aligned}$$

Thus S_n is a martingale relative to the filtration induced by the stochastic sequence $X_1, X_2, \ldots$. We can also show that it is a martingale relative to the filtration induced by $S_1, S_2, \ldots$ itself. It is of interest to ask (ala the law of large numbers) if S_n converges as $n \to \infty$ to some random variable S_∞. We shall find martingale theory provides a framework in which these questions have easy answers.

Example. *As in the last example, consider a sequence of independent non-negative random variables $X_1, X_2, \ldots$, but with mean 1. If we define $M_0 = 1$, $M_n = X_1 \ldots X_n$, then if we set $\Sigma_n = \sigma(X_1, \ldots, X_n)$, then*

$$\mathbf{E}(M_{n+1}|\Sigma_n) = M_n \mathbf{E}(X_{n+1}|\Sigma_n) = M_n \mathbf{E}(X_{n+1}) = M_n$$

so M is a martingale, provided that each M_n is in $L^1(\Omega)$. Again, an interesting question to ask is whether these values settle down to some random quantity M_∞ as $n \to \infty$.

Example. *If X is an integrable random variable, and Σ is any filtration, then the tower law gives*

$$\mathbf{E}[\mathbf{E}[X|\Sigma_{n+1}]|\Sigma_n] = \mathbf{E}[X|\Sigma_n]$$

so the process $\mathbf{E}[X|\Sigma_n]$ is a martingale with respect to Σ_n. Now it is natural to ask whether $\mathbf{E}[X|\Sigma_n]$ converges to $\mathbf{E}[X|\Sigma_\infty]$, so that in some sense, we steadily learn all there is to know about X as n increases, rather than having a 'breakthrough' at ∞, where we can put all information together to learn X.

7.3 Optional Stopping Theorem

We begin with an extension of the 'gambling' concept of martingales.

Example. *We can consider a Martingale M_n as the total amount you have made in a series of games, given that you have a 'unit stake' on each interval. If we instead have a changing stake $C_1, C_2, \ldots$ each game, then the total amount of money we make is*

$$(C \bullet M)_n = \sum_{k=1}^{n} C_k(M_k - M_{k-1})$$

It is easy to verify that, if each C is constant, then $C \bullet M$ is also a Martingale, so changing our bets 'doesn't make the system less fair'. More generally, if $C_1, C_2, \ldots$ is a **previsible process**, *in the sense that each C_n is Σ_{n-1} measurable, and each C is in $L^\infty(\Omega)$, then $C \bullet M$ is a Martingale, or if $C \in L^q(\Omega)$ where each $M_n - M_{n-1} \in L^p(\Omega)$. We call $(C \bullet M)$ the* **discrete Itô integral**, *or* **martingale Transform** *of C with respect to M, which reflects the continuous time integrals we will see later. If M is only a (super/sub) martingale, then $(C \bullet M)$ will be a (super/sub) martingale as well.*

Probability was created to analyze gambling games like the one above, and martingale theory was to analyze one particular area of gambling theory. In the 18th century, a strategy was discovered which could be applied to 'beat' certain gambling games, guaranteeing a profit whenever it was applied. It became known as the martingale. Let's consider the strategy in

it's simplest implementation, when gambling on a flip of a coin. We take a series of $\{-1,1\}$ valued independent Bernoulli trials $X_1, X_2, \ldots$, and define $M_n = X_1 + \cdots + X_n$ to be the unit stakes turnout of a best against these coin flips. We then consider the martingale $C \bullet M$, where M is a martingale, and when $C_k = 2^k$, so our bet 'doubles' each time. If $X_n = 1$, then

$$(C \bullet M)_n = \sum_{k=1}^{n} 2^k X_k = 2^n + \sum_{k=1}^{n} 2^k X_k \geqslant 1$$

Thus, if we bet along this strategy until the first time that $X_n = 1$, in which case we 'stop' the betting process, we will always come out with at least a unit profit. More rigorously, if we define the 'stopping time' $T = \inf\{k : X_k = 1\}$, then $\mathbf{P}(T < \infty) = 1$, and $(C \bullet M)_T \geqslant 1$, so we've found a guaranteed way to beat the system! The Martingale soon became all the rage in the 18th century. Casanova was one of many famous figures known to apply the strategy to his own games. The key problem with the strategy is that it assumes one is able to bet with an infinite amount of money. 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, if you don't have much cash, the strategy is somewhat risky, and if you have a lot of cash, you have a lot to lose if you stop betting.

At first, this seems to contradict the fact that for any series of previsible L^1 stakes C, $C \bullet M$ is a martingale. We can now define the stakes

$$C_n^{(T)} = C_n \chi_{n \leqslant T}$$

which model the outcomes of the bet n steps into the future, if we stop betting at time T. This is still a previsible process, because C_n is Σ_{n-1} adapted, and the event $\{n \leqslant T\}$ is the complement of the event $\{T < n\} = \{T \leqslant n-1\}$, which is Σ_{n-1} adapted because it depends only on the values of the variables $X_1, \ldots, X_{n-1}$. Since $C_n^{(T)}$ is bounded by C_n, it is certainly integrable, and so we conclude that $C^{(T)} \bullet M$ is a martingale, so= that in particular, for each n,

$$\mathbf{E}[(C^{(T)} \bullet M)_n] = \mathbf{E}[(C^{(T)} \bullet M)_0] = 0$$

and therefore from the perspective of finite time, the bet is still 'fair'. This argument shows that this remains true if T is any $\{0, 1, \ldots, \infty\}$ valued process such that the event $\{T \leqslant n\}$ is Σ_n adapted, and we call such a T a

stopping time. We have the equality

$$(C^{(T)} \bullet M)_n = (C \bullet M)_{T \wedge n}$$

which is essentially a calculation of the next result.

Theorem 7.1. *If M is a (sub/super) martingale, and T is a stopping time, then the* **stopped process** $M_n^T = M_{T \wedge n}$ *is a (sub/super)martingale.*

Proof. $M^T = (C \bullet M)$, where $C = \chi_{n \leqslant T}$, and the last argument applies. □

So paradoxically, going back to the martingale it now seems like the betting strategy corresponding to T and C is fair, because at each finite time point, the average amount of money in the gambler's pocket is zero. The difference between these two calculation is if we now consider the limit

$$\lim_{n \to \infty} (C^{(T)} \bullet M)_n = (C^{(T)} \bullet M)_\infty = (C \bullet M)_T$$

which exists pointwise because we almost surely stop betting at a finite amount of time (since $\mathbf{P}(T < \infty) = 1$), then we know $(C \bullet M)_T \geqslant 1$, so the process cannot be extended to be a martingale 'at ∞'. There are many examples of stopping times upon which we cannot extend the stopped martingale at ∞.

Example. *Consider a simple reflecting random walk M on* $\mathbf{N}$ *starting at 0. Then M is a martingale. If we consider the stopping time $T = \inf\{n : M_T = 1\}$, then it is easy to see $\mathbf{P}(T < \infty) = 1$, and even though $\mathbf{E}(M_n^T) = \mathbf{E}(M_0^T) = 0$, we find $M_T = 1$.*

The optional stopping theorem says that, provided we limit the behaviour of a martingale and its stopping time, the martingale still behaves well as a martingale in a limit, so that the bet is still fair asymptotically. One such limitation we can place on the theorem is that the variation of the martingale between steps is bounded. This is the reason why limits are put on poker and blackjack tables, so you can't bet an unbounded amount of money anymore. Effectively, the casino restricts the martingales you can choose to play, so that regardless of the 'stopping time' you choose to play while gambling, the martingale strategy won't work. If we write $M_T = M_{T \wedge n} + \mathbf{I}(T > n)M_T - \mathbf{I}(T > n)M_n$, then we obtain the result provided that $\mathbf{I}(T > n)M_T$ and $\mathbf{I}(T > n)M_n$ both become suitably small in expectation. We can do this either by bounding T's behaviour at ∞ or bounding M's behaviour.

Theorem 7.2 (The Optional Stopping Theorem)**.** *If M is a supermartingale, and T is a stopping time, then provided one of three conditions hold*

- *T is bounded.*
- *M is bounded, and T is almost surely finite.*
- *$\mathbf{E}(T) < \infty$, and the Martingale increments are bounded, so that there is a universal K such that $|M_{n+1} - M_n| \leqslant K$ almost surely.*

Then $M_T \in L^1(\Omega)$ and $\mathbf{E}(M_T) \leqslant \mathbf{E}(M_0)$. If M is a submartingale, then we find $\mathbf{E}(M_T) \geqslant \mathbf{E}(M_0)$, and if M is a martingale, then $\mathbf{E}(M_T) = \mathbf{E}(M_0)$.

Proof. We know that $M_{T\wedge n}$ is integrable for all n, because

$$\mathbf{E}(|M_{T\wedge n}|) = \mathbf{E}(\mathbf{E}(|M_{T\wedge n}||T))$$

and

$$\mathbf{E}(|M_{T\wedge n}||T = k) = \|M_{k\wedge n}\|_1 \leqslant \sup_{l\leqslant n} \|M_l\|_1$$

is in $L^\infty(\Omega)$. If $T \leqslant N$, then $M_T = M_{T\wedge N}$ is part of the stopped process corresponding to M, and therefore the required result holds obviously. Otherwise, $M_T = \lim_{n\to\infty} M_n^T$, and since the values M are bounded, we can apply the dominated convergence theorem to obtain the required property. For the third condition, we note that

$$|M_{T\wedge n} - M_0| = \left|\sum_{k=1}^{T\wedge n} (M_k - M_{k-1})\right| \leqslant KT$$

so provided $\mathbf{E}(T) < \infty$, the dominated convergence result applies. □

Example. *Consider the fair Gambler's ruin, where we start with N_0 units of money, and play a fair game until we go bust, or we end up with N_1 units of money. That is, we consider a martingale M with $M_0 = N_0$, and consider the time T as which we either first go broke, or exceed N_1 units of money, so $T = \inf\{n : M_n \leqslant 0$ or $M_n \geqslant N_1\}$. Provided the increments of the martingale are bounded (which occurs in most realistic gambling scenarios), and the stopping time has $\mathbf{P}(T < \infty) = 1$, which occurs in practice almost always, we can apply the optional stopping theorem to conclude that*

$$N_1\mathbf{P}(M_T = N_1) = \mathbf{E}(M_T) = \mathbf{E}(M_0) = N_0$$

so $\mathbf{P}(M_T = N_1) = N_0/N_1$.

In the example above, it is normally very easy to argue that $\mathbf{P}(T < \infty) = 1$, using the principle that 'whatever has a reasonable chance of happening, will almost surely happen'. The particular application is described in the next lemma.

Lemma 7.3. *Suppose T is a stopping time such that for some integer N and some $\varepsilon > 0$, we have $\mathbf{P}(T \leqslant n + N|\Sigma_n) > \varepsilon$ almost surely. Then $\mathbf{E}(T) < \infty$.*

Proof. We calculate that

$$\begin{aligned}\mathbf{E}(T) &= \sum_{k=0}^{\infty} \mathbf{P}(T > k) = \sum_{k=0}^{N} \mathbf{P}(T > k) + \sum_{k=1}^{\infty} [1 - \mathbf{E}[\mathbf{P}(T \leqslant k + N|\Sigma_k)]] \\ &\leqslant \sum_{k=0}^{N} \mathbf{P}(T \leqslant k) + \sum_{k=1}^{\infty} [1 - \varepsilon]\end{aligned}$$

TODO: finish this proof. □

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

Lemma 7.4. *If $\{X_n\}$ is a family of random variables for which these is C where $X_n^2 < C$, then the X_n are uniformly integrable.*

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

7.4 Martingale Convergence Theorems

Given a process M, and $a \leqslant b$, consider a stopping time $U[a, b] = U(M, [a, b])$ which counts the number of 'upcrossings' from a to b. That is, $U_n[a, b] = m$ if there are $0 \leqslant t_1 < s_1 < \cdots < t_n < s_n \leqslant n$ with $M_{t_i} < a$, $M_{s_i} > b$. The number of upcrossings of a stochastic process represents the amount of 'variation' in your process between a and b, and it turns out that the variation of martingales is essentially constant. The idea extends from showing that 'buy low', 'sell high' doesn't help you when you're gambling against an unfair system.

Theorem 7.5 (Doob's Upcrossing Lemma)**.** *If M is a supermartingale, then*

$$(b-a)\mathbf{E}(U_n[a,b]) \leqslant \mathbf{E}[(X_n - a)^-]$$

and if M is a submartingale, then

$$(b-a)\mathbf{E}(U_n[a,b]) \leqslant \mathbf{E}[(X_n - a)^+]$$

so that n doesn't really impact the expected number of upcrossings.

Proof. Consider a previsible process C which follows the following rule-set: We wait under $M_n < a$, and then we play unit stakes until $M_n > b$, in which case we stop playing and wait for M_n to become less than a again, in which case we start playing at unit stakes, rince and repeat. More rigorously, we define $C_1 = \mathbf{I}(X_0 < a)$, and then set

$$C_n = \mathbf{I}(C_{n-1} = 1)\mathbf{I}(M_{n-1} \leqslant b) + \mathbf{I}(C_{n-1} = 0)\mathbf{I}(M_{n-1} < a)$$

Now $\|C_n\|_\infty \leqslant 1$ is bounded, and therefore $C \bullet M$ is a supermartingale. But

$$(C \bullet M)_n \geqslant (b-a)U_n[a,b] - (M_n - a)^-$$

because we make at least $b-a$ at each 'run' of unit betting, when M_n rises to a, but lose at most $(M_n - a)^-$ because we start begging when $M_k < a$. We conclude that $\mathbf{E}((C \bullet M)_0) \leqslant \mathbf{E}((C \bullet M)_0) = 0$, and so

$$(b-a)\mathbf{E}[U_n[a,b]] \leqslant \mathbf{E}[(M_n - a)^-]$$

and this is the required inequality. If M is a submartingale, then we can consider the same betting strategy, but where we bet a negative stake rather than a positive stake, and we conclude that

$$(C \bullet M)_n \leqslant (M_n - a)^+ - (b-a)U_n[a,b]$$

and so

$$0 \leqslant \mathbf{E}((C \bullet M)_n) \leqslant \mathbf{E}((M_n - a)^+) - (b-a)\mathbf{E}[U_n[a,b]]$$

giving us the other inequality. □

Corollary 7.6. *Let M be a (sub/super) martingale bounded in $L^1(\Omega)$. If $a < b$, and if we define $U_\infty[a,b] = \lim U_n[a,b]$, then $\mathbf{P}(U_\infty = \infty) = 0$.*

Proof. Using the monotone convergence theorem, the upcrossing lemma provides a bound

$$(b-a)\mathbf{E}U_\infty[a,b] = \lim_{n\to\infty}(b-a)U_n[a,b] \leqslant \sup \mathbf{E}[(X_n - a)^-] < |a| + \sup \|X_n\|_1$$

and so $U_\infty[a,b] \in L^1(\Omega)$, hence $\mathbf{P}(U_\infty[a,b] = \infty) = 0$. The proof for submartingales is essentially the same. □

If we consider the countable set of all the pairs $p < q$, then we can conclude that $\mathbf{P}(\forall p < q \in \mathbf{Q}^+ : U_\infty[p,q] = \infty) = 0$, and since every interval $[a,b]$ contains an interval of the form $[p,q]$, we conclude that almost surely, for all $a < b$, $U_\infty[a,b] \neq \infty$.

Theorem 7.7 (Doob's Martingale Convergence)**.** *Let M be a (sub/super) martingale bounded in $L^1(\Omega)$. Then almost surely, $M_\infty = \lim_{n\to\infty} M_n$ exists in $L^1(\Omega)$, and is finite. If we define, for definiteness,*

$$M_\infty = \limsup M_n$$

then M_∞ will also be $\Sigma_\infty = \bigcup \Sigma_n$ measurable.

Proof. If $M_n(\omega)$ does not converge, then $\liminf M_n(\omega) < \limsup M_n(\omega)$. But this means we can find $a < b$ such that

$$\liminf M_n(\omega) < a < b < \limsup M_n(\omega)$$

and M_n must oscillate between a and b infinitely often, so $U_\infty[a,b](\omega) = \infty$. We have shown the set of all ω for which there exists $[a,b]$ with $U_\infty[a,b](\omega) = \infty$ is a set of probability 0, which means that M_n converges almost surely. Fatou's lemma implies that

$$\mathbf{E}|M_\infty| = \mathbf{E}(\liminf |M_n|) \leqslant \liminf_{n\to\infty} \mathbf{E}|M_n| \leqslant \sup \mathbf{E}|M_n| < \infty$$

So M_∞ is finite almost surely. □

Corollary 7.8. *If M is a non-negative super martingale, then $M_\infty = \lim M_n$ exists almost surely.*

Proof. If M is a supermartingale, then

$$\|M_n\|_1 = \mathbf{E}(M_n) = \mathbf{E}(\mathbf{E}(M_n|\Sigma_0)) \leqslant \mathbf{E}(M_0) = \|M_0\|_1$$

so any such martingale is bounded. □

Corollary 7.9 (Lévy's 'Upward' Theorem)**.** *If M is a supermartingale bounded in $L^1(\Omega)$, then $M_n \to M_\infty$ in $L^1(\Omega)$ if and only if M is uniformly integrable, and then $\mathbf{E}(M_\infty|\Sigma_n) \leqslant M_n$ almost surely. If M is a uniformly integrable submartingale, then we also obtain L^1 convergence of M_n to M_∞, and $\mathbf{E}(M_\infty|\Sigma_n) \geqslant M_n$. For martingales, we find $\mathbf{E}(M_\infty|\Sigma_n) = M_n$. This means we can think of M as being a martingale with time indices $\mathbf{N} \cup \{\infty\}$.*

Proof. We rely on the fact that for any stochastic process X, $X_n \to X$ in L^1 if and only if $X_n \to X$ in probability and X_n is uniformly integrable. Since pointwise almost sure convergence implies convergence in probability, it suffices to show that if $\|M_n - M_\infty\| \to 0$, then $\mathbf{E}(M_\infty|\Sigma_n) \leqslant M_n$. We note that for $E \in \Sigma_n$,

$$\int_E \mathbf{E}(M_m|\Sigma_n)d\mathbf{P} \leqslant \int_E M_n d\mathbf{P}$$

We then let $m \to \infty$ to conclude

$$\int_E \mathbf{E}(M_\infty|\Sigma_n)d\mathbf{P} \leqslant \int_E M_n d\mathbf{P}$$

And this shows $\mathbf{E}(M_\infty|\Sigma_n) \leqslant M_n$ almost surely. The same argument holds for submartingales. □

This proof leads to a simple proof of the 0-1 law, which is a foundational result in basic probability theory.

Corollary 7.10 (Kolmogorov's 0-1 Law)**.** *Let $X_1, X_2, \dots$ be a sequence of independent random variables, and set $\Sigma_n = \sigma(X_{n+1}, X_{n+2}, \dots)$, and set $\Sigma_\infty = \lim \Sigma_n$. Then for any $E \in \Sigma_\infty$, $\mathbf{P}(E) = 0$ or 1.*

Proof. If $E \in \Sigma_\infty$, then $\chi_E, \chi_E, \dots$ is a martingale, because for any n,

$$\mathbf{E}(\chi_E|\Sigma_n) = \chi_E$$

For each n, χ_E is independent of all of the σ algebras Σ_n, because

$$\{F : F \in \sigma(X_n, X_{n+1}, \dots, X_N) \text{ for some } N\}$$

form a π system generating Σ_n, and $\sigma(X_n, X_{n+1}, \dots, X_N)$ and Σ_{N+1} are independent, so F is independent to E. But we know

$$\chi_E = \mathbf{E}(\chi_E|\Sigma_\infty) = \lim_{n\to\infty} \mathbf{E}(\chi_E|\Sigma_n) = \lim_{n\to\infty} \mathbf{E}(\chi_E) = \mathbf{P}(E)$$

almost surely, and therefore $\mathbf{P}(E) = 0$ or $\mathbf{P}(E) = 1$. □

The next result is crucial for the theory of continuous time martingales, because it enables us to descend from discrete results about martingales to 'limits' of discrete results. Rather than discussing the behaviour of martingales on $\mathbf{N}$ as they tend to ∞, we discuss the behaviour of martingales on $-\mathbf{N}$ as they go 'back in time' to $-\infty$.

Theorem 7.11 (Lévy Doob Downward Theorem)**.** *Consider a supermartingale $M_0, M_{-1}, \ldots$ with respect to a filtration $\Sigma_0 \supset \Sigma_{-1} \supset \ldots$. If we assume $\sup \mathbf{E}(M_n) < \infty$, then the process M is uniformly integrable, the limit $M_{-\infty} = \lim M_n$ exists almost surely, we have convergence in $L^1(\Omega)$, and $\mathbf{E}(M_n|\Sigma_{-\infty}) \leqslant M_{-\infty}$ for all n, where $\Sigma_{-\infty} = \bigcap \Sigma_n$. Similar results for submartingales hold if $\inf \mathbf{E}(M_n) > -\infty$, and if M is a martingale, provided $\inf \mathbf{E}(M_n)$ and $\sup \mathbf{E}(M_n)$ are both finite, we can conclude that $\mathbf{E}(M_n|\Sigma_{-\infty}) = M_{-\infty}$.*

Proof. We first prove the uniform integrality property. Let $\varepsilon > 0$ be given. The supermartingale property implies that $\mathbf{E}(M_n) \leqslant \mathbf{E}(M_m)$ if $m \leqslant n$, so the expectation increases as m decreases. Since $\sup \mathbf{E}(M_n) \to \infty$, the values $\mathbf{E}(M_n)$ decrease to some finite value as n decreases, and we may assume that there is k such that $\mathbf{E}(M_n) \leqslant \mathbf{E}(M_k) + \varepsilon$ for all $n \leqslant k$. But now this implies that for a fixed λ, using the supermartingale property, that

$$\begin{aligned}\int_{|M_n|>\lambda} |M_n| &= \mathbf{E}(M_n) - \int_{M_n<-\lambda} M_n - \int_{M_n\leqslant\lambda} M_n \\ &\leqslant [\mathbf{E}(M_k) + \varepsilon] - \int_{M_n<-\lambda} M_k - \int_{M_n\leqslant\lambda} M_k \\ &= \int_{|M_n|>\lambda} |M_k| + \varepsilon\end{aligned}$$

Since $M_k \in L^1(\Omega)$, there exists $\delta > 0$ such that if $\mathbf{P}(E) < \delta$, then $\int_E |M_k| < \varepsilon$. If we can prove that, uniformly in n, $\mathbf{P}(|M_n| > \lambda) < \delta$ for large enough λ and n, then the proof will be complete. Applying Markov's inequality gives

$$\mathbf{P}(|M_n| > \lambda) \leqslant \frac{\mathbf{E}|M_n|}{\lambda} = \frac{\mathbf{E}(M_n) + 2\mathbf{E}(M_n^-)}{\lambda} \leqslant \frac{\sup \mathbf{E}(M_n) + 2\mathbf{E}(M_0^-)}{\lambda}$$

where we have used the fact that M_n^- is a submartingale. Letting λ be large enough gives the required result. Because M_n is uniformly integrable, it is bounded in $L^1(\Omega)$, and essentially the same proof of convergence for

supermartingales at ∞ works here, because we have the upcrossing result that

$$(b-a)\mathbf{E}(U_\infty[a,b]) \leqslant \mathbf{E}[(M_0-a)^-]$$

except that we no longer have any dependence on M_n for large negative values of n, so no boundedness condition is required. □

The strong law of large numbers appears as an immediate corollary.

Corollary 7.12. *Let $X_1, X_2, \ldots$ be i.i.d, integrable random variables of mean μ. If S_n denotes the sum of the first n variables, then $n^{-1}S_n \to \mu$ almost surely and in $L^1(\Omega)$.*

Proof. Set $\Sigma_n = \sigma(S_n, S_{n+1}, \ldots) = \sigma(S_n, X_{n+1}, X_{n+2}, \ldots)$. For any $m \leqslant n$, $\mathbf{E}(X_m|\Sigma_n) = n^{-1}S_n$. This is clear by Fubini's theorem, because if we consider the cumulative distributions $F_{X_1,\ldots,X_n}$, then the variables X_i all have a common cumulative distribution F, and $F_{X_1,\ldots,X_n}(x) = F(x_1)F(x_2)\ldots F(x_n)$. Since $S_n = \sum X_i$, we have

$$\mathbf{P}(S_n \leqslant t) = \int_{\sum x_i \leqslant t} dF(x)$$

and so for each t, i, j, we find

$$\int_{S_n \leqslant m} X_i = \int_{\sum x_i \leqslant m} x_i dF(x) = \int_{S_n \leqslant m} x_j dF(x) = \int_{S_n \leqslant m} X_j$$

and we therefore conclude $\mathbf{E}(X_i|\Sigma_n) = \mathbf{E}(X_j|\Sigma_n)$ almost everywhere. We can then use the fact that $S_n = \sum X_i$ to conclude that $S_n = \mathbf{E}(S_n|\Sigma_n) = \sum \mathbf{E}(X_i|\Sigma_n) = n\mathbf{E}(X_i|\Sigma_n)$. This means that for $m \leqslant n$,

$$\mathbf{E}(m^{-1}S_m|\Sigma_n) = n^{-1}S_n$$

so if we invert time, and look at $m^{-1}S_m$ as indexed on $(\infty, 0]$, then $m^{-1}S_m$ is a supermartingale relative to Σ_m. We calculate that $\mathbf{E}(m^{-1}S_m) = \mu < \infty$, so the last theorem implies that $A = \lim n^{-1}S_n$ exists almost surely and in $L^1(\Omega)$. If we define $S = \limsup n^{-1}S_n$, then $S \in \Sigma_\infty = \bigcap \Sigma_n \subset \bigcap \Delta_n$, where $\Delta_n = \sigma(X_n, \ldots)$. By Kolmogorov's 0-1 law, we know that $\sum \mathbf{P}(S \in [n, n+1)) = 1$, so there is an interval I_0 of length 1 such that $S \in I_0$ almost surely. If we break I_0 up into two intervals, we conclude that there is an interval I_1 of length 2^{-1} such that $S \in I_1$ almost surely. Performing this

process repeatedly, we find a decreasing series of intervals I_k of length 2^{-k} with $S \in I_k$ almost surely. This means $S \in \bigcap I_k$ almost surely, implying $\bigcap I_k$ is non-empty, and since the diameters of I_k decrease to 0, $\bigcap I_k$ can consist of only a single number m, so $S = m$ almost surely. But this means $m = \mathbf{E}(S) = \lim \mathbf{E}(S_m) = \mu$, so $S = \mu$ almost surely. □

7.5 Martingale Inequalities

Doob's upcrossing lemma allows us to prove that Martingales do not really have 'too much variation', and this gives convergence results at ∞ and $-\infty$. His further submartingale and L^p inequalities enable us to prove that Martingales are also 'essentially bounded'.

Theorem 7.13 (Doob's Submartingale Inequality)**.** *Let M be a submartingale. Then if $\lambda, n > 0$, then*

$$\lambda \mathbf{P}\left(\sup_{k \leqslant n} M_k \geqslant \lambda\right) \leqslant \int_{(\sup_{k \leqslant n} M_k) \geqslant \lambda} M_n$$

Proof. Let $E = \{(\sup_{k \leqslant n} M_k) \geqslant \lambda\}$. Then we can write E as a disjoint union of sets $E_0, E_1, \ldots, E_n$, where $E_i = \{M_0, \ldots, M_{i-1} < \lambda, M_i \geqslant \lambda\}$. Now $E_k \in \Sigma_k$, and therefore

$$\int_{E_k} M_n \geqslant \int_{E_k} M_k \geqslant \lambda \mathbf{P}(E_k)$$

and we may now sum over all k. □

Corollary 7.14. *If M is a non-negative submartingale, then we can conclude*

$$\lambda \mathbf{P}\left(\sup_{k \leqslant n} M_k \geqslant \lambda\right) \leqslant \mathbf{E}(M_n)$$

Notice that this result is independent of the step number n, which likely indicates we can obtain a uniform result when the M_n are bounded in $L^1(\Omega)$.

Lemma 7.15. *If M is a martingale, c is convex, and $c(M_n) \in L^1(\Omega)$ for each n, then $c(M)$ is a submartingale.*

Proof. Applying Jensen's inequality, we conclude that

$$\mathbf{E}(c(M_n)|\Sigma_m) \geqslant c(\mathbf{E}(M_n|\Sigma_m)) = c(M_m)$$

We needed that $c(M_n) \in L^1(\Omega)$ to apply this result. □

Lemma 7.16. *If X and Y and non-negative random variables such that for every $\lambda > 0$,*

$$\lambda \mathbf{P}(X \geqslant \lambda) \leqslant \int_{X \geqslant \lambda} Y$$

then for $p > 1$ with conjugate q, we have $\|X\|_p \leqslant q\|Y\|_p$.

Proof. Obviously, we can calculate that

$$L = \int_0^\infty p\lambda^{p-1}\mathbf{P}(X \geqslant \lambda)d\lambda \leqslant \int_0^\infty p\lambda^{p-2}\int_{X\geqslant\lambda} Y d\mathbf{P} d\lambda = R$$

By Tonelli's theorem, we find

$$L = \int_0^\infty p\lambda^{p-1}\mathbf{P}(X \geqslant \lambda)d\lambda = \int\int_0^X p\lambda^{p-1} = \int X^p = \mathbf{E}(X^p)$$

$$R = \int_0^\infty p\lambda^{p-2}\int_{X\geqslant\lambda} Y d\mathbf{P} d\lambda = \int Y \int_0^X p\lambda^{p-2} = \int \frac{p}{p-1} Y X^{p-1} = \mathbf{E}(qYX^{p-1})$$

But now we apply Holder's inequality to conclude that

$$\mathbf{E}(X^p) \leqslant \mathbf{E}(qYX^{p-1}) \leqslant q\|Y\|_p \|X^{p-1}\|_q = q\|Y\|_p \mathbf{E}(X^p)^{1/q}$$

Hence $\|X\|_p \leqslant q\|Y\|_p$. □

Theorem 7.17 (Doob's L_p inequality)**.** *Let $p > 1$, and let q denote it's conjugate. Let M be a non-negative submartingale bounded in $L^p(\Omega)$. Then $M^* := \sup M \in L^p(\Omega)$, and $\|M^*\|_p \leqslant q \sup \|M_n\|_p$. Also M_∞ is in $L^p(\Omega)$ and $M_n \to M_\infty$ in $L^p(\Omega)$. If $M = |N|$ for some martingale N bounded in $L^p(\Omega)$, then $M_\infty = |N_\infty|$ almost surely.*

Proof. Define $M_n^* = \sup_{k\leqslant n} M_n$. Now we can apply convexity to conclude that M_n^p is a submartingale, and then Doob's submartingale inequality gives

$$\lambda \mathbf{P}(M_n^* \geqslant \lambda) \leqslant \int_{M_n^* \geqslant \lambda} M_n$$

and the second lemma implies that $\|M_n^*\|_p \leqslant q\|M_n\|_p$. This implies the M_n^* are bounded in $L^p(\Omega)$ also, and monotone convergence shows that $M^* \in L^p(\Omega)$, with the required inequality. Since M_n is also bounded in $L^1(\Omega)$, we conclude that M_∞ exists. Since the variables M_n are bounded pointwise by M^*, we know that M_∞ is also bounded by M^* pointwise, and so

$$|M_n - M_\infty|^p \leqslant 2^p |M^*|^p$$

we may apply dominated convergence to conclude that $M_\infty \in L^p(\Omega)$, so $\|M_\infty\|_p = \lim \|M_n\|_p$.

If $M_n = |N_n|$, then N_n is a submartingale bounded in $L^p(\Omega)$, then M_n is surely bounded in $L^p(\Omega)$, so $M_n \to M_\infty$ almost surely pointwise. But we can use Lévy's upward thoerem to conclude that $N_n \to N_\infty$ almost surely, and we therefore conclude by taking absolute values pointwise that $M_\infty = |N_\infty|$ almost surely. □

Doob's L_p inequality shows that $L^p(\Omega)$ bounded martingales are restricted in motion incredibly well. This contrasts Brownian motion, where $\sup B_n = \infty$ almost surely; in this case, the B_n are not bounded in any $L^p(\Omega)$, even in $L^1(\Omega)$.

As in most of analysis, the nicest estimates we can get for martingales occur in the L^2 theory. We know that the conditional expectation operator is an orthogonal projection onto the subspace of measurable functions. If M is an L^2 martingale, then

$$\mathbf{E}(M_n - M_m|\Sigma_m) = 0$$

so $M_n - M_m$ is perpendicular to the space of Σ_m measurable functions for all $n \geqslant m$. This means that

$$M_n = M_0 + \sum_{k=1}^{n} (M_k - M_{k-1})$$

expressed M_n as the sum of orthogonal random variables, and so

$$\mathbf{E}(M_n^2) = \mathbf{E}(M_0^2) + \sum_{k=1}^{n} \mathbf{E}((M_k - M_{k-1})^2)$$

so the 'quadratic variation' of the process directly measures the square sum of M_n.

Theorem 7.18. *A martingale M is bounded in $L^2(\Omega)$ if and only if*

$$\mathbf{E}(M_0^2) + \sum_{k=1}^{\infty} \mathbf{E}((M_k - M_{k-1})^2) < \infty$$

and then $M_n \to M_\infty$ in $L^2(\Omega)$.

The next result says that an adapted process can be reduced to a process adapted 'one step behind' if we subtract a martingale.

Theorem 7.19 (Doob Decomposition)**.** *If X_n is an adapted process of integrable random variables, then X has a decomposition*

$$X_n = X_0 + M_n + A_n$$

where M_n is a martingale null at zero, and A is a previsible process null at zero. This decomposition is unique modulo indistinguishability. The process X_n is a submartingale if and only if A is an almost surely increasing process, and a supermartingale if A is an almost surely decreasing process.

Proof. Assume without loss of generality that $X_0 = 0$. Note that if we had such a decomposition, then

$$\mathbf{E}(X_1|\Sigma_0) = \mathbf{E}(M_1|\Sigma_0) + \mathbf{E}(A_1|\Sigma_0) = M_0 + A_1 = A_1$$

so we can set $A_1 = \mathbf{E}(X_1|\Sigma_0)$ almost surely, and $M_1 = X_1 - A_1$. More generally, we find

$$\mathbf{E}(X_{n+1}|\Sigma_n) = \mathbf{E}(M_{n+1}|\Sigma_n) + \mathbf{E}(A_{n+1}|\Sigma_n) = M_n + A_{n+1}$$

This allows us to recursively define M_n and A_{n+1} uniquely (almost surely). All that remains is to verify M and A have the required properties. It is clear that if M_n is Σ_n measurable, then A_{n+1} is Σ_n measurable, so A_{n+1} is previsible, and this follows because $M_n = X_n - A_n$ is the sum of Σ_n measurable variables. Next, we check the martingale property, verifying that

$$\mathbf{E}(M_{n+1}|\Sigma_n) = \mathbf{E}(X_{n+1}|\Sigma_n) - \mathbf{E}(A_{n+1}|\Sigma_n) = \mathbf{E}(X_{n+1}|\Sigma_n) - A_{n+1} = M_n$$

and this completes the proof. □

There is a variant of this theorem in the theory of continuous time martingales, but it is much more difficult, allowing us to break submartingales into the sum of local martingales and previsible increasing processes. We will address it when we touch on the theory of stochastic integration theory.

7.6 Quadratic Variation

The ideas of quadratic variation, which we touched on in the L^2 theory of martingales, lead to a host of useful inequalities. Given a martingale M in $L^2(\Omega)$ null at zero, then M^2 is a submartingale, and we can consider the Doob decomposition $M^2 = N + \langle M \rangle$, where N is a martingale, and $\langle M \rangle$ is an increasing previsible process. We let $\langle M \rangle_\infty = \lim \langle M \rangle_n$. Since $\mathbf{E}(M_n^2) = \mathbf{E}(\langle M \rangle_n^2)$, we know that M is bounded in $L^2(\Omega)$ if and only if $\mathbf{E}(\langle M \rangle_\infty) < \infty$, so on average M has 'finite quadratic variation'. What's more, since

$$\langle M \rangle_{n+1} - \langle M \rangle_n$$

Chapter 8

Continuous Time Regularity

The Daniell Kolmogorov theorem enables us to construct discrete time processes with any consistant distribution. While the theorem does enable us to construct distributions with any transition probability, in continuous time processes we require additional regularity hypotheses that the Daniell Kolmogorov theorem is unable to provide. Let's consider an example.

Example. *Let T be a parameter set, $m : T \to \mathbf{R}$, and $V : T \times T \to \mathbf{R}$ a symmetric non-negative-definite function, such that for any function $f : S \to \mathbf{R}$, where S is a finite subset of T,*

$$\sum_{r,s \in S} f(r) V(r,s) f(s) \geqslant 0$$

The elentary theory of Gaussian distributions implies that for any finite subset S, there exists a unique measure μ_S such that

$$\int_{\mathbf{R}^S} \exp\left(i \sum_{s \in S} \theta(s) f(s) \right) d\mu_S(f) = \exp\left(i \sum_{s \in S} \theta(r) m(r) - \frac{1}{2} \sum_{r,s \in S} \theta(r) V(r,s) \theta(s) \right)$$

We also know that the measures μ_S are compatible, and so the Daniell-Kolmogorov theorem enables us to construct the Gaussian process on the index set T with mean μ and covariance function V. In particular, if $T = [0, \infty)$, $m(t) = 0$, and $V(s,t) = s \wedge t$, then we can construct a process which has the exact same finite dimensional distributions as Brownian motion should have, but we are unable

to guarantee anything about the continuity properties of the paths of the process. Indeed, since $C[0,\infty)$ is not a Borel subset of $\mathbf{R}^{[0,\infty)}$, we cannot calculate the probability that a particular path of the process is continuous. Completion also won't help us here, because if $E \subset C[0,\infty)$ is a Borel measurable subset of $\mathbf{R}^{[0,\infty)}$, then we find that since the elements of E can only be described at countably many points, that $E = \varnothing$, so $\mathbf{P}_*(C[0,\infty)) = 0$.

Example. *If Ω is a σ finite measure space with measure λ over a σ algebra Σ, such that every singleton is measurable. We would like to construct* **Poisson set functions***, which are 'Σ indexed' processes $\Lambda : \Omega^\Sigma \to \mathbf{Z}^+ \cup \{\infty\}$, such that*

- *If $E \in \Sigma$, then $\Lambda(E)$ has a Poisson distribution with parameter $\lambda(B)$.*
- *If $E_1, \ldots, E_n \in \Sigma$ are disjoint, then $\Lambda(E_1), \ldots, \Lambda(E_n)$ are independent random variables.*
- *If E and F in Σ are disjoint, then $\Lambda(E \cup F) = \Lambda(E) + \Lambda(F)$ almost surely.*
- *For each ω, $\Lambda(\cdot)(\omega)$ is a measure over Σ.*

If S is a finite subset of Σ, then we can easily specify the desired law defining $\{\Lambda(E) : E \in S\}$, and these are consistant, so that the Daniell Kolmogorov theorem guarantees the existence of a process satisfying the first three properties above. But the Daniell Kolmogorov theorem doesn't say anything about the fourth property, and so we need to do extra work to guarantee this property holds. This is because the law defining the property also doesn't say anything about where the probability that Λ is a measure, because the subset of Ω^Σ of functions which define measures is not a measurable subset unless Σ is countable, which only occurs in the most trivial of cases. Again, if we consider a measurable subset E of Ω^Σ containing only measures, then we find that $E = \varnothing$, so we cannot complete the space to obtain good results.

One way to clarify this problem is to look at the ways that two stochastic processes with the same law can differ from one another. We say a process X is a **modification** of a process Y if, for every fixed time $t \in T$, $X_t = Y_t$ almost surely. This means exactly that X and Y have the same finite distributions. We say X and Y are **indistinguishable** if the inner probability of the set $A = \{\omega : X_t(\omega) = Y_t(\omega) \text{ for all } t\}$ is equal to 1, so A is measurable in the completion of the Σ algebra on Ω, and $\mathbf{P}(A) = 1$ in the completion. Indistinguishability preserves the properties of stochastic processes we need in the theory of continuous time.

Example. *Suppose that a Brownian motion B exists. Let U be an independent random variable to B uniformly distributed on $[0,1]$, and if we consider the set $A = \{x : B_{U(\omega)} = 0\}$, define*

$$\tilde{B}_t(\omega) = \begin{cases} B_t(\omega) & t \neq U \\ 1 & t = U, \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Then for a fixed t, $\mathbf{P}(\tilde{B}_t = B_t) \geqslant \mathbf{P}(t \neq U) = 1$, so B and $\tilde{B}$ are modifications of one another, but B and $\tilde{B}$ are not indistinguishable, because for any sample point ω, our definition guarantees $\tilde{B}_{U(\omega)} \neq B_{U(\omega)}$. And we find that this modification doesn't preserve continuity, because the paths of $\tilde{B}_t$ are always discontinuous.

The indistinguishability problem only occurs in the continuous time setting, because it is easy to verify that countable processes are modifications of one another if and only if they are indistinguishable.

Lemma 8.1. *There exists a process X into E^T with law $\mathbf{P}$ such that $X(\omega) \in A$ for all ω, for some subset $A \subset E^T$, if and only if $\mathbf{P}^*(A) = 1$.*

Proof. We rely on the fact that if $\mathbf{P}^*(A) = 1$, then for any measurable set E, $\mathbf{P}^*(A \cap E) = \mathbf{P}(E)$, and the law defined over E^T descends to a probability measure on A, with Σ algebra consisting of sets of the form $A \cap E$, where E is measurable. This means we can use A as the sample space with which to construct our stochastic process, which is just given by the projection functions $\pi_t : A \to E$, for $t \in T$. On the other hand, if there is a process X like above, and $A \subset F$, for some Borel measurable F, then $\mathbf{P}(F) = \mathbf{P}(X^{-1}(F)) = \mathbf{P}(\Omega) = 1$, so $\mathbf{P}^*(A) = 1$. □

We can never really calculate $\mathbf{P}^*(A)$ for any set A independently of constructing a process like above, but the lemma has a relaxing quality. We shall find, however, that by making heavy use of the law defining certain stochastic processes, we can find modifications of a given process with nice properties, like the continuity of Brownian motion. This is what we set out to do in this chapter. Indeed, if we can modify a process on $[0,\infty)$ such that *all sample paths are right continuous*, then the modification/indistinguishibility problem disappears.

Theorem 8.2. *If X and Y are two right continuous processes with values in a Hausdorff space, then X and Y are modifications of one another if and only if they are indistinguishable.*

Proof. The theorem essentially follows because

$$\{\omega : X_t(\omega) = Y_t(\omega) \text{ for all } t\} = \bigcap_{p \in \mathbf{Q}^+} \{\omega : X_p(\omega) = Y_p(\omega)\}$$

because we can take limits from above to prove equality at all points if we have equality on rational points, and the right hand side is the countable intersection of sets we know to be measurable in Ω, hence the left hand side is measurable in Ω also. If X and Y are modifications of one another, then each event in the intersection is a set of probability 1, so the countable intersection also has probability 1. □

8.1 Regularity of Martingales

We recall that a **continuous time martingale** on $[0, \infty)$ with respect to a filtration Σ is a Σ adapted process X such that for $t \geqslant s$, $\mathbf{E}[X_t | \Sigma_s] = X_s$. We define continuous time supermartingales and submartingales in an analogous manner. The asymptotics of martingales will play a role in showing that the paths of the process $\{X_p : p \in \mathbf{Q}^+\}$ can be 'regularized' to right continuous functions with left limits, or **cádlág**, so that in the forthcoming analysis of continous time martingales, we can always assume martingales are regularized.

Consider a (non-random) function $f : \mathbf{Q}^+ \to \mathbf{R}$. We start our analysis by considering the properties of f which guarantee that we can extend f to a right continuous function with left limits, in which case we call f **regularizable**. The trick is that if

$$\lim_{p \downarrow t} f(p) \qquad \lim_{q \uparrow t} f(q)$$

exist for all real values $t \geqslant 0$, and are finite, we can define $g(t) = \lim_{p \downarrow t} f(p)$, and then g will be right continuous with left limits. As with martingale convergence, we consider the upcrossing values $U_n[a, b]$ to be the maximum n such that we can find rational numbers

$$0 \leqslant p_1 < q_1 < p_2 < \cdots < p_n < q_n \leqslant n$$

where $f(p_i) < a$, $f(q_i) > b$. If the choice of n is unbounded, we let $U_n[a,b] = \infty$.

Lemma 8.3. *f is regularizable if and only if for any integer n and rational numbers $a < b$, we have*

$$\sup\{|f(q)| : q \in \mathbf{Q}^+ \cap [0,n]\} < \infty \qquad U_n[a,b] < \infty$$

Proof. TODO □

Corollary 8.4. *If $\{X_q : q \in \mathbf{Q}^+\}$ is a real valued stochastic process, then*

$$E = \{\omega : q \mapsto Y_q(\omega) \text{ is regularizable}\}$$

is measurable.

Proof. We have exibited conditions such that $Y(E) \subset \mathbf{R}^{[0,\infty)}$ is described as a countable intersection of measurable cylinders in $\mathbf{R}^{[0,\infty)}$, which therefore are measurable, and this also means the inverse image of the set is measurable, which is E. □

Lemma 8.5. *Let X be a supermartingale, fix $t \in [0,\infty)$ and $q_1 > q_2 > \ldots$ is a decreasing sequence of rationals which tend to t, then X_{q_i} converges pointwise almost every and in the L^1 norm.*

Proof. The sequence $X_{q_1}, X_{q_2}, \ldots$ is a reverse supermartingale, and the expectation is upper bounded, because we calculate $\mathbf{E}(X_{q_i}) < \mathbf{E}(X_t)$, which immediately gives the result by the Lévy Doob downward theorem. □

If we vary the sequence of rationals in the above lemma to a new sequence r_i, then Y_{q_i} and Y_{r_i} will still tend to the same value because we can interlace the two sequences, and the lemma above still implies convergence.

Theorem 8.6 (Doob Regularity Theorem)**.** *If X is a supermartingale, then X is regularizable almost surely, and if we define*

$$Y_t = \begin{cases} \lim_{q \downarrow t} X_q(\omega) & : q \mapsto X_q(\omega) \text{ is regularizable} \\ 0 & : \text{otherwise} \end{cases}$$

Then Y is a cádlág process.

Proof. We need only show that for a fixed n, $a < b \in \mathbf{Q}^+$, that

$$\mathbf{P}(\sup\{|X_q(\omega)| : q \in \mathbf{Q}^+ \cap [0,n]\} < \infty) = 1$$

$$\mathbf{P}(U_n[a,b] < \infty) = 1$$

where $U_n[a,b]$ is the upcrossing lemma applied to $Y|_{\mathbf{Q}^+}$. If $D_1, D_2, \ldots$ are a sequence of finite subsets of $\mathbf{Q}^+ \cap [0,n]$, each containing 0 and n, and with $D_m \uparrow \mathbf{Q}^+ \cap [0,n]$, then for a fixed λ, we know that

$$\begin{aligned}
&\mathbf{P}(\sup\{|X_q| : q \in \mathbf{Q}^+ \cap [0,n]\} > 3\lambda) \\
&\quad = \lim \mathbf{P}(\sup\{|X_q| : q \in D_m\}) \\
&\quad \leqslant \frac{4\mathbf{E}|X_0| + 3\mathbf{E}|X_n|}{\lambda}
\end{aligned}$$

Letting $\lambda \to \infty$ gives the first result. By the upcrossing lemma, if $U_n^{D_m}[a,b]$ denotes the number of upcrossings just over the finite set D_m, we know

$$\mathbf{E}[U_n[a,b]] = \lim_{m\to\infty} \mathbf{E}[U_n^{D_m}[a,b]] \leqslant \frac{\mathbf{E}|X_n| + |a|}{b-a}$$

This is where the step-independent result of the upcrossing lemma is *integral*, and therefore $U_n[a,b] \in L^1(\Omega)$, so $\mathbf{P}(U_n[a,b] < \infty) = 1$. □

You might think that we are done with the discussion, but there is one 'irregularity' with the Doob regularity lemma.

Example. *Suppose* $\Omega = \{\pm 1\}$, $\mathbf{P}(\pm 1) = 1/2$, *and* $\Sigma_t = \{\varnothing, \Omega\}$ *for* $t \leqslant 1$, *and* $\Sigma_t = 2^\Omega$ *for* $t > 1$. *Suppose that for* $\omega \in \Omega$,

$$X_t(\omega) = \omega \chi_{\{t>1\}}(\omega)$$

Then X is a martingale relative to the filtration Σ_t, *and it's regularization is*

$$Y_t(\omega) = \omega \chi_{\{t\geqslant 1\}}(\omega)$$

Note that Y_1 *is not* Σ_1 *measurable, so Y cannot possibly still be a martingale relative the filtration. Moreover,* $\mathbf{P}(X_1 = Y_1) = 0$, *so Y isn't a modification of X either, so it seems the Doob's regularity theorem doesn't work!*

The first problem we erase is that Y might not be adapted to the required filtration. However, by it's construction, we can fix this by enlarging our filtrations 'infinitisimally'. By construction, Y_t will be adapted to the *partial augmentation*

$$\Sigma_{t+} = \lim_{u \downarrow t} \Sigma_u$$

If we assume our filtration is right continuous, in the sense that $\Sigma_{t+} = \Sigma_t$ for all t, then the problem 'almost' doesn't occur. The only non Σ_t-measurable set involved in the construction is the set

$$\{\omega : t \mapsto X_t(\omega) \text{ is regularizable}\}$$

which is a subset of the σ algebra $N(\Sigma_\infty)$ of Σ_∞ measurable subsets with probability 0 or 1, we will also need to assume that $N(\Sigma_\infty) \subset \Sigma_t$ for each t. Thus it makes sense to define the *partial augmentation* $\Sigma'_t = \sigma(\Sigma_{t+}, N(\Sigma_\infty))$.

Theorem 8.7. *The process Y is a supermartingale relative to Σ', and Y is a modification of X if and only if the map $t \mapsto Y_t$ is a right continuous map into $L^1(\Omega)$, in the sense that $\lim_{s \downarrow t} \|Y_t - Y_s\|_1 = 0$ for all $t \geqslant 0$.*

Proof. Fix $0 \leqslant t < s$. Suppose $s > q_1 > q_2 > \cdots \to t$. It is easy to verify that $\mathbf{E}(X_s|\Sigma_{q_n}) \leqslant X_{q_n}$, considering $Z_n = \mathbf{E}(X_s|\Sigma_{q_n})$ as a reverse *martingale*, the Lévy-Doob downward theorem for martingales allows us to conclude that $\mathbf{E}(X_s|\Sigma_{t+}) \leqslant Y_t$. Since $N(\Sigma_\infty)$ is independent of trivially independent of every other σ algebra, we find $\mathbf{E}(X_s|\Sigma'_t) = \mathbf{E}(X_s|\Sigma_{t+}) \leqslant Y_t$. Now the Doob downward theorem guarantees that if we let s be rational, and then let $s \downarrow t$, then the converge will be in $L^1(\Omega)$, and so

$$\mathbf{E}(Y_s|\Sigma'_t) = \lim_{p \downarrow s} \mathbf{E}(X_p|\Sigma'_t) \leqslant Y_t$$

hence we have shown Y_s is a supermartingale. It now follows from the convergence definition of Y that Y is right continuous in $L^1(\Omega)$, and since we know that if $q_n \downarrow t$ then $X_{q_n} \to Y_t$ in L^1, it follows that X is a modification of Y at t if and only if X is right continuous. □

Even with this theorem, we aren't exactly satisfied by the regularity theorem, because it turns out that the filtration Σ' does not have rich enough class of stopping times for the continuous time theory. Thus we also suppose the *usual conditions*, which require that the σ algebra over

the probability space is complete, each Σ_t contains all **P** null sets, and Σ_t is right continuous. This subsumes the partial augmentation considered above. We let Σ^* denote the smallest filtration larger than Σ satisfying the usual conditions. It can be obtained by first enlarging the Δ sigma algebra over the sample space to a complete sigma algebra Δ^*, set N to be the class of null sets in Δ^* then setting

$$\Sigma_t^* = \bigcap_{s>t} \sigma(\Sigma_s, N) = \sigma(\Sigma_{t+}, N)$$

Since Σ' differs from the usual augmentation only by **P** null sets, the independence properties of conditional expectation guarantee that the theorem above still holds. To summarize our discussion, the regularity theorem guarantees that we take a supermartingale X with respect to a filtration Σ, and *regularize* to a cádlág supermartingale Y with respect to the filtration Σ^*.

Theorem 8.8. *If Ω with a filtration Σ satisfying the usual conditions, and X is a supermartingale, then X has a cádlág modification Y if and only if the map $t \mapsto \mathbf{E}(Y_t)$ from $[0,\infty)$ to $\mathbf{R}$ is right continuous, and then Y is a cádlág supermartingale.*

Proof. From the supermartingale property of X, we know that for $s > t$, $\mathbf{E}(X_s|\Sigma_t) \leqslant X_t$. Applying the regularity theory allows us to construct Y, which we know is also a supermartingale with respect to Σ, except that Y might not be a version of X. Since $X_p \to Y_t$ in L^1 if $p \downarrow t$ in $L^1(\Omega)$, then we obtain

$$Y_t = \mathbf{E}(Y_t|\Sigma_t) = \lim_{p \downarrow t} \mathbf{E}(X_p|\Sigma_t) \leqslant X_t$$

If the map $t \mapsto \mathbf{E}(X_t)$ is right continuous, then since $X_p \to Y_t$ in $L^1(\Omega)$, we conclude that $\mathbf{E}(Y_t) = \lim_{p \downarrow t} \mathbf{E}(X_p) = \mathbf{E}(X_t)$, and this shows $Y_t = X_t$ almost everywhere. On the other hand, if X has a cádlág modification then it is trivial to verify the expectation is right continuous. □

Lemma 8.9. *If X is a right continuous supermartingale with respect to a filtration Σ, then X is a supermartingale with respect to Σ^*.*

Proof. TODO □

Chapter 9

Continuous Time Markov Processes

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

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

9.1 Poisson Processes

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

What assumptions do we make about the 'counting process' $\{Y_t\}_{t\in[0,\infty)}$. Firstly, the counter shuold be increasing: the total number of people who

have entered the store should not decrease over time. Secondly, to simplify things, we shall assume that the average number of customers arriving is constant, and that the number of customers arriving at disjoint intervals are independant of one another. This is a Poisson process.

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

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

$$\mathbf{P}(X_t = 1) = \lambda \Delta t + o(t) \tag{9.1}$$
$$\mathbf{P}(X_t = 0) = 1 - \lambda \Delta t + o(t) \tag{9.2}$$
$$\mathbf{P}(X_t > 1) = o(t) \tag{9.3}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Consider the following calculation

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

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

9.2 Continuous Time Markov Process

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

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

$$P_{u,v}(x,z) = \sum_t P_{u,t}(x,y) P_(t,v)(y,z) \tag{9.4}$$

We shall also assume a continuity requirement that

$$\lim_{j \to i^+} \mathbf{P}(X_j = x | X_i = y) = \delta_{x,y} \tag{9.5}$$

A process is **time-homogenous** if

$$P_{u,t}(x,y) = P_{t-u,0}(x,y) \tag{9.6}$$

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 (9.4) and (9.6) that $P_{t+s} = P_t P_s$, so that $\{P_t\}$ forms a commutative monoid.

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

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

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

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

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

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

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

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

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

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

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

By the theory of linear differential equations, this means

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

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

Example. *Consider a Markov chain with infinitisimal generator*

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

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

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

Hence

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

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

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

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

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

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

9.3 Birth and Death Processes

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

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

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

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

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

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

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

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

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

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

Which we solved recursively.

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

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

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

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

This can be rewritten

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

By induction,

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

Hence

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

And thus the chain is transient if and only if

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

Chapter 10

Brownian Motion

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

10.1 Brownian Motion is a Martingale

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

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

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

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

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

10.2 Brownian Motion is a Gaussian Process

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

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

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

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

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

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

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

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

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

by continuity on $t \neq 0$,

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

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

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

Lemma 10.1. *We have* $\mathbf{P}(\sup B_t = \infty, \inf B_t = -\infty) = 1$.

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

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

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

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

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

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

10.3 Brownian Motion is a Markov Process

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

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

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

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

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

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

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

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

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

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

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

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

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

Chapter 11

Stochastic Calculus

Our goal in this chapter will be to make sense of the integral

$$(H \bullet X)_t = \int_0^t H_s dX_s$$

where H, X, and $H \bullet X$ will all be continuous time stochastic proceses. This equation can also be written in the 'differential form' $d(H \bullet X) = HdX$. The most well known integral of this form is known as the Itô integral, after it's creator, and generalizes the martingale $(C \bullet M)$ we studied in the discrete time setting, which represented the overall profit of a series of bets, where the stakes can be adjusted given information available directly before the bet is placed. In the discrete case, C was a previsible process, and M was a martingale, a submartingale, or a supermartingale. In the continuous time case, H will also be a previsible process, in the sense that we can determine the values of H given knowledge known 'infinitisimally before' each time step, and X was be a semimartingale, which includes the class of continuous time, submartingales, martingales, and supermartingales. It models a totalling of a series of infinitisimal bets made against the given stakes X. Like for the Lebesgue integral, we shall build up the Itô integral for the simplest class of integrands, and then construct the general integral by taking the appropriate limits.

Given two stopping times $S \leqslant T$, and a bounded, Σ_S measurable function Z, we will begin by constructing the integral of the function

$$Z(S,T](t,\omega) = Z(\omega)\chi_{(S(\omega),T(\omega)]}(t)$$

For a fixed ω, Z is constant between any two stopping times, which represents a strategy which bets a constant amount between times S and T, and so it makes sense to make the 'obvious definition'

$$\int_0^t Z(S,T]dX = Z(X_{T\wedge t} - X_{S\wedge t})$$

for any given integrator X. The reason that we insist S and T are stopping times, and that Z is bounded and Σ_S measurable implies the following result.

Lemma 11.1. *If M is a uniformly integrable cádlág martingale, then $Z(S,T]\bullet M$ is a uniformly integrable martingale.*

Proof. TODO □

Essentially, stochastic integration theory consists of trying to extend this result to as general a class of functions as we can get, by exploiting the properties of continuous time martingales to their fullest extent.

11.1 Previsibility

Recall that in the discrete setting, we defined a betting scheme C as a random variable adapted to one time step before the values of the bet are revealed. In continuous time, a process should therefore be previsible if it can be predicted 'infinitisimally' into the past. A left continuous process is the perfect candidate for a process of this form, because we can take limits on the left to approximate the next bet 'immediately' before the bet is revealed. The right definition makes sure that we have a suitable algebra of previsible processes. The **previsible σ algebra** on $(0,\infty)\times\Omega$ generated by Σ is defined to be the smallest σ algebra on $(0,\infty)\times\Omega$ such that every adapted cáglád process is measurable. A process on $(0,\infty)$ is called **previsible** if it is measurable as a map from $(0,\infty)\times\Omega$ to $\mathbf{R}$.

Lemma 11.2. *If $S \leqslant T$ are stopping times, and Z is a bounded Σ_S measurable functions, then $Z(S,T]$ is a previsible process.*

Proof. The process $Z(S,T]$ is certainly cáglád, so we need only verify that it is adapted to Σ. But now $Z(S,T] = \lim Z[S_n,T_n)$, where $S_n = S + n^{-1}$ and

$T_n = T + n^{-1}$, and since $Z \in \Sigma_{S_n}$ because $S \leqslant S_n$, it suffices to prove that $Z[S_n, T_n)$ is adapted, and this follows because

$$\{Z[S_n, T_n)_t \in E\} = \{Z \in E\} \cap \{S_n \leqslant t\} \cap \{T_n > t\}$$

and $\{Z \in E\} \cap \{S_n \leqslant t\}$ is in Σ_t, and $\{T > t\}$ is Σ_t measurable. □

We call any finite sum of processes of the form $Z(S, T]$ a *bounded elementary integrand*, a class often denoted by $b\mathcal{E}$. It is rather messy to show (combinatorially reordering the finite sums by taking mins and maxees) that any bounded elementary integrand H may be written in the 'increasing form' $Z_1(S_1, T_1] + \cdots + Z_n(S_n, T_n]$ where $S_1 \leqslant T_1 \leqslant \cdots \leqslant S_n \leqslant T_n$, and $Z_i \in b\Sigma_{S_i}$. Once we have written the theorem in this form, we can unambiguously define the integral

$$\int_0^t H dX = \sum_{k=1}^{n} Z_k[X_{T_k \wedge t} - X_{S_k \wedge t}]$$

The monotone class theorem will come in handy in extending this integral, and one fact that makes this easy is the following.

Lemma 11.3. *The smallest σ algebra containing all bounded element functions is equal to the previsible σ algebra.*

Proof. It is obvious that each element of $b\mathcal{E}$ is previsible, so it remains to show every bounded cáglád adapted process X is $\sigma(b\mathcal{E})$ measurable, and this follows because

$$X = \lim_{k\to\infty} \lim_{n\to\infty} \sum_{i=2}^{nk} X_{\frac{i-1}{n}} \left(\frac{i-1}{n}, \frac{i}{n} \right]$$

and if X is also adapted, then $X_{(i-1)n^{-1}} \in b\Sigma_{(i-1)n^{-1}}$. □

We immediately infer the following monotone class theorem, which enables us to show that the Itô integral is defined for all previsible processes.

Lemma 11.4. *If V is a vector space of bounded processes with parameter set $(0, \infty)$, and*

- *Constant functions are in V.*

- *If H_n is a sequence of elements in V which converge uniformly on $(0,\infty)\times\Omega$ to a function H, then H is in V.*
- *If H_n is a uniformly bounded sequence of nonnegative elements of V and $H_n \uparrow H$, then H is in V.*
- *V contains every bounded elementary function.*

Then V contains every bounded previsible process.

By linearity, $H \bullet X$ is also a uniformly integrable martingale.

11.2 Finite Variation Processes

A **finite variation process** null at zero is an adapted cádlág process X such that each path $t \mapsto X_t(\omega)$ is of finite variation, and $X_0 = 0$. Thus for each t and ω, the variation

$$V_X(t,\omega) = \int_{(0,t]} |dX_s(\omega)| = \sup \sum_{k=1}^{n} |X_{s_k}(\omega) - X_{s_k-1}(\omega)|$$

is finite, where the supremum is taken over all partitions $0 = s_0 < s_1 < \cdots < s_n = 1$. We write FV_0 for the space of finite variation processes null at 0. In this case, if H is a bounded $B(0,\infty)\times\Sigma$ measurable process, then we can easily define

$$\left(\int_0^t H dX\right)(\omega) = \int_0^t H(s,\omega) dX_s(\omega)$$

as the normal Lebesgue Stieltjes integral. An IV_0 process will be a FV_0 process X such that $\|X\|_V = \mathbf{E}V_X(\infty,\omega) < \infty$.

Theorem 11.5. *If H is a bounded, previsible process, and M is a martingale in IV_0, then $H \bullet M$ is a martingale in IV_0.*

Proof. TODO □

11.3 Localization

The boundedness and integrality assumptions we used to conclude on the regularity of the integrals of finite variation processes is too stringent ot be practical. To obtain a more useful result, we must relax the hypothesis of that theorem to a 'localize' version. Of course, then conclusion of the theorem then only holds locally, in some sense.

Consider reducing a global equation $d(H \bullet X) = HdX$ on $(0,\infty)$ to the 'local' equation, that $d(H \bullet X) = HdX$ on $(0,T]$, where T is a stopping time (which we can view as 'local time'). With this idea, given a stopping time T and a process H on $(0,\infty)$, it is natural to introduce the process $H(0,T]_t = \chi_{(0,T]}(t)H_t$ which represents the adjustment to the bet H where we immediately stop betting at time T. Note that if H is previsible, then so is $H(0,T]$. Similarily, if X is a process on $[0,\infty)$, it is natural to introduce the process X^T such that,

$$X^T(t,\omega) = \begin{cases} X(t,\omega) & : 0 \leqslant t \leqslant T(\omega) \\ X(T(\omega),\omega) & : t > T(\omega) \end{cases}$$

which formally means that

$$dX^T(t,\omega) = \begin{cases} dX(t,\omega) & : 0 \leqslant t \leqslant T(\omega) \\ 0 & : t > T(\omega) \end{cases}$$

so that we 'close off all bets' at time T. We can now define localization as saying that the equation

$$(H \bullet X)^T = H(0,T] \bullet X^T$$

holds for stopping time T. Note that now we are forced to view X as a process, rather than just a 'measure' dX.

Let us begin by defining how we localize integrands. Let $\mathcal{L}$ be a family of previsible process with the property that if $H \in \mathcal{L}$, then $H(0,T]$ is in $\mathcal{L}$ for every stopping time T. We say a process H on $(0,\infty)$ is in the localization $l\mathcal{L}$ of the vector space if there exists a sequence of stopping times $T_1 \leqslant T_2 \leqslant \ldots$ with $T_i \uparrow \infty$ such that $H(0,T_n] \in \mathcal{L}$ (we say the T_i 'reduces' H into $\mathcal{L}$). Then $l\mathcal{L}$ is a space stable under the localization $H \mapsto H(0,T]$. If $\mathcal{L}$ is the space of all bounded, previsible processes, then $l\mathcal{L}$ is called the space of all *locally bounded previsible processes*.

Lemma 11.6. *If H is an adapted cáglád process with $\limsup_{t\downarrow 0}|H_t| < \infty$, then H is a locally bounded previsible process.*

Proof. Let $T_n = \inf\{t : |H_t| > n\}$. Then $H(0, T_n]$ is cáglád and bounded, hence bounded and previsible. □

On the other hand, we now define the localization of integrators. Let $\mathcal{L}_0$ be a family of adapted, cádlág processes null at zero, such that if $X \in \mathcal{L}_0$, then $X^T \in \mathcal{L}_0$ for any stopping time T, and we then say $\mathcal{L}_0$ is stable under stopping. A process is in $\mathcal{L}_{0,\text{loc}}$ if there exists a sequence of stopping times $T_1 \leqslant T_2 \leqslant \dots$ with $T_n \to \infty$ such that for all n, $X^{T_n} \in \mathcal{L}_0$. As with integrands, we say the T_i are a reducing sequence for X.

Example. *We let $\mathcal{M}_0$ denote the class of all martingales null at zero, $UI\mathcal{M}_0$ the class of uniformly integrable martingales null at zero, $FV\mathcal{M}_0$ the space of all finite variation martingales null at zero, and $IV\mathcal{M}_0$ the space of all integrable variation martingales null at zero. Each of these spaces are stable under stopping, and give us the corresponding spaces of* **local martingales**.