Stochastic Processes

Jacob Denson

November 27, 2017

Table Of Contents

I Discrete Time Stochastic Processes **3**

1 Finite Markov Chains **4**

1.1 Asymptotics of Markov chains 8
1.2 Irreducibility 9
1.3 Irreducibility and Potential Functions 10
1.4 Existence of a Stationary Distribution 11
1.5 Perron-Frobenius 14
1.6 Aperiodicity and Irreducibility 18
1.7 Periodicity and Average State Distributions 20
1.8 Stopping Times 22

2 Countable-State Markov Chains **27**

2.1 General Properties 27
2.2 Recurrence and Transience 27

3 Harmonic Functions **31**

4 Branching Processes **33**

4.1 The Distribution of the n'th Generation 34
4.2 Mean Population Size 35
4.3 Probability of Extinction 36
4.4 Martingales and Branching Asymptotics 39

5 Reversibility **45**

6 Discrete Time Martingales **48**

6.1 Filtrations 48
6.2 Martingales 50

6.3 Optional Stopping Theorem 51
6.4 Martingale Convergence Theorems 60
6.5 Martingale Inequalities 68
6.6 Martingales in $L^2(\Omega)$ 75
6.7 Optional Sampling Theorem 79

II Continuous Time Stochastic Processes 84

7 Continuous Time Regularity 85
7.1 Regularity of Martingales 89
7.2 Regularization of Brownian Motion and the Markov Property 94
7.3 Kolmogorov's Continuity Criterion 96

8 Continuous Time Martingales 97
8.1 Continuous Time Convergence Theorems 98

9 Stopping Times 102
9.1 The Debút Theorem 107
9.2 The Pre Stopping Time σ algebra 107

10 Markov Processes 108
10.1 Poisson Processes 109
10.2 Continuous Time Markov Process 111
10.3 Birth and Death Processes 114

11 Brownian Motion 116
11.1 Construction of Brownian Motion 117
11.2 The Wiener Measure 122
11.3 Basic Properties 123
11.4 Local Properties of Brownian Motion 125
11.5 Brownian Motion is a Markov Process 128

12 Stochastic Calculus 131
12.1 Previsibility 133
12.2 Finite Variation Processes 137
12.3 Localization 138
12.4 Stochasting Integration By Localization 141
12.5 Local Martingales 142

12.6 Semimartingales as Integrators 144
12.7 Stochastic Integration in L^2 145
12.8 The Kunita Watanabe Inequalities 149

13 Appendix: Conditional Expectations 150
13.1 Classical Conditioning 151
13.2 Kolmogorov's Realization 152
13.3 General Conditional Expectations 154
13.4 Properties of the Conditioning Operator 156
13.5 Conditional Probabilities 159
13.6 Independence and Conditional Expectation 162

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. The theory of stochastic processes is devoted to constructing and studying the various stochastic processes which occur in pure mathematics and the sciences.

Example. *To model the uncertainty of weather, we may take a stochastic process with state space $S = \{sunny, rainy\}$. For $n \in \mathbf{N}$, we may model the weather by a random variable X_n modelling the weather on a certain day n. Then $\{X_n : n \in \mathbf{N}\}$ 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_t be the value of a certain stock at time t, for $t \in [0,\infty)$. This is a continuous time random process, because the values are indexed over time, and the states are also continuous.*

Example. *To estimate the cumulative density function of an independant and identically distributed sample $X_1,\dots,X_n \sim F$, we can take the estimate*

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

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.

There are two main ways of thinking about stochastic proceses. On one hand, fixing t, we can think of each X_t as an individual random variable, and then use classical methods such as expectation and convergence theorems to understand how the variables X_t behave over time. The other way of studying the process is to fix ω, and to think of the process as a random trajectory $X_t(\omega)$ changing over time. Taken to its extreme, we can view stochastic processes as a S^T valued random variable. This is most powerful in the study of continuous time stochastic processes, where we can think of a stochastic process as a random curve in space.

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. So let's get to it!

Part I

Discrete Time Stochastic Processes

Chapter 1

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 $X_1, X_2, \dots$ 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 $x_1, \dots, x_n, x_{n+1}$,

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

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,*

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

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

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

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

This shows that $\{Y_k\}$ satisfies the Markov property, 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 $X_0, X_1, \dots$ 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

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

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

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 Carátheódory 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*

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

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*

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

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 $\{0,1,\dots,n\}$, where each integer is connected to adjacent integers. The transition probabilities are given by*

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

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

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.

1.1 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 average amount of time spent in each state converges 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.

1.2 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 $n \to \infty$.

1.3 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 1.1. *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 1.2. *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.

1.4 Existence of a Stationary Distribution

Given a state x on a Markov process $X_0, X_1, \dots$, 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 1.3. *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 1.4. *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 $X_0, \dots, X_k$, 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 $X_0, \dots, X_k$. 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

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

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.

1.5 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 1.5 (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 $g = \min(g_1, \dots, g_d)$, 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 $w = (|v_1|, \dots, |v_n|)$, 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.*

1.6 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 1.6. *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, $P_{ii}^m > 0$ for all i. If $j \neq i$, there is some k for which $P_{ij}^k > 0$. Then, for large enough m, $P_{ij}^m > 0$, since

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

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

Corollary 1.7. *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 $x_0, x_1, \dots$ to $x_1, \dots$ is measure preserving under measure induced by the random variables $X_0, X_1, \dots$. 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*

$$P = \begin{array}{c c} & \begin{array}{cc} x & y \end{array} \\ \begin{array}{c} x \\ y \end{array} & \left[\begin{array}{cc} 1-p & p \\ q & 1-q \end{array} \right] \end{array}$$

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.

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

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

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.

1.8 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 $\{X_0, X_1, \ldots\}$ is a $\mathbf{Z} \cup \{\infty\}$ valued random variable τ, such that, if we know the values of $X_0, \ldots, X_n$, we can tell if $\tau = n$. Rigorously, $\mathbf{I}(\tau = n)$ is a function of the $X_0, \ldots, X_n$.

A stopping time basically encapsulates a decision process. After observing the $X_0, \ldots, X_n$, 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 $\{X_0, X_1, \ldots\}$ 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 $\mathbf{I}(\tau_s = n) = \mathbf{I}(X_0 \neq s, \ldots, X_{n-1} \neq s, X_n = s)$, this is a stopping time.

Example. *Let $\{X_0, X_1, \ldots\}$ 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 $t_1 < \dots < t_n < \tau$,

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

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

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

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

So the process is strongly Markov. □

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_{ij}^k is the probability that $X_k = x_j$ given $X_0 = x_i$, so $(\sum_{k=0}^n Q^k)_{ij}$ is the expected number of visits to x_j from time epoch n starting from x_i. Taking $n \to \infty$, we find the expected number of visits to the state before hitting a recurrent state is $(I - Q)_{ij}^{-1}$. If we sum up row i, we get the expected number of states before hitting a recurrent state starting from i.

We can also use this method in an irreducible chain to find the expected time to reach a state x_j starting at x_i, for $i \neq j$. We modify the Markov process by making it impossible to leave x_j once it has been entered. This makes all other states transient. Then the expected number of visits before 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 2

Countable-State Markov Chains

2.1 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 complicated, since in a single 'recurrence class' of infinite size, it may still be very rare for a state to return to itself.

2.2 Recurrence and Transience

We call a state **recurrent** if the markov chain is almost certain to return to every state it starts at 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. Note that, just because every states returns to itself with probability one, does not imply that we always return to every state infinitely many times, because there are infinitely many states.

There is a feasible way to 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

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

FINISH HERE and the walk is recurrent for $d \leqslant 2$, 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 3

Harmonic Functions

We've already seen some work on harmonic functions on a finite state space. Here we address some of the theory when the state space is countable. Suppose that we are given a time homongenous Markov transition operator $P : E \times E \to [0,1]$. We have seen that P operates on probability measures μ on E on the right to give the probability measure μP one step into the future in the Markov process starting with distribution μ, and that if $f : E \to \mathbf{R}$, then

$$(Pf)(x) = \sum P(x,y)f(y)$$

Gives the expected value of f one step into the future starting at x. We call a function **harmonic** when $Pf = f$, **subharmonic** if $Pf \geqslant f$, and **superharmonic** if $Pf \leqslant f$.

If $X_1, X_2, \dots$ is a Markov chain induced from P by any initial measure μ, and f is a superharmonic function, then $f(X_n)$ is a supermartingale, because

$$\mathbf{E}[f(X_n)|X_1,\dots,X_{n-1}] = \sum f(X_{n-1},y)f(y) = (Pf)(X_{n-1}) \leqslant f(X_{n-1})$$

In particular, if f is non-negative, and if the markov chain is irreducible recurrent, in the sense that the stopping times $T^x = \min\{n > 0 : X_n = x\}$ of visiting each state in the markov chain are finite almost surely regardless of the intiail distribution, then we can apply the optional stopping theorem for non-negative supermartingales to conclude that if $X_0 = x$, then

$$f(y) = \mathbf{E}[f(X_{T^y})] \leqslant \mathbf{E}[f(X_0)] = f(x)$$

Since x and y were arbitrary, we conclude that f must be a constant function. Conversely, suppose every non-negative harmonic function on a Markov chain is constant. If we let $A(x,y) = \mathbf{P}_x(T^y < \infty)$ denote the probability that we will eventually reach y starting at x, then we find

$$A(x,y) = P(x,y) + \sum_{z \neq y} P(x,z)A(z,y) \geqslant \sum P(x,z)A(z,y)$$

It follows from this that the function A is a non-negative superharmonic function, hence A is constant, $A(x,y) = C \in [0,1]$. But now we find

$$C = A(x,y) = P(x,y) + C \sum_{z \neq y} P(x,z) = P(x,y) + C[1 - P(x,y)]$$

so $CP(x,y) = P(x,y)$, and since we cannot have $P(x,y) = 0$ for all x,y, we conclude $C = 1$, which means exactly that the Markov chain is irreducible recurrent. Thus the class of superharmonic functions on the process has sufficient information to tell us whether a process is irreducible recurrent. The theory of harmonic functions in stochastic processes explores how strong the correspondence is between a given process and the harmonic functions on that process.

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 $X_0, X_1, \dots$, where the initial generation X_0 produces the offspring X_1, which produces the offspring X_2, and so on, through the ages. Each time epoch 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 $\rho * \dots * \rho$ is the distribution of a sum of k independent random variables $Y_1, \dots, Y_k$ distributed according to ρ, and so by independence

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

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 $Y^1, \dots, Y^k$, with $Y^i_0 = 1$. 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

$$\pi_n(k) = \mathbf{P}(X_n = 0) = \mathbf{P}(Y^1_n = 0, \dots, Y^k_n = 0) = \prod \mathbf{P}(Y^i_n = 0) = \pi_n^k$$

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

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

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

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

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 $\{1, 2, \dots, \infty\}$, 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 $Y_1, Y_2, \dots$ 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 $X_0, X_1, \dots, X_n$ are the first few frames of the movie, then $(X_0, \dots, X_n)$ is distributed identically to $(X_n, \dots, X_0)$. We have

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

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

Definition. A Markov chain is **reversible** if there is a measure 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 $(X_0, X_1, \dots, X_n)$ is identically distributed to $(X_n, \dots, X_0)$, 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 $(X_0,\dots,X_N)$ 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 $(X_N,\dots,X_0)$, 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 $Z_1, \dots, \infty$ 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

Discrete Time Martingales

6.1 Filtrations

As a stochastic process evolves, we observes the states it take, giving us more and more information into the future of the process. We can model this 'increase' in information in a measure theoretic manner, by a **filtration**, which is just an increasing family of σ algebras

$$\Sigma = \{\Sigma_0 \subset \Sigma_1 \subset \dots\}$$

Each σ algebra is the set of information available at a given time, which should increase as we observe more and more states of the process. A process X is adapted to a filtration Σ if X_i is Σ_i measurable. This means that observations of the values of X_i are available at time i in the filtration. In fact, without a filtration, it would be fairly difficult to think of the indices $\{0, 1, \dots\}$ as time, because we have no way to judge the relationship between the sequence of random variables.

Example. *Let $X_0, X_1, \dots$ be a stochastic process. The* **natural filtration** *corresponding to X is the filtration $\Sigma_n = \sigma(X_0, \dots, X_n)$, which represents the information given over time if we recieve the values $X_0, X_1, \dots$ sequentially.*

Filtrations have been underlying all our work on Markov chains this far. If we set Σ_n to be the natural filtration on a process X, the Markov property says that for each state s,

$$\mathbf{P}(X_{n+1} = s | \Sigma_n) = \mathbf{P}(X_{n+1} = s | X_n)$$

So a Markov chain is a stochastic process at which the current value of the chain is the only useful information in a filtration to predict the next value of the process. We didn't really need abstract filtrations in the analysis of finite state Markov chains, but now that we know what a filtration is, we can describe the time dependant properties of the process as a relationship between filtrations.

The natural filtration is not the only filtration which occurs in probability theory, though it is often the natural filtration to use, hence the name. The point is that a filtration should contain all the relevant statistical information we have about a scenario. For instance, the payout in a particular game of poker is certainly important information to keep in mind to come up with future strategies, but one also needs to keep in mind the behaviour of the other players, which enable us to predict a player's future behaviour. A filtration modelling a game of poker needs to take all this information into account, rather than just the payout of individual games.

Example. *Consider a family $X_1, X_2, \dots$ of fair $\{0,1\}$ valued Bernoulli random variables, and the corresponding sums $S_n = \sum_{k \leqslant n} X_k$. Consider an independent random walk Y with $\{\pm 1\}$ fair Bernoulli increments, and then consider the process*

$$Z_n = Y_n + \lfloor S_n / m \rfloor$$

which mimics the random walk Y, except that when S_n reaches certain increments the random walk drifts one step upwards. If we are able to observe the random variables X_i as well as the process Z_n, then the natural filtration to use in this situation is

$$\Sigma_n = \sigma(X_1, X_2, \dots, X_n, Z_1, \dots, Z_n) = \sigma(X_1, \dots, X_n, Y_1, \dots, Y_n)$$

which contains much more information then the natural filtration with respect to the Z_n, because now with the values X_i we can anticipate when the next 'shift' will happen.

We shall often have interest in analyzing the limiting σ algebra $\Sigma_\infty = \lim \Sigma_n$, which represents all the information we could ever be able to obtain by looking at the stochastic process. If a stochastic process X is adapted to a filtration Σ, then all variables in the process are Σ_∞ measurable, so to understand these variables it suffices to look at their action on Σ_∞. For statistical interest, given a random variable X, we can consider the sequence

$$\mathbf{E}(X|\Sigma_0), \mathbf{E}(X|\Sigma_1), \mathbf{E}(X|\Sigma_2), \dots$$

and we can measure how 'smoothly' we are given information about X by discussing the convergence of $\mathbf{E}(X|\Sigma_n)$ to $\mathbf{E}(X|\Sigma_\infty)$. If the sequence converges nicely, this means we consistantly learn all we need to know about the function.

6.2 Martingales

We wish to discuss stochastic processes X representing a series of 'fair bets' over time. Each point of time in the process represents the amount of money in a gambler's pocket at a certain time. If we have a certain amount of money at time i, then we should expect to have, on average, the same amount of money in the future, given that all our bets were fair. This should be stronger than the simple condition that the $\mathbf{E}(X_i)$ are all equal to one another, because we are able to observe how much money is in our pocket at any moment in time. A **discrete integrable martingale** with respect to a filtration $\Sigma = \{\Sigma_0, \Sigma_1, \dots\}$ is a Σ adapted process $M = \{M_0, M_1, \dots\}$ consisting of integrable random variables such that, $\mathbf{E}(M_n|\Sigma_m) = M_m$ for $m \leqslant n$. Though it is helpful to think of these processes as modelling gambling, they seem to crop up in almost every aspect of probability theory. Note that, by the tower rule, to verify a process M is a martingale, it suffices to show that $\mathbf{E}(M_{n+1}|\Sigma_n) = M_n$.

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

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

Thus S_n is a martingale relative to the filtration induced by X. Because

$$\sigma(S_1, \dots, S_n) = \sigma(X_1, \dots, X_n)$$

S_n is also a martingale relative to the natural filtration.

Example. *As in the last example, consider a sequence of independent non-negative integrable random variables $X_1, X_2, \dots$, but now with mean 1 rather than 0. If we define $M_0 = 1$, $M_n = X_1 \dots X_n$, then*

$$\mathbf{E}(M_{n+1}|X_1, \dots, X_n) = M_n \mathbf{E}(X_{n+1}|X_1, \dots, X_n) = M_n \mathbf{E}(X_{n+1}) = M_n$$

so M is a martingale; because the product of two integrable independent random variables is also integrable, this guarantees M is also integrable. It is also a martingale with respect to the natural filtration on the M_n, though this time it is possible that

$$\sigma(M_1,\dots,M_n) \neq \sigma(X_1,\dots,X_n)$$

The intuition being that if $X_i = 0$ at some point, then $M_i = \dots = M_n = 0$, so we do not obtain as much information about $X_{i+1},\dots,X_n$.

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.

It will be helpful to consider more general processes than just ‘fair bet’ processes. 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 is consistantly in the player's favour, and a supermartingale is consistanly a game against the player. The names are often confusing, for we would imagine supermartingales to increase, and submartingales to decrease. Unfortunately, we are stuck with these names. One can remember this by remembering su*b*martingales increase, and su*p*ermartingales decrease. Like with Martingales, 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.

6.3 Optional Stopping Theorem

We begin our analysis of martingales with an extension of the ‘gambling’ interpretation. First, note that we can reexpre the conditions guaranteeing a process M is a martingale to the equation

$$\mathbf{E}[M_{n+1} - M_n|\Sigma_n] = 0$$

The values of $M_{n+1} - M_n$ can be seen as the payout on a particular ‘game’ in a series of games in a casino. These ‘games’ are fair, because the average amount of money we make is 0, given all possible information about the previous games that we can obtain.

We can obtain a family of martingales from a single martingale M by changing the stakes made on game n from a unit amount to an arbitrary random stake C_n which doesn't 'cheat' by looking into the outcome of a game. This condition is specified by saying C is **previsible**, so each C_n is Σ_{n-1} measurable. The total amount of money we make by game n is then

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

It is easy to verify that if each C is constant, then $C \bullet M$ is a Martingale. More generally, if C_n is previsible, and is in $L^q(\Omega)$ whenever $M_k - M_{k-1} \in L^p(\Omega)$, then $C \bullet M$ will be a integrable martingale. We call $C \bullet M$ the **discrete Itô integral** or **martingale transform** of C with respect to M. $C \bullet M$ also inherits the property of being a submartingale or a supermartingale, if M has this property.

Example. *Suppose we enter a casino with an initial monetary stake X_0, and we play a series of independant games $Y_1, Y_2, \dots$ with $\mathbf{P}(Y_i = 1) = p$, $\mathbf{P}(Y_i = -1) = q$. To be realistic, our stakes in a particular gambling game must be bounded by the amount of money at a particular time. In particular, this means that if we choose stakes C_k, and then consider*

$$X_n = X_0 + (C \bullet Y)_n$$

Then $0 \leqslant C_n \leqslant X_n$ holds pointwise. Obviously, our aim is to choose our betting strategy C_n to maximize the amount of profit we make. If we gamble too much of our money away at once, we may lose the ability to make profit in the future by limiting the stakes we can make, so the betting strategies must balance risk and reward to obtain an optimal value. To make things nice mathematically, we attempt to maximize the expected value of the 'interest rate'

$$\log(X_N/X_0) = \log(X_N) - \log(X_0)$$

Note that if we set our filtration to be $\Sigma_n = \sigma(Y_1, \dots, Y_n)$, and enforce that C is previsible with respect to this filtration, then

$$\mathbf{E}[X_{n+1}|\Sigma_n] = X_n + (p-q)C_{n+1}$$

If $p < q$, then X_n will always be a submartingale, and so Jensen's inequality guarantees that, since log *is a concave increasing function,*

$$\mathbf{E}[\log(X_{n+1})|\Sigma_n] \leqslant \log(\mathbf{E}[X_{n+1}|\Sigma_n]) \leqslant \log(X_n)$$

Thus the expected value $\log(X_n)$ *cannot increase, and choosing not to play guarantees that the expected interest rate is maximized, because it won't decrease from* $\log(X_0)$. *Similarily, if* $p = q$, *then* $\log(X_n)$ *will be a supermartingale, and choosing not to play will maximize our expected interest rate. If* $p > q$, *things are more interesting. We calculate that*

$$\mathbf{E}[\log(X_{n+1})|\Sigma_n] = p\log(X_n + C_{n+1}) + q\log(X_n - C_{n+1})$$

We now attempt to choose $\alpha > 0$ *such that*

$$p\log(X_n + C_{n+1}) + q\log(X_n - C_{n+1}) \leqslant X_n + \log\alpha$$

It then follows that $\log X_n - n\log\alpha$ *is a supermartingale, so*

$$\mathbf{E}[\log(X_n/X_0)] \leqslant n\log\alpha$$

bounds the growth of the martingale. Since $0 \leqslant C_{n+1} \leqslant X_n$, *we calculate using elementary calculus that*

$$\max_{0\leqslant y\leqslant x} p\log(x+y) + q\log(x-y) = 2p^p q^q x$$

which is attained when $y = (p-q)x$. *It follows that the bound about is attained if* $\alpha = 2p^p q^q$, *and so*

$$\mathbf{E}[\log(X_n/X_0)] \leqslant n(\log 2 + p\log p + q\log q)$$

we call $\log 2 + p\log p + q\log q$ *the entropy of the game. It follows that in order to maximize the average interest rate of the game, we should set* $C_n = (p-q)X_{n-1}$. *It is interesting to note that in this game, the optimal strategy is invariant of time; we need not change our gambling strategy near the end, or near the beginning to maximize the amount of money we make.*

Probability was created to analyze gambling games, and martingale theory was invented 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 $\{\pm 1\}$ valued independent Bernoulli trials $X_1, X_2, \dots$, and define $M_n = X_1 + \dots + X_n$ 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 where $C_k = 2^k$, so our bet 'doubles' each time. If $X_n(\omega) = 1$, then

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

Thus, if we consider a betting strategy relative to C where, as soon as we win a single bet, we leave the casino, then we always come out with a unit profit. More rigorously, if we define the 'stopping time' T to be the first time n that $X_n = 1$ (a random quantity), then $\mathbf{P}(T < \infty) = 1$, and $(C \bullet M)_T \geqslant 1$. This means we've found a guaranteed way to beat the casino! The Martingale was all the rage among gamblers in the 18th century. Casanova was one of many famous figures known to apply the strategy to his games.

At first, the martingale strategy seems to contradict the fact that for any series of previsible stakes C, $C \bullet M$ is a martingale. However, the problem with the martingale analysis is that we are 'changing time units'. Consider the modified stakes

$$C_n^{(T)} = \begin{cases} C_n & n \leqslant T \\ 0 & n > T \end{cases}$$

which model the outcomes of the bet n steps into the future if we stop betting at time T. The modified stakes are still previsible, because $\{n \leqslant T\}$ is the complement of the event $\{T < n\}$, which is Σ_{n-1} adapted because it depends only on the variables $X_1, \dots, X_{n-1}$. Since the modified stakes are bounded by C_n, we conclude that $C^{(T)} \bullet M$ is still 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 $\{0, 1, \dots, \infty\}$ valued random variable 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 implies the next result.

Theorem 6.1. *If T is a stopping time, then M^T inherits the property of being a submartingale or a supermartingale from M.*

Paradoxically, 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 the two analyses is that we began by analyzing the random variable

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

and even though this limits holds, we find

$$\mathbf{E}((C \bullet M)_0) = 0 < 1 = \mathbf{E}((C \bullet M)_T)$$

so we cannot extend $C^{(T))} \bullet M$ to be a martingale 'at ∞', or rather, we cannot extend $C \bullet M$ to be a martingale at T. This is one of many examples where the expectation operator does not play as nicely as we would like with respect to pointwise convergence. Here is another example.

Example. *Consider a simple absorbing random walk M on $\mathbf{N}$ starting at 1. Then M is a martingale. If we consider the stopping time T to be the first time that $M_T = 0$, then one can calculate that $\mathbf{P}(T < \infty) = 1$, and even though $\mathbf{E}(M^T_n) = \mathbf{E}(M^T_0) = 1$, and $M^T_n \to M_T$ pointwise, we find $M_T = 0$, so we cannot extend the stopped martingale to a martingale in the limit.*

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. From the discussion above, the optional stopping theorem fits in the family of results, such as the monotone convergence theorem, dominated convergence theorem, and so on, which say that expectation does play nicely with pointwise convergence given certain additional conditions. One such condition we could place is that the stakes C_n are bounded. This is the reason why limits are placed on poker and blackjack tables. 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 give you an asymptotic edge over the casino. The essential part of the analysis governing the optional stopping theorem is to write

$$M_T = M^T_n + \mathbf{I}(T > n)(M_T - M_n)$$

Provided $\mathbf{I}(T > n)M_T$ and $\mathbf{I}(T > n)M_n$ are small in expectation for large n, we can obtain the result. We can do this either by bounding T's behaviour at ∞ or bounding M's behaviour. We say 'The Optional Stopping Theorem' holds for a stopping time T and a stochastic process M if M_T is integrable, and one of the following cases holds

- M is a supermartingale, and $\mathbf{E}[M_T] \leqslant \mathbf{E}[M_0]$.
- M is a submartingale, and $\mathbf{E}[M_T] \geqslant \mathbf{E}[M_0]$.
- M is a martingale, and $\mathbf{E}[M_T] = \mathbf{E}[M_0]$.

The theorem provides a list of conditions which cause the theorem to hold.

Theorem 6.2 (The Optional Stopping Theorem)**.** *Let T is a stopping time and M a stochastic process. The optional stopping theorem holds in the following cases:*

- $\mathbf{E}[M_T] < \infty$, $\mathbf{P}(T < \infty)$*, and*
$$\lim_{n \to \infty} \int_{\{T > n\}} |M_n| = 0$$
In particular, this covers the case where T is bounded, or M is bounded and T is finite almost surely.
- $\mathbf{E}(T) < \infty$*, and the increments of the process are bounded.*
- *M is a non-negative supermartingale, and* $\mathbf{P}(T < \infty) = 1$.
- *M is uniformly integrable, and* $\mathbf{P}(T < \infty) = 1$*, and* $M_T \in L^1(\Omega)$.

Proof. Let us start with the first case. We have noted that the optional stopping theorem holds provided $\mathbf{E}[\mathbf{I}(T > n)(M_T - M_n)] \to 0$ as $n \to \infty$. Since M_T is integrable, $\mathbf{E}[\mathbf{I}(T > n)M_T] \to 0$ since $\mathbf{P}(T > n) \to 0$. The integrality condition guarantees $\mathbf{E}[\mathbf{I}(T > n)M_n] \to 0$. This completes the first part of the proof. To address the next case, we set K to be a constant such that $|M_{n+1} - M_n| \leqslant K$ almost surely, and we then find

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

almost surely. Thus $|M_n^T|$ is bounded by an $L^1(\Omega)$ function $|M_0| + KT$, so we can again apply the dominated convergence theorem, since $M_n^T \to M_T$

pointwise, to conclude the optional stopping theorem holds. If M_n is a non-negative supermartingale, then we can apply Fatou's lemma to M_n^T to conclude

$$\mathbf{E}[M_T] = \mathbf{E}[\liminf M_n^T] \leqslant \liminf \mathbf{E}[M_n^T] = \mathbf{E}[M_0]$$

so the proof is easy. Finally, if M is uniformly integrable, and $\mathbf{P}(T < \infty) = 1$, then $\mathbf{P}(T > n) \to 0$, and by uniform integrability, we conclude

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

Provided $\mathbf{E}|M_T| < \infty$, we obtain the required convergence. □

Example. *Consider a variant of the gambler's ruin problem, where we start with A units of money, and play a fair game until we go bust, or stop playing when we end up with B units of money. In particular, let's consider the problem where the games are fair $\{\pm 1\}$ valued Bernoulli trials, so the amount of money in the gambler's pocket is modelled by the martingale*

$$M_n = A + X_1 + \cdots + X_n$$

We let T be the first time that $M_T = 0$ or $M_T = B$. Now M_n might not be a bounded process, but M_n^T is a bounded process, because $0 \leqslant M_n^T \leqslant B$, and since $\mathbf{P}(T < \infty) = 1$, we can apply the optional stopping theorem to M_n^T to conclude that

$$B\mathbf{P}(M_T = A) = \mathbf{E}[M_0^T] = \mathbf{E}[M_0] = A$$

hence the probability that a gambler actually succeeds in making the money he set out to get is A/B. We can also consider the martingale

$$N_n = M_n^2 - n$$

Now N^T is not necessarily a bounded martingale, but by general properties of Markov processes on a finite state space, we know there is $0 < \rho < 1$ such that $\mathbf{P}(T > n) = O(\rho^n)$, and since we have $|N_n| \leqslant \max(n, B^2 - n) = O(n)$, we conclude

$$\int_{\{T > n\}} |N_n| = O(n\rho^n) \to 0$$

and therefore we can apply the optional stopping theorem to conclude that

$$A^2 = \mathbf{E}[N_0] = \mathbf{E}[N_T] = \mathbf{E}[M_T^2] - \mathbf{E}[T] = (A/B)B^2 - \mathbf{E}[T]$$

Rearranging this equation gives $\mathbf{E}[T] = A(B - A)$, which is the expected time before the gambler will finish playing.

Example. *Kolmogorov's theorem tells us that a monkey typing randomly on a typewriter will almost surely type any particular sequence of random characters. Using the optional sampling theorem, we can calculate the expected time until the monkey types a given sequence. The classic example is the word "ABRACADABRA". Consider a bookie, which places bets on a particular character being typed at* $26 : 1$ *odds, until the monkey finishes typing "ABRACADABRA", in which case the bookie immediately closes, and no-one can continue betting. Consider a man who enters the bookie at time* n *with a single dollar, and makes successive bets on the next character being A, B, R, A, and so on, until the bet is lost, in which case the man has no money, and leaves. Of course, if the casino closes, the man has to leave anyway, but gets to keep all the money he's earnt so far. Formally, we let* T *be the stopping time where "ABRACADABRA" is finished being spelt, or the time where the man loses his bet. If* M *denotes the amount of money in the man's pocket at a given time, then* M *is a martingale.* T *is bounded, so we can apply the optional stopping theorem to conclude that*

$$\mathbf{E}[M_T] = \mathbf{E}[M_0] = 1$$

but there are only a couple situations where the gambler leaves with any money at all

- *The bookie closes before the man can start betting, so* $M_T = 1$ *where* $T < n$.
- *The man bets on A being the character the monkey types at time* n, *and he wins, but the monkey finishes type "ABRACADABRA" at this instance and the man has to leave, so* $M_T = 26$, *and* $T = n$.
- *The man wins his fourth bet that "ABRA" will be spelt consecutively starting at time* n, *but the monkey finishes spelling "ABRACADABRA" at the last A, so* $M_T = 26^4$, *and* $T = n + 4$.
- *The man wins all of his bets, so he leaves with* 26^{10} *dollars, and* $T = n+1$.

If we set $p_n = \mathbf{P}(T = n)$, *then we conclude*

$$1 = \mathbf{E}[M_T] = \mathbf{P}(T < n) + 26p_n + 26^4 p_{n+4} + 26^{10} p_{n+10}$$

Thus for any $n \geqslant 1$,

$$\mathbf{P}(T \geqslant n) = 26p_n + 26^4 p_{n+4} + 26^{10} p_{n+10}$$

If we sum this equation over all values of n, we conclude that

$$\mathbf{E}[T] = \sum_{n=1}^{\infty} 26p_n + 26^4 p_{n+4} + 26^{10} p_{n+10} = 26 + 26^4 + 26^{10}$$

The same technique can be used to calculate the expected time we wait until a monkey types any sequence of characters, where we determine which prefixes of the sequence are also suffixes of the word. Regardless, one can argue that the expected waiting time for a sequence of length n is bounded below by 26^n, and bounded above by $(1+1/25)26^n$, so we'll have to wait exponential time in the number of characters we want the monkey to type.

Example. *Consider a random walk $S_n = X_1 + \dots + X_n$ given by fair independent $\{\pm 1\}$ valued Bernoulli random varibles $X_1, X_2, \dots$, with $S_0 = 0$. We now show how one can use generating function methods in combination with martingale methods to calculate the distribution of the stopping time T, which is the first time that $S_T = 1$. Note that if X is Bernoulli, then*

$$\mathbf{E}[e^{tX}] = \frac{e^t + e^{-t}}{2} = \cosh(t)$$

so $\mathbf{E}[(\mathrm{sech}\, t)e^{tX}] = 1$, and therefore the sequence of random variables

$$M_n^t = (\mathrm{sech}\, t)^n e^{tS_n}$$

is a martingale, with $M_0^t = 1$. If $t > 0$, then $e^{tS_{T \wedge n}} \leqslant e^t$, and if $T = \infty$, then $M_n^t \to 0$ as $n \to \infty$ since $S_n \leqslant 0$ for all n, hence if we define $M_T = 0$ if $T = \infty$, and M_T canonically otherwise, then the optional stopping theorem gives

$$1 = \mathbf{E}[M_T^t] = \mathbf{E}[(\mathrm{sech}\, t)^T] e^t$$

Hence $\mathbf{E}[(\mathrm{sech}\, t)^T] = e^{-t}$. As $t \to 0$, $(\mathrm{sech}\, t)^T \to 1$ pointwise monotonely if $T < \infty$, and $(\mathrm{sech}\, t)^T \to 0$ if $T = \infty$ monotonely, hence the monotone convergence theorem guarantees that

$$1 = \mathbf{P}(T < \infty)$$

Now if $\alpha = \mathrm{sech}\, t$, we conclude using the binomial theorem that

$$\mathbf{E}(\alpha^T) = \sum \alpha^n \mathbf{P}(T = n) = e^{-t} = \frac{1 - \sqrt{1 - \alpha^2}}{\alpha} = \sum_{m=1}^{\infty} (-1)^{m+1} \binom{1/2}{m} \alpha^{2m-1}$$

Hence $\mathbf{P}(T = 2m-1) = (-1)^{m+1} \binom{1/2}{m}$.

In most applications of the optional stopping theorem, we want to bound the expected value of T. This uses the probabilistic principle that 'whatever has a reasonable chance of happening, will almost surely happen infinitely often'. The particular application is described in the next lemma.

Lemma 6.3. *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

$$\mathbf{P}(T > N) = 1 - \mathbf{P}(T \leqslant n + N) \leqslant 1 - \varepsilon$$

By induction, we calculate that $\mathbf{P}(T > kN) \leqslant (1-\varepsilon)^k$, by checking that

$$\begin{aligned} \mathbf{P}(T > (k+1)N) &= \mathbf{P}(T > kN) - \mathbf{P}((k+1)N \geqslant T > kN) \\ &\leqslant (1-\varepsilon)^k - \mathbf{E}(\mathbf{P}[(k+1)N \geqslant T > kN | \Sigma_{kN}]) \\ &\leqslant (1-\varepsilon)^k - \mathbf{E}\left(\chi_{\{T > kN\}} \mathbf{P}[(k+1)N \geqslant T | \Sigma_{kN}]\right) \\ &\leqslant (1-\varepsilon)^k - \varepsilon \mathbf{P}(T > kN) = (1-\varepsilon)^{k+1} \end{aligned}$$

It follows that

$$\begin{aligned} \mathbf{E}(T) &= \sum_{n=1}^{\infty} \mathbf{P}(T \geqslant n) \\ &= \sum_{n=1}^{N} \mathbf{P}(T \geqslant n) + \sum_{m=1}^{\infty} N \mathbf{P}(T > nK) \\ &\leqslant \sum_{n=1}^{N} \mathbf{P}(T \geqslant n) + N/\varepsilon < \infty \end{aligned}$$

and this bound completes the proof. □

6.4 Martingale Convergence Theorems

In many cases, we want to analyze the asymptotic properties of Martingales over time. In this section, we obtain powerful convergence results about these martingales. Here are some examples where this is useful.

Example. *If X_i are independent with mean 0, then $S_n = \sum X_i$ is a martingale, and 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_∞. Similarily, if X_i have mean one, we want to analyze whether $M_n = X_1 \dots X_n$ settles down to some random variable M_∞, which is a 'product form' of the law of large numbers.*

Example. *Given a random variable X and a filtration Σ_n, we want to ask whether the martingale*

$$\mathbf{E}[X|\Sigma_0], \mathbf{E}[\Sigma_1], \dots$$

converges to $\mathbf{E}[X|\Sigma_\infty]$, which means 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.

We start by proving that if a martingale doesn't 'spread out' too much, then it must actually converge over time to a finite value, because the conditional expectations restrict the martingale to having smaller and smaller spreads over time. Doob's upcrossing lemma provides the first result along this manner, and we consider it's consequences to its natural conclusion. 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 $0 \leqslant t_1 < s_1 < \dots < t_n < s_n \leqslant n$ 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. It turns out that the variation of martingales is essentially time invariant, if the 'spread' of the martingale is time invariant. We obtain the proof by showing that 'buy low', 'sell high' doesn't help you when you're gambling against an unfair system.

Theorem 6.4 (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 expected spread over time impacts the expected number of upcrossings, rather than the time epoch n itself.

Proof. Consider a previsible process C which follows the following ruleset: 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 6.5. *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 $\mathbf{P}(U_\infty[a,b] = \infty) = 0$.*

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 6.6 (Doob's Upward Convergence Theorem). *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 $M_\infty = \limsup M_n$, 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

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

So M_∞ is integrable, and finite almost surely. □

Corollary 6.7. *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. □

It would be wise before moving on to review the concept of the uniform integrability of random variables, since it is one of the keystones from moving the pointwise convergence of M_n to M_∞ in the previous theorems to convergence in the L^1 norm. We recall the main point that a sequence of random variables X_n converges in L^1 to X if and only if the variables converge in probability, and the variables are uniformly integrable. The next theorem is essentially obvious given the standard results on uniform integrability.

Corollary 6.8. *If M is a (sub/super) martingale bounded in $L^1(\Omega)$, then $M_n \to M_\infty$ in $L^1(\Omega)$ if and only if M is uniformly integrable. It then follows that $\mathbf{E}(M_\infty|\Sigma_n) \leqslant M_n$ almost surely if M is a supermartingale, and $\mathbf{E}(M_\infty|\Sigma_n) \geqslant M_n$ for submartingales. This means we can think of M as being a martingale with time indices $\mathbf{N} \cup \{\infty\}$.*

Proof. Since pointwise almost sure convergence implies convergence in probability, if M_n is uniformly integrable, then the result we mentioned above implies $M_n \to M_\infty$ in $L^1(\Omega)$, and the only part of this theorem left to prove is that $\mathbf{E}(M_\infty|\Sigma_n) \leqslant M_n$ if $M_n \to M_\infty$ in the L^1 norm. 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 next theorem shows that random variables in L^1 are 'smoothly learnable'.

Lemma 6.9. *If $X \in L^1(\Omega)$ is a random variable, then, as Σ ranges over all subsigma algebras of the probability space, $\{\mathbf{E}(X|\Sigma)\}$ is a uniformly integrable family.*

Proof. By the conditional form of Jensen's inequality, we conclude that $|\mathbf{E}(X|\Sigma)| \leqslant \mathbf{E}(|X||\Sigma)$ almost everywhere. Now for every $K > 0$, $|\mathbf{E}(X|\Sigma)| > K$ is Σ measurable, so

$$\int_{|\mathbf{E}(X|\Sigma)|>K} |\mathbf{E}(X|\Sigma)| \leqslant \int_{|\mathbf{E}(X|\Sigma)|>K} \mathbf{E}(|X||\Sigma) = \int_{|\mathbf{E}(X|\Sigma)|>K} |X|$$

Now since

$$\mathbf{E}\left[|\mathbf{E}(X|\Sigma)|\right] \leqslant \mathbf{E}\left[\mathbf{E}(|X||\Sigma)\right] = \mathbf{E}|X|$$

We can apply Markov's inequality to conclude

$$\mathbf{E}|X| \geqslant K\mathbf{P}(|\mathbf{E}(X|\Sigma)| > K)$$

So if we choose K big enough that $\mathbf{E}|X|/K < \delta$, then $\mathbf{P}(|\mathbf{E}(X|\Sigma)| > K) < \delta$ for all σ algebras Σ, and since we know that, because X is in $L^1(\Omega)$, for every ε, there is δ such that if $\mathbf{P}(E) < \delta$, then $\int_E |X| < \varepsilon$, and this shows

$$\int_{|\mathbf{E}(X|\Sigma)|>K} |\mathbf{E}(X|\Sigma)| < \varepsilon$$

Uniform integrability has been verified. □

Theorem 6.10. *If $X \in L^1$ and Σ is any filtration, then $\mathbf{E}[X|\Sigma_n] \to \mathbf{E}[X|\Sigma_\infty]$, where $\Sigma_\infty = \sigma(\bigcup \Sigma_n)$.*

Proof. We know $M_n = \mathbf{E}[X|\Sigma_n]$ is a uniformly integrable martingale, which implies M_n converges in L_1 to M_∞, which is Σ_∞ measurable. Since $\bigcup \Sigma_n$ is a π system generating Σ_∞, we verify that for each $E \in \Sigma_n$,

$$\int_E M_\infty = \lim_{m \to \infty} \int_E \mathbf{E}[X|\Sigma_m]$$

Now $\int_E \mathbf{E}[X|\Sigma_m] = \int_E X$ for $m \geqslant n$, so that we have verified M_∞ is a conditional expectation for X over a π system generating Σ_∞, and so $M_\infty = \mathbf{E}[X|\Sigma_\infty]$. □

This proof leads to a simple martingale proof of the 0-1 law, which is a foundational result in basic probability theory. The proof of the Kolmogorov theorem is essentially the same as the standard proof, except that some of the statements can be buried into facts about the convergence of martingales.

Corollary 6.11 (Kolmogorov's 0-1 Law). *Let $X_1, X_2, \dots$ be a sequence of independent random variables, and then let*

$$\Sigma_n = \sigma(X_{n+1}, \dots) \quad \Sigma_\infty = \lim \Sigma_n$$

Then for any $E \in \Sigma_\infty$, $\mathbf{P}(E) = 0$ or 1.

Proof. If $E \in \Sigma_\infty$, then $\mathbf{E}(\chi_E|\Sigma_n) = \chi_E$ is a uniformly integrable martingale. Furthermore, 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 the χ_E are obviously uniformly integrable and converge in probability, so the upward theorem guarantees

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

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 back in time. 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 6.12 (Lévy Doob Downward Theorem)**.** *Consider a supermartingale $M = \{M_0, M_{-1}, \dots\}$ with respect to a filtration $\Sigma = \{\Sigma_0 \supset \Sigma_{-1} \supset \dots\}$. 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 and in the L^1 norm, and $\mathbf{E}(M_n | \Sigma_{-\infty}) \leqslant M_{-\infty}$ for all n, where $\Sigma_{-\infty} = \bigcap \Sigma_n$. Similar results for backward time 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 6.13. *Let $X_1, X_2, \dots$ be i.i.d, integrable random variables of mean μ. If $S_n = X_1 + \dots + X_n$, then $n^{-1}S_n \to \mu$ almost surely and in $L^1(\Omega)$.*

Proof. Set $\Sigma_n = \sigma(S_n, S_{n+1}, \dots) = \sigma(S_n, X_{n+1}, X_{n+2}, \dots)$. For any $m \leqslant n$, $\mathbf{E}(X_m|\Sigma_n) = n^{-1}S_n$. 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 $\Delta_n = \sigma(X_n, \dots)$. By Kolmogorov's 0-1 law, we know $\sum \mathbf{P}(S \in [n, n+1)) = 1$, 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 x, so $S = x$ almost surely. But this means

$$x = \mathbf{E}(S) = \lim \mathbf{E}(S_m) = \mu$$

so $S = \mu$ almost surely. □

6.5 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 bounded in L^p are also pointwise bounded in a nice way.

Theorem 6.14 (Doob's Submartingale Inequality). *Let M be a submartingale, and define $M_n^* = \sup_{k \leqslant n}$. Then if $\lambda, n > 0$, then*

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

Proof. Let $E = \{M_n^* \geqslant \lambda\}$. Then we can write E as a disjoint union of sets $E_0, E_1, \dots, E_n$, where $E_i = \{M_0, \dots, M_{i-1} < \lambda, M_i \geqslant \lambda\}$. 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 6.15. *If M is a non-negative submartingale, then we can conclude*

$$\lambda \mathbf{P}\left(M_n^* \geqslant \lambda\right) \leqslant \mathbf{E}(M_n)$$

Notice that the submartingale inequality is independent of the step number n, which indicates we can obtain a uniform result for $M^* = M_\infty^*$ when the M_n are bounded in $L^1(\Omega)$. It is easy to obtain probabilistic results from the theorem above, but we will show that M^* actually has L^p bounds as well.

Lemma 6.16. *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. □

The most well know consequence of this result is Kolmogorov's inequality, which allows us to bound the deviation of the averages of zero-mean random variables.

Corollary 6.17. *If $X_1, X_2, \dots$ is a sequence of independent, zero-mean finite variance random variables, with $\sigma_k^2 = \mathbf{V}(X_k)$. If $S_n = X_1 + \dots + X_n$, and $V_n = \mathbf{V}(S_n) = \sigma_1^2 + \dots + \sigma_n^2$, then for $\lambda > 0$, if we consider the values $S_n^* = \max_{k \leqslant n} S_k$, then*

$$\lambda^2 \mathbf{P}(S_n^* \geqslant \lambda) \leqslant V_n$$

Proof. Apply the submartingale inequality to S^2. □

Now we move on to the general class of theorems again, this time establishing results about M^* in the L^p spaces.

Lemma 6.18. *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 6.19 (Doob's L_p inequality)**.** *Let $p > 1$, q it's conjugate, and M a non-negative submartingale bounded in $L^p(\Omega)$. Then $M^* \in L^p(\Omega)$, 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

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

The lemma we just proved 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 then apply dominated convergence to conclude that $M_\infty \in L^p(\Omega)$. If $M_n = |N_n|$, and 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 really restricted in motion for $p > 1$. This contrasts a random walk on the integers, which dances back and forth infinitely without settling down to a finite value.

Corollary 6.20 (Kakutani's Product Martingale Theorem). *Let $X_1, X_2, \dots$ be independent non-negative random variables of mean one, and consider the non-negative martingale $M_n = X_1 \dots X_n$. Then M is non-negative, so M_∞ exists pointwise almost surely, and the following are equivalent:*

(i) $\mathbf{E}(M_\infty) = 1$.

(ii) $M_n \to M_\infty$ *in* $L^1(\Omega)$.

(iii) M *is uniformly integrable.*

(iv) $\prod \mathbf{E}(X_n^{1/2}) > 0$.

(v) $\sum 1 - \mathbf{E}(X_n^{1/2}) < \infty$

If any one of these statements fails, then $M_\infty = 0$ almost surely.

Proof. Suppose (iv) holds, and set

$$N_n = \frac{(X_1 \dots X_n)^{1/2}}{\mathbf{E}(X_1^{1/2}) \dots \mathbf{E}(X_n^{1/2})}$$

Then N_n is a martingale, and

$$\mathbf{E}(N_n^2) = \frac{1}{\left[\mathbf{E}(X_1^{1/2}) \dots \mathbf{E}(X_n^{1/2})\right]^2} \leqslant \left(\frac{1}{\prod \mathbf{E}(X_n^{1/2})}\right)^2 < \infty$$

so N is bounded in $L^2(\Omega)$, and therefore N_n converges in L^2 to N_∞. Doob's L^2 inequality gives

$$\mathbf{E}\left(M^*\right) \leqslant \mathbf{E}\left((N^*)^2\right) \leqslant 2 \sup \mathbf{E}[N_n^2] < \infty$$

so M is dominated by an L^1 function, and is therefore uniformly integrable, hence properties (i), (ii), and (iii) hold. On the other hand, if $\prod \mathbf{E}(X_n^{1/2}) = 0$, then N is still a non-negative martingale, so $\lim N_n$ exists, but in order for this to be true we must have $M_n^{1/2} \to 0$ almost surely, hence $M_\infty = 0$ almost surely This means that (i), (ii), and (iii) cannot hold. The equivalence of (iv) and (v) follows from the simple bounds $1 - x \leqslant e^{-x}$ and $\log(1-x) \leqslant 1 - x$. □

Example. *This theorem has important consequences in the theory of statistical likelihood. Given two probability measures P and Q on the same sample space, if Q is absolutely continuous with respect to P, then we can use the Radon Nikodym theorem to find a non-negative random variable dQ/dP, which we call the likelihood ratio of Q given P. P is absolutely continuous with respect to Q if $dQ/dP > 0$ almost surely, and in this case we say P and Q are equivalent, and we actually find $dP/dQ = 1/(dQ/dP)$. We can then make sense of*

$$\int \sqrt{dP dQ} := \int \sqrt{dP/dQ}\, dQ = \int \sqrt{dQ/dP}\, dP$$

The **Hellinger distance** *$H(P,Q)$ between two equivalent distributions P and Q is defined such that*

$$H^2(P,Q) = 1 - \int \sqrt{dP dQ} = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2$$

which quantifies the similarity between the two distributions. We will use this concept to draw a link between Kakutani's theorem and the theory of the likelihood ratio in statistics.

Given a family of probability density functions $f_1, f_2, \dots$ and $g_1, g_2, \dots,$ we can consider the unique probability distributions P and Q on $\mathbf{R}^{\mathbf{N}}$ such that the random variables $X_i(x) = x(i)$ are continuous independent random variables with densities f_i (for P) and g_i (for Q). This is the canonical model to consider a sequence of independent random variables with two possible distributions.

If Q is absolutely continuous with respect to P, we can consider the Radon Nikodym derivative dQ/dP, and if

$$E = E_1 \times \dots \times E_n \times \mathbf{R}^{\mathbf{N}-\{1,\dots,n\}}$$

$$r_i = g_i/f_i \qquad Y_i = r_i(X_i)$$

then

$$\begin{aligned}
\int_E (dQ/dP) dP &= \int_E dQ \\
&= \int_{E_1 \times \dots \times E_n} dX_*(Q) \\
&= \int_{E_1 \times \dots \times E_n} g_1(x_1) \dots g_n(x_n)\, dx_1 \dots dx_n \\
&= \int_{E_1 \times \dots \times E_n} r_1(x_1) \dots r_n(x_n)\, f(x_1) \dots f(x_n)\, dx_1 \dots dx_n \\
&= \int_{E_1 \times \dots \times E_n} r_1(x_1) \dots r_n(x_n)\, dX_*(P) \\
&= \int_E r_1(X_1) \dots r_n(X_n) dP = \int_E Y_1 \dots Y_n
\end{aligned}$$

It follows that if $\Sigma_n = \sigma(X_1, \dots, X_n)$, then $\mathbf{E}[dQ/dP|\Sigma_n] = Y_1 \dots Y_n$. Each Y_i is independent with mean 1 with respect to P, so in particular, this means that $\mathbf{E}[dQ/dP|\Sigma_n]$ is a uniformly integrable product martingale with respect to P. Conversely, if the martingale $M_n = Y_1 \dots Y_n$ is uniformly integrable with respect to P, then M_∞ exists almost surely, and $\mathbf{E}(M_\infty|\Sigma_n) = M_n$ for all n, in which case the measures

$$E \mapsto Q(E) \qquad E \mapsto \int_E M_\infty dP$$

are equal on the π system of finite cylinders, and hence they agree everywhere, so M_∞ is a version of the Radon Nikodym derivative dQ/dP. In particular, this means Q is absolutely continuous with respect to P. Kakutani's theorem therefore implies that Q is absolutely continuous with respect to P if and only if

$$\prod \mathbf{E}(Y_n^{1/2}) = \prod \int \sqrt{f_n(x) g_n(x)}\, dx > 0$$

which is equivalent to saying

$$\sum \int \left(\sqrt{f_n(x)} - \sqrt{g_n(x)} \right)^2 dx = 2 \sum H^2(f_n, g_n) < \infty$$

Because this condition is clearly symmetric in P and Q, we can repeat the argument above to conclude that Q is absolutely continuous with respect to P if and only if P is absolutely continuous with respect to Q.

Suppose that we now take X_n to be identically distributed random variables with respect to P and Q, so $f_n = f$ and $g_n = g$ for all n. The Kakutani condition established above guarantees that Q is equivalent to P if and only if $f = g$ almost everywhere, and then we conclude that $P = Q$. Moreover, Kakutani's theorem implies that if $Q \neq P$, then $M_n \to 0$ almost surely with respect to P. This is the consistancy result for the likelihood ratio test in statistics. To recall, suppose we have a sequence of i.i.d variables $X_1, X_2, \dots$ with an unknown common distribution, and we have a hypothesis that the common distribution either corresponds to the density function f or the density function g. We can then calculate the statistic

$$M_n = \frac{g(X_1) \dots g(X_n)}{f(X_1) \dots f(X_n)} = Y_1 \dots Y_n$$

The results above show that if $P \neq Q$, then $M_n \to 0$ almost surely with respect to P, and $M_n \to \infty$ almost surely with respect to Q. It therefore follows that if we decide that P is the correct distribution if $M_n < \alpha$, and we decide that Q is the correct distribution if $M_n > \alpha$, for any particular $\alpha > 0$, which is known as the likelihood ratio test, then we will have proved the test's consistancy, because it is guaranteed correct for large enough n.

Doob's L^p theorem also leads to a novel proof of the Radon Nikodym theorem using Martingales, which indicates that Martingales can be used in mathematical analysis rather than just statistics. The idea is that the martingale proof allows us to take the right limits of the solution to the Radon-Nikodym theorem in discrete spaces, in which the theorem is obvious.

Theorem 6.21. *Suppose that (Ω, P) is a probability space with a separable σ algebra. If Q is a finite measure which is absolutely continuous with respect to another measure P, then there is $X \in L^1(\Omega, P)$ with*

$$\int_E dQ = \int_E X dP$$

for all measurable sets E.

Proof. Suppose that the σ algebra of Ω is generated by the sets $E_1, E_2, \dots$. Consider the finite algebra $\Sigma_n = \sigma(E_1, \dots, E_n)$, which form a filtration. Because Σ_n is finite, it is easy to construct a function X_n such that if $F \in \Sigma_n$, then

$$\int_F dQ = \int_F X dP$$

by just making the function constant over each atom, normalized so that the measures balance out when integrated. Because of how the finite algebras are split up when Σ_n increases to Σ_{n+1}, we conclude that X_n is a martingale with respect to Σ_n. Because the X_n are non-negative, we can apply the convergence theorem to conclude that $X_\infty = \lim X_n$ exists almost everywhere with respect to P. If we could verify that the X_n were uniformly integrable, then we would obtain that X_∞ is the L^1 limit of the X_n, so that the measures

$$E \mapsto \int_E X_\infty dP \qquad E \mapsto Q(E)$$

agree on the π system which is the union of the Σ_n, and hence everywhere, so we have found a Radon Nikodym derivative. By absolute continuity, it is easy to argue that for any $\varepsilon > 0$, there is $\delta > 0$ such that if $P(E) < \delta$, $Q(E) < \varepsilon$, and let $K > 0$ be such that $Q(\Omega) < \delta K$. Then

$$\mathbf{P}(X_n > K) \leqslant \frac{\mathbf{E}(X_n)}{K} = \frac{Q(\Omega)}{K} < \delta$$

and this gives uniform integrability. □

Using the Radon Nikodym result for separable σ algebras allows us to obtain the result for all σ finite measure spaces, but the theorem is tricky and unenlightening. The argument generalizes to σ finite measures on general measure spaces. Statistically, it means the argument says that we can 'learn' the Radon Nikodym derivative in a separable space by successively observing $E_1, E_2, \dots$, and adapting the conditional expectation of the derivative accordingly (the values X_n in our proof).

6.6 Martingales in $L^2(\Omega)$

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

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

expresses 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 expectation of the **quadratic variation** $[M]_n = \sum (M_n - M_{n-1})^2$ directly measures the average square sum of the M_n up to $\mathbf{E}[M_0]$.

Theorem 6.22. *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 $L^2(\Omega)$.

Example. *We know that the harmonic series $\sum 1/n$ diverges, whereas the alternating harmonic series $\sum (-1)^n/n$ converges. If we let $X_1, X_2, \dots$ denote independent fair $\{-1,1\}$ Bernoulli random variables, what is the probability that $\sum X_i/n$ converges? The finite sums $S_1, S_2, \dots$ certainly form a martingale, and we find that*

$$\mathbf{E}[S_i^2] = \sum \frac{\mathbf{V}(X_i)}{n^2} = \frac{1}{4} \sum \frac{1}{n^2} < \infty$$

so $S_n \to S_\infty$ almost surely, and $S_\infty \in L^2(\Omega)$ so the sum is finite almost surely as well. More generally, if we consider $a_i > 0$, then the corresponding sequence $\sum a_i X_i$ converges almost surely if $\sum a_i^2 < \infty$. On the other hand, if $\sum a_i^2 = \infty$, then the finite sums of $\sum a_i X_i$ oscillate infinitely, because the quadratic variation $[S_n] \to \infty$.

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

Theorem 6.23 (Doob Decomposition)**.** *If X_n is an adapted process consisting 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 are forced to set $A_1 = \mathbf{E}(X_1 | \Sigma_0)$, and then $M_1 = X_1 - A_1$. 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. 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. □

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, known as the **predictable quadratic variation**.

Theorem 6.24. *The predictable quadratic variation satisfies the following:*

- *M is bounded in $L^2(\Omega)$ if and only if $\mathbf{E}\langle M \rangle_\infty < \infty$.*
- $\lim M_n(\omega)$ *exists for almost all ω with $\langle M \rangle_\infty(\omega) < \infty$.*
- *If M has uniformly bounded increments, then $\langle M \rangle_\infty(\omega) < \infty$ for almost all ω for which* $\lim M_n(\omega)$ *exists.*

Proof. Since $\mathbf{E}[M_n^2] = \mathbf{E}\langle M \rangle_n$, the first property is obvious. To obtain the other two properties, we consider stopping times. For each $k \in \mathbf{N}$, let S_k be the first time that $\langle M \rangle_{S_k} > k$. It is easy to see that $\langle M \rangle^{S_k}$ is previsible, because $\langle M \rangle$ is. Since

$$(M^{S_k})^2 - \langle M \rangle^{S_k} = (M^2 - \langle M \rangle)^{S_k}$$

is a martingale, and $\langle M \rangle^{S_k}$ is an increasing, previsible process, we conclude $\langle M^{S_k} \rangle = \langle M \rangle^{S_k}$, and since the process $\langle M \rangle^{S_k}$ is bounded by k, we conclude M^{S_k} is bounded in the L^2 norm, and hence $\lim M^{S_k}$ exists almost everywhere pointwise and in L^2, and in particular, we conclude $\lim M_n$ exists on almost every point of the set $\{S_k = \infty\}$. But

$$\{\langle M \rangle_\infty < \infty\} = \bigcup \{S_k = \infty\}$$

and we obtain the second result. To obtain the third, suppose that

$$\mathbf{P}(\langle M \rangle_\infty = \infty, M^* < \infty) > 0$$

Then for some $\lambda > 0$, $\mathbf{P}(T_\lambda = \infty, \langle M \rangle_\infty = \infty) > 0$ where T_λ is the stopping time, which is the first time where $|M_{T_\lambda}| > \lambda$. Now by the martingale property,

$$\mathbf{E}(M^2_{T_\lambda \wedge n} - \langle M \rangle_{T_\lambda \wedge n}) = 0$$

and M^{T_λ} is bounded by $K + \lambda$, if the increments of M are bounded by K, so

$$\mathbf{E}\langle M \rangle_{T_\lambda \wedge n} \leqslant (K + \lambda)^2$$

But if $\langle M \rangle_m(\omega) \to \infty$ as $m \to \infty$, we conclude that $\langle M \rangle_{T_\lambda \wedge n) \to \infty}$ if $T_\lambda(\omega) = \infty$, and this occurs on a set of positive measure by assumption, hence $\mathbf{P}(T_\lambda = \infty) = 0$. If $\lim M_n(\omega)$ exists, then $M^*(\omega) < \infty$, so $\langle M \rangle_\infty < \infty$ almost surely. □

Recall that the quadratic variation measures how well bounded the square norm of the sequence of random variables in M is in much the same way that the predictable quadratic variation $\langle M \rangle_n$ does.

Theorem 6.25. *If M is a martingale in L^2 null at zero, then*

$$M^2 - [M] = C \bullet M$$

where $C_n = 2M_{n-1}$. We conclude $M^2 - [M]$ is a martingale (but $[M] \neq \langle M \rangle$ is not necessarily true, because $[M]$ is not a previsible process). If M is bounded in L^2, then $M^2 - [M]$ is uniformly integrable.

Proof. We calculate that

$$M_n^2 - [M]_n = M_n^2 - \sum_{k=1}^n (M_k - M_{k-1})^2 = 2 \sum_{k=1}^n M_k M_{k-1}$$

We apply Doobs' L^2 inequality to conclude that $M^* \in L^2$, and since

$$M^2 - [M] \leqslant (M^*)^2 + [M]_\infty \in L^1(\Omega)$$

Hence $M^2 - [M]$ is dominated by a random variable in $L^1(\Omega)$, and is therefore uniformly integrable. □

It is important to note that by construction,

$$\begin{aligned}\langle M\rangle_n &= \mathbf{E}[M_n^2 - N_n|\Sigma_{n-1}]\\ &= \mathbf{E}[M_n^2|\Sigma_{n-1}] - N_{n-1}\\ &= \mathbf{E}[M_n^2 - M_{n-1}^2|\Sigma_{n-1}] + M_{n-1}^2 - N_{n-1}\\ &= \mathbf{E}[M_n^2 - M_{n-1}^2|\Sigma_{n-1}] + \langle M\rangle_{n-1}\end{aligned}$$

Hence

$$\langle M\rangle_n - \langle M\rangle_{n-1} = \mathbf{E}[M_n^2 - M_{n-1}^2|\Sigma_{n-1}]$$

measures the conditional expected variation of the M_n. This gives the formula

$$\mathbf{E}[M_n^2 - M_{n-1}^2|\Sigma_{n-1}] = \mathbf{E}[(M_n - M_{n-1})^2|\Sigma_{n-1}] = \mathbf{E}[[M]_n - [M]_{n-1}|\Sigma_{n-1}]$$

giving a relationship between the two 'bracket processes'.

6.7 Optional Sampling Theorem

The optional sampling theorem generalizes the optional stopping theorem to cases where we do not 'immediately stop', but we instead take two stopping times S and T, take all information known about S when it stops, and see if the average value of X_T is different from $\mathbf{E}[X_S]$. If $S = n$ is deterministic, this is essentially the optional stopping theorem.

To formally see what the optional sampling theorem says, we have to define what the 'information known when S is stopped' is. It must of course be a σ algebra, and we define it to be the set

$$\Sigma_T = \bigcap_{n=0}^{\infty} \{E : \{T \leqslant n\} \cap E \in \Sigma_n\}$$

which is equivalent to saying a set E is in Σ_T if $E \cap \{T = n\} \in \Sigma_n$ for all n. Thus knowing Σ_T is equivalent to knowing the occurence of all events which are known at time n if T is stopped at time n.

Lemma 6.26. *Let S and T be stopping times.*

(i) If X is a Σ adapted process, then X_T is Σ_T measurable.

(ii) *If $S \leqslant T$ then $\Sigma_S \subset \Sigma_T$.*

(iii) *$\Sigma_{S \wedge T} = \Sigma_S \cap \Sigma_T$. Information determined at both times S and T is also determined at $S \wedge T$.*

(iv) *If $E \in \Sigma_{S \vee T}$, then $E \cap \{S \leqslant T\} \in \Sigma_T$.*

(v) *$\Sigma_{S \vee T} = \sigma(\Sigma_S, \Sigma_T)$. Thus information known at time $S \vee T$ is exactly the information known at time S combined with the information known at time T.*

Proof. For (i),

$$\{X_T \leqslant t, T = n\} = \{X_n \leqslant t, T = n\} \in \Sigma_n$$

To obtain (ii), we note that if $E \cap \{S \leqslant n\} \in \Sigma_n$ for all n, then

$$E \cap \{T \leqslant n\} = (E \cap \{S \leqslant n\}) \cap \{T \leqslant n\} \in \Sigma_n \cap \Sigma_n$$

(ii) implies $\Sigma_{S \wedge T} \subset \Sigma_S \cap \Sigma_T$, and if $E \cap \{S \leqslant n\}, E \cap \{T \leqslant n\} \in \Sigma_n$ for all n, then

$$E \cap \{S \wedge T \leqslant n\} = (E \cap \{S \leqslant n\}) \cap (E \cap \{T \leqslant n\}) \in \Sigma_n$$

so $E \in \Sigma_{S \wedge T}$. For (iv),

$$\begin{aligned} E \cap \{S \leqslant T\} \cap \{T \leqslant n\} &= E \cap \{S \vee T \leqslant n\} \cap \{S \leqslant T \leqslant n\} \\ &= E \cap \{S \vee T \leqslant n\} \cap \bigcup_{k=0}^{n} (\{S \leqslant k\} \cap \{T = k\}) \in \Sigma_n \end{aligned}$$

It is obvious that $\sigma(\Sigma_S, \Sigma_T) \subset \Sigma_{S \vee T}$, and (iv) gives that if $E \in \Sigma_{S \vee T}$, then

$$E = (E \cap \{S \leqslant T\}) \cup (E \cap \{S \geqslant T\}) \in \Sigma_T \cup \Sigma_S \subset \sigma(\Sigma_S, \Sigma_T)$$

and this gives the final result. □

Theorem 6.27 (Optional Sampling Theorem)**.** *Let $0 \leqslant S \leqslant T \leqslant \infty$ be stopping times.*

- *If M be a uniformly integrable martingale, then $\mathbf{E}[M_T | \Sigma_S] = M_S$, so M is of 'class (D)' in that the family $\{X_T : T \text{ is a stopping time}\}$ is uniformly integrable.*

- *If X is a uniformly integrable supermartingale, and consider the Doob decomposition $X = X_0 + M - A$. Then $\mathbf{E}(A_\infty) < \infty$, M is uniformly integrable, X is also of class (D), and*

$$\mathbf{E}[X_T|\Sigma_S] \leqslant X_S$$

for all $0 \leqslant S \leqslant T \leqslant \infty$. Essentially the same is true for submartingales.

Proof. For the first part, suppose that $0 \leqslant S \leqslant T \leqslant k$ for some finite k. Then M_T and M_S are in L^1, because they are dominated by $|M_1| + \cdots + |M_k|$. Let $E \in \Sigma_S$, and define the stochastic process

$$C_n(\omega) = \mathbf{I}(\omega \in E, S(\omega) < n \leqslant T(\omega))$$

Then C is a previsible process, because it is the indicator function

$$E \cap \{S < n\} \cap \{n \leqslant T\} = E \cap \{S \leqslant n-1\} \cap \{T \leqslant n-1\}^c \in \Sigma_{n-1}$$

It follows that

$$0 = \mathbf{E}[(C \bullet M)_k] = \int_E M_T - M_S$$

so the integral of M_T and M_S agree on all Σ_S measurable subsets, and this gives $\mathbf{E}[M_T|\Sigma_S] = M_S$. In general, given $S \leqslant T$, we have proved the theorem for $S \wedge k$ and $T \wedge k$ for each k, so applying the theorem for $T \wedge k \leqslant k$, we find that, since we also have L^1 convergence $M_k \to M_\infty$, and also $\Sigma_k \supset \Sigma_{T\wedge k}$, that

$$M_{T\wedge k} = \mathbf{E}[M_k|\Sigma_{T\wedge k}] = \mathbf{E}[\mathbf{E}[M_\infty|\Sigma_k]|\Sigma_{T\wedge k}] = \mathbf{E}[M_\infty|\Sigma_{T\wedge k}]$$

Since M is uniformly integrable, we obtain L^1 convergence of $M_{T\wedge k}$ to M_T, and Lévy's upward theorem gives

$$M_T = \lim_{k\to\infty} M_{T\wedge k} = \lim_{k\to\infty} \mathbf{E}[M_\infty|\Sigma_{T\wedge k}] = \mathbf{E}\left[M_\infty \middle| \lim_{k\to\infty} \Sigma_{T\wedge k}\right]$$

Note, however, that $\lim_{k\to\infty} \Sigma_{T\wedge k}$ (the finishing point of many falacious proofs). Of course, we have $\lim_{k\to\infty} \Sigma_{T\wedge k} \subset \Sigma_T$. Suppose that $E \in \Sigma_T$. Then we have $E \cap \{T \leqslant k\} \in \Sigma_{T\wedge k}$, hence

$$E \cap \{T < \infty\} = \lim_{k\to\infty} E \cap \{T \leqslant k\} \in \lim_{k\to\infty} \Sigma_{T\wedge k}$$

Thus

$$\int_{E\cap\{T<\infty\}} M_T = \int_{E\cap\{T<\infty\}} M_\infty$$

But of course, we know tautologically that

$$\int_{E\cap\{T=\infty\}} M_T = \int_{E\cap\{T=\infty\}} M_\infty$$

and this gives $\int_E M_T = \int_E M_\infty$ for all $E \in \Sigma_T$, hence $\mathbf{E}[M_\infty|\Sigma_T] = M_T$. But, swapping out S for T, if M is uniformly integrable, M^T is also uniformly integrable, and we have proven $\mathbf{E}[M_T|\Sigma_S] = \mathbf{E}[M^T_\infty|\Sigma_S] = M_S$. That M is of class (D) is now obvious, because

$$\{M_T : T \text{ is a stopping time}\} = \{\mathbf{E}[M_\infty|\Sigma_T] : T \text{ is a stopping time}\}$$

If $X = X_0 + M - A$ is a uniformly integrable supermartingale, then

$$\mathbf{E}(A_n) = \mathbf{E}(X_0) - \mathbf{E}(X_n)$$

is bounded in L^1, and therefore

$$\mathbf{E}(A_\infty) = \lim_{n\to\infty} \mathbf{E}(A_n) < \infty$$

Since X is uniformly integrable, it follow fairly easily that, since the expectation of the A_n are bounded, that M is uniformly integrable. But then this means that we can apply the first case of the optional stopping theorem to conclude that

$$\{M_T : T \text{ is a stopping time}\}$$

is uniformly integrable. But $\{A_T : T \text{ is a stopping time}\}$ is also uniformly integrable, because the family is uniformly bounded. But this implies

$$\{X_T = X_0 + M_T - A_T\}$$

is also a uniformly integrable family. It now follows fairly simply that

$$\begin{aligned}\mathbf{E}[X_T|\Sigma_S] &= X_0 + \mathbf{E}[M_T|\Sigma_S] - \mathbf{E}[A_T|\Sigma_S]\\ &\leqslant X_0 + M_S - A_S\end{aligned}$$

because $A_T \geqslant A_S$. □

If X is a uniformly integrable supermartingale, then X has a unique **Riesz decomposition** $X = Y + Z$, where Y is a martingale and Z is a *potential*, in the sense that Z is a non-negative uniformly integrable supermartingale with $Z_\infty = 0$ almost surely. Z is then of class (D), and of course, we can set $Z_n = \mathbf{E}(A_\infty - A_n | \Sigma_n)$. The proof is a fairly easy exercise.

Theorem 6.28. *If X is a non-negative supermartingale, and $S \leqslant T$ are stopping times, then $\mathbf{E}(X_T|\Sigma_S) \leqslant X_S$. For non-negative martingales, there is NO almost sure equality.*

Proof. We calculate that $\mathbf{E}(X_{T\wedge n}|\Sigma_{S\wedge n}) \leqslant X_{S\wedge n}$, as in the standard proof of the optional stopping theorem. Essentially the same proof again gives $\mathbf{E}(X_{T\wedge n}|\Sigma_S) \leqslant X_{S\wedge n}$, and then we can take Fatou's lemma. □

These results relate to a 'commutivity property' of the expectation operator.

Theorem 6.29. *If T is a stopping time, define the operator $\mathbf{E}_T$ from $L^1(\Omega, \Sigma)$ to $L^1(\Omega, \Sigma_T)$ by defining $\mathbf{E}_T(X) = \mathbf{E}(X|\Sigma_T)$. Then for any S and T,*

$$\mathbf{E}_S \mathbf{E}_T = \mathbf{E}_T \mathbf{E}_S = \mathbf{E}_{S\wedge T}$$

Proof. We repeatedly apply the optional sampling theorem for uniformly integrable martingales. Given $X \in L^1(\Omega)$, consider $X_n = \mathbf{E}(X|\Sigma_n)$. The optional sampling theorem gives $\mathbf{E}(X|\Sigma_T) = X_T$. Similarily, $X_{T\wedge n}$ is a family of uniformly integrable martingales tending to X_T, as is the family $\mathbf{E}(X_T|\Sigma_n)$, and so we know $\mathbf{E}(X_T|\Sigma_n) = X_{T\wedge n}$ by the optional sampling theorem. Finally, applying the optional sampling theorem to $X_{T\wedge n}$, now proven uniformly integrable, we conclude $\mathbf{E}(\mathbf{E}(X|\Sigma_T)|\Sigma_S) = X_{T\wedge S}$. □

Part II

Continuous Time Stochastic Processes

Chapter 7

Continuous Time Regularity

In physical phenomena and statistical experiments, one can only really obtain the finite dimensional distributions of a stochastic processes, which we call it's **law**. The Daniell Kolmogorov theorem provides a very satisfying conclusion to questions on whether all laws can be modelled by stochastic processes. In the case of discrete time, this is sufficient to describe almost all the useful information about a stochastic process. However, in continuous time, the uncountability of the time set causes certain paradoxes to occur. Given that we only really care about the finite dimensional distributions of a stochastic process, it makes technical sense to choose a stochastic process with the required law, but which is as mathematically pleasant as possible. This chapter is devoted to showing that for a large class of laws, we can construct stochastic processes with certain continuity properties. The next two examples indicate two examples where we require this regularity.

Example. *Let T be a parameter set, $m : T \to \mathbf{R}$ a function, 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 elementary theory of Gaussian distributions implies that for any finite sub-

set 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 X on the index set T with mean μ and covariance function V. If $T = [0,\infty)$, $\mu(t) = 0$ for all t, and $V(t,s) = t \wedge s$, then the Gaussian process constructed is almost a Brownian motion, in the sense that all the finite dimensional distributions of X satisfy the properties of Brownian motion, except that X may not have continuous sample paths. Since $C[0,\infty)$ is not a measurable subset of $\mathbf{R}^{[0,\infty)}$, we cannot even calculate $\mathbf{P}(X \in C[0,\infty))$ to conclude that the system is 'almost surely' continuous. Completion also won't help us here, because if $E \subset C[0,\infty)$ is measurable in $\mathbf{R}^{[0,\infty)}$, then E is only specified at finitely many points, so the only way it can avoid discontinuous functions is if $E = \varnothing$. Similarily, if $C[0,\infty) \subset E$, then E is only specified at countably many points, so $E = \mathbf{R}^{[0,\infty)}$, so $\mathbf{P}_*(C[0,\infty)) = 0$, and $\mathbf{P}^*(C[0,\infty)) = 1$. Thus we need a further theory to allow us to 'modify' X in such a way that all it's sample paths are continuous.

Example. *Suppose 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*

$$\Lambda : \Sigma \to (\mathbf{N} \cup \{\infty\})^\Omega = \mathbf{N}_\infty^\Omega$$

such that

- *If $E \in \Sigma$, then $\Lambda(E)$ is a $\mathbf{N}_\infty$ valued random variable which has a Poisson distribution with parameter $\lambda(E)$.*
- *If $E_1, \dots, E_n \in \Sigma$ are disjoint, then $\Lambda(E_1), \dots, \Lambda(E_n)$ 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 the first three properties specify the desired law of $\{\Lambda(E) : E \in S\}$ uniquely, and these laws are consistant, so that the Daniell Kolmogorov theorem guarantees the existence of a unique probability distribution $\mathbf{P}$ on $\mathbf{N}_\infty^\Sigma$ such that the function $\Lambda(n,E) = n$ satisfies the three properties above. But the Daniell Kolmogorov theorem doesn't say anything about the fourth property. It is difficult to correct this, because the subset of $\mathbf{N}_\infty^\Sigma$ consisting of all functions which form measures is not a measurable subset of $\mathbf{N}_\infty^\Sigma$ in all but the most trivial of cases.

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 outer probability of the set $A = \{\omega : X_t(\omega) = Y_t(\omega) \text{ for all } t\}$ is equal to 1, so we can descend to a subspace Ω_0 of Ω upon which $X_t = Y_t$. Indistinguishability preserves the properties of stochastic processes we need in the theory of continuous time.

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

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

Then for a fixed t, $\mathbf{P}(\tilde{B}_t = B_t) = \mathbf{P}(t \neq X) = 1$, 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 $\tilde{B}_{X(\omega)} \neq B_{X(\omega)}$, so

$$\mathbf{P}^*(\{\omega : \tilde{B}_t(\omega) = B_t(\omega) \textit{ for all } t\}) = \mathbf{P}(\varnothing) = 0$$

Note that $\tilde{B}$ is discontinuous, so modifications do not preserve the continuity of sample paths.

The problems of indistinguishability occur only 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 7.1. *There exists a process X into S^T with a given law $\mathbf{P}$ such that $X(\omega) \in S_0$ for all ω, for some subset $S_0 \subset S$, if and only if $\mathbf{P}^*(S_0^T) = 1$.*

Proof. We rely on the fact that if $\mathbf{P}^*(S_0^T) = 1$, then for any measurable set $E \subset S^T$, $\mathbf{P}^*(S_0^T \cap E) = \mathbf{P}(E)$, and the law $\mathbf{P}$ defined over S^T descends to a probability measure on S_0^T, with Σ algebra consisting of sets of the form $S_0^T \cap E$, where E is measurable in S^T, and where the law is defined by

$$\mathbf{P}(S_0^T \cap E) = \mathbf{P}^*(S_0^T \cap E) = \mathbf{P}(E)$$

This means that if $\mathbf{P}^*(S_0^T) = 1$, then we can make S_0^T into a measure space, and the projection function $\pi_t : S_0^T \to S$ given by $\pi_t(x) = x_t$ form a stochastic process with law $\mathbf{P}$, because

$$\mathbf{P}(\pi_t \in F) = \mathbf{P}^*(F \cap A^T) = \mathbf{P}(F)$$

On the other hand, if there is a process X with law $\mathbf{P}$ and $X(\omega) \in S_0$ for all ω, and $S_0^T \subset E$ for some measurable $E \subset S^T$, then

$$\mathbf{P}(E) = \mathbf{P}(X^{-1}(E)) = \mathbf{P}(\Omega) = 1$$

so $\mathbf{P}^*(S_0^T) = 1$. □

These is never really a feasible way to calculate $\mathbf{P}^*(S_0^T)$ for any set S independent of using this lemma, but the lemma tells us that the law $\mathbf{P}$ can tell us about the values that all modifications of a process can take. In particular, this means we can study all the possible modifications of a stochastic process by understanding it's law. Indeed, we shall find out that if the law of a stochastic process defines a martingale over $[0, \infty)$ then under very slight regularity conditions there is a **càdlàg** version of the law, that is, a process such that $t \mapsto X_t(\omega)$ is right continuous with left limits for all values ω. In this case, the modification/indistinguishibility problem disappears.

Theorem 7.2. *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. We calculate that

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

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. $\square$

If we restrict ourselves to stochastic processes X such that $t \mapsto X_t(\omega)$ is càdlàg for all ω, then X can be viewed as a measurable function into the family $D[0,\infty)$ of càdlàg functions, which has the σ algebra induced from $\mathbf{R}^{[0,\infty)}$. We shall find that $D[0,\infty)$ is a very useful space to analyze stochastic processes, because it contains most of the measurable sets we want to analyze. In particular, $C[0,\infty)$ is a measurable subset of $D[0,\infty)$, because

$$C[0,\infty) = \bigcap_{p \in \mathbf{Q}^+} \left\{ f \in D[0,\infty) : \lim_{q \to p} f(q) = f(p) \right\}$$

where we let q range only over rational numbers. This means that it makes sense to ask whether a càdlàg process is almost surely continuous, and indeed, we can redefine a Brownian motion to be a càdlàg process which is almost surely continuous.

7.1 Regularity of Martingales

The definition of a martingale, supermartingale, and submartingale generalizes incredibly easily to continuous time. A filtration Σ on $[0,\infty)$ is just a $[0,\infty)$ indexed increasing family of σ algebras, and a **continuous time martingale** on $[0,\infty)$ with respect to Σ is a Σ adapted process X such that for $s \leqslant t$, $\mathbf{E}[X_t|\Sigma_s] = X_s$. 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 **càdlàg** , so that in the forthcoming analysis of continous time martingales, we can always assume martingales are regularized in a nice fahsion (we can also consider left continuous functions with right limits, know as **cáglád**, and the regularity proofs are the same).

We start by discussing when 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

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

exists for all real values $t \geqslant 0$, and is finite, then 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

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

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

Lemma 7.3. *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. If $\sup\{|f(q)| : q \in \mathbf{Q}^+ \cap [0,n]\} = \infty$ for some n, then we can choose a monotonic family $q_1, q_2, \dots$ with $|f(q_k)| > k$ converging to some value, and then the required limits at this value. Similarly, if $U_n[a,b] = \infty$, then we can set $t^* = \inf\{t : U_t[a,b] = \infty\}$, and if $U_{t*}[a,b] = \infty$, then we can obtain an infinite oscillating family on the left converging to t^* breaking the limit properties of t^*, or $U_{t*}[a,b] < \infty$, and then we can find an infinite oscillating family on the right converging to t^*. On the other hand, if f does not have left or right limits at a point, and the values of $|f(q)|$ are bounded on the interval at this point, then we can find an infinitely oscillating family for some a, b, by setting a to be the lim inf of the values around this point, and b the lim sup around the point. □

Corollary 7.4. *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 $X(E) \subset \mathbf{R}^{[0,\infty)}$ 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 7.5. *Let X be a supermartingale, fix $t \in [0,\infty)$ and $q_1 > q_2 > \dots$ is a decreasing sequence of rationals which tend to t, then X_{q_i} converges pointwise almost everywhere and in the L^1 norm, to a function invariant of the particular sequence q_i.*

Proof. The sequence $X_{q_1}, X_{q_2}, \dots$ 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. Interlacing two decreasing sequences of rationals shows the convergence is invariant of the particular values chosen. □

The same results hold if X is a submartingale, but now we need to lower bound the infinum of $\mathbf{E}[X_{q_i}]$, and X_t can be used for this purpose.

Theorem 7.6 (Doob Regularity Theorem)**.** *If X is a supermartingale, then X is regularizable almost surely, and if we define*

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

Then Y is a càdlàg process.

Proof. Because of our discussion, 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 number applied to $Y|_{\mathbf{Q}^+}$. If $D_1, D_2, \dots$ 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

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

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

There are only a few things left to do before we can make the switch from studying X to studying Y. First, we must argue that Y is a modification of X. Second, we need to conclude that Y is a (sub/super) martingale with respect to the filtration of X, and this is where we run into irregularities.

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 $\omega \in \Omega$,*

$$X_t(\omega) = \begin{cases} \omega : t > 1 \\ 0 : t \leqslant 1 \end{cases}$$

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

$$Y_t(\omega) = \begin{cases} \omega : t \geqslant 1 \\ 0 : t < 1 \end{cases}$$

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

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 that $\Sigma_{t+} = \Sigma_t$ for all t, then the problem 'almost' doesn't occur. The only possibly non Σ_t-measurable set involved in the construction is then

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

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 7.7. *For any supermartingale X, 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 $s > q_1 > q_2 > \dots \to t$. 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 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 convergence of X_s to Y_t will be in the L^1 norm, and so

$$\mathbf{E}(Y_s|\Sigma'_t) = \lim_{s \downarrow t} \mathbf{E}(X_s|\Sigma'_t) \leqslant Y_t$$

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 to work. It is often useful to assume the *usual conditions*, which require that the total σ algebra over the probability space is complete, each Σ_t contains all $\mathbf{P}$ 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^*, setting $N(\Delta)$ to be the class of null sets in Δ^* then setting

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

Since Σ' differs from Σ^* only by $\mathbf{P}$ null sets, the independence properties of conditional expectation guarantee that the theorem above still holds if we swap Σ' with Σ^*. To summarize our discussion, the regularity theorem guarantees that we can take a supermartingale X with respect to a filtration Σ, and *regularize* it to a càdlàg supermartingale Y with respect to the filtration Σ^*.

Theorem 7.8. *If Ω with a filtration Σ satisfying the usual conditions, and X is a supermartingale, then X has a càdlàg 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 càdlàg supermartingale with respect to Σ.*

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 càdlàg modification then it is trivial to verify the expectation is right continuous. □

Lemma 7.9. *If X is a right continuous supermartingale with respect to a filtration Σ, then X is a supermartingale with respect to Σ^*.*

Proof. Suppose $0 \leqslant t < s$. We may assume that Σ has all the completion properties required of Σ^*, except for right continuity, by adding all the required null sets, which are independent of any conditional expectation. If $s \geqslant s_1 > s_2 > \dots \to t$, then the supermartingale property implies that $\mathbf{E}[X_s|\Sigma_{s_i}] \leqslant X_{s_i}$. The Doob downward theorem guarantees that the left hand side implies that

$$\mathbf{E}[X_s|\Sigma_{t+}] = \lim \mathbf{E}[X_s|\Sigma_{s_i}] \leqslant \lim X_{s_i} = X_t$$

and $\Sigma_{t+} = \Sigma^*$ if Σ contains all the null sets required. □

7.2 Regularization of Brownian Motion and the Markov Property

There are still a couple regularity problems we need to fix, that can be best displayed in an example. Consider the pre-Brownian motion X on

$\mathbf{R}^{[0,\infty)}$ constructed by the Daniell Kolmogorov theorem. With respect to the natural filtration

$$\Sigma_t = \sigma(Y_s : s \leqslant t)$$

Then X is a martingale with respect to Σ_t. The martingale is also right continuous in L^1, so we can construct a right continuous martingale Y with respect to the Σ^*. But now various questions are raised. With respect to Σ_t, X is a Markov process, in the sense that

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

Now that Σ^* can look infinitisimally into the future, is it still true that the Markov property holds? If Σ is already right continuous, this is fine, because then Σ^* only contains null sets which we can't gain any information from. But now we need a further analysis.

Theorem 7.10. *For a pre-Brownian motion X,*

- $\sigma(X_{t+s} - X_t : s \geqslant 0)$ *is independent of* Σ_{t+}.
- $\Sigma_{t+} \subset \sigma(\Sigma_t, N(\Sigma_\infty))$.

The first property says that looking infinitisimally ahead does not destroy independence, and that Σ_{t+} doesn't even look ahead of t, because every element of Σ_{t+} differs from a Σ_t set by a null set.

Proof. For any ε, $X_{t+s+\varepsilon} - X_{t+\varepsilon}$ is independent of $\Sigma_{t+\varepsilon/2}$. If $E \in \Sigma_{t+}$, and f is a bounded continuous function on $\mathbf{R}$, then we conclude

$$\begin{aligned}&\mathbf{E}[f(X_{t+s_1+\varepsilon} - X_{t+\varepsilon}, \dots, X_{t+s_n+\varepsilon} - X_{t+\varepsilon}); E]\\ &\quad = \mathbf{P}(E)\mathbf{E}[f(X_{t+s_1+\varepsilon} - X_{t+\varepsilon}, \dots, X_{t+s_n+\varepsilon} - X_{t+\varepsilon})]\end{aligned}$$

Letting $\varepsilon \to 0$, and applying the bounded convergence theorem, we conclude that

$$\mathbf{E}[f(X_{t+s_1} - X_t, \dots, X_{t+s_n} - X_t); E] = \mathbf{P}(E)\mathbf{E}[f(X_{t+s_1} - X_t, \dots, X_{t+s_n} - X_t)]$$

We can then apply the monotone class theorem to show that this equation holds for every Borel function f, whence in particular we conclude that the sigma algebra $\sigma(X_{t+s_1} - X_t, \dots, X_{t+s_n} - X_t)$ is independent of Σ_{t+}. But this means that Σ_{t+} is independent of every event E in the π system of events which are in $\sigma(X_{t+s_1} - X_t, \dots, X_{t+s_n} - X_t)$ for some finite choice of $s_1, \dots, s_n$, and this π system generates the entire σ algebra.

To prove the second property, we note that

$$\Sigma_\infty = \sigma(\Sigma_t, \sigma(X_{t+s} - X_t : s \geqslant 0))$$

and therefore Σ_∞ is generated by a π system of sets of the form $E_t \cap F_t$, where $E \in \Sigma_t$, and $F \in \sigma(X_{t+s} - X_t : s \geqslant 0)$. Let Y be bounded and Σ_{t+} measurable, and consider $Y - \mathbf{E}[Y|\Sigma_t]$. It suffices to show that $Y = \mathbf{E}[Y|\Sigma_t]$ almost surely, for then Y differs from a Σ_t measurable function on a Σ_∞ measurable nullset, so Y is $\sigma(\Sigma_t, N(\Sigma_\infty))$ measurable. Since Y is Σ_{t+} measurable, we know that $Y - \mathbf{E}[Y|\Sigma_t]$ is independent of $\sigma(X_{t+s} - X_t : s \geqslant 0)$, and so

$$\mathbf{E}[Y - \mathbf{E}[Y|\Sigma_t]; E_t \cap F_t] = \mathbf{P}(F_t)\mathbf{E}[Y - \mathbf{E}[Y|\Sigma_t]; E_t] = 0$$

This means that the integral of $Y - \mathbf{E}[Y|\Sigma_t]$ vanishes over every Σ_∞ measurable set, and since $Y - \mathbf{E}[Y|\Sigma_t]$ is Σ_∞ measurable, we conclude $Y = \mathbf{E}[Y|\Sigma_t]$ almost surely. □

The consequence of this is that if we regularize X to a càdlàg pre-Brownian motion Y, then $\sigma(Y_{t+u} - Y_t)$ is independent of Σ_t^*, and $\Sigma_t^* = \sigma(\Sigma_t, N(\Sigma))$.

7.3 Kolmogorov's Continuity Criterion

Chapter 8

Continuous Time Martingales

The class of martingales is most easy to extend to the case of continuous time. However, it is not without it's difficulties. First of all, we find that we need some kind of regularity on our martingales in order to obtain convergence results. In our case, the natural choice on our processes is *càdlàg* . A process X indexed over some interval is càdlàg if, for all sample points ω, the map $t \mapsto X_t(\omega)$ is right continuous, with left limits at all points (the term càdlàg stands for the french "continue à droite, limite à gauche"). Once this is assumed, almost all the main results of discrete martingale theory extend to continuous time. What's more, we will show that almost all martingales can be modified to a càdlàg process without changing the finite dimensional distributions of the process. Of course, rather than a discrete filtration $\Sigma = \{\Sigma_0 \subset \Sigma_1 \subset \dots\}$, we now consider filtrations Σ indexed over a continuous time interval, such that if $s \leqslant t$, $\Sigma_s \subset \Sigma_t$. A martingale M is then exactly a Σ adapted process such that $\mathbf{E}(X_t|\Sigma_s) = X_s$. Submartingales are supermartingales are defined in the exact same way.

Example. *Let $X_1,\dots,X_n$ be independant and identically distributed random variables with a common CDF F where β is the first point where $F(\beta) = 1$ (we can have $\beta = \infty$). Then we define the* **empirical distribution** *by*

$$\widehat{F}(t) = \frac{\#\{k : X_k \leqslant t\}}{n}$$

which is an estimate of F given that we observe the values X_i. Then $\widehat{F}$ is a *càdlàg process. If we define*

$$A_t = \sqrt{n}\left(\widehat{F}(t) - F(t)\right)$$

$$M_t = \frac{A_t}{1 - F(t)} \qquad B_t = A_t + \int_{-\infty}^{t} M_s \, dF(s) \qquad V_t = B_t^2 - \widehat{F}(t)$$

One calculates that M_t, B_t, and V_t are all càdlàg martingales on $(-\infty, \beta)$.

8.1 Continuous Time Convergence Theorems

To obtain the basic convergence results, we have need of a continuous variant of the upcrossing lemma which will guarantee we can obtain limits at all points. This is where the step-number invariant properties of the upcrossing lemma becomes integral to our argument. In order to extend the upcrossing lemma, we must restrict ourselves to a countable subset of time indices. In our proofs, we then use the continuity of our martingales to prove asymptotic results. If X is a stochastic process defined on $[0,t]$, define the upcrossing number $U_t[a,b] = U_t([a,b];X)$ to be the supremum over all n such that there exists *rational numbers*

$$0 \leqslant p_1 < q_1 < \dots < p_n < q_n \leqslant n$$

with $X_{p_i} \leqslant a$, and with $X_{q_i} \geqslant b$.

Lemma 8.1 (Continuous Time Upcrossing Lemma)**.** *If X is a supermartingale, and $t \in \mathbf{Q}$, then*

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

and if X is a submartingale,

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

Note that there are no continuity restrictions here since we are working over rational points of the process.

Proof. Let $D_1, D_2, \dots$ be an increasing family of finite subsets of $\mathbf{Q}^+ \cap [0,t]$ with $\lim D_i = \mathbf{Q}^+ \cap [0,t]$. For each finite set $D_i = \{t_1 < \dots < t_n < t\}$, we can apply the upcrossing lemma in discrete time to conclude that

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

We then just let $D_t \to \infty$, since $U_{D_t}[a,b] \to U_t[a,b]$ monotonely. □

In particular, if X is a sub or supermartingale bounded in L^1, then by monotone convergence we conclude

$$\mathbf{E}[U_\infty[a,b]] \leqslant \frac{\sup \mathbf{E}|X_t| + |a|}{b-a}$$

so in particular, $U_\infty[a,b]$ is finite almost surely. Letting a and b range over all elements of $\mathbf{Q}^+$ gives that almost surely, for all a, b, $U_\infty[a,b]$ is finite.

Theorem 8.2 (Doob's Convergence Theorem)**.** *Let X be a càdlàg sub or super martingale which is bounded in $L^1(\Omega)$. Then $X_\infty = \lim X_t$ exists almost surely, and can be chosen to be Σ_∞ measurable (take $X_\infty = \limsup X_t$, for instance).*

Proof. By right continuity, the almost sure condition makes sense, because it is equal to the limit over rational points tending to ∞. Since $t \mapsto X_t(\omega)$ is right continuous,

$$\liminf_{t\to\infty} X_t(\omega) = \liminf_{p\to\infty} X_p(\omega) \quad \limsup_{s\to\infty} X_t(\omega) = \limsup_{q\to\infty} X_q(\omega)$$

where p and q range over $\mathbf{Q}^+$. If we are able to choose $a, b \in \mathbf{Q}^+$ such that

$$\liminf_{p\to\infty} X_p(\omega) < a < b < \limsup_{q\to\infty} X_q(\omega)$$

then $U_\infty[a,b] = \infty$, implying that except on a null set, $\lim X_p(\omega)$ exists. □

Theorem 8.3 (Doob's Convergence Theorem)**.** *If X is a L^1 a uniformly integrable supermartingale then $X_t \to X_\infty$ in $L^1(\Omega)$, and $\mathbf{E}[X_\infty|\Sigma_t] \leqslant X_t$. If X is a submartingale, then $\mathbf{E}[X_\infty|\Sigma_t] \geqslant X_t$. If X is an L^1 bounded martingale, then $\mathbf{E}[X_\infty|\Sigma_t] = X_t$, and $X \to X_\infty$ in L^1 if and only if X is uniformly integrable (this is NOT true for sub and supermartingales).*

Proof. This theorem is proven exactly as it was for discrete martingales. The only new thing here is that if X is a martingale, then $X \to X_\infty$ in L^1 if and only if X is uniformly integrable, which follows because we then have $X_t = \mathbf{E}[X_\infty|\Sigma_t]$, and so the family $\{X_t\} = \{\mathbf{E}[X_\infty|\Sigma_t]\}$ is uniformly integrable. □

Example. *Recall the processes A_t, M_t, B_t, and V_t that emerge from the theory of empirical distributions. If $F(t) > 1 - 1/n$, then $A_t > 0$ only when*

$$\widehat{F(t)} = 1$$

so

$$\mathbf{E}(A_t^+) = (1-F(t))^{n+1} \qquad \mathbf{E}(A_t^-) = \mathbf{E}(A_t^+ - A_t) = (1-F(t))^{n+1}$$

Thus $\mathbf{E}|A_t| = 2(1-F(t))^{n+1}$, and so $\mathbf{E}|M_t| = 2(1-F(t))^n$ is bounded, so we conclude M_t converges almost surely as $t \to \beta$, so almost surely,

$$\widehat{F}(t) = F(t) + \Theta(1-F(t))$$

Now

$$\mathbf{E}[B_t^2] = \lim_{s \to -\infty} \mathbf{E}[\widehat{F}(t)] + \mathbf{E}[B_s^2 - \widehat{F}(s))] = \lim_{s \to -\infty} F(t) - F(s) + \mathbf{E}[B_s^2]$$

Because $F(s) \to 0$, and since $B_s^2 \to 0$ pointwise, such that for $F(s) \leqslant 1/2$,

$$|B_s^2| \leqslant \left(2\sqrt{n} + 2\sqrt{n}\frac{F(s)}{1-F(s)}\right)^2 = \frac{4n}{[1-F(s)]^2} \leqslant 8n$$

By the DCT, we conclude $\mathbf{E}[B_s^2] \to 0$, thus $\mathbf{E}[B_t^2] = F(t) \leqslant 1$, so B_t is an L^2 bounded càdlàg martingale, and therefore converges pointwise almost surely and in L^2 as $t \to \beta$. Since $\widehat{F}(t) \to 1$ almost surely as $t \to \beta$, and $F(t) \to 1$, we conclude that the function $s \mapsto M_s$ is in $L^1(F)$ almost surely, and

$$\mathbf{E}\left(\int_{-\infty}^{\beta} M_s dF(s)\right) = \lim_{s \to -\infty} \mathbf{E}[B_s] = 0$$

In the case where the X_i are chosen from a distribution on $[0,1]$, the A_t converge to a Brownian bridge on $[0,1]$, and B_t to a Brownian motion.

The next result again follows exactly from the discrete case, this time from applying the continuous upcrossing lemma instead of the discrete upcrossing lemma.

Theorem 8.4 (Doob's Downward Theorem)**.** *If X is a càdlàg supermartingale on $(0,\infty)$, and $\sup \mathbf{E}(X_t) < \infty$, then $X_0 = \lim_{t \downarrow 0} X_t$ exists almost surely and in L^1 and $\mathbf{E}(X_t|\Sigma_{0+}) \leqslant X_0$. If X is a submartingale and $\inf \mathbf{E}(X_t) > -\infty$ the same holds.*

This lemma implies the continuous version of Lévy's upward theorem, provided that we are using a filtration Σ such that the random variables $\mathbf{E}[X|\Sigma_t]$ can be chosen simultaneously to be right continuous. Later on, we shall find that this is possible provided that the Σ_t are right continuous, in the sense that for each t, $\Sigma_{t+} = \bigcap_{s>t} \Sigma_s = \Sigma_t$, and that each Σ_t contains all the null sets of Σ_∞. The downward convergence theorem then gives us the result for free.

Theorem 8.5 (Upward Theorem)**.** *If X is integrable, and if $\mathbf{E}[X|\Sigma_t]$ is a càdlàg choice of conditional expectations, then $\mathbf{E}[X|\Sigma_t] \to \mathbf{E}[X|\Sigma_\infty]$ almost surely and in L^1.*

Now we consider the standard L^p results in the continuous setting.

Theorem 8.6. *If X is a non-negative càdlàg submartingale, then*

$$\lambda \mathbf{P}\left(\sup_{s \leqslant t} X_s \geqslant \lambda\right) \leqslant \mathbf{E}\left[X_t ; \sup_{s \leqslant t} X_s \geqslant \lambda\right] \leqslant \mathbf{E}[X_t]$$

Proof. Let $D_1, D_2, \ldots$ be an increasing sequence of finite subsets of $[0,t]$ whose limit is dense. Then for each ω,

$$\sup_{0 \leqslant s \leqslant t} X_s(\omega) = \lim_{n \to \infty} \sup_{s \in D_n} X_s(\omega)$$

It follows that

$$\left\{\sup_{0 \leqslant s \leqslant t} X_s > \lambda\right\} = \lim_{n \to \infty} \left\{\sup_{s \in D_n} X_s > \lambda\right\}$$

(Note the importance of the strict inequality $>$ rather than $\geqslant$). We now just apply the discrete version of this result on D_n, and let $n \to \infty$. □

Theorem 8.7 (Doob's L^p Inequality)**.** *Let $p > 1$, $q = p^*$. If X is a non-negative càdlàg submartingale bounded in L^p, then $X^* = \sup X_t$ is in L^P, and*

$$\|X^*\|_p \leqslant q \sup \|X_t\|_p$$

The process X is therefore uniformly dominated by X^*. This means X_∞ exists almost surely, $X_t \to X_\infty$ in L^p, and

$$\|X_\infty\|_p = \sup \|X_t\|_p = \lim \|X_t\|_p$$

If $X = |M|$, where M is a càdlàg martingale bounded in L^p, then M_∞ exists almost surely in L^p and $\mathbf{E}(M_\infty|\Sigma_t) = M_t$.

Chapter 9

Stopping Times

We now consider the natural equivalent of a discrete stopping time. A $[0,\infty]$ valued random variable T is a **stopping time** with respect to a filtration Σ if $\{T \leqslant t\} \in \Sigma_t$ for each $t \leqslant \infty$. As should be expected by the presence of it's own chapter, stopping times in continuous time form a much deeper theory than discrete stopping times.

Example. *Let X_t be a Σ adapted process taking values in some measurable space S. If E is some measurable subset of S, then we define the* **début** *or* **first entrance time** $D_E = \inf\{t \geqslant 0 : X_t \in E\}$, *and the* **hitting time** $H_E = \inf\{t > 0 : X_t \in E\}$. *If the process was discrete, then D_E and H_E would always be stopping times, because*

$$\{D_E \leqslant n\} = \bigcup_{k=0}^{n} \{X_t \in E\} \quad \{H_E \leqslant n\} = \bigcup_{k=1}^{n} \{X_t \in E\}$$

Note, however, that this trick doesn't work in continuous time, because then we would have to take an uncountable intersection, which isn't necessarily measurable. This is the first example of the difficulty of continuous stopping times.

Note that if Σ is a filtration, we can consider the filtration Σ_{t^+}, which contains information infinitisimally ahead of the process. A Σ_{t^+} stopping time is thus *almost* determined less than or equal to t at time t, but requires an infinitisimal motion ahead of time to determine whether the stopping time is less than or equal to t exactly. In particular, T is a Σ_{t^+} stopping time if and only if, for every $t \leqslant \infty$,

$$\{T < t\} \in \Sigma_t$$

Because then

$$\{T \leqslant t\} = \lim_{\varepsilon \to 0} \{T < t + \varepsilon\} \in \lim \Sigma_{t+\varepsilon} = \Sigma_{t^+}$$

conversely, if T is a Σ_{t^+} stopping time, then

$$\{T < t\} = \lim_{\varepsilon \to 0} \{T \leqslant t - \varepsilon\} \in \lim \Sigma_{(t-\varepsilon)^+} \subset \Sigma_{t^-}$$

so equality and inequality are irrelevant here.

Lemma 9.1. *If E is a closed subset of a metric space, and X is a continuous Σ adapted process, then the first entrance time is a stopping time with respect to $\{\Sigma_t\}$, and the hitting time is a stopping time with respect to $\{\Sigma_{t+}\}$.*

Proof. Note by the continuity of X, we can write $D_E = \min\{t \geqslant 0 : X_t \in E\}$. Let d denote the distance function in the metric space. Then, for each ω, the map $(t,\omega) \mapsto d(X_t(\omega), E)$ is a function continuous in t for each fixed ω, and Σ_t measurable for each fixed t. By continuity, and the fact that E is closed, we conclude $D_E \leqslant t$ if and only if there is $u \in [0,t]$ with $X_u \in E$, hence we can find rational $p_1, p_2, \dots$ with $p_i \to u$, and thus $d(X_{p_i}, E) \to 0$. This means

$$\begin{aligned}\{D_E \leqslant t\} &= \{\inf\{d(X_p,E) : p \in \mathbf{Q}^+ \cap [0,t]\} = 0\} \\ &= \bigcap_{n=1}^{\infty} \bigcup_{p \in \mathbf{Q}^+ \cap [0,t]} \{d(X_p, E) \leqslant 1/n\} \in \Sigma_t\end{aligned}$$

Now $H_E \leqslant t$ for $t > 0$ if and only if for small enough $\varepsilon > 0$,

$$\inf\{d(X_p,E) : p \in \mathbf{Q}^+ \cap [\varepsilon, t]\} = 0$$

Thus

$$\{H_E \leqslant t\} = \liminf_{n \to \infty} \lim_{m \to \infty} \bigcup_{p \in \mathbf{Q}^+ \cap [1/n, t]} \{d(X_p, E) \leqslant 1/m\} \in \Sigma_t$$

and this gives H_E Σ_{t^+} measurability, because

$$\{H_E = 0\} = \lim_{t \to 0} \{H_E \leqslant t\} \in \Sigma_{0^+}$$

Note we shouldn't expect H_E to be Σ_t measurable, because if $X_0 \in E$, we need to determine the values of the process infintitisimally into the future to determine if the process moves away from E, or remains in E. □

Lemma 9.2. *If X is a right continuous Σ adapted process with values in a topological space S, then the first entrance time into an open set E is a Σ_{t^+} stopping time.*

Proof. Right continuity implies that

$$\{D_E < t\} = \bigcup_{p<t} \{X_p \in E\} \in \Sigma_{t^-}$$

so D_E is a Σ_{t^+} stopping time. Even if X is continuous, we cannot expect D_E to be a Σ_t stopping time, because normally X_{D_E} is only on the boundary of E, and we need to look into the future to determine whether X enters the set E, or darts away from the process. □

These are essentially the only interesting stopping times we can obtain on a continuous stochastic process without assuming extra completeness criteria on our filtration Σ. It is why in most continuous time scenarios we assume 'the usual conditions' on our filtration.

Given a set E, we define the **first approach time** $A_E(u)$ to be u if $X_u \in E$, or if $X_u \notin E$, and

$$A_E(u) = \inf\{t > u : X_t \text{ or } X_{t^-} \in E\}$$

This time will play a key role in the construction of stopping times to càdlàg processes.

Lemma 9.3. *If X is a càdlàg process taking values in a metric space, and E is a compact subset, then $A_E(u)$ is a Σ_t stopping time.*

Proof. Note that if $t < u$, then $\{A_E(u) \leqslant t\} = \varnothing$. Now $\{A_E(u) = u\}$ is true if $X_u \in E$, or if there is a sequence $u_1 > u_2 > \dots \to u$ with $X_{u_i} \in E$ for all i, or there is a sequence with $X_{u_i^-} \in E$. In both cases, we conclude that because E is closed, and X is right continuous, $X_u \in E$. Thus

$$\{A_E(u) = u\} = \{X_u \in E\} \in \Sigma_u$$

Now assume $t > u$. Then, by the last argument, $A_E(u) = t$ if $X_t \in E$ or $X_{t^-} \in E$, and so

$$\{A_E(u) \leqslant t\} = \{X_t \in E\} \cup \{\inf\{X_p : p \in \mathbf{Q}^+ \cap [u,t]\} = 0\}$$

and this set is easily verified to be in Σ_t. □

Lemma 9.4. *If $S_1 \leqslant S_2 \leqslant \cdots \to S$, where each S is a Σ_t stopping time, then S is a Σ_t stopping time. If $S_1 \geqslant S_2 \geqslant \cdots \to S$, then S is a Σ_{t+} stopping time.*

Proof. We calculate that in the monotone increasing case,

$$\{S \leqslant t\} = \bigcap_{n=1}^{\infty} \{S_n \leqslant t\} \in \Sigma_t$$

and in the monotone decreasing case,

$$\{S \leqslant t\} = \bigcap_{m=1}^{\infty} \limsup_{n \to \infty} \{S_n \leqslant t + 1/m\}$$

and these sets converge to an event in Σ_{t^+}. □

Theorem 9.5. *If X is a càdlàg process in a separable metric space E, and K is a compact subset of E, then D_K is a Σ_t stopping time provided each Σ_t contains all probability zero events, and the overlying probability space is complete.*

Proof. Consider the following sequence of stopping times. Set $S_1 = A_K(0)$. We now perform some ordinal arithmetic. For ordinals η, define

$$S_{\eta+1} = A_K(S_\eta) \quad S_{\lim \eta} = \lim S_\eta$$

These limits exist because S is verified to be increasing. If S_η is a stopping time, then essentially the same proof as for S_η shows $S_{\eta+1}$ is a stopping time. Since the S_η are monotone increasing, this also means that $S_{\lim \eta}$ is a Σ_t stopping time, provided that $\lim \eta$ is a *countable ordinal*. Now the first approach time is *almost* the Debút time, except that it can be tricked by processes that appear to enter E, then jump away at the last second. However, our discussion of the first approach time tells us that if $S_\eta = S_{\eta+1}$, then $X_{S_\eta} \in E$, and so $D_K = S_\delta$, where δ is the first ordinal with $S_\delta = S_{\delta+1}$. We know that δ is always a countable ordinal, because the S_η can only have countably many jumps. One calculates that for any (countable or uncountable) ordinal η,

$$S_{\eta+1} = \sum_{\nu \leqslant \eta} (S_{\nu+1} - S_\nu)$$

and since $S_{\eta+1}$ is finite, this implies only countably many of the $S_{\nu+1} - S_\nu$ are non-zero, and if we take the limit of this countable set of ordinals, the limit will be a countable ordinal S_ν with $S_{\nu+1} = S_\nu$. The problem is that the set of countable ordinals is uncountable, so even though

$$S = \lim_{\nu \text{ countable}} S_\nu$$

we cannot conclude S is a Σ_t measurable, because the limit is on an uncountable set. Define the sequence

$$c_\eta = \mathbf{E}e^{-S_\eta} \qquad c = \inf c_\eta$$

We can then find a sequence $\eta_1 \leqslant \eta_2 \leqslant \dots$ with $c_{\eta_k} \leqslant c + 1/k$. It follows that, because the sequence c_η is decreasing, if $\nu = \lim \eta_k$, then $c_\nu = c$. Since the limit of a sequence of countable ordinals is countable, ν is countable, and since $c_\nu = c$, $S_\nu = \sup_\eta S_\eta = S$ almost surely. But this means that, provided Σ_t contains all probability zero events in the probability space, that

$$\{S \leqslant t\} = \{S_\nu \leqslant t\} \cap \{S = S_\eta\} \cup E$$

where E is a subset of $\{S \neq S_\eta\}$, which occurs with probability zero. By completeness of the probability space, we conclude E is an event of probability zero, and hence in Σ_t. □

This is why in most of the theory of continuous time stochastic processes, we need to assume that our filtrations are complete in the sense of the proof. In order to assume hitting times are also stopping times, we also need to assume our filtrations are right continuous. These are called the **usual conditions** on a filtration. They also appear in the regularity theory of martingales, which makes them even more integral to the theory.

We would hope that, if Σ_t is a filtration, the filtration Π_t which is the smallest such filtration which satisfies the usual conditions, only superficially gives more information to the filtration. Dynkin's theorem shows this is true, except for the presense of right continuity, which enables us to look infinitisimally into the future.

Theorem 9.6. *If T is a Π_t stopping time, then there exists a Σ_{t^+} stopping time S with $T = S$ almost surely. Furthermore, Π_S is the smallest σ algebra containing Σ_{S+} and all null events in the completion of the probability space.*

9.1 The Debút Theorem

9.2 The Pre Stopping Time σ algebra

Given T, we define the pre σ algebra Σ_T as the set of events E such that $E \cap \{T \leqslant t\} \in \Sigma_t$ for each t.

Lemma 9.7. *The following hold for all stopping times S and T,*

- *If $S \leqslant T$, $\Sigma_S \subset \Sigma_T$.*
- *$\Sigma_{S \wedge T} = \Sigma_S \cap \Sigma_T$.*
- *If $E \in \Sigma_{S \vee T}$, $E \cap \{S \leqslant T\} \in \Sigma_T$.*
- *$\Sigma_{S \vee T} = \sigma(\Sigma_S, \Sigma_T)$.*
- *If T is a Σ_{t^+} stopping time, and if we define $\Sigma_{T^+} = (\Sigma_+)_T$, then $E \in \Sigma_{T^+}$ if and only if $E \cap \{T < t\} \in \Sigma_t$.*

Proof. The proof of the first two properties is exactly the same as for discrete martingales, as is the fourth once the third is proved, and this is proven by a slight modification of the discrete argument. If $E \in \Sigma_{S \vee T}$, then

$$\begin{aligned} E \cap \{S \leqslant T\} \cap \{T \leqslant n\} &= E \cap \{S \vee T \leqslant n\} \cap \{S \leqslant T \leqslant n\} \\ &= E \cap \{S \vee T \leqslant n\} \cap \left[\{S = T \leqslant n\} \cup \bigcup_{\substack{q \leqslant n \\ q \in \mathbf{Q}^+}} (\{S \leqslant q\} \cap \{q \leqslant T \leqslant n\}) \right] \end{aligned}$$

and this last set is easily seen to be in Σ_n. The last point follows from the same reasoning we had about Σ_{t^+} stopping times. □

Chapter 10

Markov Processes

We can use the idea of a filtration to introduce the concept of a continuous time Markov process, which should be a process whose future depends only on the exact present state. A process X with time index T valued on some state space S (a measure space) with respect to a filtration Σ and a family probability distributions $\mathbf{P}^x$ on S^T for each state $x \in S$, is **Markov**, if for any bounded, measurable function $f : S \to \mathbf{R}$, and $t, s \in T$,

$$\mathbf{E}^x(f(X_{t+s})|\Sigma_t) = (P_s f)(X_t)$$

where $\mathbf{E}^x$ is the expectation taken with respect to $\mathbf{P}^x$, and $\{P_t\}$ is a **transition semigroup** on S, such that

- For each $x \in S$, $P_t^x = P_t(x, \cdot)$ is a measure on S with $P_t(x, S) \leqslant 1$.
- For each measurable $E \subset S$, the map $x \mapsto P_t(x, E)$ is measurable.
- The integral equation

$$P_{t+s}(x, E) = \int_E \frac{dP_t^x}{d\mathbf{P}}(y) P_s^y(y)\, dy$$

Alternatively, recalling that the dual of the space $B(S)$ of bounded, measurable functions on S is the space of bounded, finitely additive measures on S, we can define a transition semigroup to be a family of positive bounded operators on $B(S)$ with operator norm $\leqslant 1$, and the Chapman Kolmogorov equation can then be reexpressed as saying $P_t \circ P_s = P_{t+s}$.

10.1 Poisson Processes

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.

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

$$\mathbf{P}(X_t = 1) = \lambda \Delta t + o(t) \tag{10.1}$$

$$\mathbf{P}(X_t = 0) = 1 - \lambda \Delta t + o(t) \tag{10.2}$$

$$\mathbf{P}(X_t > 1) = o(t) \tag{10.3}$$

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.

10.2 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

$$P_{u,v}(x,z) = \sum_t P_{u,t}(x,y) P_(t,v)(y,z) \tag{10.4}$$

We shall also assume a continuity requirement that

$$\lim_{j \to i^+} \mathbf{P}(X_j = x | X_i = y) = \delta_{x,y} \tag{10.5}$$

A process is **time-homogenous** if

$$P_{u,t}(x,y) = P_{t-u,0}(x,y) \tag{10.6}$$

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 (10.4) and (10.6) 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 $x_1, \dots, x_n$. Define a matrix P by $P_{i,j} = \alpha(x_i, x_j)$ for $i \neq j$, and $A_{i,i} = -\alpha(x_i)$. We call A the 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 $\beta = -\tilde{A}^{-1}1$.

10.3 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,

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

Hence

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

And thus the chain is transient if and only if

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

Chapter 11

Brownian Motion

Brownian motion is one of fundamental continuous stochastic processes, modeling random continuous motion. It owes it's name to a British botanist who observed the motion in 1827 while studying the motion of pollens in a liquid. The hitting of a grain of pollen by smaller molecules in the liquid leads to the pollen undergoing sporadic motion over time. In 1905, The mathematical properties of this motion were introduced by Einstein, who used Brownian motion to prove the existence of atoms. The main properties of a Brownian motion are that

- The motion of the particle has independent increments.
- The motion is continuous.
- The motion is time homogenous, in the sense that the distribution of the change in position on an interval depends only on the length of the interval.

These are simple assumptions for random motion to satisfy. Surprisingly, they uniquely specify the qualitative properties of the motion. If we let $B : \Omega \times [0, \infty) \to \mathbf{R}$ be any such motion, then it follows from an argument from the central limit theorem that there exists a vector μ, and a positive semidefinite matrix Σ such that for all $s, t \geqslant 0$, $B_{t+s} - B_t$ is normally distributed with mean $s\mu$, and with covariance matrix $s\Sigma\Sigma^T$. We say B is a Brownian motion with *drift* μ and *diffusion marix* Σ. If $\tilde{B}$ is a Brownian motion with no drift and the identity as the diffusion matrix, then one can verify that $\mu t + \Sigma\tilde{B}$ is a Brownian motion with respect to μ and Σ, and so to understand Brownian motion, it suffices to understand this basic family of motions, and we call such a process a **standard Brownian motion**. More formally, we say B is a (standard) Brownian motion if for *almost all* ω, $t \mapsto B(\omega, t)$ is continuous, $B_{t+s} - B_t$ is independent of $\sigma(B_u : u \leqslant s)$ for all $t, s \geqslant 0$, and is $N(0, s)$ distributed.

11.1 Construction of Brownian Motion

Though Einstein used his physical intuition to determine the mathematical properties of Brownian motion, it was Norbert Wiener that put Brownian motion on a rigorous footing, and gave the first construction giving the existence of Brownian motion. Indeed, just because we describe the properties that Brownian motion *should* have does not tells us that a stochastic process should exist with these properties. Consider the example

Example. *Suppose we wish to model white noise, which is a continuous, random, normally distributed error. This is of importance in signal processing, for instance. Such a process should have continuous sample paths, such that the family of values at each point are independent and $N(0,1)$ distributed. Unfortunately, no such process exists. Let $X : \Omega \to \mathbf{R}^{[0,\infty)}$ be any mean zero Gaussian process with covariance function $\Gamma(t,s) = \mathbf{I}(t = s)$. It X was $\Omega \times [0,\infty)$ measurable, with respect to the Lebesgue σ algebra on $[0,\infty)$, then we find*

$$\int_0^t \mathbf{E}|X_s^2| \; ds = t$$

so we would conclude $X \in L^2(\Omega \times [0,t]) \subset L^1(\Omega \times [0,t])$, hence by Fubini's theorem the random variable

$$Y_t = \int_0^t X_s \; ds$$

is well defined and finite almost everywhere on Ω, and by Fubini's theorem,

$$\begin{aligned} \mathbf{E}[Y_t^2] &= \mathbf{E}\left[\left(\int_0^t X_s \; ds \right)^2 \right] = \mathbf{E}\left[\int_0^t \int_0^t X_s X_u \; ds \; du \right] \\ &= \int_0^t \int_0^t \mathbf{E}[X_s X_u] \; ds \; du = \int_0^t \int_0^t \mathbf{E}[X_s]\mathbf{E}[X_u] = 0 \end{aligned}$$

which implies $Y_t(\omega) = 0$ for almost every $\omega \in \Omega$. But this means that for almost every $\omega \in \Omega$, $Y_t(\omega) = 0$ for all $t \in \mathbf{Q}$, and for any such ω, we conclude that $X_s(\omega) = 0$ for almost every s (because $X(\omega)$ integrates to zero over a basis of intervals), which contradicts the fact that $\mathbf{E}[X_t^2] = 1$ for every t. This means that X cannot even be jointly measurable in time and sample space, and therefore it definitely cannot be a continuous process. Thus white noise cannot exist.

It is a delicate discussion to determine whether a process exists that models a particular physical phenomenon. The Daniell-Kolmogorov theorem guarantees the existence of processes with certain finite dimensional distributions, but does not guarantee that these processes have continuous sample paths. The construction of continuous Brownian motion we give was introduced by Ciesielski, but is based on the basic idea of Wiener to represent Brownian motion as a random Fourier series.

Theorem 11.1 (Wiener)**.** *Brownian motion exists.*

Proof. Our discussion above indicates that, in order to construct any Brownian motion, it suffices to construct a standard Brownian motion. If we can construct a Brownian motion in one dimension, we can construct it in multiple dimensions by taking independent one dimensional Brownian motions as coordinates. Furthermore, to construct Brownian motion on $[0,\infty)$, it suffices to construct it on $[0,1]$, because we can then obtain the general motion by patching together independent Brownian motions on each interval. If we let $D_n = \{k2^{-n} : 0 \leqslant k \leqslant 2^n\}$ denote the *nth order dyadic numbers* in $[0,1]$, then we will iteratively define B_t *exactly* for all $t \in D_n$, then take $n \to \infty$ and interpolate betwen dyadic numbers to obtain the values of the process at all points.

We begin by taking a probability space Ω upon which countably many independent standard normal random variables Z_p, where we let p range over elements of D. We start our construction by defining $B_0 = 0$, and $B_1 = Z_1$. We will assume B is defined over D_{n-1}, and then define B on D_n. At all iterations of the construction, the following invariants should hold:

- For $s < t \in D_n$, $B_t - B_s$ is normally distributed with mean zero and variance $t - s$, independent of the family $\{B_s - B_r : r < s \in D_n\}$.

- The families $\{B_t : t \in D_n\}$ and $\{Z_t : t \in D - D_n\}$ are independent.

This is clearly true for our construction of B_t on D_0. For $t \in D_n - D_{n-1}$, we interpolate between values to the left and right of t, writing

$$B_t = \frac{B_{t-1/2^n} + B_{t+1/2^n}}{2} + 2^{-(n+1)/2} Z_t$$

Then $\{B_t : t \in D_n\}$ is independent of $\{Z_t : t \in D - D_n\}$ because

$$\sigma(B_t : t \in D_n) \subset \sigma(\sigma(B_t : t \in D_{n-1}), \sigma(Z_t : t \in D_n - D_{n-1}))$$

and the involved families are by assumption all independent of one another. We can write

$$B_t = \frac{B_{t+1/2^n} - B_{t-1/2^n}}{2} + B_{t-1/2^n} + 2^{-(n+1)/2} Z_t$$

each of these random variables is normally distributed and independent of each another, so B_t is normally distributed with mean zero and variance $1/2^{n+1} + (t - 1/2^n) + 1/2^{n+1} = t$. Finally, we need to verify that

$$\{B_t - B_{t-1/2^n}, B_{t+1/2^n} - B_t : t \in D_n - D_{n-1}\}$$

is a family of independent random variables, from which, by taking sums, it will follow that $\{B_t - B_s : t, s \in D_n\}$ is an independent family. Since they're normally distributed, it suffices to verify pairwise independence. Note that

$$B_{t+1/2^n} - B_t = \frac{B_{t+1/2^n} - B_{t-1/2^n}}{2} - 2^{-(n+1)/2} Z_t$$

if $t \in D_n - D_{n-1}$, and

$$B_t - B_{t-1/2^n} = \frac{B_{t+1/2^n} - B_{t-1/2^n}}{2} + 2^{-(n+1)/2} Z_t$$

We know that

$$\begin{pmatrix} B_{t+1/2^n} - B_{t-1/2^n} \\ Z_t \end{pmatrix} \sim N\left(0, \begin{pmatrix} 2^{1-n} & 0 \\ 0 & 1 \end{pmatrix}\right)$$

and since

$$\begin{pmatrix} B_{t+1/2^n} - B_t \\ B_t - B_{t-1/2^n} \end{pmatrix} = \begin{pmatrix} 1/2 & -2^{-(n+1)/2} \\ 1/2 & 2^{-(n+1)/2} \end{pmatrix} \begin{pmatrix} B_{t+1/2^n} - B_{1-1/2^n} \\ Z_t \end{pmatrix}$$

which is now seen to be normally distributed with mean zero and covariance matrix

$$\begin{pmatrix} 1/2 & -2^{-(n+1)/2} \\ 1/2 & 2^{-(n+1)/2} \end{pmatrix} \begin{pmatrix} 2^{1-n} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ -2^{-(n+1)/2} & 2^{-(n+1)/2} \end{pmatrix} = \begin{pmatrix} 1/2^n & 0 \\ 0 & 1/2^n \end{pmatrix}$$

so the two variables are independent of one another. Verifying the other intervals are pairwise independent is easy, because they are each defined in terms of two families of random variables which are independent of one another. This means our new iteration satisfies the required properties, and so we have constructed Brownian motion on the Dyadic numbers.

Since the Dyadic numbers are dense in $[0,1]$, for each ω, there is at most one way to define $B_t(\omega)$ for all $t \in [0,1]$ continuously. We show that this is possible for almost all ω. If this is possible, then we can use our knowledge of the distributions of Brownian motion over the dyadics to verify the required properties of the motion over all real numbers. Indeed, if $s < t$, then by continuity, $B_t - B_s = \lim B_{t_i} - B_{s_i}$, where $t_i \uparrow t$ and $s_i \downarrow s$ are dyadic approximations of t and s. But if $u \in D$ is greater than all the t_i and s_i, then by the independence properties, we already know

$$B_{t_i} = \mathbf{E}[B_u | B_s : s \leqslant t_i] \quad B_{s_i} = \mathbf{E}[B_u | B_s : s \leqslant s_i]$$

and so the $B_{t_i} - B_{s_i}$ are members of a uniformly integrable family, hence $B_{t_i} - B_{s_i}$ converges to $B_t - B_s$ in L^1, and so we conclude that $B_t - B_s$ is $N(0, t-s)$ distributed. Independence of intervals follows from similar considerations.

All that remains is to show that we can almost surely extend our definition of Brownian motion to $[0,1]$ continuously. Let f_n be the continuous function defined by

$$f_n(t) = \begin{cases} 2^{-(n+1)/2} Z_t & t \in D_n - D_{n-1} \\ 0 & t \in D_{n-1} \end{cases}$$

and then linearly interpolating between these values ot define f_n on $[0,1]$. Then for $t \in D_n$, we know

$$B(t,\omega) = \sum_{k=0}^{n} f_n(t,\omega) = \sum_{k=0}^{\infty} f_n(t,\omega)$$

we now use this formula to define, for every $t \in [0,1]$, the value of $B(t,\omega)$. Since each f_n is continuous, f_n is $[0,\infty) \times \Omega$ measurable, and we conclude that f is $[0,\infty) \times \Omega$ measurable. We now show that almost always, the f_n sum uniformly to f, and therefore f is almost surely continuous. First, we use the elementary normal distribution bound

$$\mathbf{P}(Z_t \geqslant C\sqrt{n}) \leqslant \exp(-nC^2/2)$$

to determine that

$$\sum_{n=0}^{\infty} \sum_{t \in D_n} \mathbf{P}(Z_t \geqslant C\sqrt{n}) \leqslant \sum_{n=0}^{\infty} 2^n \exp(-nC^2/2)$$

This series converges as soon as $C > \sqrt{2\log 2}$, so the Borel-Cantelli lemma implies that for almost surely, $Z_t < C\sqrt{n}$ except for finitely many t. But if this ineuqality holds for all $t \in D - D_{n-1}$, then

$$\|f_n\|_\infty < C\sqrt{n}2^{-n/2}$$

so that for almost surely, f_n are absolutely summable in the $\|\cdot\|_\infty$ norm to some function f, which must be continuous. □

Remark. Let $B : [0,1] \to \Omega \to \mathbf{R}$ be any Brownian motion. Define f_0 on $[0,1]$ by setting $f_0(0) = B_0$ and $f_0(1) = B_1$, and then applying linear interpolation. We then inductively define f_n by linearly interpolating the values

$$f_n(t) = B_t - \sum_{k<n} f_k(t)$$

for $t \in D_n$. Then one can verify that $f_n(t) = 0$ for $t \in D_{n-1}$, and $f_n(t)$ is normally distributed with mean zero and variance $2^{-(n+1)}$ for $t \in D_n - D_{n-1}$. For $t \in D$, we therefore know that

$$B_t = \sum_{k=0}^{\infty} f_k(t)$$

arguing similarly to Ciesielski's construction of Brownian motion, we can argue that for the f_k are absolutely summable almost surely, and therefore almost surely, for all t, $B_t = \sum_{k=0}^{\infty} f_k(t)$. Thus any Brownian motion can be thought of as a random Fourier series, ala Ciesielski.

There is another way to construct Brownian motion, which relies on the regularity theory of continuous time martingales, and the construction of Gaussian processes on any index set. Note that any standard Brownian motion B is a Gaussian process, with mean zero, and, $t \geqslant s$, covariance

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

Thus for any t and s, $\text{Cov}(B_t, B_s) = t \wedge s$. Using the general theory of Gaussian processes, we can construct a Gaussian process X with $\mathbf{E}[X_t]$ and $\mathbf{E}[X_t X_s] = t \wedge s$. In particular, this means that if $s \leqslant s' \leqslant t \leqslant t'$,

$$\mathbf{E}[(X_{t'} - X_t)(X_{s'} - X_s)] = s' - s' - s + s = 0$$

so disjoint intervals have covariance zero, and so by properties of Gaussian distributions, for any disjoint intervals $t_1 \leqslant s_1 \leqslant \dots \leqslant t_n \leqslant s_n$, the random variables $X_{s_i} - X_{t_i}$ are independent of one another. In particular, this means that $X_s - X_t$ is independent of $\sigma(X_u : u \leqslant t)$. Thus X satisfies all the properties of Brownian motion, except that X may not be continuous. This requires an application of Kolmogorov's continuity theorem, which guarantees that a random process X has a continuous modification if there is $\alpha, \beta > 0$ such that

$$\mathbf{E}|X_{t+s} - X_t|^\alpha \lesssim s^{1+\beta}$$

In our situation, we calculate

$$\mathbf{E}|B_{t+s} - B_t|^4 = 3s^2$$

this means that we can find a version B of X where all paths are continuous. This completes the abstract construction of Brownian motion.

11.2 The Wiener Measure

The measurable space $\mathbf{R}^{[0,\infty)}$, which is defined to be the weakest topology such that the projection functions $f_t(X) = X_t$ are measurable, does not contain enough measurable sets to do probability theory on continuous time stochastic processes. Indeed, we were forced to define Brownian motion to have continuous paths *for all* ω rather than just for almost all ω because the set $C[0,\infty)$ is not a measurable subset of $\mathbf{R}^{[0,\infty)}$ (measurable subsets of $\mathbf{R}^{[0,\infty)}$ depend on only countable many projections, whereas $C[0,\infty)$

depends on the properties of the function on every point). However, if B is a Brownian motion, then because it is continuous at all points, we can define a σ algebra on $C[0,\infty)$ by setting $E \cap C[0,\infty)$ to be a measurable subset for each measurable subset E of $\mathbf{R}^{[0,\infty)}$, and then if B is a Brownian motion, the map $\omega \mapsto B(\omega)$ is measurable. This σ algebra turns out to be equivalent to the Borel σ algebra with respect to the topology of uniform convergence on compact sets. The induced measure $\mathbf{W}(E) = \mathbf{P}(B^{-1}(E))$ is known as the **Wiener measure**. Wiener's theorem says that this measure is the unique one satisfying the properties of Brownian motion, which is obvious given our discussion.

11.3 Basic Properties

Once Brownian motion in one dimension has been constructed, it is easy to construct many other interesting processes with similar properties to Brownian motion.

Example. *If B is a one-dimensional Brownian motion, then we can use it to construct a probability space upon which we have n independent Brownian motions $B^1,\dots,B^n$, and we call the corresponding stochastic process $(B^1,\dots,B^n)$ an* **n dimensional Brownian motion**.

Example. *If $x \in \mathbf{R}$ is a random variable, and B_t is a Brownian motion, then $x + B_t$ is known as a* **Brownian motion started at** *x. We can also consider n dimensional Brownian motions started at $x \in \mathbf{R}^n$.*

Example. *The process (t,B) is known as the* **heat process** *in $\mathbf{R}$. If B is instead an n dimensional Brownian motion, we get the heat process in $\mathbf{R}^n$.*

Example. *The process $B_t - tB_1$ on $[0,1]$ is an example of a* **Brownian bridge**, *which models a variant of a Brownian motion which starts and returns to the origin after a unit time. Whereas a Brownian motion is just 'tied down' at the origin, a Brownian bridge is also tied down at 1, hence we think of it like a bridge, with pylons tied down at different points. It is a continuous Gaussian process with constant mean zero and covariance*

$$\mathbf{E}[(B_t - tB_1)(B_s - sB_1)] = s(1-t)$$

where $s \leqslant t$.

Example. *The process $O_t = e^{-t} B_{e^{2t}}$ is an example of the* **Ornstein-Uhlenbeck process**, *which is a modification of Brownian motion which has a tendency to move back to the origin at time ∞.*

One way we can understand Brownian motion is by determining the transformations of time and space which do not affect the law of the motion. This gives us invariance properties we can use in proofs. Let B be an d dimensional Brownian motion. Then we can obtain other examples of Brownian motion in the following cases:

- (Time Homogeneity) For $s \geqslant 0$, $B_{t+s} - B_s$ is a Brownian motion independent of $\sigma(B_u : u \leqslant s)$.
- (Symmetry): For any $T \in O_n(\mathbf{R})$, $T(B_t)$ is a Brownian motion. In particular, $-B_t$ is a Brownian motion.
- (Scaling): For every $c \neq 0$, cB_{t/c^2} is a Brownian motion. This is one of the most important properties of Brownian motion, for it implies a particular sample of the motion will have a fractal property to it.
- (Time-Inversion) $tB_{1/t}$ is a Brownian motion, if we define $0 \cdot B_\infty = 0$.

these properties are most easily verified by using the fact that Brownian motion is precisely a Gaussian process on $[0,\infty)$ with mean zero, and with covariance $\text{Cov}(B_t, B_s) = t \wedge s$, and with continuous sample paths. The only tricky part is to verify that $tB_{1/t} \to 0$ almost surely as $t \to 0$, and this follows because by continuity of B,

$$\mathbf{P}\left(\lim_{t \downarrow 0} tB_{1/t} = 0\right) = \mathbf{P}\left(\lim_{n \to \infty} \lim_{m \to \infty} \bigcap_{p \in \mathbf{Q}^+ \cap (0,1/m]} |pB_{1/p}| < 1/n\right)$$

Because the set on the right hand side is a measurable subset of $\mathbf{R}^{(0,\infty)}$, and the law of $tB_{1/t}$ and B_t agrees on $(0,\infty)$, we conclude this is equal to

$$\mathbf{P}\left(\lim_{n \to \infty} \lim_{m \to \infty} \bigcap_{p \in \mathbf{Q}^+ \cap (0,1/m]} |B_p| < 1/n\right) = \mathbf{P}\left(\lim_{t \downarrow 0} B_t\right) = 1$$

and this gives almost sure continuity. One example of the time invariance symmetry is a form of the 'law of large numbers' for Brownian motion, since we may think of B_t as the the integral of independent and identically distributed infinistisimals.

Theorem 11.2 (The Law of Large Numbers)**.** *Almost surely, $B_t = o(t)$.*

Proof. By the time inversion symmetry, the process $X_{1/t} = B_t/t$ is a continuous modification of B, and so

$$\mathbf{P}\left(\lim_{t\to\infty} B_t/t = 0\right) = \mathbf{P}\left(\lim_{t\to 0} X_t = 0\right) = \mathbf{P}\left(\lim_{t\to 0} B_t = 0\right) = 1$$

hence the limit almost surely tends to zero. □

More interestingly, we can go the other direction, and bound how Brownian motion behaves near the origin. Of course, by time homogeneity, this bounds how Brownian motion moves at any time point. If $f : [0,1] \to \mathbf{R}$ is a continuous function, then it is uniformly continuous, and therefore there exists a function $m : [0,1] \to \mathbf{R}$ with $m(s) \to 0$ as $s \to 0$ and for any $s \in [0,1]$,

$$\sup_{0\leqslant t\leqslant 1-s} \left| \frac{f(t+s) - f(t)}{m(s)} \right| \leqslant 1$$

we call m a **modulus of continuity** for f. Since Brownian motion is continuous, there obviously exists a random modulus of continuity for the motion, but surprisingly, we can find a deterministic modulus of continuity, which turns out to be $m(s) = \sqrt{2s\log(1/s)}$.

Lemma 11.3. *s*

Theorem 11.4 (Lévy's Modulus of Continuity Theorem)**.** *Almost surely, for any $s \in [0,1]$,*

$$\lim_{s\to 0} \sup_{0\leqslant t\leqslant 1-s} \left| \frac{f(t+s) - f(t)}{\sqrt{2s\log(1/s)}} \right| = 1$$

so for small values of s, B_s is well approximated by $\sqrt{2s\log(1/s)}$.

Proof. s □

11.4 Local Properties of Brownian Motion

Here, we use the scale invariance properties of Brownian motion to determine how Brownian motion looks locally. First, a generalization of the Kolmogorov continuity criterion.

Lemma 11.5. *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

$$\mathbf{P}(0 < Z < \varepsilon) = \mathbf{P}(0 < Z(N/\varepsilon) < N) = \mathbf{P}(0 < Z < N)$$

Letting $\varepsilon \to 0$, and $N \to \infty$ gives

$$\mathbf{P}(0 < Z < \infty) = 0$$

Now applying independence, we find

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

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. □

Corollary 11.6. *For any* $a \in \mathbf{R}$, $\{t : B_t(\omega) = a\}$ *is almost surely not bounded above. Thus every* a *is a recurrent state of the stochastic process.*

We can actually obtain a much stronger theorem, by using similar properties, which guarantees Brownian motion is almost surely not Hölder continuous.

Theorem 11.7. $\limsup B_t/\sqrt{t} > 0$ *almost surely.*

Proof. Note that

$$\mathbf{P}\left(\limsup_{t \to \infty} B_t/\sqrt{t} \leqslant 0\right) \leqslant \mathbf{P}\left(\limsup_{n \to \infty} B_n/\sqrt{n} \leqslant 0\right) = \mathbf{P}(E)$$

Consider the σ algebra

$$\Sigma_n = \sigma(B_{n+1} - B_n, B_{n+2} - B_{n+1}, \dots) = \sigma(B_k - B_n : k \geqslant n)$$

We claim that for any m, $E \in \Sigma_m$. Since the $B_{m+1} - B_m$ are independent of one another, the Kolmogorov zero-one law guarantees that either $\mathbf{P}(E) = 0$

or $\mathbf{P} = 1$. Fix ω, and suppose that for any $a > 0$, the inequality $B_n(\omega) \leqslant a\sqrt{n}$ holds for large enough n. Then

$$B_n(\omega) - B_m(\omega) \leqslant \left(a - \frac{B_m(\omega)}{\sqrt{n}} \right) \sqrt{n}$$

Since $B_m(\omega)$ is finite, eventually, we can guarantee $B_n(\omega) - B_m(\omega) \leqslant 2a\sqrt{n}$. Conversely, if, for any $a > 0$, we can guarantee that $B_n(\omega) - B_m(\omega) \leqslant a\sqrt{n}$, then for large enough n, $B_m(\omega) \leqslant a\sqrt{n}$, and then $B_n(\omega) \leqslant 2a\sqrt{n}$. We conclude

$$E = \left\{ \limsup_{n\to\infty} \frac{B_n - B_m}{\sqrt{n}} \leqslant 0 \right\} \in \Sigma_m$$

The theorem is complete if we can show $\mathbf{P}(E) < 1$. By Brownian scaling, for any $c > 0$,

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{B_n}{\sqrt{n}} \leqslant 0 \right) = \mathbf{P}\left(\limsup_{n\to\infty} c\frac{B_{n/c^2}}{\sqrt{n}} \leqslant 0 \right) = \mathbf{P}\left(\limsup_{n\to\infty} \frac{B_{n/c^2}}{\sqrt{n}} \leqslant 0 \right)$$

and now for any $a > 0$,

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{B_{n/c^2}}{\sqrt{n}} \leqslant 0 \right) \leqslant \mathbf{P}\left(\liminf_{n\to\infty} \left\{ \frac{B_{n/c^2}}{\sqrt{n}} \leqslant a \right\} \right) \leqslant \lim_{n\to\infty} \mathbf{P}\left(\frac{B_{n/c^2}}{\sqrt{n}} \leqslant a \right)$$

Since $B_{n/c^2}/\sqrt{n} \sim N(0, 1/c^2)$, we can let $c \to 0$ so that the right hand side tends to $1/2$. This shows the $\mathbf{P}(E) \leqslant 1/2 < 1$, so the Kolmogorov zero one law implies $\mathbf{P}(E) = 0$. □

Corollary 11.8. *Brownian motion is almost surely nowhere locally α Hölder continuous if $\alpha > 1/2$.*

Proof. For $x \geqslant 0$, consider

$$\mathbf{P}\left(\sup_{t,s\in[0,x]} \frac{|B_t - B_s|}{|t-s|^\alpha} < \infty \right)$$

By Brownian scaling, for any $c > 0$, this is equal to

$$\mathbf{P}\left(\sup_{t,s\in[0,x]} \frac{c|B_{t/c^2} - B_{s/c^2}|}{|t-s|^\alpha} < \infty \right) = \mathbf{P}\left(c^{1-2\alpha} \sup_{t,s\in[0,x/c^2]} \frac{|B_t - B_s|}{|t-s|^\alpha} < \infty \right)$$

as $c \to 0$ □

One reason Brownian motion is nice to study is that it is an example of almost every stochastic process, and in particular it is an example of these stochastic process upon which calculations are feasible. For instance, Brownian motion is a martingale.

Theorem 11.9. *A Brownian motion B is a martingale with respect to the natural filtration $\Sigma_t = \sigma(B_s : s \leqslant t)$, as is $B_t^2 - t$.*

Proof. Each B_t is in L^1, because it is $N(0,t)$ distributed. We find that

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

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

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

and rearranging this gives the Martingale property for $B_t^2 - t$. □

Once the theory of stochastic integrals 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. If we define a Brownian motion with respect to a filtration Σ as a Σ adapted Brownian motion such that $\mathbf{E}[B_t - B_s|\Sigma_s] = 0$, then B is a Σ adapted martingale.

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

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

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 12

Stochastic Calculus

Our goal in this chapter will be to make sense of a stochastic integral

$$(H \bullet X)_t = \int_0^t H_s dX_s$$

where H, X, and $H \bullet X$ will all be stochastic proceses on $(0, \infty)$. This equation can also be written in the 'differential form' $d(H \bullet X) = HdX$. We shall take Σ as the underlying filtration, which we will assume satisfies the usual conditions, so that all martingales can be assumed càdlàg (which we do from now on, unless otherwise started), and we have a sufficiently large family of stopping times.

The most used interpretation of the stochastic integral is known as the Itô integral, after it's creator, generalizing the discrete time process

$$(H \bullet X)_n(\omega) = \sum_{k=1}^n H_k(\omega)[X_k(\omega) - X_{k-1}(\omega)]$$

which represents the amount of money made after n discrete games, where the value obtain for a unit bet is given by the discrete $X_k(\omega) - X_{k-1}(\omega)$. In the discrete case, this process was most interesting if the 'integrand' H was previsible, and X was a martingale, a submartingale, or a supermartingale. In the continuous time case, H will also be a previsible process, under a suitable continuous reformulation of the previsibility condition, and X will be a semimartingale, which includes the class of all continuous time sub, super, and just plain martingales. We note that the stochastic processes which form our integrands will, unlike in the basic analysis of continuous time martingales, be indexed on $(0,\infty)$.

The discrete time formula is most easily generalized if the integrand-is constant over a certain interval. If Z is a random variable, and $S \leqslant T$ are stopping times, we define

$$Z(S,T](t,\omega) = Z(\omega)\chi_{(S(\omega),T(\omega)]}(t)$$

which represents a continuous bet of value $Z(\omega)$ made starting just after time $S(\omega)$, and ending at time $T(\omega)$ (the intervals chosen are made in order to make $Z(S,T]$ cáglád , which will be important later). In order for this random variable to 'not cheat', we should know the value of Z at time S. That is, Z should be a Σ_S measurable random variable, where X is Σ adapted. If the Itô integral correctly modelled a continuous time bet, it should be true that

$$\int_0^t Z(S,T]_s dX_s = Z\int_{S\wedge t}^{T\wedge t} dX_s = Z[X_{T\wedge t} - X_{S\wedge t}]$$

and we take this as a *definition* of the Itô integral in this very simple case.

Lemma 12.1. *If M is a UI martingale, and Z is bounded and Σ_S measurable, then the integral $Z(S,T] \bullet M$ is a UI martingale.*

Proof. By the definition of $Z(S,T] \bullet M$, it suffices to show that for each stopping time R, $Z[M_T - M_S]_R = Z[M_{T\wedge R} - M_{S\wedge R}]$ is integrable and has mean zero. The fact this random variable is integrable follows from the fact that $\|Z\|_\infty < \infty$, and because each $M_{T\wedge R}$ and $M_{S\wedge R}$ is integrable by the optional stopping theorem. Now we know

$$\mathbf{E}[M_\infty|\Sigma_{T\wedge R}] = M_{T\wedge R} \quad \mathbf{E}[M_\infty|\Sigma_{S\wedge R}] = M_{S\wedge R}$$

If $E \in \Sigma_S$, the function

$$T_E(\omega) = \begin{cases} T(\omega) & \omega \in E \\ \infty & \text{otherwise} \end{cases} \qquad S_E(\omega) = \begin{cases} S(\omega) & \omega \in E \\ \infty & \text{otherwise} \end{cases}$$

is a stopping time, and

$$\mathbf{E}(M_{T_E\wedge R} - M_{S_E\wedge R}) = \mathbf{E}(M_{T\wedge R} - M_{S\wedge R}; E)$$

and we can again apply the optional stopping theorem to conclude

$$\mathbf{E}(M_{T_E \wedge R} - M_{S_E \wedge R}) = \mathbf{E}[M_0 - M_0] = 0$$

Thus $\mathbf{E}[M_{T \wedge R} - M_{S \wedge R} | \Sigma_S] = 0$, and by the monotone class theorem, for any bounded, Σ_S measurable function Z,

$$\mathbf{E}(Z[M_{T \wedge R} - M_{S \wedge R}]) = 0$$

and this completes the proof. □

Stochastic integration theory essentially generalizes this lemma using the techniques of monotone class methods and localization, in tandem with using this lemma to extend the stochastic integral to a *much* larger family of integrands and integrals.

12.1 Previsibility

Recall that in the discrete setting, we defined a betting scheme C as a random variable adapted 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 enlarges this definition only slightly. 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 $\mathbf{R}$ with respect to the previsible sigma algebra.

Lemma 12.2. *If $S \leqslant T$ is a stopping times, and Z is bounded and Σ_S measurable, 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. Now for each t, and for each Borel $E \subset \mathbf{R}$, if $0 \notin E$,

$$\{Z(S,T]_t \in E\} = \{Z \in E\} \cap \{S < t \leqslant T\} = \{Z \in E\} \cap \{S < t\} \cap \{t \leqslant T\}$$

because Z is Σ_S measurable, $\{Z \in E\} \in \Sigma_S$, so

$$\{Z \in E\} \cap \{S < t\} = \lim_{s \uparrow t} \{Z \in E\} \cap \{S \leqslant s\} \in \Sigma_t$$

and $\{t \leqslant T\}^c = \{T < t\} \in \Sigma_t$, so $\{Z(S,T]_t \in E\}$ is Σ_t measurable. The proof is completed by noting that

$$\begin{aligned}\{Z(S,T]_t = 0\} &= \{Z = 0\} \cup \{t \leqslant S\} \cup \{t < T\}\\ &= \{Z = 0, S < t\} \cup \{t \leqslant S\} \cup \{t < T\}\end{aligned}$$

the first set is Σ_t measurable, and the other sets are also easily verified to be Σ_t measurable by their relation to stopping times. □

We can obtain a more general class of previsible processes by taking finite sums of these basic betting strategies. The set of finite sums of these processes will be known as the set of **bounded elementary integrands**, which we will sometimes denote by $\mathcal{E}$.

Lemma 12.3. *Any bounded elementary integrand can be written as*

$$\sum_{i=0}^{n} Z_i(S_i, T_i]$$

where $S_1 \leqslant T_1 \leqslant \dots \leqslant S_n \leqslant T_n$, and the Z_i are bounded and Σ_{S_i} measurable.

Proof. We prove this theorem by induction. Consider a bounded elementary integrand

$$X = \sum_{k=0}^{n+1} A_k(L_k, R_k]$$

By induction on n (with the base case being obvious), we can write

$$\sum_{k=0}^{n} A_k(L_k, R_k] = \sum_{k=0}^{m} Z_k(S_k, T_k]$$

It remains to 'slot' $A_{n+1}(L_{n+1}, R_{n+1}]$ into the process to complete the decomposition of X. We have a decomposition

$$\begin{aligned}&\sum_{k=0}^{m} Z_k(S_k, T_k] + A_{n+1}(L_{n+1}, R_{n+1}]\\ &\quad= \sum_{k=0}^{m} (Z_k + A_{n+1})(S_k \vee L_{n+1}, T_k \wedge R_{n+1}]\\ &\quad\quad+ \sum_{k=0}^{m} B_k(S_k \wedge L_{n+1}, S_k \vee L_{n+1}] + \sum_{k=0}^{m} C_k(T_k \wedge R_{n+1}, T_k \vee R_{n+1}]\end{aligned}$$

Where

$$B_k = \begin{cases} Z_k & S_k \leqslant L_{n+1} \\ A_{n+1} & S_k > L_{n+1} \end{cases}$$

$$C_k = \begin{cases} Z_k & T_k \geqslant R_{n+1} \\ A_{n+1} & T_k < R_{n+1} \end{cases}$$

Now Z_k is Σ_{S_k} measurable, and A_{n+1} is $\Sigma_{L_{n+1}}$ measurable, so since Σ_{S_k} and $\Sigma_{L_{n+1}}$ are subalgebras of $\Sigma_{S_k \vee R_{n+1}}$, we conclude that $Z_k + A_{n+1}$ is $\Sigma_{S_k \vee L_{n+1}}$ measurable. To verify that B_k is $\Sigma_{S_k \wedge L_{n+1}}$ measurable, it suffices to show that B_k is Σ_{S_k} and $\Sigma_{R_{n+1}}$ measurable. But

$$B_k = Z_k \mathbf{I}(S_k \leqslant L_{n+1}) + A_{n+1} \mathbf{I}(L_{n+1} < S_k)$$

The map $t \mapsto A_{n+1}\mathbf{I}(L_{n+1} < t)$ is left continuous and Σ_t adapted, because if $0 \notin E$,

$$\{A_{n+1}\mathbf{I}(L_{n+1} < t) \in E\} = \{A_{n+1} \in E, L_{n+1} < t\} \in \Sigma_t$$

and

$$\{A_{n+1}\mathbf{I}(L_{n+1} < t) = 0\} = \{A_{n+1} = 0, L_{n+1} < t\} \cup \{L_{n+1} \geqslant t\} \in \Sigma_t$$

It follows that the process is progressive, so if we stop the process at time S_k, we get a Σ_{S_k} measurable random variable $A_{n+1}\mathbf{I}(S_k > R_{n+1})$. Similarily, $t \mapsto Z_k\mathbf{I}(S_k \leqslant t)$ is right continuous and Σ_t adapted, so we can stop at R_{n+1} to get a $\Sigma_{R_{n+1}}$ measurable random variable $Z_k\mathbf{I}(S_k \leqslant L_{n+1})$. The same strategy allows us to show $\mathbf{I}(S_k \leqslant L_{n+1})$ and $\mathbf{I}(L_{n+1} < S_k)$ are both Σ_{S_k} and $\Sigma_{L_{n+1}}$ measurable, and putting all these measurability results tells us that B_k is both Σ_{S_k} and $\Sigma_{L_{n+1}}$ measurable. The result for C_k is similar. □

Once we have written a bounded elementary integrand H in the form $Z_1(S_1, T_1] + \dots + Z_n(S_n, T_n]$, where $S_1 \leqslant T_1 \leqslant \dots \leqslant S_n \leqslant T_n$, we can unambiguously define the integral

$$\int_0^t H dX = \sum_{k=1}^n \int_0^t Z_k(S_k, T_k] dX = \sum_{k=1}^n Z_k[X_{T_k \wedge t} - X_{S_k \wedge t}]$$

This is because if $Z_1(S_1, T_1] + \dots + Z_n(S_n, T_n] = 0$, then because each $(S_i, T_i]$ is a disjoint interval, we conclude $Z_1 = \dots = Z_n = 0$. By linearity, is H is a bounded elementary integrand, and M is a uniformly integrable martingale, then $H \bullet M$ is also a uniformly integrable martingale. The monotone class theorem will come in handy in extending this integral to all previsible processes, and one fact that makes this easy is the following.

Lemma 12.4. *The smallest σ algebra on $(0,\infty) \times \Omega$ such that all bounded elementary functions are measurable is equal to the previsible σ algebra.*

Proof. Let Π be the smallest σ algebra on $(0,\infty) \times \Omega$ such that all bounded elementary functions are measurable. It is obvious that each bounded elementary integrand is previsible, so it remains to show every bounded cáglád adapted process is measurable with respect to the, 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 this gives the required result. □

The following expresses the MCT using the result above.

Corollary 12.5. *If V is a subspace of bounded functions jointly measurable on $(0,\infty) \times \Omega$ which contains constant processes and all bounded elementary integrands, is closed under uniform convergence, and closed under monotone convergence, in the sense that if the H_n are uniformly bounded and non-negative, and $H_n \uparrow H$, then H is in V. Then V contains every bounded previsible process on $(0,\infty) \times \Omega$.*

Here is a simple, but important, corollary, which should be intuitive.

Theorem 12.6. *The previsible σ algebra is generated by the family of sets*

$$\{(t,\infty) \times E : E \in \Sigma_t\}$$

Proof. It suffices to show that if $\Pi = \sigma((t,\infty) \times E : E \in \Sigma_t)$, then all $Z(S,T]$ are Π measurable. Now if a and b are non-random, and $E \in \Sigma_a$, then

$$\chi_E(a,b] = \mathbf{I}((a,\infty) \times E) - \mathbf{I}((b,\infty) \times E)$$

which is the difference of two Π measurable processes. But now if $S \leqslant T$ are stopping times, and $E \in \Sigma_S$, then

$$\chi_E(S,T](t) = \lim_{n\to\infty} \sum_{m=0}^{\infty} \chi_{E_{nm} \cap E}(m/2^n, m+1/2^n]$$

where $E_{nm} = \{(m-1)/2^n \leqslant S < m/2^n \leqslant T\}$. Now we just calculate that

$$E \cap E_{nm} = E \cap \{S < m/2^n\} \cap \{m/2^n \leqslant T\} \cap \{(m-1)/2^n \leqslant S\} \in \Sigma_{m/2^n}$$

so each term in the sum is Π measurable, hence the sum is Π measurable, and then $\chi_E(S,T]$ is the limit of Π measurable functions. We then apply the monotone class theorem to conclude that $Z(S,T]$ is Π measurable for all bounded Σ_S measurable Z. □

12.2 Finite Variation Processes

A **finite variation process** is an adapted process X such that each path $t \mapsto X_t(\omega)$ is of finite variation. 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 finite partitions of $[0,1]$. If, in addition, the integrated variation norm

$$\|X\|_V = \mathbf{E}(V_X(\infty)) = \lim_{t\to\infty} \mathbf{E}(V_X(t)) < \infty$$

we say X has **integrable variation**. In the case X is a finite variation process, it follows that for any H, it makes sense to define

$$\int_0^t H_s dX_s$$

pointwise, for each sample point, as the Lebesgue Stieltjes integral.

Theorem 12.7. *If H is a bounded, previsible process, and M is a martingale with integrable variation, null at zero, then $H \bullet M$ is a martingale with integrable variation.*

Proof. Let V be the class of all bounded previsible processes H such that $H \bullet M$ is a martingale. Since we have

$$\sup_{t\in(0,\infty)} \left| \int_0^t H_s dM_s \right| \leqslant \int_0^\infty |H_s\ dM_s| \leqslant \|H\|_\infty V_M(0,\infty)$$

we have bounded $H \bullet M$ pointwise by an integrable random variable, and so $H \bullet M$ is uniformly integrable for all bounded H. It is automatically because of it's definition in terms of a Stieltjes integral. We also have shown that V contains all the bounded elementary functions. If H is a constant, then

$$\int_0^t H_s dM_s = H \int_0^t dM_s = H(M_t - M_0)$$

so $H \bullet M$ is obviously a martingale as a difference of two martingales. Suppose that H_n is a uniformly bounded sequence of elements of V, with $H_n \to H$ pointwise. Then the dominated convergence theorem implies that for each fixed t, $(H_n \bullet M)_t \to H \bullet M$ in $L^1(\Omega)$, and it follows from this that $H \bullet M$ is a martingale, by the regularity of conditional expectations under L^1 convergence. In follows from the monotone class variant we established for bounded previsible processes that $H \bullet M$ is a martingale for all bounded, previsible H. □

12.3 Localization

The boundedness and integrality assumptions we used to conclude on the regularity of the integrals of finite variation processes is too stringent to be practical for the processes we encounter in practice. Brownian motion is almost surely never of finite variation, and we often want to consider integrals $\int_0^t X_s dB_s$, so more work is needed.

To begin, we must relax the hypothesis of that theorem to a 'localized' version, in the sense of locally integrable functions. Of course, then the conclusion of the theorem above will only hold locally. 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 a 'local time' which stops before the condition becomes too erratic. Given a stopping time T and a process H on $(0, \infty)$, it is natural to introduce the process $H(0, T]$ 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 consider the stopped process X^T. We can interpret this formally as saying $dX_t^T = \chi_{(0,T]} dX_t$, so that we 'close off all bets' at time T – surely, for all the integrands we have considered,

$$\int_0^t H_s dX_s^T = \int_0^t \chi_{(0,T]} H_s dX_s$$

We then say the differential equation $d(H \bullet X) = HdX$ holds on $(0,T]$ if

$$(H \bullet X)^T = H(0,T] \bullet X^T$$

Given some family Γ of processes, such that if $H \in \Gamma$, then $H(0,T] \in \Gamma$ for all stopping times T, then we say another process H is **locally** in Γ if there is a sequence of almost surely increasing stopping times $T_1, T_2, \dots$ such that $T_k \to \infty$ almost surely, and $H(0,T_n] \in \Gamma$ for all n. We say the T_n **reduces** H into Γ. The most important example occurs when Γ is the space of bounded previsible processes, in which we say a process locally in Γ is a **locally bounded and previsible** process.

Lemma 12.8. *If H is an adapted cáglád process with $\limsup_{t \downarrow 0} |H_t| < \infty$ almost surely, then H is a locally bounded previsible process.*

Proof. Let $T_n = \inf\{t > 0 : |H_t| \geqslant n\}$. Then T_n is a Σ_t adapted stopping time (with $\Sigma_0 = \lim \Sigma_t$), and the growth condition on H near 0 shows that for any sample point ω, $\limsup_{t \downarrow 0} |H_t(\omega)| = \alpha < \infty$, so if $n > \alpha$, $T_n(\omega) \neq 0$, and if $T_n(\omega) \to \beta > 0$ as $n \to \infty$, then we can find $u_n \geqslant T_n(\omega)$ with $u_n - T_n(\omega) \leqslant 1/n$ and $|H_{u_n}(\omega)| > n$, and then $u_n \to \beta$, hence the sequence either contains a left or right convergent subsequence, and this subsequence breaks the cáglád property of H. Thus $T_n \to \infty$. □

We note that if H is locally, locally in a family Γ, in the sense that there are $T_i \uparrow \infty$ with $H(0,T_i]$ locally in Γ, then we can find $\lim_{j \to \infty} U_{ij} \to \infty$ with $H(0,T_i](0,U_{ij}] = H(0, T_i \wedge U_{ij}]$ in Γ. If we set choose $k(n)$ such that

$$\mathbf{P}(U_{nk(n)} \wedge T_n \geqslant T_n - 2^{-n}) = \mathbf{P}(U_{nk(n)} \geqslant T_n - 2^{-n}) \geqslant 1 - 2^{-n}$$

and then let $R_n = \inf\{U_{mk(m)} : m \geqslant n\}$. Since

$$\sum_{n=1}^{\infty} \mathbf{P}(T_n \wedge U_{nk(n)} \wedge T_n < T_n - 2^{-n}) < \infty$$

the Borel-Cantelli theorem implies that, almost surely, for large enough n we have $T_n \wedge U_{nk(n)} \geqslant T_n - 2^{-n}$, so almost surely, for large enough n we find

$$R_n \geqslant \inf\{T_m - 2^{-m} : m \geqslant n\} \geqslant \inf\{T_m : m \geqslant n\} - 2^{-n}$$

so $R_n \to \infty$ almost surely, and localize H into Γ. We never need to consider second order locality, because this reduces into first order locality.

On the other hand, we can also localize integrators. Let Γ be a family of adapted processes, such that Γ is stable under stopping, in the sense that if $X \in \Gamma$, $X^T \in \Gamma$ for every stopping time T; we say a process X is locally in Γ if there is an almost surely increasing family $T_1, T_2, \dots$ with $T_i \to \infty$ almost surely such that $X^{T_n} \in \Gamma$ for all n. The most important versions of this process is where Γ is the family of all martingales, in which case we say X is a **local martingale**. Similarily, we can define locally uniform integrable martingales, local integrable variation martingales, and locally finite variation martingales. Since the condition of having finite variation is a local condition, we find that the class of locally finite variation processes are just processes with finite variation. First, of course, we must prove that these spaces are stable under stopping.

Lemma 12.9. *If U, V are bounded stopping times, and M is a martingale, then* $\mathbf{E}[M_U | \Sigma_V] = M_{U \wedge V}$.

Proof. Suppose $U, V \leqslant K$. The martingale $M^U = (M^K)^U$ is uniformly integrable, since M^K is uniformly integrable, and so the optional sampling theorem gives

$$\mathbf{E}[M_U | \Sigma_V] = \mathbf{E}[M_\infty^U | \Sigma_V] = (M^U)_V = M_{U \wedge V}$$

and this gives the result. This result also shows that the operators $\mathbf{E}(\cdot | \Sigma_U)$ and $\mathbf{E}(\cdot | \Sigma_V)$ commute with one another, with composition $\mathbf{E}(\cdot | \Sigma_{U \wedge V})$. □

Theorem 12.10. *The space of martingales, uniformly integrable martingales, finite variation martingales, and integrable variation martingales are stable under stopping.*

Proof. Fix some stopping time T. Since X^T is a martingale if X is, this shows the space of martingales is stable under stopping. If X is uniformly integrable, then $X_t^T = \mathbf{E}[X_\infty^T | \Sigma_t]$ is uniformly integrable as t ranges. If X has finite variation, it is trivial to see that X^T has finite variation because the paths of X^T are just the paths of X cut short, so

$$V_{X^T}(0, t) = V_X(0, T \wedge t) \leqslant V_X(t, \omega)$$

This is also sufficient to show that integrable variation martingales are stable under stopping. □

Lemma 12.11. *Every local martingale is a local uniformly integrable martingale. Every locally finite variation martingale locally has integrable variation.*

Proof. For the first point, it suffices to show every martingale is locally uniformly integrable, and this follows because for each $k \in \mathbf{N}$, M^k is uniformly integrable because

$$M_t^k = M_{t \wedge k} = \mathbf{E}[M_k | \Sigma_{t \wedge k}]$$

and it is easy to see this is uniformly integrable. For the second case, because of the first case, we may prove that every *uniformly integrable* finite variation martingale locally has integrable variation. Let $V_M(t)$ denote the variation of M on $(0, t]$. Define $T_n = \inf\{t > 0 : V_M(t) > n\}$. Then T_n is the first entry time into an open set of an adapted process, so it is a stopping time, and

$$V_M(T_n) \leqslant n + |M(T_n) - M(T_n-)| \leqslant n + |M(T_n)| + |M(T_n-)|$$

But $|M(T_n-)| \leqslant n$, and $\mathbf{E}|M(T_n)| = \mathbf{E}|M(\infty)|$, because of the optional sampling theorem, so M^{T_n} has integrable variation. □

12.4 Stochasting Integration By Localization

Let's summarize our setup so far. We have a family Λ of integrands, and a collection Γ of integrals, and we want to find a map $(H, X) \mapsto H \bullet X$ into some space Θ which represents the stochastic integral. The main result is that if Γ and Θ are stable under stopping, and Λ is stable under bet stopping, and we know ahead of time that

$$\int_0^T H_s dX_s = \int_0^\infty H(0, T]_s dX_s^T$$

In our first case, Γ is the set of bounded, previsible processes, and Λ and Θ the set of martingales with integral variation, in which the local equation obviously holds pointwise. If H is a process which is only locally in Λ, and X is only locally in Γ, then we can find a sequence of stopping times T_n which simulatenously reduce H and X, and we can consider the integrals $H(0, T_n] \bullet X^{T_n}$. We can then *define* the process $H \bullet X$ such that

$$\int_0^{t \wedge T_n} H dX = \int_0^t H(0, T_n] dX^{T_n}$$

holds, and this uniquely defines the integral at all points because $T_n \to \infty$ pointwise almost surely, and then $H \bullet X$ will then be locally in Θ for the stopping times T_i. This is well defined, because, first of all

$$\int_0^{t \wedge T_{n-1}} H(0, T_n] dX^{T_n} = \int_0^t H(0, T_{n-1}] dX^{T_{n-1}}$$

so the stopping time integrals agree with one another, and if $S_1 \leqslant S_2 \dots$ is another reducing sequence for H and X, then

$$\int_0^{t \wedge S_n} H(0, T_n] dX^{T_n} = \int_0^{\infty} H(0, T_n \wedge S_n] dX^{T_n \wedge S_n} = \int_0^{t \wedge T_n} H(0, S_n] dX^{S_n}$$

and this means exactly that the two definitions agree with one another. As long as the integral we began with was bilinear, it is obvious that the new locally defined integral we have here is also bilinear.

In the first case we considered, $H \bullet X$ is just the stochastic Lebesgue-Stieltjes integral, because the local definition agrees with the normal stochastic Lebesgue-Stieltjes integral. Since every locally finite variation process also locally has integrable variation, then our reasoning implies the following result.

Theorem 12.12. *If H is locally bounded and previsible and M is locally a finite variation martingale, and $H \bullet M$ is the Stieltjes integral, then $H \bullet M$ locally has finite variation.*

Frequently, we can only define H up to indistinguishable. This could potentially cause problems with stochastic integration theory. We say that a process is **evanescent** if, for almost all sample points, for every t, $X_t = 0$. Provided that we can show that all evanescent processes are able to be integrated, and if H is evanescent, then the stochastic integral of $H \bullet X$ is evanescent, then our stochastic integrals can be defined up to indistinguishability, and so this will cause no problems.

12.5 Local Martingales

A **local martingale** on $[0, \infty)$ is a process M such that M_0 is Σ_0 measurable, and $M_t - M_0$ is locally a martingale. Note here that M_0 might not even be an integrable variable. In this case, we define

$$\int_0^t HdM = \int_{(0,t]} H_s d(M_s - M_0)$$

when the appropriate definition of stochastic integration is possible (right now, we only know this integral if $M_s - M_0$ locally has paths of finite variation). It is obvious that every martingale is a local martingale.

Lemma 12.13 (Stochastic Fatou Lemma). *If M is a nonnegative local martingale such that $\mathbf{E}M_0 < \infty$, then M is a supermartingale.*

Proof. Let T_n be a reducing sequence for $M_t - M_0$. Then we find that

$$M_t^{T_n} = (M_t^{T_n} - M_0) + M_0$$

is integrable, and for $t > s$,

$$\begin{aligned}\mathbf{E}[M_t^{T_n}|\Sigma_s] &= \mathbf{E}[M_0|\Sigma_s] + \mathbf{E}[M_t^{T_n} - M_0|\Sigma_s]\\ &= M_0 + M_{s\wedge T_n} - M_0 = M_{s\wedge T_n}\end{aligned}$$

Each M_t is integrable, because

$$\mathbf{E}M_t = \mathbf{E}(M_t - M_0) + \mathbf{E}(M_0)$$

We can then apply the normal Fatou's lemma to conclude that, since $M_{t\wedge T_n} \to M_t$ pointwise, that

$$\mathbf{E}[M_t] \leqslant \liminf \mathbf{E}[M_t^{T_n}] = \mathbf{E}M_0 < \infty$$

and so the M_t are integrable, and we can then apply the conditional expectation version of conditional expecation to conclude

$$\mathbf{E}[M_t|\Sigma_s] \leqslant \liminf \mathbf{E}[M_{t\wedge T_n}|\Sigma_s] = \liminf M_{s\wedge T_n} = M_s$$

so M_t is a supermartingale. □

Example. *Let B be some Brownian motion in $\mathbf{R}^3$ with $|B_0| = 1$, and consider $M_t = 1/|B_t|$, with. The sequence $T_n = \inf\{t : |M_t| > n\}$ reduces M into the class of local martingales, and for $t > s$,*

$$\mathbf{E}[M_{t\wedge T_n}|\Sigma_s] = M_{s\wedge T_n}$$

but $\mathbf{E}[M_t|\Sigma_s] < M_s$. One phenomenon which prevents this is that the variables $M_{t\wedge T_n}$ are not uniformly integrable. The $T_n \to \infty$ almost surely because, in three dimensions, Brownian motion almost surely never touches any particular point, in this case the origin.

12.6 Semimartingales as Integrators

A **semimartingale** is an adapted, process X which may be written as

$$X = X_0 + M + A$$

where M is a local martingale null at zero, and A is a process null at 0 with paths of finite variation. The concept of semimartingales was originally arrived from by ad hoc methods, as the stochastic integral was gradually extended to be defined on finite variation processes, then to local martingales, and then we can define

$$\int_0^t H_s dX_s = \int_0^t H_s dM_s + \int_0^t H_s dA_s$$

where the last integral is just the stochastic Stieltjes integral. Deliberately avoiding the discussion of uniqueness, the problem of defining the integral of semimartingales requires us only to define the integral of local martingales. We shall find that if X is a semimartingale, and H is locally bounded and previsible, then $H \bullet X$ is also a semimartingale. What's more, if $f \in C^\infty(\mathbf{R}^n)$, and X is a supermartingale in $\mathbf{R}^n$, in the sense that each X^i is a semimartingale, then $f(X)$ is also a semimartingale. Even better, almost every process you encounter in practice is a semimartingale, so the fact that we can treat these processes as integrands is the best feature of the integration theory.

Semimartingales are about the most general process we can consider as integrands, because of the following result. Call a process X an **integrator** if the following condition holds: For any sequence $H_n = \sum Z_i^n (S_i^n, T_i^n]$ of bounded, elementary integrands with $\|H_n\|_\infty \to 0$, then for all t, the random variables

$$\sum Z_i^n \left(X_{T_i^n} - X_{S_i^n} \right)$$

converges to 0 in probability. This is about the weakest result we could use to build a sufficient theory of integration, because the convergence of the H_n is the strongest possible, and the convergence of the 'integrals' the weakest possible. Using the theory of vector-valued measures, we can define the stochastic integral of integrators, and using these ideas, work up to the following theorem, which we won't exposit here, but indicate that semimartingales are the right place for stochastic integration.

Theorem 12.14. *X is an integrator if and only if X is a semimartingale.*

So let's get going with defining the integrals of semimartingales!

12.7 Stochastic Integration in L^2

If a process N is of finite variation, it is easy to define the stochastic integral with respect to dN. However, almost all interesting stochastic processes are not of finite variation, and we can use randomness to show that integrals of certain simple functions converge *in probability* to a random variable, and the nicest place that this theory operates is in the space of martingales which are bounded in L^2.

We let M^2 denote the vector space of martingales which are bounded in the L^2 norm. Then Doob's L^2 convergence theorem guarantees that M^2 is in bijection with the square integrable Σ_∞ measurable functions by the map $M \mapsto M_\infty$, since we can recover a process M from M_∞ by the conditional expectations $M_t = \mathbf{E}[M_\infty|\Sigma_t]$. We use this correspondence to establish a topological structure on M^2, which makes M^2 into a Hilbert space.

Suppose M^k is a Cauchy sequence in M^2, so that M^k_∞ is Cauchy in the square integrable norm. Because $M^k_\infty \to M_\infty$ in the L^2 norm, we find $M^k_t = \mathbf{E}[M^k_\infty|\Sigma_t]$ converges in L^2 to $M_t = \mathbf{E}[M_\infty|\Sigma_t]$. Now $M^k_t - M_t$ is a martingale bounded in L^2, and so Doob's L^2 inequality implies

$$\mathbf{E}[[(M^k - M)^*_\infty]^2] \leqslant 4\mathbf{E}[(M^k_\infty - M_\infty)^2]$$

Therefore $(M^k - M)^*_\infty \to 0$ in L^2, and therefore there is a subsequence k_n such that, almost surely, $M^{k_n} \to M$ uniformly.

Lemma 12.15. *The space cM^2 of continuous M^2 martingales is closed in M^2.*

Proof. The discussion above shows that a sequence of martingales converges uniformly almost surely, and the uniform limit of continuous functions is continuous. □

We say two processes $M, N \in M^2$ are **weakly orthogonal** if they are orthogonal in L^2, so $\mathbf{E}[M_\infty N_\infty] = 0$, and **strongly orthogonal** if for every stopping time T, $\mathbf{E}[M_T N_T] = 0$.

Lemma 12.16. *M and N are strongly orthogonal iff MN is a UI martingale.*

Proof. This is trivial by the characterization of UI martingales as progressive martingales for which $\mathbf{E}[M_T N_T] = 0$. □

A subspace of M^2 is known as **stable** if it is closed and stable under stopping. For instance, cM^2 and M_0^2 are both stable under stopping. We are going to decompose every martingale in M_0^2 into a continuous martingale in cM_0^2 and a martingale in the complement of the direct sum, dM_0^2 (it's discontinuous part). This follows from the next theorem.

Theorem 12.17. *If N_0 is a stable subspace of M_0^2, then the space $N_0^\perp$ is a stable subspace of M_0^2 and every $Z \in N_0^\perp$ is strongly orthogonal to every $N \in N_0$, and every $M \in M_0^2$ decomposes uniquely as $M = N + Z$, with $N \in N_0$ and $Z \in N_0^\perp$.*

Proof. $N_0^\perp$ is trivially closed. Now suppose $Z \in N_0^\perp$. For each stopping time T, we want to show $\mathbf{E}[Z_T N_\infty) = 0$ for every $N \in N_0$. Now $N^T \in N_0$, so $\mathbf{E}[Z_\infty N^T] = 0$. But now

$$\begin{aligned}\mathbf{E}[Z_\infty N_T] &= \mathbf{E}[\mathbf{E}[Z_\infty N_T | \Sigma_T]] = \mathbf{E}[N_T \mathbf{E}[Z_\infty | \Sigma_T]] \\ &= \mathbf{E}[N_T Z_T] = \mathbf{E}[\mathbf{E}[N_\infty | \Sigma_T] Z_T] = \mathbf{E}[\mathbf{E}[N_\infty Z_T | \Sigma_T]] = \mathbf{E}[N_\infty Z_T]\end{aligned}$$

which has proven the claim. □

Given a martingale M, we let $M = M^c + M^d$, where $M^c \in cM_0^2$ and $M^d \in dM_0^2$. We note that using similar techniques to above, we can show that for any stopping time T, $(M^T)^c = (M^c)^T$.

Theorem 12.18. *For any stopping time T, $(M^T)^c = (M^c)^T$.*

Proof. TODO □

Example. *Let*

$$H = \sum_{i=1}^n Z_{i-1}(T_{i-1}, T_i]$$

be a bounded elementary function, where $0 \leqslant T_0 \leqslant \dots \leqslant T_n$. Then we know that for any martingale M, $H \bullet M$ is a uniformly integrable martingale. Now we show that if $M \in M^2$, then $H \bullet M \in M^2$, because, firstly, if we define $\Delta_i = M_{T_i} - M_{T_{i-1}}$, then for $i < j$,

$$\mathbf{E}[Z_{i-1} Z_{j-1} \Delta_i \Delta_j | \Sigma_{T_{j-1}}] = Z_{i-1} Z_{j-1} \Delta_i \mathbf{E}[\Delta_j | \Sigma_{T_{j-1}}] = 0$$

and so

$$\mathbf{E}[(H \bullet M)_\infty^2] = \mathbf{E}\left[\left(\sum_{i=1}^n Z_{i-1}\Delta_i\right)^2\right] = \sum_{i=1}^n \mathbf{E}[Z_{i-1}^2 \Delta_i^2]$$

Thus $H \bullet M \in M^2$.

The ideas above will generalize to the more general class of stochastic integrals, once we introduce the **quadratic variation** of a process. We begin with a lemma which is proved similarly to the example above.

Lemma 12.19. *Consider a uniformly bounded simple process*

$$H = \sum_{i=1}^\infty Z_{i-1}(T_{i-1}, T_i]$$

where $0 \leqslant T_0 \leqslant T_1 \leqslant T_2 \leqslant \dots$, and $\sup|Z_k| \leqslant C$. If we define $H \bullet M$ in the obvious fashion, then $H \bullet M \in M_0^2$, and

$$\mathbf{E}(H \bullet M)_\infty^2 \leqslant C^2 \mathbf{E}[M_\infty^2]$$

Proof. TODO □

Theorem 12.20. *Let $M \in M_0^2$. Then there is a unique increasing process $[M]$, known as the* **quadratic variation***, null at 0, such that*

- *$M^2 - [M]$ is a uniformly integrable martingale.*
- *$\Delta[M] = (\Delta M)^2$.*

If we set $4[M,N] = [M+N] - [M-N]$, then $[M,N]$ is the unique finite variation process null at 0 such that $MN - [M,N]$ is a UI martingale, and $\Delta[M,N] = (\Delta M)(\Delta N)$. We call $[M,N]$ the **quadratic covariation***.*

Proof. TODO □

We shall find that this is true for all local martingales locally bounded in L^2, which will enable us to construct a stochastic integral for a more general class of integrands, such as Brownian motion. For now, we construct the stochastic integral in L^2. Let $M \in M^2$. If $H = \sum Z_{i-1}(T_{i-1}, T_i]$, the $H \bullet M$ is an L^2 bounded martingale. Since $M^2 - [M]$ is a uniformly integrable martingale, then for any stopping times $S \leqslant T$,

$$\mathbf{E}[(M_T - M_S)^2 | \Sigma_S] = \mathbf{E}[M_T^2 - M_S^2 | \Sigma_S] = \mathbf{E}([M]_T - [M]_S | \Sigma_S)$$

So we obtain the crucial identity

$$\mathbf{E}\left(\left(\int_0^\infty H_s dM_s\right)^2\right) = \mathbf{E}\left(\sum Z_{i-1}^2([M]_{T_i} - [M]_{T_{i-1}})\right) = \mathbf{E}\int_0^\infty H_s^2 d[M]_s$$

We define the L^2 norm on previsible processes with respect to M by the norm

$$\|H\|_M = \left(\mathbf{E}\int_0^\infty H_s^2 d[M]_s\right)^{1/2}$$

The space $L^2(M)$ of H with $\|H\|_M < \infty$ is of course, a Hilbert space containing the bounded elementary functions. The map I mapping a bounded elementary H to $(H \bullet M)_\infty$ given by stochastic integration is therefore an isometry. This means that stochastic integration extends uniquely to be an isometry on the vector space all processes approximatable in the $L^2(M)$ norm by bounded elementary integrands. But one verifies that this class satisfies the monotone class theorem, which is the family of all bounded previsible processes. Thus we obtain the Itô integral for all such processes. We summarize this in a theorem.

Theorem 12.21 (Kunita, Watanabe). *Let $M \in M_0^2$, and $H \in L^2(M)$. Then the Itô integral $H \bullet M$ is the isometry of $L^2(M)$ into the space of square integrable Σ_∞ measurable functions, obtained from extending the isometry on basic elementary integrands, and in particular,*

$$\mathbf{E}\left[\left(\int_0^\infty H_s dM_s\right)^2\right] = \mathbf{E}\left[\int_0^\infty H_s^2 d[M]_s\right]$$

We can further extend the integral by applying localization theory, in the way we already mentioned.

Theorem 12.22. *Fix $M \in M^2$. Then all bounded previsible processes are in $L^2(M)$, and for any $H \in L^2(M)$ and stopping time T,*

- $(H \bullet M)^T = H(0,T) \bullet M = H \bullet M^T$

- $\left(\int_0^t HdM\right)^2 - \int_0^t H_s^2 d[M]_s$ *is a uniformly integrable martingale.*

- *If H and K are bounded previsible processes, then $H\bullet(K\bullet M) = (HK)\bullet M$.*

- $\Delta(H \bullet M) = H\Delta M$.

- *If M is continuous, $H \bullet M$ is also continuous.*

Proof. TODO □

Because of this lemma, if H is locally bounded and previsible, and M is locally in M^2, then we can define $H \bullet M$, which is also locally in M_0^2. Furthermore, if M is also continuous, then $H \bullet M$ is continuous. It is not at all clear that this agrees with the stochastic finite variation integral, though we will have to wait some time before we can do this This isn't really important to the theory we develop, however.

12.8 The Kunita Watanabe Inequalities

The construction of the stochastic integral constructed above is the preferred one, because it is a good starting point for stochastic calculus theory. Nonetheless, it is not as elegant as prior definitions. The Kunita-Watanabe theory of L^2 integrals relates to the following inequalities.

Theorem 12.23. *s*

Theorem 12.24. *Let M be a square integrable martingale, and let $M \in L^2(M)$. The $\int_0^\infty H_s dM_s$ is the unique square integrable Σ_∞ random variable such that for every $N \in M^2$,*

$$\mathbf{E}\left(N_\infty \int_0^\infty HdM\right) = \mathbf{E}\left(\int_0^\infty Hd[M,N]\right)$$

Chapter 13

Appendix: 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.*

13.1 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.

13.2 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 13.1 (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 $Y_1, Y_2, \dots$ 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.

13.3 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 13.2. *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.

13.4 Properties of the Conditioning Operator

Let Δ denote the overlying σ algebra of our probability space. Since $\mathbf{E}(X|\Sigma)$ is unique up to a set of measure zero, the many to one map from $\mathcal{L}^1(\Omega, \Delta)$ to $\mathcal{L}^1(\Omega, \Sigma)$ descends to an *operator* from $L^1(\Omega, \Delta)$ to $L^1(\Omega, \Sigma)$. 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 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 13.3. *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. □

Since $L^1(\Omega, \Sigma)$ embeds in $L^1(\Omega, \Delta)$, we can consider the conditional expectation as an operator from $L^1(\Omega)$ to itself, except for one important point: just because two functions agree almost everywhere with respect to Δ, this does not imply they either are both measurable with respect to Σ or both not measurable, because Σ does not necessarily contain all the null sets in Σ. Thus if we consider the conditional operator on $L^1(\Omega)$, then we can only conclude that the equivalence class of $\mathbf{E}(X|\Sigma)$ contains Σ measurable versions of the conditional expectation.

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 13.4. *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 13.5. *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 $Y_1 \leqslant Y_2 \leqslant \dots$, 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. □

13.5 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 $E_1, E_2, \dots$ 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

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

Thus $\mathbf{P}(A \cap B | \mathcal{B}[0,1]) = \chi_B/2$ almost surely for each set B. Since $\mathcal{B}[0,1]$ 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

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

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.

13.6 Independence and Conditional Expectation

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

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

for any Borel set $A_1, \dots, A_n$. 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 $\Sigma_1, \dots, \Sigma_n$ is independent if

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

if $A_1 \in \Sigma_1, \dots, A_n \in \Sigma_n$, 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 13.6. *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 13.7. *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 $X_1, X_2, \dots$, 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*

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

Note that $X_{n+1}, X_{n+2}, \dots$ 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

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

so $\mathbf{E}(X_i|S_n) = \mathbf{E}(X_j|S_n)$. Now, using the fact that $S_n = \sum \mathbf{E}(X_i|S_n) = n\mathbf{E}(X_i|S_n)$, we see that $\mathbf{E}(X_i|S_n) = S_n/n$. This calculation leads to a very nice proof of the strong law of large numbers in the theory of discrete martingales, as we will find in the next chapter.