Stat433/833 Lecture Notes

Stochastic Processes

Jiahua Chen

Department of Statistics and Actuarial Science

University of Waterloo

Key Words: σ-field, Brownian motion, diffusion process, ergordic, finite dimensional distribution, Gaussian process, Kolmogorov equations, Markov property, martingale, probability generating function, recurrent, renewal theorem, sample path, simple random walk, stopping time, transient,

Stat 433/833 Course Outline

- We use the formula for final grade:

$$25\%\text{assignment} + 25\%\text{midterm} + 50\%\text{final}.$$

- There will be 4 assignments.
- Office hours: _____ to be announced.
- Office: MC6007 B, Ex 35506, jhchen@uwaterloo.ca

Textbook and References

The text book for this course is **Probability and Random Processes** by Grimmett and Stirzaker.

It is NOT essential to purchase the textbook. This course note will be free to download to all students registered. We may not be able to cover all the materials included.

Stat433 students will be required to work on fewer problems in assignments and exams.

The lecture note reflects the instructor's still in-mature understanding of this general topic, which were formulated after reading pieces of the following books. A request to reserve has been sent to the library for books with call numbers.

The References are in Random Order.

1. *Introduction to probability models.* (QA273 .R84) Sheldon M. Ross. Academic Press.

2. *A Second course in stochastic processes.* S. Karlin and H. M. Taylor. Academic Press.

3. *An introduction to probability theory and its applications.* Volumes I and II. W. Feller, Wiley.

4. *Convergence of probability measures.* Billingsley, P. Wiley.

5. *Gaussian random processes.* (QA274.4.I2613) I.A. Ibragimov, and Rozanov. Springer-Verlag, New York.

6. *Introduction to stochastic calculus with applications.* Fima C. Klebaner. Imperial College Press.

7. *Diffusions, Markov processes and martingales.*(QA274.7.W54) L. C. G. Rogers and D. Williams. Cambridge.

8. *Probability Theory, independence, interchangeability, martingale.* Y. S. Chow and H. Teicher. Springer, New York.

Contents

Chapter 1

Review and More

1.1 Probability Space

A probability space consists of three parts: sample space, a collection of events, and a probability measure.

Assume an experiment is to be done. The set of all possible outcomes is called **Sample Space**. Every element ω of Ω is called a sample point. Mathematically, the sample space is merely an arbitrary set. There is no need of a corresponding experiment.

A probability measure intends to be a function defined for all subsets of Ω. This is not always possible when the probability measure is required to have certain properties. Mathematically, we settle on a collection of subsets of Ω.

Definition 1.1

A collection of sets $\mathcal{F}$ is a σ-field (algebra) if it satisfies

1. The empty set $\phi \in \mathcal{F}$,
2. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} \in \mathcal{F}$ (closed under **countable** union).
3. If $A \in \mathcal{F}$, then its complement $A^c \in \mathcal{F}$.

$\diamondsuit$

Let us give a few examples.

Example 1.1

1. The simplest σ-field is $\{\phi, \Omega\}$.
2. A simplest non-trivial σ-field is $\{\phi, A, A^c, \Omega\}$.
3. A most exhaustive σ-field is $\mathcal{F}$ consists of all subsets of Ω.

$\diamondsuit$

It can be shown that if A and B are two sets in a σ-field, then the resulting sets of all commonly known operations between A and B are members of the same σ-field.

Axioms of Probability Measure

Given a sample space Ω and a suitable σ-field $\mathcal{F}$, A probability measure P is a mapping from $\mathcal{F}$ to R (set of real numbers) such that:

1. $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$;
2. $P(\Omega) = 1$;
3. $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for any $A_i \in \mathcal{F}$, $i = 1, 2, \ldots$ which satisfy $A_i A_j = \phi$ whenever $i \neq j$.

Mathematically, the above definition does not rely on any hypothetical experiment. A probability space is given by $(\Omega, \mathcal{F}, P)$. The above axioms lead to restrictions on $\mathcal{F}$. For example, suppose $\Omega = [0, 1]$ and $\mathcal{F}$ is the σ-field that contains all possible subsets. In this case, if we require that the probability of all closed intervals equal their lengthes, then it is impossible to find such a probability measure satisfying Axiom 3.

When the sample space Ω contains finite number of elements, the collection of all subsets forms a σ-field $\mathcal{F}$. Define $P(A)$ as the ratio of sizes of A and of Ω, then it is a probability measure. This is the classical definition of probability.

The axioms for a probability space imply that the probability measure has many other properties not explicitly stated as axioms. For example, since $P(\phi \cup \phi) = P(\phi) + P(\phi)$, we must have $P(\phi) = 0$.

Axioms 2 and 3 imply that

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c).$$

Hence, $P(A^c) = 1 - P(A)$.

For any two events A_1 and A_2, we have

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2).$$

In general,

$$P(\cup_{i=1}^n A_i) = \sum P(A_i) - \sum_{i_1<i_2} P(A_{i_1} A_{i_2}) + \cdots, +(-1)^{n+1} P(\cap_{i=1}^n A_i).$$

Lemma 1.1 *Continuity of the probability measure.*

Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $A_1 \subset A_2 \subset A_3 \subset \cdots$ is an increasing sequence of events and

$$A = \lim_{n\to\infty} \cup_{i=1}^n A_i = \lim_{n\to\infty} A_n,$$

then

$$P(A) = \lim_{n\to\infty} P(A_n).$$

PROOF: Let $B_n = A_n - A_{n-1}$, $n = 1, 2, 3, \ldots$ with $A_0 = \phi$. Then $A_n = \cup_{i=1}^n B_i$, and $A = \cup_{i=1}^\infty B_i$. Notice that $B_1, B_2, \ldots$ are mutually exclusive, and $P(B_n) = P(A_n) - P(A_{n-1})$. Using countable additivity,

$$\begin{aligned} P(A) &= \sum_{i=1}^\infty P(B_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^n P(B_i) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

This completes the proof. ♢

1.2 Random Variable

Definition 1.1

A random variable is a map $X : \Omega \to R$ such that $\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in R$. $\diamondsuit$

Notice that $\{X \leq x\}$ is more often the notation of an event than the notation for the outcome of an experiment.

The cumulative distribution function (c.d.f.) of a random variable X is $F_X(t) = P(X \leq t)$. It is known that a c.d.f. is a non-decreasing, right continuous function such that

$$F(-\infty) = 0, \quad F(\infty) = 1.$$

The random behaviour of X is largely determined by its distribution through c.d.f. At the same time, there are many different ways to characterise the distribution of a random variable. Here is an incomplete list:

1. If X is absolutely continuous, then the probability density function (p.d.f.) is $f(x) \geq 0$ such that

   $$F(x) = \int_{-\infty}^{x} f(t)dt.$$

   For almost all x, $f(x) = dF(x)/dx$.

2. If X is discrete, taking values on $\{x_1, x_2, \ldots\}$, the probability mass function of X is

   $$f(x_i) = P(X = x_i) = F(x_j) - F(x_j-)$$

   for $i = 1, 2, \ldots$, where $F(a-) = \lim_{x \to a, x<a} F(x)$.

3. The moment generating function is defined as

   $$M(t) = E\{\exp(tX)\}.$$

4. The characteristic function of X is defined as

$$\phi(t) = E\{\exp(\mathbf{i}tX)\}$$

where $\mathbf{i}$ is such that $\mathbf{i}^2 = -1$. The advantage of the characteristic function is that it exists for all X. Otherwise, the moment generating function is simpler as it is not built on the concepts of complex numbers.

5. The probability generating function is useful when the random variable is non-negative integer valued.

6. In survival analysis or reliability, we often use survival function

$$S(x) = P(X > x) = 1 - F(x)$$

where we assume $P(X \geq 0) = 1$.

7. The hazard function of X is

$$\lambda(x) = \lim_{\Delta \to 0+} \frac{P\{x \leq X < x + \Delta | X \geq x\}}{\Delta}.$$

It is also called "instantaneous failure rate" as it is the rate at which the individual fails in the next instance given it has survived so far.

When X is absolutely continuous,

$$\lambda(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx} \log S(x).$$

Hence, for $x \geq 0$,

$$S(x) = \exp\{-\int_0^x \lambda(t)dt\}.$$

Let us examine a few examples to illustrate the hazard function and its implication.

Example 1.2

(1) When X has exponential distribution with density function $f(x) = \lambda \exp(-\lambda x)$, it has constant hazard rate λ.

(2) When X has Weibull distribution with density function

$$f(x) = \lambda\alpha(\lambda x)^{\alpha-1} \exp\{-(\lambda x)^{\alpha}\}$$

for $x \geq 0$. The hazard function is given by

$$\lambda(x) = \lambda\alpha(\lambda x)^{\alpha-1}.$$

(a) When $\alpha = 1$, it reduces to exponential distribution. Constant hazard implies the item does not age.

(b) When $\alpha > 1$, the hazard increases with time. Hence the item ages.

(c) When $\alpha < 1$, the hazard decreases with time. The longer the item has survived, the more durable this item becomes. It ages negatively.

(3) Human mortality can be described by a curve with bathtub shape. The death rate of new borns are high, then it stabilizes. After certain age, we becomes vulnerable to diseases and the death rate increases.

1.3 More σ-fields and Related Concepts

Consider a random variable X defined on the probability space $(\Omega, \mathcal{F}, P)$. For any real number such as $-\sqrt{2}$, the sample points satisfying $X \leq -\sqrt{2}$ form a set which belongs to $\mathcal{F}$. Similarly $X \geq 1.2$ is also a set belongs to $\mathcal{F}$. This claim extends to the union of these two events, the intersection of these two events, and so on.

Putting all events induced by X as above together, we obtain a collection of events which will be denoted as $\sigma(X)$. It can be easily argued that $\sigma(X)$ is the smallest sub σ-field generated by X.

Example 1.3

1. The σ-field of a constant random variable;
2. The σ-field of an indicator random variable;
3. The σ-field of a random variable takes only two possible values.

If X and Y are two random variables, we may define the conditional cumulative distribution function of X given by $Y = y$ by

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

This definition works if $P(Y = y) > 0$. Otherwise, if X and Y are jointly absolutely continuous, we make use of their joint density to compute the conditional density function.

One purpose of introducing conditional density function is for the sake of computing conditional expectation. Because we have to work with conditional expectation more extensively later, we hope to define conditional expectation for any pairs of random variables.

If Y is an indicator random variable I_A for some event A, such that $0 < P(A) < 1$, then $E\{X|Y = 0\}$ and $E\{X|Y = 1\}$ are both well defined as above. Further, the corresponding values are the average sizes (expected values) of X over the ranges A^c and A. Because of this, the conditional expectation of X given Y is expected sizes of X over various ranges of Ω partitioned according to the size of Y. When Y is as simple as an indicator random variable, this partition is also simple and we can easily work out these numbers conceptually. When Y is an ordinary random variable, the partition of Ω is $\sigma(Y)$. The problem is: $\sigma(Y)$ contains so many events in general, and also they are not mutually exclusive. Thus, there is no way to present the conditional expectation of X given Y by enumerating them all.

The solution to this difficulty in mathematics is to define $Z = E\{X|Y\}$ as a measurable function of Y such that

$$E\{ZI_A\} = E\{XI_A\}$$

for any $A \in \sigma(Y)$.

We may realize that this definition does not provide any concrete means for us to compute $E\{X|Y\}$. In mathematics, this definition has to be backed up by an existence theorem that such a function is guaranteed to exist. At the same time, when another random variable differs to Z only by a zero-probability event, that random variable is also a conditional expectation of X given Y. Thus, this definition is unique only up to some zero-probability event.

Since Z is a function of Y, it is also a random variable. In particular, by letting $A = \Omega$, we get

$$E[E\{X|Y\}] = E\{X\}.$$

A remark here is: the conditional expectation is well defined only if the expectation of X exists. You may remember a famous formula:

$$Var(X) = Var[E\{X|Y\}] + E[Var\{X|Y\}].$$

Note that definition of the conditional expectation relies on the σ-field generated by Y only. Thus, if $\mathcal{G}$ is a σ-field, we simply define $Z = E\{X|\mathcal{G}\}$ as a $\mathcal{G}$ measurable function such that

$$E\{ZI_A\} = E\{XI_A\}$$

for all $A \in \mathcal{G}$. Thus, the concept of conditional expectation under measure theory is in fact built on σ-field.

Let us go over a few concepts in elementary probability theory and see what they mean under measure theory.

Recall that if X and Y are independent, then we have

$$E\{g_1(X)g_2(Y)\} = E\{g_1(X)\}E\{g_2(Y)\}$$

for any functions g_1 and g_2. Under measure theory, we need to add a cosmestic condition that these two functions are measurable, plus these expectations exist. More rigorously, the independence of two random variables is built on the independence of their σ-fields. It can be easily seen that the σ-field generated by $g_1(X)$ is a sub-σ-field of that of X. One may then see what $g_1(X)$ is independent of $g_2(Y)$.

Example 1.4 *Some properties.*

1. If X is $\mathcal{G}$ measurable, then

$$E\{XY|\mathcal{G}\} = XE\{Y|\mathcal{G}\}.$$

2. If $\mathcal{G}_1 \subset \mathcal{G}_2$, then

$$E[E\{X|\mathcal{G}_2\}|\mathcal{G}_1] = E\{X|\mathcal{G}_1\}.$$

3. If $\mathcal{G}$ and $\sigma(X)$ are independent of each other, then

$$E\{X|\mathcal{G}\} = E\{X\}.$$

Most well known elementary properties of the mathematical expectation remain valid.

1.4 Multivariate Normal Distribution

A group of random variables X have joint normal distribution if their joint density function is in the form of

$$C\exp(-xAx^\tau - 2bx^\tau)$$

where A is positive definite, and C is a constant such that the density function has total mass 1. Note that x and b are vectors.

Being positive definite implies that there exists an orthogonal decomposition of A such that

$$A = B\Lambda B^\tau$$

where Λ is diagonal matrix with all element positive and $BB^\tau = I$, the identity matrix.

With this knowledge, it is seen that

$$\int \exp(-\frac{1}{2}xAx^\tau)dx = \int \exp(-\frac{1}{2}y\Lambda y^\tau)dy = \frac{(2\pi)^{n/2}}{|A|^{1/2}}.$$

Hence when $b = 0$, the density function has the form

$$f(x) = [(2\pi)^{-n}|A|]^{1/2}\exp(-\frac{1}{2}xAx^\tau).$$

Let X be a random vector with the density function given above, and let $Y = X + \mu$. Then the density function of Y is given by

$$\{(2\pi)^{-n}|A|\}^{1/2}\exp\{-\frac{1}{2}(y-\mu)A(y-\mu)'\}.$$

It is more convenient to use $V = A^{-1}$ in most applications. Thus, we have the definition as follows.

Definition 1.1 *Multivariate normal distribution*

$X = (X_1, \ldots, X_n)$ has multivariate normal distribution (written as $N(\mu, V)$), if its joint density function is

$$f(x) = \{(2\pi)^n |V|\}^{-1/2} \exp\{-\frac{1}{2}(x-\mu)V^{-1}(x-\mu)^\tau\}$$

where V is positive definite matrix. $\diamondsuit$

If X is multivariate normal $N(\mu, V)$, then $E(X) = \mu$ and $Var(X) = V$.

If X(length n) is $N(\mu, V)$ and D is an $n \times m$ matrix of rank $m \leq n$, then $Y = XD$ is $N(\mu D, D^\tau V D)$.

More general, we say a random vector $X = (X_1, \ldots, X_n)$ has multivariate normal distribution when Xa^τ is a normally distributed random variable for all $a \in R^n$.

A more rigorous and my favoured definition is: if X can be written in the form $AY + b$ for some (non-random) matrix A and vector b, and Y is a vector of iid standard normally distributed random variables, then X is a multinormally distributed random vector.

We now list a number of well known facts about the multivariate distributions. If you find that you are not familiar with a large number of them, it probably means some catch up work should be done

1. Suppose X has normal distribution with mean μ and variance σ^2. Its characteristic function is

   $$\phi(t) = \exp\{\mathbf{i}\mu t - \frac{1}{2}\sigma^2 t^2\}.$$

   Its moment generating function is

   $$M(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}.$$

2. If X has standard normal distribution, then all its odd moments are zero, and its even moments are

   $$EX^{2r} = (2r-1)(2r-3)\cdots 3 \cdot 1.$$

3. Let $\Phi(x)$ and $\phi(x)$ be the cumulative distribution function and the density function of the standard normal distribution. It is known that

$$\left\{\frac{1}{x} - \frac{1}{x^3}\right\}\phi(x) \leq 1 - \Phi(x) \leq \frac{1}{x}\phi(x)$$

for all positive x. In particular, when x is large, they provide a very accurate bounds.

4. Suppose that $X_1, X_2, \ldots, X_n$ are a set of independent and identically distributed standard normal random variables. Let $X_{(n)} = \max\{X_i, i = 1, \ldots, n\}$. Then $X_{(n)} = O_p(\sqrt{\log n})$.

5. Suppose Y_1, Y_2 are two independent random variables such that $Y_1 + Y_2$ are normally distributed, then Y_1 and Y_2 are jointly normally distributed. That is, they are multivariate normal.

6. A set of random variables have multivariate normal distribution if and only if all their linear combinations are normally distributed.

7. Let $X_1, \ldots, X_n$ be i.i.d. normal random variables. The sample mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and the sample variance $S_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are independent. Further, $\bar{X}_n$ has normal distribution and S_n^2 has chisquare distribution with $n-1$ degrees of freedom.

8. Let X be a vector having multivariate normal distribution with mean μ and covariance matrix Σ, and A be a non-negative definite symmetric matrix. Then, XAX^τ has chi-squared distribution when

$$A\Sigma A\Sigma A = A\Sigma A.$$

1.5 Summary

We did not intend to give you a full account of probability theory learned in courses preceding this one. Neither we try to lead you to the world of measure theory based, advanced, and widely regarded as useless rigorous probability theory. Yet, we did introduce the concept of probability space, and why a dose of σ-field is needed in doing so.

In the later half of this course, we need the σ-field based definitions of independence, conditional expectation. It is also very important to get familiar with multivariate normal random variables. These preparations are not sufficient, but will serve as starting point. Also, you may be happy to know that that the pressure is not here yet.

Chapter 2

Simple Random Walk

Simple random walk is an easy object in the family of stochastic processes. At the same time, it shares many properties with more complex stochastic processes. Understanding simple random walk helps us to understand abstract stochastic processes.

The building block of a simple random walk is a sequence of iid random variables $X_1, \ldots, X_n, \ldots$ such that

$$P(X_1 = 1) = p, \quad P(X_1 = -1) = 1 - p = q.$$

We assume that $0 < p < 1$, otherwise, the simple random walk becomes trivial.

We start with $S_0 = a$ and define $S_{n+1} = S_n + X_{n+1}$ for $n = 1, 2, \ldots$. It mimics the situation of starting gambling with an initial capital of $\$a$, and the stake is \$1 on each trial. The probability of winning a trial is p and the probability of losing is $q = 1 - p$. Trials are assumed independent.

When $p = q = 1/2$, the random walk is symmetric.

Properties:

A simple random walk is

(i) time homogeneous:

$$P\{S_{n+m} = j | S_m = a\} = P\{S_n = j | S_0 = a\}.$$

(ii) spatial homogeneous:

$$P\{S_{n+m} = j + b | S_m = a + b\} = P\{S_n = j | S_0 = a\}.$$

(iii) Markov:

$$P\{S_{n+m} = j | S_0 = j_0, S_1 = j_1, \cdots, S_m = j_m\} = P\{S_{n+m} = j | S_m = j_m\}.$$

♢

Example 2.1 *Gambler's ruin*

Assume that start with initial capital $S_0 = a$, the game stops as soon as $S_n = 0$ or $S_n = N$ for some n, where $N \geq a \geq 0$. Under the conditions of simple random walk, what is the probability that the game stops at $S_n = N$?

SOLUTION: A brute force solution for this problem is almost impossible. The key is to view this probability as a function of initial capital a. By establishing a relationship between the probabilities, we will get a difference equation which is not hard to solve. ♢

2.1 Counting sample paths

If we plot $(S_i, i = 0, 1, \ldots, n)$ against $(0, 1, \ldots, n)$, we obtain a path connecting $(0, S_0)$ and (n, S_n). We call it a sample path. Assume that $S_0 = a$ and $S_n = b$. There might be many possible sample paths leading from $(0, a)$ to (n, b).

Property 2.1

The number of paths from $(0, a)$ to (n, b) is

$$N_n(a, b) = \binom{n}{\frac{n+b-a}{2}}$$

when $(n + a - b)/2$ is a positive integer, 0 otherwise. ♢

Property 2.2

For a simple random walk, and when $(n+a-b)/2$ is a positive integer, all sample paths from $(0,a)$ to (n,b) have equal probability $p^{(n+b-a)/2}q^{(n+a-b)/2}$ to occur. In addition,

$$P(S_n = b|S_0 = a) = \binom{n}{\frac{n+b-a}{2}} p^{(n+b-a)/2}q^{(n+a-b)/2}.$$

♢

The proof is straightforward.

Example 2.2

It is seen that

$$P(S_{2n} = 0|S_0 = 0) = \binom{2n}{n} p^n q^n$$

for $n = 0, 1, 2, \ldots$. This is the probability when the random walk returns to 0 at trial $2n$.

When $p = q = 1/2$, we have

$$u_{2n} = P(S_{2n} = 0|S_0 = 0) = \binom{2n}{n} (1/2)^{2n}.$$

Using Stirling's approximation

$$n! \approx \sqrt{2\pi n}(n/e)^n$$

for large n. The approximation is good even for small n. We have

$$u_{2n} \approx \frac{1}{\sqrt{n\pi}}.$$

Thus, $\sum_n u_{2n} = \infty$. By renewal theorem to be discussed, $S_n = 0$ is a recurrent event when $p = q = 1/2$. ♢

Property 2.3 *Reflection principle*

Let $N_n^0(a,b)$ be the number of sample paths from $(0,a)$ to (n,b) that touch or cross the x-axis. Then, when $a, b > 0$,

$$N_n^0(a,b) = N_n(-a,b).$$

PROOF: We Simply count the number of paths. The key step is to establish the one-to-one relationship. ♢

The above property enables us to show the next interesting result.

Property 2.4 *Ballot Theorem*

The number of paths from (0, 0) to (n, b) that do not revisit the axis is

$$\frac{|b|}{n} N_n(0, b).$$

PROOF: Notice that such a sample path has to start with a transition from $(0, 0)$ to $(1, 1)$. The number of paths from $(1, 1)$ to (n, b), such that $b > 0$, which do not touch x-axis is

$$\begin{aligned} N_{n-1}(1, b) - N^0_{n-1}(1, b) &= N_{n-1}(1, b) - N_{n-1}(-1, b) \\ &= \binom{n-1}{\frac{n-b}{2}} - \binom{n-1}{\frac{n-b-2}{2}} \\ &= \frac{|b|}{n} N_n(0, b). \end{aligned}$$

♢

What does this name suggest? Suppose that Micheal and George are in a competition for some title. In the end, Micheal wins by b votes in n casts. If the votes are counted in a random order, and $A =$"Micheal leads throughout the count", then

$$P(A) = \frac{\frac{b}{n} N_n(0, b)}{N_n(0, b)} = \frac{b}{n}.$$

Theorem 2.1

Assume that $S_0 = 0$ and $b \neq 0$. Then

(i) $P(S_1 S_2 \cdots S_n \neq 0, S_n = b | S_0 = 0) = \frac{|b|}{n} P(S_n = b | S_0 = 0)$.

(ii) $P(S_1 S_2 \cdots S_n \neq 0) = n^{-1} E(|S_n|)$. ♢

PROOF: The first part can be proved by counting the number of sample paths which do not touch x-axis.

The second part is a consequence of the first one. Just sum over the probability of (i) for possible values of S_n. ♢

Theorem 2.2 *First passage through b at trial n (Hitting time problem).*

If $b \neq 0$ and $S_0 = 0$, then

$$f_n(b) = P(S_1 \neq b, \cdots, S_{n-1} \neq b, S_n = b | S_0 = 0) = \frac{|b|}{n} P(S_n = b).$$

PROOF: If we reverse the time, the sample paths become those who reach b in n trials without touching the x-axis. Hence the result. $\diamondsuit$

The next problem of interest is how high S_i have attained before it settles at $S_n = b$. We assume $b > 0$. In general, we often encounter problems of finding the distribution of the extreme value of a stochastic process. This is an extremely hard problem, but we have a kind of answer for the simple random walk.

Define $M_n = \max\{S_i, i = 1, \ldots, n\}$.

Theorem 2.3

Suppose $S_0 = 0$. Then for $r \geq 1$,

$$P(M_n \geq r, S_n = b) = \begin{cases} P(S_n = b) & \text{if } b \geq r \\ (q/p)^{r-b} P(S_n = 2r - b) & \text{if } b < r \end{cases}$$

Hence,

$$\begin{aligned} P(M_n \geq r) &= P(S_n \geq r) + \sum_{c=-\infty}^{r-1} (q/p)^{r-b} P(S_n = 2r - b) \\ &= P(S_n = r) + \sum_{c=r+1}^{\infty} [1 + (q/p)^{c-r}] P(S_n = c) \end{aligned}$$

When $p = q = 1/2$, it becomes

$$P(M_n \geq r) = 2P(S_n \geq r + 1) + P(S_n = r).$$

PROOF: When $b \geq r$, the event $M_n \geq r$ is a subset of the event $S_n = b$. Hence the first conclusion is true.

When $b < r$, we may draw a horizontal line $y = r$. For each sample path belongs to $M_n \geq r$, $S_n = b$, we obtain a partial mirror image: it retains the part until the sample path touches the line of $y = r$, and completes it with the mirror image from there. Thus, the number of sample paths is the

same as the number of sample paths from 0 to $2r - b$. The corresponding probability, however, is obtained by exchanging the roles of $r - b$ pairs of p and q.

The remaining parts of the theorem are self illustrative. ♢

Let μ_b be the mean number of visits of the walk to the point b before it turns to its starting point. Suppose $S_0 = 0$. Let $Y_n = I(S_1 S_2 \cdots S_n \neq 0, S_n = b)$. Then $\sum_{n=1}^{\infty} Y_n$ is the number of times it visits b. Notice that this reasoning is fine even when the walk will never return to 0.

Since $E(Y_n) = f_b(n)$, we get

$$\mu_b = \sum_{n=1}^{\infty} f_b(n).$$

Thus, in general, the mean number of visit is less than 1. When $p = q = 0.5$, all states are recurrent. Therefore $\mu_b = 1$ for any b. ♢

Suppose a perfect coin is tossed until the first equalisation of the accumulated numbers of heads and tails. The gambler receives one dollar every time that the number of heads exceeds the number of tails by b. This fact results in comments that the "fair entrance fee" equals 1 independent of b.

My remark: how many of us thinks that this is against their intuition?

Theorem 2.4 *Arc sine law for last visit to the origin.*

Suppose that $p = q = 1/2$ and $S_0 = 0$. The probability that the last visit to 0 up to time $2n$ occurred at time $2k$ is given by

$$P(S_{2k} = 0)P(S_{2n-2k} = 0).$$

♢

PROOF: The probability in question is

$$\begin{aligned} \alpha_{2n}(2k) &= P(S_{2k} = 0)P(S_{2k+1} \cdots S_{2n} \neq 0 | S_{2k} = 0) \\ &= P(S_{2k} = 0)P(S_1 \cdots S_{2n-2k} \neq 0 | S_0 = 0). \end{aligned}$$

We are hence asked to show that

$$P(S_1 \cdots S_{2n-2k} \neq 0 | S_0 = 0) = u_{2n-2k}.$$

Since $S_0 = 0$ is part of our assumption, we may omit the conditional part. We have

$$\begin{aligned} P(S_1 \cdots S_{2m} \neq 0 | S_0 = 0) &= \sum_{b \neq 0} P(S_1 \cdots S_{2m} \neq 0, S_{2m} = b) \\ &= \sum_{b \neq 0} \frac{|b|}{m} P(S_{2m} = b) \\ &= 2 \sum_{b=1}^{m} \frac{2b}{2m} P(S_{2m} = 2b) \\ &= 2 \left(\frac{1}{2}\right)^{2m} \sum_{b=1}^{m} \left[\binom{2m-1}{m+b-1} - \binom{2m-1}{m+b} \right] \\ &= \binom{2m}{m} \left(\frac{1}{2}\right)^{2m} \\ &= u_{2m}. \end{aligned}$$

♢

The proof shows that

$$P(S_1 S_2 \ldots S_{2m} \neq 0) = P(S_{2m} = 0).$$

Denote, as usual, $u_{2n} = P(S_{2n} = 0)$ and $\alpha_{2n}(2k) = P(S_{2k} = 0, S_i \neq 0, i = 2k+1, \ldots, 2n)$. Then the theorem says

$$\alpha_{2n}(2k) = u_{2k} u_{2n-2k}.$$

As $p = q = 1/2$ is a simple case, we have accurate approximation for u_{2k}.

It turns out that $u_{2k} \approx (\pi k)^{-1/2}$ and so $\alpha_{2n}(2k) \approx \{\pi[k(n-k)]\}^{-1/2}$. If we let T_{2n} be the time of the last visit to 0 up to time $2n$, the it follows that

$$\begin{aligned} P(T_{2n} \leq 2xn) &\approx \sum_{k \leq xn} \{\pi[k(n-k)]\}^{-1/2} \\ &\approx \int_0^x \frac{1}{\pi[u(1-u)]^{1/2}} du \\ &= \frac{2}{\pi} \arcsin(\sqrt{x}). \end{aligned}$$

That is, the limiting distribution has a density function described by $\arcsin(\sqrt{x})$.

(1) Since $p = q = 1/2$, one may think the balance of heads and tails should occur very often. Is it true? After $2n$ tosses with n large, the chance that it never touches 0 after trial n is 50% which is surprisingly large.

(2) The last time before trial $2n$ when the simple random walk touches 0 should be closer to the end. This result shows that it is symmetric about midpoint n. It is more likely to be at the beginning and near the end.

Why is it more likely at the beginning? since it started at 0, it is likely to touch 0 again soon. Once it wandered away from 0, touching 0 becomes less and less likely.

Why is it also likely occur near the end: if for some reason the simple random walk returned to 0 at some point, it becomes more likely to visit 0 again in the near future. Thus, it pushes the most recent visit closer and closer to the end.

(3) If we count the number of k such that $S_k > 0$, what kind of distribution it has? Should it be more likely to be close to n? It turns out that it is either very small or very large.

Thus, if you gamble, you may win all the time, or lose all the time even though the game is perfectly fair in the sense of probability theory.

Let us see how the result establishes (3).

Property 2.5 *Arc sine law for sojourn times.*

Suppose that $p = 1/2$ and $S_0 = 0$. The probability that the walk spends exactly $2k$ intervals of time, up to time $2n$, to the right of the origin equals $u_{2k}u_{2n-2k}$. ♢

PROOF: Call this probability $\beta_{2n}(2k)$. We are asked to show that $\beta_{2n}(2k) = \alpha_{2n}(2k)$.

We use mathematical induction.

The first step is to consider the case when $k = n = m$.

In our previous proofs, we have shown that for symmetric random walk

$$u_{2m} = P(S_1 \cdots S_{2m} \neq 0).$$

Since these sample paths can belong to one of the two possible groups: always above 0 or always below 0. Hence,

$$u_{2m} = 2P(S_1 > 0, \ldots, S_{2m} > 0).$$

On the other hand, we have

$$\begin{aligned}
& P(S_1 > 0, S_2 > 0, \cdots, S_{2m} > 0) \\
= \; & P(S_1 = 1, S_2 \geq 1, \ldots, S_{2m} \geq 1) \\
= \; & P(S_1 - 1 = 0, S_2 - 1 \geq 0, \ldots, S_{2m} - 1 \geq 0) \\
= \; & P(S_1 = 1 | S_0 = 0) P(S_2 - 1 \geq 0, \ldots, S_{2m} - 1 \geq 0 | S_1 - 1 = 0) \\
= \; & \frac{1}{2} P(S_2 \geq 0, S_3 \geq 0, \ldots, S_{2m} \geq 0 | S_1 = 0) \\
= \; & \frac{1}{2} P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2m-1} \geq 0 | S_0 = 0) \\
= \; & \frac{1}{2} P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2m-1} \geq 0, S_{2m} \geq 0 | S_0 = 0) \\
= \; & \frac{1}{2} \beta_{2m}(2m)
\end{aligned}$$

where the last equality comes from the fact that $S_{2m-1} \geq 0$ implies $S_{2m} \geq 0$ as it is impossible for $S_{2m-1} = 0$. Consequently, since $u_0 = 1$, we have shown that

$$\alpha_{2m}(2m) = u_{2m} = \beta_{2m}(2m).$$

Let us now make an induction assumption that

$$\alpha_{2n}(2k) = \beta_{2n}(2k)$$

for all n with $k = 0, 1, \ldots, m-1$ and $k = n, n-1, \ldots, n-(m-1)$. Note the reason for the validity of the second sets of k is symmetry. Our induction task is to show that the same is true when $k = m$.

By renewal equation,

$$u_{2m} = \sum_{r=1}^{m} u_{2m-2r} f_{2r}.$$

Using idea similar to renewal equation, we also have

$$\beta_{2n}(2m) = \frac{1}{2} \sum_{r=1}^{m} f_{2r} \beta_{2n-2r}(2m-2r) + \frac{1}{2} \sum_{r=1}^{n-m} f_{2r} \beta_{2n-2r}(2m)$$

since the walk will start with either a positive move or a negative move, it will touch 0 again at some point $2r$ in another $2n-2r$ trials (it could happen that $2r = 2n$).

So, with renewal equation, and induction assumption,

$$\begin{aligned}\beta_{2n}(2m) &= \frac{1}{2}\sum_{r=1}^{m} f_{2r}u_{2m-2r}u_{2n-2m} + \frac{1}{2}\sum_{r=1}^{n-m} f_{2r}u_{2m-2r}u_{2n-2m-2r} \\ &= \frac{1}{2}[u_{2n-2m}u_{2m} + u_{2m}u_{2n-2m}] \\ &= u_{2m}u_{2n-2m}.\end{aligned}$$

The induction is hence completed. ♢

2.2 Summary

What have we learned in this chapter? One observation is that all sample paths starting and ending at the same location have equal probability to occur. Making use of this fact, some interesting properties of the simple random work are revealed.

One such property is that the simple random work is null recurrent when $p = q$ and transient otherwise. Let us recall some detail; by counting sample paths, it is possible to provide an accurate enough approximation to the probability of entering 0.

The reflection principle allows us to count the number of paths going from one point to another without touching 0. This result is used to establish the Ballot Theorem. In old days, there were enough nerds who believed that this result is intriguing.

We may not believe that the hitting time problem is a big deal. Yet when it is linked with stock price, you may change your mind. If you find some neat estimates of such probabilities for very general stochastic processes, you will be famous. In this course, we provide such a result for the simple random work. There is a similar result for Brownian motion to be discussed.

Similarly, arc sine law also have its twin in Brownian motion. Please do not go away and stay tuned.

Chapter 3

Generating Functions and Their Applications

The idea of generating functions is to transform the function under investigation from one functional space to another functional space. The new functional space might be more convenient to work with.

If X is a non-negative integer valued random variable, then we call

$$G_X(s) = E(s^X)$$

the probability generating function of X.

Advantages? It can be seen that moments of X equal derivatives of $G_X(s)$ at $s = 1$.

When X and Y are independent, then the probablity generating function of $X + Y$ is

$$G_{X+Y}(s) = G_X(s)G_Y(s).$$

Theorem 3.1

If $X_1, X_2, \ldots$ is a sequence of independent and identically distributed random variables with common generating function $G_X(s)$ and $N(\geq 0)$ is a random variable which is independent of the X_i's and has generating function $G_N(s)$, then

$$S = X_1, +X_2 + \cdots + X_N$$

has generating function given by

$$G_S(s) = G_N(G_X(s)).$$

♢

The proof can be done through conditioning on N. When $N = 0$, we assume that $S = 0$.

Definition 3.2

The joint probability generating function of two random variables X_1 and X_2 taking non-negative integer values, is given by

$$G_{X_1,X_2}(s_1, s_2) = E[s_1^{X_1} s_2^{X_2}].$$

♢

Theorem 3.2

X_1 and X_2 are independent if and only if

$$G_{X_1,X_2}(s_1, s_2) = G_{X_1}(s_1)G_{X_2}(s_2).$$

♢

The idea of generating function also applies to a sequence of real numbers. If $\{a_n\}_{n=0}^{\infty}$ is a sequence of real numbers, then

$$A(s) = \sum_n a_n s^n$$

is called the generating function of $\{a_n\}_{n=0}^{\infty}$ when it converges in a neighborhood of $s = 0$.

Suppose that $\{a_n\}_{n=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are two sequence of real numbers, and their generating functions exists in a neighborhood of $s = 0$. Define

$$c_n = \sum_{i=0}^{n} a_i b_{n-i}$$

for $n = 0, 1, \ldots$. Then

$$C(s) = A(s)B(s)$$

where $A(s), B(s)$ and $C(s)$ are generating functions of $\{a_n\}_{n=0}^{\infty}$, $\{b_n\}_{n=0}^{\infty}$ and $\{c_n\}_{n=0}^{\infty}$ respectively.

3.1 Renewal Events

Consider a sequence of trials with outcomes $X_1, X_2, \ldots$. We do not require X_i's be independent of each other. Let λ represent some property which, on the basis of the outcomes of the first n trials, can be said unequivocally to occur or not to occur at trial n. By convention, we suppose that λ has just occurred at trial 0, and E_n represents the "event" that λ occurs at trial n, $n = 1, 2, \ldots$.

Roughly speaking, a property is a renewal event if the waiting time distribution for the next occurrence remains the same, and is independent of the past, given that it has just occurred. Thus, the process renewals itself each time when the renewal event occurs. It re-sets the clock back to time 0.

Let λ represent a renewal event and as before define the *lifetime sequence* $\{f_n\}$ where $f_0 = 0$ and

$$f_n = P\{\lambda \text{ occurs for the first time at trial } n\}, \quad n = 1, 2, \ldots.$$

In like manner, we define the renewal sequence u_n, where $u_0 = 1$ and

$$u_n = P\{\lambda \text{ occurs at trial } n\}, \quad n = 1, 2, \ldots.$$

Let $F(s) = \sum f_n s^n$ and $U(s) = \sum u_n s^n$ be the generating functions of $\{f_n\}$ and $\{u_n\}$. Note that

$$f = \sum f_n = F(1) \leq 1$$

because f has the interpretation that λ recurs at some time in the sequence. Since the event may not occur at all, it is possible for f to be less than 1. Clearly, $1 - f$ represents the probability that λ never recurs in the infinite sequence of trials. When $f < 1$, the probability that λ occurs finite number of times only is 1. Hence, we say that λ is **transient**. Otherwise, it it **recurrent**.

For a recurrent renewal event, $F(s)$ is a probability generating function. The mean inter-occurrence time is

$$\mu = F'(1) = \sum_{n=0}^{\infty} n f_n.$$

If $\mu < \infty$, we say that λ is **positive recurrent**. If $\mu = \infty$, we say that λ is **null recurrent**.

Finally, if λ can occur only at $n = t, 2t, 3t, \ldots$ for some positive integer $t > 1$, we say that λ is **periodic** with period t. More formally, let $t = g.c.d.\{n : \ f_n > 0\}$. (g.c.d. stands for the greatest common divisor). If $t > 1$, the recurrent event λ is said to be periodic with period t. If $t = 1$, λ is said to be **aperiodic**.

For a renewal event λ to occur at trial $n \geq 1$, either λ occurs for the first time at n with probability $f_n = f_n u_0$, or λ occurs for the first time at some intermediate trial $k < n$ and then occurs again at n. The probability of this event is $f_k u_{n-k}$. Notice that $f_0 = 1$, we therefore have

$$u_n = f_0 u_n + f_1 u_{n-1} + \cdots + f_{n-1} u_1 + f_n u_0, \quad n = 1, 2, \ldots.$$

This equation is called **renewal equation**. Using the typical generating function methodology, we get

$$U(s) - 1 = F(s)U(s).$$

Hence

$$U(s) = \frac{1}{1 - F(s)} \quad or \quad F(s) = 1 - \frac{1}{U(s)}.$$

Theorem 3.3

The renewal event λ is

1. transient if and only if $u = \sum u_n = U(1) < \infty$,
2. recurrent if and only if $u = \infty$,
3. periodic if $t = g.c.d.\{n : u_n > 0\}$ is greater than 1 and aperiodic if $t = 1$.
4. null recurrent if and only if $\sum u_n = \infty$ and $u_n \to 0$ as $n \to \infty$.

The following is the famous renewal theorem.

Theorem 3.4 *(The renewal theorem).*

Let λ be a recurrent and aperiodic renewal event and let

$$\mu = \sum n f_n = F'(1)$$

be the mean inter-occurrence time. Then

$$\lim_{n\to\infty} u_n = \mu^{-1}.$$

The proof of the Renewal Theorem is rather involved. We will put it as a section here. To the relief of many, this section will only for the sake of those who are nerdy enough in mathematics.

3.2 Proof of the Renewal Theorem

I am using my own language to restate the results of Feller (1968, page 335).

First, the result can be stated without the renewal event background.

Equivalent result: *Let $f_0 = 0, f_1, f_2, \ldots$ be a sequence of non-negative numbers such that $\sum f_n = 1$, and 1 is the greatest common divisor of these n for which $f_n > 0$. Let $u_0 = 1$ and*

$$u_n = f_0 u_n + f_1 u_{n-1} + \cdots + f_n u_0$$

for all $n \geq 1$.

Then $u_n \to \mu^{-1}$ as $n \to \infty$ where $\mu = \sum n f_n$ (and $\mu^{-1} = 0$ when $\mu = \infty$).
$\diamondsuit$

Note this definition of u_n does not apply to the case of $n = 0$. Otherwise, the u-sequence would be a convolution of f-sequence and itself.

The implication of being aperiodic is as follows. Let A be the set of all integers for which $f_n > 0$, and denote by A^+ the set of all positive linear combinations

$$p_1 a_1 + p_2 a_2 + \cdots + p_r a_r$$

of numbers $a_1, \ldots, a_r$ in A.

Lemma 3.1

There exists an integer N such that A^+ contains all integers $n > N$. $\diamondsuit$

In other words, if the greatest common divisor of a set of positive integers is 1, then the set of their linear combinations contains all but finite number positive integers. To reduce the burden in math respect, this result will not be proved here. Some explanation will be given in class.

The next lemma has been re-stated in different words.

Lemma 3.2

Suppose we have a sequence of sequences: $[\ \{(a_{m,n})_{m=1}^{\infty}\}_{n=1}^{\infty}]$ such that $0 \leq a_{m,n} \leq 1$ for all m, n. There exists a sequence $\{n_i : i = 1, 2, \ldots\}$ such that $\lim_{i\to\infty} a_{m,n_i}$ exists for each m. $\diamondsuit$

Again, the proof will be omitted. This result is also used when proving a result about the relationship between convergence in distribution and the convergence of characteristic functions.

Lemma 3.3

Let $\{w_n\}_{n=-\infty}^{\infty}$ be a doubly infinite sequence of numbers such that $0 \leq w_n \leq 1$ and

$$w_n = \sum_{k=1}^{\infty} f_k w_{n-k}$$

for each n. If $w_0 = 1$ then $w_n = 1$ for all n. $\diamondsuit$

To be more explicit, the sequence f_n is assumed to have the same property as the sequence introduced earlier.

PROOF: Since all numbers involved are non-negative and $w_n \leq 1$, we have that for all n

$$w_n = \sum_{k=1}^{\infty} f_k w_{n-k} \leq \sum f_k = 1.$$

Applying this conclusion to $n = 0$ with the condition $w_0 = 1$, we have $w_{-k} = 1$ whenever $f_k = 0$.

Recall the definition of A, the above conclusion can be restated as

$$w_{-a} = 0 \ \text{ whenever } \ a \in A.$$

For each $a \in A$, using the same argument:

$$w_n = \sum_{k=1}^{\infty} f_k w_{n-k} \leq \sum f_k = 1$$

to $n = a$, we find $w_{-a-k} = 1$ whenever both $a, k \in A$.

It can then be strengthened to $w_{-a} = 1$ for any $a \in A^+$ by induction.

An implication is then: there exists an N such that

$$a_{-N} = 1, a_{-N-1} = 1, a_{-N-2} = 1, \ldots.$$

Using

$$w_n = \sum_{k=1}^{\infty} f_k w_{n-k} \leq \sum f_k = 1$$

again for $n = -N + 1$, we get $w_{-N+1} = 1$.

Repeat it, we get $w_{-N+2} = 1$. Repeat it again and again, we find $w_n = 1$ for all n. ♢

Finally, we prove the renewal theorem.

Proof of Renewal Theorem: Our goal is to prove the limit of u_n exists and equals μ^{-1}. Since the items in this sequence are bounded, a subsequence can be found such that this subsequence converges to some number η. Assume this η is the upper limit.

Notationally, let us denote it as $u(i_v) \to \eta$ as $v \to \infty$.

Let $u_{n,v}$ be a double sequence such that for each v, $u_{n,v} = u_{n+i_v}$ when $n + i_v \geq 0$; and $u_{n,v} = 0$ otherwise. In other words, it shifts the original sequence by i_v positions toward left, and further adding zeroes to make a double sequence.

For example, suppose $i_5 = 10$, then we have

$$\ldots, u_{-2,5} = u_8, u_{-1,5} = u_9, u_{0,5} = u_{10}, u_{1,5} = u_{11}, \ldots.$$

If we set $n = 0$ and let v increases, the resulting sequence is

$$\ldots, u(0,0) = u(i_0), u_(0,1) = u(i_1), u_(0,2) = u(i_2), \ldots.$$

Hence, $\lim_{v\to\infty} u_{0,v} = \eta$.

For the sequence with $n = \pm 1$, they look like

$$\ldots, u(1,0) = u_{i_0+1}, u(1,1) = u_{i_1+1}, u(1,2) = u_{i_2+1}, \ldots;$$
$$\ldots, u(-1,0) = u_{i_0-1}, u(-1,1) = u_{i_1-1}, u(-1,2) = u_{i_2-1} \ldots;$$

with the additional claus that if $n < -i_v$, $u(n,v) = 0$.

In summary, for each $n = 0, \pm 1, \pm 2, \ldots$, we have a sequence $u(n,v)$ in v. By the earlier lemma, it is possible to find a subsequence in v, say $v_j, j = 1, 2, \ldots$ such that

$$\lim_{j\to\infty} u_{n,v_j}$$

exists for each n. Call this limit w_n.

Since for all $n > -v$, we have

$$u(n,v) = \sum_{k=1}^{\infty} f_k u(n-k,v).$$

Taking limit along the path of $v = v_j$, we get

$$w_n = \sum_{k=1}^{\infty} f_k w_{n-k}.$$

By the other lemma, we then get $w_n = \eta$ for all n.

After such a long labour, we have only achieve that the limit of each $u_{n,v_j} = u_{v_j+n}, j = 1, 2, \ldots$ are the same. What we wanted, however, is that the limit of u_n exists and equal μ^{-1}.

To aim a bit low, let us show that $\eta = \mu^{-1}$. This is easy when $\mu = \infty$. Let

$$\rho_k = f_{k+1} + f_{k+2} + \cdots$$

for $k = 0, 1, \ldots$, we have $sum\rho_k = \mu$. We should have seen it before, but if you do not remember, it is a good opportunity for us to redo it with generating function concept. Another note is that

$$1 - \rho_k = f_1 + f_2 + \cdots + f_k.$$

Let us now list the defining relations as follows:

$$\begin{aligned}
u_1 &= f_1 u_0; \\
u_2 &= f_2 u_0 + f_1 u_1; \\
u_3 &= f_3 u_0 + f_2 u_1 + f_1 u_2; \\
\ldots &= \ldots; \\
u_N &= f_N u_0 + f_{N-1} u_1 + f_{N-2} u_2 + \cdots + f_1 u_{N-1}.
\end{aligned}$$

Adding up both sides, we get

$$\sum_{k=1}^{N} u_k = (1-\rho_N)u_0 + (1-\rho_{N-1})u_1 + (1-\rho_{N-2})u_2 + \cdots + (1-\rho_1)u_{N-1}.$$

This is the same as

$$u_N = (1-\rho_N)u_0 - \rho_{N-1}u_1 - \rho_{N-2}u_2 - \cdots - \rho_1 u_{N-1}.$$

Noting that $u_0 = 1$, we have

$$\rho_N u_0 + \rho_{N-1}u_1 + \cdots + \rho_1 u_{N-1} + \rho_0 u_N = 1. \tag{3.1}$$

Recall that

$$\lim_{j\to\infty} u_{n,v_j} = w_n = \eta$$

for each n. At the same time, $u_{n,v_j} = u_{v_j+n}$ is practically true for all n (or more precisely for all large n). Let $N = v_j$, and let $j \to \infty$, (3.1), implies

$$\eta \times \{\sum \rho_k\} = 1.$$

That is, either $\eta = \mu^{-1}$ when μ is finite, or 0 otherwise.

Even with this much trouble, our proof is not complete yet. By definition, η is the upper limit of u_n. Is it also THE limit of u_n? If its limit exists, then the answer is yes. If $\mu = \infty$, then answer is also yes because the lower limit cannot be smaller than 0. Otherwise, we have more work to do.

If the limit of u_n does not exist, it must has a subsequence whose limit exists and smaller than η. Let it be $\eta_0 < \eta$. Being the upper limit of u_n, it implies that given any small positive quantity ϵ, for all large enough n, $u_n \le \eta + \epsilon$. Let us examine the relationship:

$$\rho_N u_0 + \rho_{N-1}u_1 + \cdots + \rho_1 u_{N-1} + \rho_0 u_N = 1$$

again. By letting some u_k replace by a larger value 1, it becomes

$$\sum_{k=0}^{N-r-1} \rho_{N-k} + \sum_{k=N-r}^{N-1} \rho_{N-k}u_k + \rho_0 u_N \ge 1.$$

By including more terms in the first summation, it becomes

$$\sum_{k=r+1}^{\infty} \rho_k + \sum_{k=N-r}^{N-1} \rho_{N-k} u_k + \rho_0 u_N \geq 1.$$

Replacing u_k by a larger value $\eta + \epsilon$ which is true for large k, we have

$$\sum_{k=r+1}^{\infty} \rho_k + (\eta + \epsilon) \sum_{k=1}^{r} \rho_k + \rho_0 u_N \geq 1.$$

Further, since $\sum_{k=0}^{\infty} \rho_k = \mu$, we get

$$\sum_{k=r+1}^{\infty} \rho_k + (\eta + \epsilon)(\mu - \rho_0) + \rho_0 u_N \geq 1.$$

Re-organize terms slightly, we get

$$\sum_{k=r+1}^{\infty} \rho_k + (\eta + \epsilon)\mu + \rho_0(u_N - \eta - \epsilon) \geq 1.$$

At last, let N go to infinite along the subsequence with limit η_0, we get

$$\sum_{k=r+1}^{\infty} \rho_k + (\eta + \epsilon)\mu + \rho_0(\eta_0 - \eta - \epsilon) \geq 1.$$

Since this result applies to any r and ϵ, let $r \to \infty$ and $\epsilon \to 0$, we get

$$\eta\mu + \rho_0(\eta_0 - \eta) \geq 1.$$

Because $\eta\mu = 1$, we must have $\eta_0 - \eta \geq 0$. Combined with the fact that η is the upper limit, we must have $\eta_0 = \eta$ or there will be a contradiction.

Since the limit of any subsequence of u_n must be the same as η, the limit of u_n exists and equals η.

Thus, we finally proved the Renewal Theorem.

3.3 Properties of Random Walks by Generating Functions

.

3.3.1 Quick derivation of some generating functions

The generating functions for certain sequences:

For "first returning to 0": $F(s) = 1 - (1 - 4pqs^2)^{1/2}$.

For "returning to 0": $U(s) = (1 - 4pqs^2)^{-1/2}$.

For "first passage of 1": $\Lambda(s) = (2qs)^{-1}[1 - (1 - 4pqs^2)^{1/2}]$.

Consequences:

$$P(\text{ever returns to 0}) = 1 - |p - q|.$$

$$P(\text{ ever rearches 1}) = \min(1, p/q).$$

(We will go through a lot of details in the class)

3.3.2 Hitting time theorem

.

Hitting time theorem is more general than what has been shown. Consider the random walk such that

$$S_n = X_1 + X_2 + \cdots + X_n,$$

$$P(X \leq 1) = 1$$

and $P(X = 1) > 0$. This is so called right-continuous random walk. The random walk cannot skip over (up-ward) any state without stepping on it.

Let T_b be the first time when $S_n = b$ for some $b > 0$ and assume $S_0 = 0$. It has been shown for simple random walk that

$$f_b(n) = P(T_b = n) = \frac{b}{n} P(S_n = n).$$

Let us see why this is still true for the new random walk we have just defined.

For this purpose, define

$$F_b(z) = E(z^{T_b})$$

which is the generating function of T_b. It may happen that $P(T_b = \infty) > 0$, in which case, it is more precise to define

$$F_b(z) = \lim_{n\to\infty} \sum_{i=1}^{n} z^n P(T_b = n).$$

We also define

$$G(z) = E(z^{1-X_1}).$$

Since X_1 can be negative with large absolute values, work with generating function of $1-X_1$ makes sense. Since $1-X_1$ is non-negative, G is well defined for all $|z| \leq 1$. Please note that our G here is a bit different from G in the textbook.

The purpose of using letter z, rather than s, is to allow z to be a complex number. It seems that defining a function without a value at $z = 0$ calls for some attention. We will pretend z is just a real number for illustration.

If we ignore the mathematical subtlety, our preparations boil down to have defined generating functions of T_b and X_1. Due to the way S_n is defined, the "generating function" of S_n is determined by $[G(z)]^n$. Hence the required **hitting time theorem** may be obtained by linking $G(z)$ and $F_b(z)$.

This step turns out to be rather simple. It is seen that

$$F_b(z) = [F_1(z)]^b$$

as the random walk cannot skip a state without landing on it. We take notice that this relationship works even when $b = 0$.

Further,

$$F_1(z) = E[E(z^{T_1}|X_1)] = E[z^{1+T_{1-X_1}}|X_1] = zE\{[F_1(z)]^{1-X_1}\} = zG(F_1(z)).$$

Denote $w = F_1(z)$, this relation can then be written as

$$z = \frac{w}{G(w)}$$

That is, it is known that $w = F_1(z)$. At the same time, its relationship with w is also completely determined by the above equation. The two versions must be consistent with each other. The result to be proved is a consequence of this consistency.

It turns out there is a complex analysis theorem for inversing $z = w/G(w)$.

Theorem 3.5 *Lagranges' inversion formula*

Let $z = w/f(w)$ where $f(w)$ is an analysis function (has derivative in complex analysis sense) in a neighborhood of $w = 0$. (and a one-to-one relationship between z and w in this neighborhood.). If g is infinitely differentiable, (but I had the impression being analytic implies infinitely differentiable), then

$$g(w(z)) = g(0) + \sum_{n=1}^{\infty} \frac{z^n}{n!} \left[\frac{d^{n-1}}{du^{n-1}} [g'(u)\{f(u)\}^n] \right]_{u=0}.$$

(See Kyrala (1972), Applied functions of a complex variable. Page 51 for a close resemble of this result. Related to Cauchy integration over a closed path). ♢

Apply this result with $f(w)$ being our $G(w)$ and $g(w) = w^b$, we have

$$[F_1(z)]^b = \sum_{n=1}^{\infty} \frac{z^n}{n!} \left[\frac{d^{n-1}}{du^{n-1}} [bu^{b-1} G^n(u)] \right]_{u=o}.$$

Let us not forget that $[F_1(z)]^b$ is the "probability" generating function of T_b, and $G^n(u)/z^n$ is the probability generating function of S_n.

Matching the coefficient of z^n, we have

$$n! P(T_b = n) = \left[\frac{d^{n-1}}{du^{n-1}} [bu^{b-1} G^n(u)] \right]_{u=0}.$$

Notice the latter is $(n-1)!$ times of the coefficient of u^{n-1} in the power expansion of $bu^{b-1}G^n(u)$, which in turn, is the coefficent of u^{n-b} in the power series expansion of $bG^n(u)$, and which is the coefficient of u^{-b} in the expansion of $bu^{-n}G^n(u)$.

Now, we point out that $u^{-n}G^n(u) = Eu^{-S_n}$. That is, $(n-1)!$ times of this coefficient is $b(n-1)!P(S_n = b)$. Hence we get the result. ♢

3.3.3 Spitzer's Identity

Here we discuss the magical results of Spitzer's identity. I kind of believe that a result like this can be useful in statistics. Yet I find that it is still too complex to be practical at the moment.

Theorem 3.6

Assume that S_n is a right-continuous random walk, and let $M_n = \max\{S_i : 0 \le i \le n\}$ be the maximum of the walk up to time n. Then for $|s|, |t| < 1$,

$$\log\left(\sum_{n=0}^{\infty} t^n E(S^{M_n})\right) = \sum_{n=1}^{\infty} \frac{1}{n} t^n E(s^{S_n^+})$$

where $S_n^+ = \max\{0, S_n\}$ (which is the non-negative part of S_n). ♢

PROOF: We keep the notation $f_b(n) = P(T_b = n)$ as before. To have $M_n = b$, it has to land on b at some time j between 1 and n, and at the same time, the walk does not land on $b+1$ within another $n-j$ steps. Thus,

$$\begin{aligned} P(M_n = b) &= \sum_{j=0}^{n} P(T_b = j) P(T_1 > n - j) \\ &= \sum_{j=0}^{n} P(T_b = j) P(T_1 > n - j). \end{aligned}$$

Multiply both sides by $s^b t^n$ and sum over $b, n \ge 0$. The left hand side is

$$\sum_{n=0} t^n E(s^{M_n})$$

Recall $\sum t^n P(T_1 > n) = \frac{1-F_1(t)}{1-t}$. The right hand side is, by convolution relationship,

$$\begin{aligned} \sum_{b=0}^{\infty} s^b E(t^{T_b}) \frac{1 - F_1(t)}{1-t} &= \frac{1 - F_1(t)}{1-t} \sum_{b=0}^{\infty} s^b [F_1(t)]^b \\ &= \frac{1 - F_1(t)}{(1-t)(1 - sF_1(t))}. \end{aligned}$$

Denote this function as $D(s,t)$.

The rest contains mathematical manipulation: By Hitting time theorem,

$$nP(T_1 = n) = P(S_n = 1) = \sum_{j=0}^{n} P(T_1 = j) P(S_{n-j} = 0)$$

and we get generating function relationship as

$$tF_1'(t) = F_1(t)U(t).$$

(Recall $U(s)$ is the g.f. for returning to 0).

With this relationship, we have

$$\begin{aligned}
\frac{\partial}{\partial t}\log[1-sF_1(t)] &= \frac{-sF_1'(t)}{1-sF_1(t)} \\
&= -\frac{s}{t}F_1(t)U(t)\sum_k s^k[F_1(t)]^k \\
&= -\sum_{k=0}^{\infty}\frac{s^{k+1}}{t}[F_1(t)]^{k+1}U(t) \\
&= -\sum_{k=1}^{\infty}\frac{s^k}{t}[F_1(t)]^k U(t) \\
&= -\sum_{k=1}^{\infty}\frac{s^k}{t}[F_1(t)]^k U(t)
\end{aligned}$$

Notice

$$P(S_n = k) = \sum_{j=0}^{n} P(T_k = j)P(S_{n-j} = 0).$$

Thus, the generating function $[F_1(t)]^k U(t)$ is in fact a generating function for the sequence of $P(S_n = k)$ in n. That is,

$$[F_1(t)]^k U(t) = \sum_{n=0}^{\infty} t^n P(S_n = k).$$

Notice that the sum is over n.

It therefore implies

$$\frac{\partial}{\partial t}\log[1-sF_1(t)] = -\sum_{n=1}^{\infty} t^{n-1}\sum_{k=1}^{\infty} s^k P(S_n = k).$$

So, we get

$$\begin{aligned}
\frac{\partial}{\partial t}\log D(s,t) &= -\frac{\partial}{\partial t}\log(1-t) + \frac{\partial}{\partial t}\log[1-F_1(t)] - \frac{\partial}{\partial t}\log[1-sF_1(t)] \\
&= \sum_{i=1}^{\infty} t^{n-1}\left(1 - \sum_{k=1}^{\infty} P(S_n = k) + \sum_{k=1}^{\infty} s^k P(S_n = k)\right) \\
&= \sum_{i=1}^{\infty} t^{n-1}\left(P(S_n \le 0) + \sum_{k=1}^{\infty} s^k P(S_n = k)\right) \\
&= \sum_{i=1}^{\infty} t^{n-1} E(s^{S_n^+}).
\end{aligned}$$

Now integrate with respect to t to get the final result.

$\diamondsuit$

Remark: I do not see why this final result is more meaningful than the intermedian result. I am wondering if any of you can offer some insight.

3.3.4 Leads for tied-down random walk

Let us now go back to the simple random walk such that $S_0 = 0$ and $P(X_i = 1) = p$, $P(X_i = -1) = q = 1 - p$.

Define L_{2n} as the number of steps when the random walk was above the line of 0 (but is allowed to touch 0). We have an arc-sin law established for this random variable already. This time, we investigate its property when $S_{2n} = 0$ is given.

Theorem 3.7 *Leads for tied-down random walk.*

For the simple random walk S,

$$P(L_{2n} = 2k | S_{2n} = 0) = \frac{1}{2n+1},$$

for $k = 0, 1, \ldots, n$. $\diamondsuit$

Remark: we may place more possibility at $L_{2n} = n$. It should divide the time evenly above and below zero. This theorem, however, claims that a uniform distribution is the truth. Note also that the result does not depend on the value of p.

PROOF: Define T_0 be the first time when S is zero again. Given $T_0 = 2r$ for some r, L_{2r} can either be 0 or $2r$ as the simple random walk does not cross zero. Making use of the results on the size of λ_{2r}, and f_{2r}, it turns out the conditional distribution is placing half and half probabilities on 0 and $2r$. Thus,

$$E(s^{L_{2r}} | S_{2n} = 0, T_0 = 2r) = \frac{1}{2} + \frac{1}{2} s^{2r}.$$

Now we define $G_{2n}(s) = E[S^{L_{2n}} | S_{2n=0}$ and $F_0(s) = E(S^{T_0})$. Further, let

$$H(s,t) = \sum_{n=0}^{\infty} t^{2n} P(S_{2n} = 0) G_{2n}(s).$$

Conditioning on T_0, we have

$$\begin{aligned} G_{2n}(s) &= \sum_{r=1}^{n} E[S^{L_{2n}}|S_{2n}=0, T_0=2r]P(T_0=2r|S_{2n}=0) \\ &= \sum_{r=1}^{n} (\frac{1}{2}s^{2r})P(T_0=2r|S_{2n}=0). \end{aligned}$$

Also

$$P(T_0=2r|S_{2n}=0) = \frac{P(T_0=2r)P(S_{2n-2r}=0)}{P(S_{2n}=0)}.$$

Hence,

$$H(s,t) - 1 = \frac{1}{2}H(s,t)[F_0(t) + F_0(st)].$$

Recall that

$$F_0(t) = (1-4pqt^2)^{-1/2},$$

one obtains

$$\begin{aligned} H(s,t) &= \frac{2}{\sqrt{1-t^2}+\sqrt{1-s^2t^2}} \\ &= \frac{2[\sqrt{1-s^2t^2}-\sqrt{1-t^2}]}{t^2(1-s^2)} \\ &= \sum_{n=0}^{\infty} t^{2n}P(S_{2n}=0)\left(\frac{1-s^{2n+2}}{(n+1)(1-s^2)}\right). \end{aligned}$$

Compare the coefficient of t^{2n} in two expansions of $H(s,t)$, we deduce

$$G_{2n}(s) = \sum_{k=0}^{n}(n+1)^{-1}s^{2k}.$$

♢

3.4 Branching Process

We only consider a simple case where $Z_0 = 1$ representing a species starts from a single individual. It is assumed that each individaul will give birth of random number of offsprings independently of each other. We call it family size. We assume family sizes have the same distribution with probability

generating function $G(s)$. Let μ and σ^2 be the mean and variance of the family size.

Let the population size of the nth generation be called Z_n.

It is known that

$$E[Z_n] = \mu^n$$

$$Var(Z_n) = \frac{\sigma^2(\mu^n - 1)}{\mu - 1}\mu^{n-1}.$$

These relations can be derived from the identity:

$$G_n(t) = G(G_{n-1}(t))$$

where $G_n(t) = E(t^{Z_n})$. The same identity also implies that the probability of ultimate extinction $\eta = \lim P(Z_n = 0)$ is the smallest non-negative solution of the equation

$$x = G(x).$$

Because of this, it is known that:

1. if $\mu < 1$, $\eta = 1$;
2. if $\mu > 1$, $\eta < 1$;
3. if $\mu = 1$ and $\sigma^2 > 0$, then $\eta = 1$;
4. if $\mu = 1$ and $\sigma^2 = 0$, then $\eta = 0$.

Discussion will be given in class.

3.5 Summary

The biggest deal of this chapter is the proof of the Renewal Theorem. The proof is not much related to generating function at all. One may learn a lot from this proof. At the same time, it is okay to choose to ignore the proof completely.

One should be able to see the beauty of generating functions when handling problems related to the simple random walk and the branching process. At the same time, none of these discussed should be new to you. You may

realize that two very short sections on the simple random walk and brach-ing process are rich in content. Please take the opportunity to plant these knowledge firmly in you brain if they were not so before.

Chapter 4

Discrete Time Markov Chain

A discrete time Markov chain is a stochastic process consists of a countable number of random variables arranged in a sequence. Usually, we name these random variables as $X_0, X_1, X_2, \ldots$. To qualify as a Markov chain, its state space, S, which is the set of all possible values of these random variables, must be countable, In addition, it must have Markov property:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0) = P(X_{n+1} = j | X_n = i).$$

When this transition probability does not depend on n, we say that the Markov chain is time homogeneous.

The Markov chain discussed in this course will be assumed time homogeneous unless otherwise specified. In this case, we use notation

$$p_{ij} = P(X_1 = j | X_0 = i)$$

and $\mathbf{P}$ for the matrix with the i, jth entry being p_{ij}.

It is known that all entries of the transition matrix are non-negative, and its row sums are all 1.

Further, let $\mathbf{P}^{(m)}$ be the m-step transition matrix defined by

$$p_{ij}^{(m)} = P(X_m = j | X_0 = i).$$

Then they satisfy the Chapman-Kolmogorov equations:

$$\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)} \mathbf{P}^{(n)}$$

for all non-negative positive integers m and n. Here, we denote $\mathbf{P}^{(0)} = I$.

Let $\mu^{(m)}$ be the row vector so that its ith entry

$$\mu_i^{(m)} = P(X_m = i).$$

It can be seen that $\mu^{(m+n)} = \mu^{(m)}\mathbf{P}^n$.

Both simple random walk and branching process are special cases of discrete time Markov chain.

Example 4.1 *Markov's other chain.*

Note that the Chapman-Kolmogorov equation is the direct consequence of the Markov property. Suppose that a discrete time stochastic process has countable state space, and its transition probability matrix defined in the same way as for the Markov chain, satisfies the Chapman-Kolmogorov equation. Does it have to be a Markov chain?

In mathematics, we can offer a rigorous proof to show this is true, or offer an example a stochastic process with this property which is not a Markov chain. It turns out the latter is the case.

Let $Y_1, Y_3, Y_5, \ldots$ be a sequence of iid random variables such that

$$P(Y_1 = 1) = P(Y_1 = -1) = \frac{1}{2}.$$

Let $Y_{2k} = Y_{2k-1}Y_{2k+1}$ for $k = 1, 2, \ldots$. We hence have a stochastic process $Y_1, Y_2, Y_3, \ldots$ which has state space $\{-1, 1\}$.

Note that

$$E[Y_{2k}Y_{2k+1}] = E[Y_{2k-1}Y_{2k+1}^2] = E[Y_{2k-1}] = 0.$$

Since these random variables take only two possible values, having correlation 0 implies independence. All other non-neighbouring pairs of random variables are also independent of each other by definition.

Thus,

$$P(X_{m+n} = j | X_n = i) = P(X_{m+n} = j) = \frac{1}{2}$$

for any $i, j = \pm 1$. Thus, the m-step transition matrix is

$$\mathbf{P}^{(m)} = \mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

which is idempotent. It implies $\mathbf{P}^{(m+n)} = \mathbf{P}^{(m)}\mathbf{P}^{(n)}$ for all m and n.

That is, the Chapman-Kolmogorov equation is satisfied. However,

$$P(Y_{2k+1} = 1|Y_{2k} = -1) = \frac{1}{2}$$

but

$$P(Y_{2k+1} = 1|Y_{2k} = -1, Y_{2k-1} = 1) = 0.$$

Thus, this process does not have Markov property and is not a Markov chain. $\diamondsuit$

Although this example shows that a stochastic process with transition probability matrices satisfying the Chapman-Kolmogorov equations is not necessarily a Markov chain. When we are given a set of stochastic matrices satisfying Chapman-Kolmogorov equations, it is always possilbe to construct a Markov chain with this set of stochastic matrices as its transition probability matrices.

4.1 Classification of States and Chains

We use i, j, k and so on to denote states in state space.

If there exists an integer $m \geq 0$ such that $p_{ij}^{(m)} > 0$, we say that j is reachable from i.

If i is reachable from j and j is reachable from i, then we say that i and j communicate.

It is obvious that communication is an equivalence relationship. Thus, the state space is partitioned into classes so that any pair of states in the same class communicate with each other. Any pair of states from two different classes do not communicate.

It is usually very simple to classify the state space. In most examples, it can be done by examining the transition probability matrix of the Markov chain.

Now let us consider individual classes. Let G be a class. If G consists of all states of the Markov chain, we say that the chain is irreducible.

If for any state $i \in G$ and state $j \notin G$, $p_{ij} = 0$. Then the Markov chain can never leave class G once it enters. That is, once the value of X_n is in G

for some n, then the values of X_{n+m}'s are all in G when $m \geq 0$. A class G is closed when it has this property. Notice that given $X_n \in G$ for some n, the Markov chain $\{X_n, X_{n+1}, \ldots\}$ has effectively G as its state space. Thus, the state space is reduced when G is a true subset of the state space.

In contrast to a closed class, if there exist states $i \in G$ and $j \notin G$ such that $p_{ij} > 0$, then the class is said open.

4.2 Class Properties

It turns out that the states in the same class have some common properties. We call them class properties.

Due to Markov property, the event $X_n = i$ for any i is a renewal event. Some properties that describe renewal events are positive or null recurrent, transient, periodicity.

Definition 4.1 *Property of renewal events*

1. A renewal event is recurrent(persistent) if the probability for its future occurrence is 1, given its occurrence at time 0.

2. If a renewal event is not recurrent, then it is called transient.

3. If the expected waiting time for the next occurrence of a recurrent renewal event is finite, then the renewal event is positive recurrent. Otherwise, it is null-recurrent.

4. Let A be a renewal event and $\mathcal{T} = \{n : P(A \ occurs \ at \ n | A \ occurs \ at \ 0) > 0\}$. The period of A is d, the greatest common divisor of $\mathcal{T}$. If $d = 1$, then A is aperiodic.

$\diamondsuit$

Due to the fact that entering a state is a renewal event, some properties of renewal event can be translated easily here. Let

$$P_{ij}(s) = \sum_{i=1}^{\infty} s^n p_{ij}(n)$$

as the generating function of the n-step transition probability. Define

$$f_{ij}(n) = P(X_1 \neq j, \ldots, X_{n-1} \neq j | X_0 = i)$$

be the probability of entering j from i for the first time at time n. Let its corresponding generating function be $F_{ij}(s)$. We let

$$f_{ij} = \sum_n f_{ij}(n)$$

for the probability that the chain ever enters state j starting from i. Note that $f_{ij} = F_{ij}(1)$.

Lemma 4.1

(a): $P_{ii}(s) = 1 + F_{ii}(s)P_{ii}(s)$.
(b): $P_{ij}(s) = F_{ij}(s)P_{jj}(s)$ if $i \neq j$. ♢

PROOF: The first property is the renewal equation. The second one is the delayed renewal equation. ♢

Based on these two equations, it is easy to show the following.

Theorem 4.1

(a) State j is recurrent if and only if $\sum_j p_{jj}(n) = \infty$. Consequently, $\sum_n p_{ij}(n) = \infty$ when j is reachable from i

(b) State j is transient if and only if $\sum_j p_{jj}(n) < \infty$. Consequently, $\sum_n p_{ij}(n) < \infty$.

♢

PROOF: Using the generating function equations.

Theorem 4.2

. All properties in Definition 4.1 are class properties. ♢

PROOF: Suppose i and j are two states in the same class. State i is recurrent, and we want to show that state j is also recurrent.

Let $p_{ij}(m)$ be the m-step transition probability from i to j. Hence, $\sum_m p_{ii}(m) = \infty$. Since i and j communicate, there exist m_1 and m_2 such that $p_{ij}(m_1) > 0$ and $p_{ji}(m_2) > 0$. Hence, $p_{jj}(m+m_1+m_2) \geq p_{ji}(m_2)p_{ii}(m)p_{ij}(m_1)$. Consequently,

$$\begin{aligned} \sum_m p_{jj}(m) &\geq \sum_m p_{jj}(m+m_1+m_2) \\ &\geq p_{ji}(m_2)\{\sum_m p_{ii}(m)\}p_{ij}(m_1) \\ &= \infty. \end{aligned}$$

That is, j is also a recurrent state.

The above proof also implies that if i is transient, j cannot be recurrent. Hence, j is also transient.

We delegate the positive recurrentness proof to the future.

At last, we show that i and j must have the same period.

Define $\mathcal{T}_i = \{n : p_{ii}(n) > 0\}$ and similarly for $\mathcal{T}_j$. Let d_i and d_j be the periods of state i and j. Note that if $n_1, n_2 \in \mathcal{T}_i$, then $an_1 + bn_2 \in \mathcal{T}_i$ for any positive integers a and b. Since d_i is the greatest common divisor of $\mathcal{T}_i$, there exist integers $a_1, \ldots, a_m$ and $n_1, \ldots, n_m \in \mathcal{T}_i$ such that

$$a_1n_1 + a_2n_2 + \cdots + a_mn_m = d_i.$$

By grouping positive and negative coefficients. it is easy to see that the number of items m can reduced to 2. Thus, we assume that it is possible to find a_1, a_2 and n_1, n_2 such that

$$a_1n_1 + a_2n_2 = d_i.$$

We can then further pick non-negative coefficients such that

$$a_{11}n_1 + a_{12}n_2 = kd_i, \quad a_{21}n_1 + a_{22}n_2 = (k+1)d_i$$

for some positive integer k.

Let m_1 and m_2 be the number of steps the chain can go and return from state i to state j. Then, both

$$m_1 + m_2 + kd_i, m_1 + m_2 + (k+1)d_i \in \mathcal{T}_j.$$

Thus, we must have d_j divides d_i. The reverse is also true by symmetry. Hence $d_i = d_j$. $\diamondsuit$

4.2.1 Properties of Markov Chain with Finite Number of States

If a Markov chain has finite state space, then one of the states will be visited infinite number of times over infinite time horizon. Thus, at least one of them will be recurrent (persistent). We can prove this result rigorously.

Lemma 4.2

If the state space is finite, then at least one state is recurrent and all recurrent states are positive recurrent. ♢

PROOF: A necessary condition for a recurrent event to be transient is $u_n \to 0$ as n increases. In the context of Markov chain, it implies that state j is transient implies $p_{ij}(n) \to 0$. Because

$$1 = \sum_j p_{ij}(n),$$

we cannot have $p_{ij}(n) \to 0$ for all j when the state space is finite. Hence, at least one of them is recurrent. ♢

Intuitively, the last conclusion implies that there is always a closed class of recurrent states for a Markov chain with finite state space. Since the Markov chain cannot escape from the finite class, the average waiting time for the next visit will be finite. Thus, at least one of them is recurrent and hence positive recurrent.

We skip the rigorous proof for now.

4.3 Stationary Distribution and the Limit Theorem

A Markov chain consists of a sequence of discrete random variables. For convenience, they are regarded as non-negative integers.

For each n, X_n is a random variable whose distribution may be different from that of X_{n-1} but is involved from it. When n is large, the dependence of the distribution of X_n on that of X_0 gets weaker and weaker. It is therefore

possible that the distribution stabilizes and it may hence have a limit. This limit turns out to exist in many cases and it is related to so called stationary distribution.

Definition 4.1

The vector π is called a stationary distribution of a Markov chain if π has entries $\pi_j, j \in S$ such that

(a) $\pi_j \geq 0$ for all j and $\sum_j \pi_j = 1$;

(b) $\pi = \pi P$ where P is the transition probability matrix. $\diamondsuit$

It is seen that if the probability function of X_0 is given by a vector called α, then the probability function of X_n is given by

$$\beta(n) = \alpha \mathbf{P}^n.$$

Consequently, if the distribution of X_n is given by π, then all distributions of X_{n+m} are given by π. Hence it gains the name of stationary.

It should be noted that if the distribution of X_n is given by π, the distribution of X_{n-m} does not have to be π.

Lemma 4.3

If the Markov chain is irreducible and recurrent, there exists a positive root $\mathbf{x}$ of the equation $\mathbf{x} = \mathbf{x}P$, which is unique up to a multiplicative constant. The chain is non-null if $\sum_i x_i < \infty$, and null if $\sum_i x_i = \infty$. $\diamondsuit$

PROOF: Assume the chain is recurrent and irreducible. For any states $k, i \in S$,

$$N_{ik} = \sum_n I(X_n = i, T_k \geq n)$$

with T_k being the time of the first return to state k. Note that T_k is a well defined random variable because $p(T_k < \infty) = 1$.

Define $\rho_i(k) = E[N_{ik}|X_0 = k]$ which is the mean number of visits of the chain to state i between two successive visits of state k.

By the way, if j is recurrent, then $\mu_j = E[T_j]$ is called the mean recurrent time. It is allowed to be infinity when the summation in the definition of

expectation does not converge. When j is transient, we simply define $\mu_j = \infty$. Thus, being positive recurrent is equivalent to claim that $\mu_j < \infty$.

Note that $\rho_k(k) = 1$. At the same time,

$$\rho_i(k) = \sum_n P(X_n = i, T_k \geq n | X_0 = k).$$

It will be seen that the vector ρ with $\rho_i(k)$ as its kth component is a base for finding the stationary distribution.

We first show that $\rho_i(k)$ are finite for all k. Let

$$l_{ki} = P(X_n = i, T_k \geq n | X_0 = k),$$

the probability that the chain reaches i in n steps but with no intermediate return to the starting point k.

With this definition, we see that

$$f_{kk}(m+n) \geq l_{ki}(m) f_{ik}(n)$$

in which the right hand side is one of the sample paths taking a roundtrip from k back to k in $m+n$ steps, without visiting k in intermediate steps. Because the chain is irreducible, there exists n such that $f_{ik}(n) > 0$. Now sum over m on two sides and we get

$$\sum_m f_{kk}(m+n) \geq f_{ik}(n) \sum_m l_{ki}(m).$$

Since $\sum_m f_{kk}(m+n) < \infty$ and $f_{ik}(n) > 0$, we get $\sum_m l_{ki}(m) < \infty$. That is, $\rho_i(k) < \infty$.

Now we move to the next step. Note that $l_{ki}(1) = p_{ki}$, the one-step transition probability. Further,

$$\begin{aligned} l_{ki}(n) &= \sum_{j:j\neq k} P(X_i, X_{n-1} = j, T_k \geq n | X_0 = k) \\ &= \sum_{j:j\neq k} l_{kj}(n-1) p_{ji} \end{aligned}$$

for $n \geq 2$. Now sum over $n \geq 2$, we have

$$\begin{aligned} \rho_i(k) &= p_{ki} + \sum_{j:j\neq k} \left\{ \sum_{n\geq 2} l_{kj}(n-1) \right\} p_{ji} \\ &= \rho_k(k) p_{ki} + \sum_{j:j\neq k} \rho_j(k) p_{ji} \end{aligned}$$

because $\rho_k(k) = 1$. We have shown that ρ is a solution to $\mathbf{x}P = \mathbf{x}$.

This proves the existence part of the lemma. We prove uniqueness next.

First, if a chain is irreducible, we cannot have a non-negative solution for $\mathbf{x}P = \mathbf{x}$ with zero-components. If so, write $\mathbf{x} = (0, \mathbf{x}_2)$ and

$$P = \begin{bmatrix} P_{11} & P_{21} \\ P_{21} & P_{22} \end{bmatrix}.$$

Then, we must have $\mathbf{x}_2 P_{21} = 0$. Since all components in $\mathbf{x}_2$ are positive, we must have $P_{21} = 0$. This contradicts the irreducibility assumption.

If there are two solutions to $\mathbf{x}P = \mathbf{x}$ who are not multiplicative of each other, then we can obtain another solution with non-negative components and at least one 0 entry. This will contradicts the irreducibility assumption. (This proof works only if there are finite number of states. The more general case will be proved later).

Due to the uniqueness, for any solution $\mathbf{x}$, we must have for any state k,

$$\mu_k = c \sum x_i.$$

Thus, if k is positive recurrent, we have $\sum x_i < \infty$.

This completes the proof of the lemma. ♢

We have an immediate corollary.

Corollary 4.1

If states i and j communicate and state i is positive recurrent, then so is state j. ♢

One intermediate results in the proof of the theorem can be summarised as another lemma.

Lemma 4.4

For any state k of an irreducible recurrent Markov chain, the vector $\rho(k)$ satisfies $\rho_i(k) < \infty$ for all i, and furthermore $\rho(k) = \rho(k)P$. ♢

The above lemma shows why we did not pay much attention on when a Markov chain is positive recurrent. Once we find a suitable solution to $\mathbf{x} = \mathbf{x}P$, we know whether it is positive recurrent or not immediately.

The next theorem summary these results to give a conclusion on the existence of stationary distribution.

Theorem 4.3

An irreducible Markov chain has a stationary distribution π if and only if all the states are positive recurrent; in this case, π is the unique stationary distribution and is given by $\pi_j = \mu_j^{-1}$ for each $i \in S$, where μ_j is the mean recurrence time of j. ♢

PROOF. We have already shown that if the chain is positive recurrent, then the stationary distribution exists and is unique. If the chain is recurrent but not positive recurrent, then due to the uniqueness, there cannot be another solution $\mathbf{x}$ such that $\sum x_i < \infty$. Thus, there cannot exist a stationary distribution or it violates the uniqueness conclusion.

Now if the chain is transient, we have $p_{ij}(n) \to 0$ for all ij as $n \to \infty$. If a stationary distribution π exists, we must have

$$\pi_j = \sum_i \pi_i p_{ij}(n) \to 0$$

by dominate converges theorem. This contradicts the assumption that π is a stationary distribution.

Thus, all left to be shown is whether $\pi_j = \mu_j^{-1}$ for each $i \in S$ when the chain is positive recurrent.

Suppose X_0 has π as its distribution.. Then

$$\pi_j \mu_j = \sum_{n=1}^{\infty} P(T_j \geq n | X_0 = j) P(X_0 = j) = P(T_j \geq n, X_0 = j).$$

However, $P(T_j \geq 1, X_0 = j) = P(X_0 = j) = \pi_j$ and for $n \geq 2$,

$$\begin{aligned}
&P(T_j \geq n, X_0 = j) \\
&= P(X_0 = j, X_m \neq j \text{ for all } 1 \leq m \leq n-1) \\
&= P(X_m \neq j \text{ for } 1 \leq m \leq n-1) - P(X_m \neq j \text{ for } 0 \leq m \leq n-1) \\
&= P(X_m \neq j \text{ for } 0 \leq m \leq n-2) - P(X_m \neq j \text{ for } 0 \leq m \leq n-1) \\
&= a_{n-2} - a_{n-1}
\end{aligned}$$

with a_m defined as is.

Thus, sum over n we obtain

$$\pi_j \mu_j = P(X_0 = j) + P(X_0 \neq j) - \lim a_n = 1 - \lim a_n.$$

Since state j is recurrent,

$$a_n = P(X_m \neq j \text{ for all } 1 \leq m \neq n)$$

has limit 0. Hence we have shown $\pi_j = \mu_j^{-1}$. $\diamondsuit$.

Now we come back for the uniqueness again. We skipped its proof earlier. It turns out that the general proof is very tricky.

Suppose $\{X_n\}$ is an irreducible and recurrent Markov chain with transition probability matrix P. Let $\mathbf{x}$ be a solution to $\mathbf{x} = \mathbf{x}P$. We construct another Markov chain Y_n such that

$$q_{ij}(n) = P(Y_n = j | Y_0 = i) = \frac{x_j}{x_i} p_{ji}(n)$$

for all $i, j \in S$ and n. It is easy to verify that q_{ij} make a proper transition probability matrix. Further, it is obvious that $\{Y_n\}$ is also irreducible. By checking the sum over n, it further shows that $\{Y_n\}$ is recurrent.

Let, for $i \neq j$,

$$l_{ji}(n) = P(X_n = i, X_m \neq j \text{ for } 1 \leq m \leq n-1 | X_0 = j).$$

We show that

$$g_{ij}(n) = P(Y_n = j, Y_m \neq j, \text{ for } 1 \leq m \leq n-1 | Y_0 = i)$$

satisfies

$$g_{ij}(n) = \frac{x_j}{x_i} l_{ij}(n).$$

Let us examine the expressions of g_{ij} and l_{ji}. One is the probability of all the pathes which start from j and end up at state i before they ever visit j again. The other is the probability of all the pathes which start from i and end up at state j without being there in between. If we reverse the time of the second, then we are working with the same set of sample pathes.

For each sample path corresponding to $l_{ji}(n)$, let us denote it as

$$j, k_1, k_2, \ldots, k_{n-1}, i,$$

its probability of occurrence is

$$p_{j,k_1} p_{k_1,k_2} \cdots p_{k_{n-1},i}.$$

The reverse sample path for $g_{ij}(n)$ is $i, k_{n-1}, \ldots, k_2, k_1, j$ and its probability of occurrence is

$$\begin{aligned} & q_{i,k_{n-1}} q_{k_{n-1},k_{n-2}} \cdots q_{k_2,k_1} q_{k_1,j} \\ = \; & p_{k_{n-1},i} \cdots p_{k_1,k_2} p_{j,k_1} \times \frac{x_{k_{n-1}}}{x_i} \frac{x_{k_{n-2}}}{x_{k_{n-1}}} \cdots \frac{x_{k_1}}{x_{k_2}} \frac{x_{k_j}}{x_{k_1}} \\ = \; & \frac{x_i}{x_j} p_{k_{n-1},i} \cdots p_{k_1,k_2} p_{j,k_1} \end{aligned}$$

Since it is true for every sample path, the result is proved.

Note that $\sum_n g_{ij}(n)$ corresponds to the probability that the chain will ever enter state j and the chain is recurrent and irreducible, we have $\sum_n g_{ij}(n) = 1$. Consequently,

$$\frac{x_j}{x_i} = [\sum_n l_{ij}(n)]^{-1}$$

and hence the ratio is unique.

4.3.1 Limiting Theorem

We have seen that an irreducible Markov chain has a unique stationary distribution when all states are positive recurrent. It turns out that the limits of the transition probabilities $p_{ij}(n)$ exist when n increases, provided they are aperiodic.

When the chain is periodic, the limit does not always exist. For example, if $S = \{0, 1\}$ and $p_{12} = p_{21} = 1$, then

$$p_{11}(n) = p_{22}(n) = I(n \text{ is even}).$$

Obviously, the limit does not exist.

When the limits exist, the conclusion is neat.

Theorem 4.4

For an irreducible aperiodic Markov chain, we have that

$$p_{ij}(n) \to \mu_j^{-1}$$

as $n \to \infty$ for all i and j. $\diamondsuit$

Remarks:

1. If the chain is transient or null recurrent, then it is known that $p_{ij}(n) \to 0$ for all i and j. Since $\mu_j = \infty$ in this case, the theorem is automatically true.

2. If the chain is positive recurrent, then according to this theorem, $p_{ij}(n) \to \pi_j = \mu_j^{-1}$.

3. This theorem implies that the limit of $p_{ij}(n)$ does not depend on n. It further implies that

$$P(X_n = j) \to \mu_j^{-1}$$

irrespective of the distribution of X_0.

4. If $\{X_n\}$ is an irreducible chain with period d, then $\{X_{nd}\}$ is an aperiodic chain. (Not necessarily irreducible). It follows that

$$p_{jj}(nd) = P(Y_n = j | Y_0 = j) \to d\mu_j^{-1}$$

as $n \to \infty$.

Proof:

Case I: the Markov chain is transient. In this case, the theorem is true by renewal theorem.

Case II: the Markov chain is recurrent. We could use renewal theorem. Yet let us see another line of approach.

Let $\{X_n\}$ be the Markov chain with the transition probability matrix P under consideration. Let $\{Y_n\}$ be an independent Markov chain with the same state space S and transition matrix P with $\{X_n\}$.

Now we consider the stochastic process $\{Z_n\}$ such that $Z_n = (X_n, Y_n)$, $n = 0, 1, 2, \ldots$. Its state space is $S \times S$. Its transition probabilities are simply multiplication of original transition probabilities. That is,

$$p_{(ij)\to(kl)} = p_{ik} p_{jl}.$$

The new chain is still irreducible. If $p_{ik}(m) > 0$ and $p_{jl}(n) > 0$, then $p_{ik}(mn)p_{jl}(mn) > 0$ and so kl is reachable from ij.

The new chain is still aperiodic. It states that if the period is one, then $p_{ij}(n) > 0$ for all sufficiently large n. Thus, $p_{ij}(n)p_{kl}(n) > 0$ for all large enough n too.

Case II.1: positive recurrent. In this case, $\{X_n\}$ has stationary distribution π, and hence $\{Z_n\}$ has stationary distribution given by $\{\pi_i\pi_j : i, j \in S\}$. Thus by the lemma in the last section, $\{Z_n\}$ is also positive recurrent.

Assume $(X_0, Y_0) = (i, j)$ for some (i, j). For any state k, define

$$T = \min\{n : Z_n = (k, k)\}.$$

Due to positive recurrentness, $P(T < \infty) = 1$.

Implication? Sooner or later, X_n and Y_n will occupy the same state. If so, from that moment and on, X_{n+m} and Y_{n+m} will have the same distribution. If Y_n has π as its distribution, so will X_n.

More precisely, starting from any pair of states (i, j), we have

$$\begin{aligned} p_{ik}(n) &= P(X_n = k) \\ &= P(X_n = k, T \le n) + P(X_n = k, T > n) \\ &= P(Y_n = k, T \le n) + P(X_n = k, T > n) \\ &\le P(Y_n = k) + P(T > n) \\ &= p_{jk}(n) + P(T > n). \end{aligned}$$

Due to the symmetry, we also have

$$p_{jk}(n) \le p_{ik}(n) + P(T > n).$$

Hence

$$|p_{ik}(n) - p_{jk}(n)| \le P(T > n) \to 0$$

as $n \to \infty$. That is,

$$p_{ik}(n) - p_{jk}(n) \to 0.$$

or if the limit of $p_{jk}(n)$ exists, it does not depend on j. (Remark: $P(T > n) \to 0$ is a consequence of recurrentness).

For the existence, we have

$$\pi_k - p_{jk}(n) = \sum_{i \in S} \pi_i (p_{ik}(n) - p_{jk}(n)) \to 0.$$

Case II.2: null recurrent. We can no longer claim $P(T > n) \to 0$.

In this case, $\{X_n, Y_n\}$ may be transient or null recurrent.

If $\{X_n, Y_n\}$ is transient, then

$$P(X_n = j, Y_n = j | X_0 = i, Y_0 = i) = [p_{ij}(n)]^2 \to 0.$$

Hence, the conclusion $p_{ij}(n) \to 0$ remains true.

If $\{X_n, Y_n\}$ is null recurrent, the problem becomes a bit touchy. However, $P(T > n) \to 0$ stays. Thus, we still have the conclusion

$$p_{ik}(n) - p_{jk}(n) \to 0,$$

but we do not have a stationary distribution handy.

What we hope to show is that $p_{ij}(n) \to 0$ for all $i, j \in S$. If this is not true, then we can find a subsequence of n such that

$$p_{ij}(n_r) \to \alpha_j$$

as $r \to \infty$, and at least one of α_j is not zero.

Let F be a finite subset of S. Then

$$\sum_{j \in F} \alpha_j = \lim_{r \to \infty} p_{ij}(n_r) \leq 1.$$

This is true for any finite subset, which implies

$$\alpha = \sum_{i \in S} \alpha_j \leq 1.$$

(Recall that S is countable).

Using the idea of one-step transition, we have

$$\sum_{k \in F} p_{ik}(n_r) p_{kj} \leq p_{ij}(n_r + 1) = \sum_{k \in S} p_{ik} p_{kj}(n_r).$$

Let $r \to \infty$, we have

$$\sum_{k \in F} \alpha_k p_{kj} \leq \sum_{k \in S} p_{ik} \alpha_j = \alpha_j.$$

Let F get large, we have

$$\sum_{k \in S} \alpha_k p_{kj} \leq \alpha_j.$$

The equality has to be true, otherwise,

$$\begin{aligned}\sum_{k\in S}\alpha_k &= \sum_{k\in S}\alpha_k\sum_{j\in S}p_{kj}\\ &= \sum_{k,j\in S}\alpha_k p_{kj}\\ &= \sum_{j\in S}[\sum_{k\in S}\alpha_k p_{kj}]\\ &< \sum_{j\in S}\alpha_j\end{aligned}$$

which is a contradiction. However, when the equality holds, we would have

$$\sum_{k\in S}\alpha_k p_{kj} = \alpha_j$$

for each $j \in S$. Thus, $\{\alpha_j/\alpha, j \in S\}$ is a stationary distribution of the Markov chain. This contradicts the assumption of the null recurrentness. $\diamondsuit$

Theorem 4.5

For any aperiodic state j of a Markov chain, $p_{jj}(n) \to \mu_j^{-1}$ as $n \to \infty$. Furthermore, if i is another state, then $p_{ij}(n) \to f_{ij}\mu_j^{-1}$ as $n \to \infty$. $\diamondsuit$

Corollary 4.2

Let

$$\tau_{ij}(n) = \frac{1}{n}\sum_{m=1}^{n} p_{ij}(m)$$

be the mean proportion of elapsed time up to the nth step during which the chain was in state j, starting from i. If j is aperiodic, then

$$\tau_{ij}(n) \to f_{ij}/\mu_j$$

as $n \to \infty$. $\diamondsuit$

4.4 Reversibility

Let $\{X_n : n = 0, 1, 2, \ldots, N\}$ be a (part) of a Markov chain whose transition probability matrix is P and it is irreducible and positive recurrent so that π is its stationary distribution.

Now let us define $Y_n = X_{N-n}$ for $n = 0, 1, \ldots, N$. Suppose all X_n has distribution given by the stationary distribution π. It can be shown that $\{Y_n\}$ is also a Markov chain.

Theorem 4.6

The sequence $Y = \{Y_n : n = 0, 1, \ldots, N\}$ is a Markov chain with transition probabilities

$$q_{ij} = P(Y_{n+1} = j | Y_n = i) = \frac{\pi_j}{\pi_i} p_{ji}.$$

♢

Proof We need only verify the Markov property. Other conditions for a Markov chain are obvious.

$$\begin{aligned}
& P(Y_{n+1} = i_{n+1} | Y_k = i_k, k = 0, \ldots, n) \\
= \;& P(Y_k = i_k, k = 0, \ldots, n, n+1) / P(Y_k = i_k, k = 0, \ldots, n) \\
= \;& P(X_{N-k} = i_k, k = 0, \ldots, n, n+1) / P(X_{N-k} = i_k, k = 0, \ldots, n) \\
= \;& \frac{P(X_{N-n-1} = i_{n+1}) P(X_{N-n} = i_n | X_{N-n-1} = i_{n+1})}{P(X_{N-n} = i_n)} \\
= \;& \frac{\pi_{i_{n+1}} p_{i_{n+1}, i_n}}{\pi_{i_n}}.
\end{aligned}$$

Since this transition probability does not depend on i_k, $k = 0, \ldots, n-1$, the Markov property is verified. ♢

Although Y in the above theorem is a Markov chain, it is not the same as the original Markov chain.

Definition 4.1

Let $\{X_n : 0 \leq n \leq N\}$ be an irreducible Markov chain such that it has stationary distribution π for all n. The chain is call **reversible** if the transition matrices of X and its time-reversal; Y are the same, which is to say that

$$\pi_i p_{ij} = \pi_j p_{ji}$$

for all i, j. $\diamondsuit$

Why do we define the reversibility? One advantage is when a chain is reversible, the transitions between i and j are balanced at equilibrium. This provides us a convenient way to solve for the stationary distribution. We have also used the idea to prove the uniqueness of the solution to $\pi = \pi P$.

Theorem 4.7

Let P be the transition matrix of an irreducible chain X and suppose that there exists a distribution π such that $\pi_i p_{ij} = \pi_j p_{ji}$ for all $i, j \in S$. Then π is a stationary distribution of the chain. Furthermore, X is reversible in equilibrium.

Chapter 5

Continuous Time Markov Chain

It is more realistic to consider processes which do not have to evolve at specified epochs. However, setting up continuous time stochastic processes properly involves a lot of effort in mathematics. We now work on two special continuous time stochastic processes first to motivate the continuous time Markov chain.

5.1 Birth Processes and the Poisson Process

These processes may be called counting processes in general. We have a process $\{N(t) : t \geq 0\}$ such that

(a) $N(0) = 0$, and $N(t) \in \{0, 1, 2, \ldots\}$,

(b) if $s < t$, then $N(s) \leq N(t)$.

What other detailed properties should we place on it?

Definition 5.1

A Poisson process with intensity λ is a process $N = \{N(t) : t \geq 0\}$ taking values in $S = \{0, 1, 2, \ldots\}$ such that

(a) $N(0) = 0$; if $s < t$, then $N(s) \leq N(t)$,

$$\text{(b)} P\{N(t+h) = n+m | N(t) = n\} = \begin{cases} \lambda h + o(h) & \text{if } m = 1, \\ o(h) & \text{if } m > 1, \\ 1 - \lambda h + o(h) & \text{if } m = 0. \end{cases}$$

(c) if $s < t$, the random variable $N(t) - N(s)$ is independent of the $N(s)$.
♢

In general, we call (b) the individuality, and call (c) the independence property of the Poisson process. It is well known as these specifications imply that $N(t) - N(s)$ has Poisson distribution with mean $\lambda(t-s)$ for $s < t$.

Theorem 5.1

$N(t)$ has the Poisson distribution with parameter λt; that is to say

$$P(N(t) = j) = \frac{(\lambda t)^j}{j!} \exp(-\lambda t), \quad j = 0, 1, 2, \ldots.$$

♢

PROOF: Denote $p_j(t) = P(N(t) = j)$. The properties of the Poisson process lead to the equation

$$p_j'(t) = \lambda p_{j-1}(t) - \lambda p_j(t)$$

of $j \neq 0$; likewise

$$p_0'(t) = -\lambda p_0(t).$$

With the boundary condition $p_j(0) = I(j = 0)$, the equations can be solved and the conclusion proved. ♢

We may view a counting process by recording the arrival time of the nth event. For that purpose, we define

$$T_0 = 0, \quad T_n = \inf\{t : N(t) = n\}.$$

The inter-arrival times are the random variables $X_1, X_2, \ldots$ given by

$$X_n = T_n - T_{n-1}.$$

The counting process can be easily recontructed from X_n's too.

The following theorem is a familiar story.

Theorem 5.2

The random variables $X_1, X_2, \ldots$ are independent, each having the exponential distribution with parameter λ. ♢

PROOF: It can be shown by working on

$$P(X_{n+1} > t + \sum_{i=1}^{n} t_i | X_i = t_i, i = 1, 2, \ldots, n)$$

by making use of the property of independent increment. ♢

There is a bit problem with this proof as the event we conditioning on has zero probability. It could be made more rigorous with some measure theory results.

The Poisson process is a very satisfactory model for redioactive emissions from a sample of uranium-235 since this isotope has a half-life of 7×10^8 years and decays fairly slowly. That is, we have a constant radio-active source in a short time. For some other kinds of radio-active substances, the rate of emission should depend on the number of detected emission already.

It can be shown that $\{N(t) : t \geq 0\}$ is a Poisson process if and only if $X_1, X_2, \ldots$ are independent and identically distributed exponential random variables. If the X_i has exponential distribution with rate λ_i instead, then we have a general birth process.

Definition 5.2

A birth process with intensity $\lambda_0, \lambda_1, \ldots$ is a process $\{N(t) : t \geq 0\}$ taking values in $S = \{0, 1, 2, \ldots\}$ such that

(a) $N(0) \geq 0$; if $s < t$, then $N(s) \leq N(t)$,

$$\text{(b)} P[N(t+h) = n + m | N(t) = n) = \begin{cases} \lambda_n h + o(h) & \text{if } m = 1, \\ o(h) & \text{if } m > 1, \\ 1 - \lambda_n h + o(h) & \text{if } m = 0. \end{cases}$$

(c) if $s < t$, the random variable $N(t) - N(s)$ is independent of the $N(s)$.

♢

Here is a list of special cases:

(a) Poisson process. $\lambda_n = \lambda$ for all n.

(b) Simple birth. $\lambda_n = n\lambda$. This is the case when each living individual gives birth independently of others, and at the constant rate.

(c) Simple birth with immigration. $\lambda_n = n\lambda+\nu$. In addition to the simple birth, there is a steady source of immigration.

The differential equation we derived for the Poisson process can be easily generalized. We can find two basic sets of them. Define $p_{ij}(t) = P(N(s+t) = j|N(s) = i)$. The boundary conditions are $p_{ij}(0) = \delta_{ij} = I(i = j)$.

Forward system of equations:

$$p'_{ij}(t) = \lambda_{j-1}p_{i,j-1}(t) - \lambda_j p_{ij}(t)$$

for $j \geq i$.

Backward system of equations:

$$p'_{ij}(t) = \lambda_i p_{i+1,j}(t) - \lambda_i p_{ij}(t)$$

for $j \geq i$.

The forward equation can be obtained by computing the probability of $N(t + h) = j$ conditioning on $N(t) = i$. The backward equation is obtained by computing the probability of $N(t + h) = j$ conditioning on $N(h) = i$.

Theorem 5.3

The forward system has a unique solution, which satisfies the backward system. ♢

PROOF: First, it is seen that $p_{ij}(t) = 0$ whenever $j < i$. When $j = i$, $p_{ii}(t) = \exp(-\lambda_j t)$ is the solution. Substituting into the foward equation, we obtain the solution for $p_{i,i+1}(t)$. Repeat this procedure implies that the forward system has a unique solutions.

Using Laplace transformation reveals the structure of the solution better. Define

$$\hat{p}_{ij}(\theta) = \int_0^\infty \exp(-\theta t)p_{ij}(t)dt.$$

Then, the forward system becomes

$$(\theta + \lambda_j)\hat{p}_{ij}(\theta) = \delta_{ij} + \lambda_{j-1}\hat{p}_{i,j-1}(\theta).$$

The new system becomes easy to solve. We obtain

$$\hat{p}_{ij}(\theta) = \frac{1}{\lambda_j} \frac{\lambda_i}{\theta + \lambda_i} \frac{\lambda_{i+1}}{\theta + \lambda_{i+1}} \cdots \frac{\lambda_j}{\theta + \lambda_j}$$

for $j > i$. The uniqueness is determined by the inversion theorem for Laplace transforms.

If $\pi_{ij}(t)$'s solve the backword systems, their corresponding Laplace transforms will satisfy

$$(\theta + \lambda_i)\hat{\pi}_{ij}(\theta) = \delta_{ij} + \lambda_i \hat{\pi}_{i+1,j}(\theta).$$

It turns out that these satisfying the forward system will also satisfy the backward system here in Laplace transforms. Thus, the solution to the forward equation is also a solution to the backward equation. ♢

An implicit conclusion here is: the backward equation may have many solutions. It turns out that if there are many solutions to the backward equation, the solution given by the forward equation is the minimum solution.

Theorem 5.4

If $\{p_{ij}(t)\}$ is the unique solution of the forward system, then any solution $\{\pi_{ij}(t)\}$ of the backward system satisfies $p_{ij}(t) \leq \pi_{ij}(t)$ for all i, j, t. ♢

If $\{p_{ij}(t)\}$ is the transition probabilities of the specified birth and death process, then it must solve both forward and backward systems. Thus, the solution to the forward system must be the transition probabilities. This could be compared to the problem related to probability of ultimate extinction in the branching process. Conversely, the solution to the foward system can be shown to satisfy the Chapman-Kolmogorov equations. Thus, it is a relevent solution.

The textbook fails to demonstrate the why the proof of the uniqueness for the forward system cannot be applied to the backward system. The key is the assumption of $p_{ij}(t) = 0$ when $j < i$. When this restrict is removed, the backward system may have multiple solutions. This restriction reflects the existence of some continuous time Markov chains which have the same transition probability matrices to some degree to the birth process.

Consequently, the text book should not have made use of this restriction in proving the uniqueness of solution to the forward system.

Intuitively, we may expect that

$$\sum_{j\in S} p_{ij}(t) = 1$$

for any solutions. If so, no solutions can be larger than other solutions and hence the uniqueness is automatic. The non-uniqueness is exactly built on this observation. This constraint does not hold for some birth processes. When the birth rate increases with the population size fast enough, the population size may rearch infinity in finite amount of time.

When the solution to the forward equation

$$\sum_{j\in S} p_{ij}(t) < 1$$

for some $t > 0$ and i, it is possible then to construct another solution which also satisfies the backward system. See Feller for detailed constructions. In that case, it is possible to design a new stochastics process so that its transition probabilities are given by this solution.

What is the probabilistic interpretation when

$$\sum_{j\in S} p_{ij}(t) < 1$$

for some finite t, and i? It implies that within the period of t, the population size has jumped or exploded all the way to infinity. Consequently, infinite number of transitions must be have occurred. Recall the waiting time for the next transition when $N(t) = n$ is exponential with rate λ_n. Let $T_n = \sum_{i=1}^{n} X_i$ be the waiting time for the nth transition. Define $T_\infty = \lim_{n\to\infty} T_n$.

Definition 5.3 *Honest/dishonest*

We call the process N honest if $P(T_\infty < \infty) = 0$ and dishonest otherwise.

Lemma 5.1

Let $X_1, X_2, \ldots$ be independent random variables, X_n having the exponential distribution with parameter λ_{n-1}, and let $T_\infty = \sum_{n=1}^{\infty} X_n$. We have that

$$P(T_\infty < \infty) = \begin{cases} 0 & \text{if } \sum_{n=1}^{\infty} \lambda_n^{-1} = \infty. \\ 1 & \text{if } \sum_{n=1}^{\infty} \lambda_n^{-1} < \infty \end{cases}$$

PROOF: By definition of T_∞, we have

$$E[T_\infty] = \sum_{n=1}^{\infty} \lambda_n^{-1}.$$

Hence, when $\sum_{n=1}^{\infty} \lambda_n^{-1} < \infty$, $E[T_\infty] < \infty$ and $P(T_\infty < \infty) = 1$. This implies dishonesty.

If $\sum_{n=1}^{\infty} \lambda_n^{-1} = \infty$, it does not imply $T_\infty = \infty$ with any positive probability. If, however, $P(T_\infty < t) > 0$ for any $t > 0$, then $E[\exp(-T_\infty)] \geq \exp(-t)P(T_\infty < t) > 0$. We show that this is impossible under the current assumption. Note that

$$\begin{aligned} E[\exp(-T_\infty)] &= \lim_{N\to\infty} E \prod_{n=1}^{N} \exp(-X_n) \\ &= \lim_{N\to\infty} \prod_{n=1}^{N} E \exp(-X_n) \\ &= \lim_{N\to\infty} \prod_{n=1}^{N} (1 + \lambda_n^{-1})^{-1} \\ &= \lim_{N\to\infty} [\prod_{n=1}^{N} (1 + \lambda_n^{-1})]^{-1}. \end{aligned}$$

According to mathematical analysis result, the product with infinite terms equals infinity when $\sum_{n=1}^{\infty} \lambda_n^{-1} = \infty$. (Check its logrithm). Hence, $E[\exp(-T_\infty)] = 0$ which contridicts the assumption $E[\exp(-T_\infty)] > 0$ made earlier. $\diamondsuit$

The following theorem is an easy consequence.

Theorem 5.5

The process N is honest if and only if $\sum_{n=1}^{\infty} \lambda_n^{-1} = \infty$.

5.1.1 Strong Markov Property

The Markov property for discrete time stochastic process states: given the present ($X_n = i$), the future (X_{n+m}'s) is independent of the past (X_{n-k}'s). The time represents the present is a non-random constant. In the example of simple random walk, we often claim that once the random walk returns to 0, it renews itself. That is, the future behavior of the random walk does not depend on how the random walk got into 0 nor when it returns to 0. We may notice that this notion is a bit different from the Markov property. The "present time" is a random variable.

This example implies that the Markov property is true for some randomly selected time. The question is what kind of random variable can be used for this purpose? Suppose $\{N(t) : t \geq 0\}$ is a stochastic process and T is a random variable. If the event $T \leq t$ is completely determined by the knowledge of $\{N(s) : s \leq t\}$, then we call it a **stopping time** for stochastic process $\{N(t) : t \geq 0\}$. More rigorous definition relates to the concept of σ-field we introduce before.

Theorem 5.6 *Strong Markov Property*

Let N be a birth process and let T be a stopping time for N. Let A be an event which depends on $\{N(s) : s > T\}$ and B be an event which depends on $\{N(s) : s \leq T\}$. Then

$$P(A|N(T) = i, B) = P(A|N(T) = i)$$

for all i. ♢

Proof: In fact, this is a simple case, as $N(T)$ is a discrete random variable. The kind of event B which causes most trouble are those contains all information about the history of the process before and including time T. If that is the case, then the value of T is completely determined by B. Hence, we may write $T = T(B)$. Hence, it claims that

$$P(A|N(T) = i, B) = P(A|N(T) = i, T = T(B), B).$$

Among the three pieces of information on the right hand side, T is defined based on $\{N(s) : s \leq T(B)\}$ and is a constant when "$N(T) = i, T = T(B)$".

Hence the Markov (weak one) property allows us to ignore B itself. At the same time, the process is time homogeneous, it depends only on the fact that the chain is in state i now, not when it first reached state i. Hence, the part of $T = T(B)$ is also not informative. Hence the conclusion.

The measure theory proof is as follows. Let $H = \sigma\{N(s) : s \leq T\}$ which is the σ-field generated by these random variables. An event B containing historic information before T is simply an event in this σ-algebra. Recall the formular $E[E(X|Y)] = E(X)$. Let H play the role of Y, and $E(\cdot|N(T) = i, B)$ play the role of expectation, we ahve

$$\begin{aligned} P(A|N(T) = i, B) &= E(I(A)|N(T) = i, B) \\ &= E[E(I(A)|N(T) = i, B, H)|N(T) = i, B). \end{aligned}$$

We claim $E(I(A)|N(T) = i, B, H) = E(I(A)|N(T) = i)$ since it is H-measurable, plus T is a constant on B and the weak Markov property. Since this function is a constant in the eyes of $N(T) = i, B$, we have

$$E\{E(I(A)|N(T) = i)|N(T) = i, B\} = E(I(A)|N(T) = i).$$

This result is easy to present for discrete time Markov chain. Hence, if you cannot understand the above proof, work on the discrete time example in the assignment will help. ♢

Example 5.1

Consider the birth process with $N(0) = I > 0$. Define $p_n(t) = P(N(t) = n)$. Set up the forward system and solve it when $\lambda_n = n\lambda$. ♢

5.2 Continuous time Markov chains

Let $X = \{X(t) : t \geq 0\}$ be a family of random variables taking values in some countable state space S and indexed by the half-line $[0, \infty)$. As before, we shall assume that S is a subset of non-negative integers.

Definition 5.1

The process X satisfies the Markov property of

$$P(X(t_n) = j|X(t_1) = i_1, \ldots, X(t_{n-1}) = i_{n-1}) = P(X(t_n) = j|X(t_{n-1}) = i_{n-1})$$

for all $j, i_1, \ldots, i_{n-1} \in S$ and any sequence $0 < t_1 < t_2 < \cdots < t_n$ of times.

A continuous time stochastic process X satisfying the Markov property is called a continuous time Markov chain.

One obvious example of continuous time Markov process is the Poisson process. The pure birth process is not a Markov process when it is dishonest. The reason is that to qualify as a Markov chain, we would have required $\{N(t) : t \geq 0\}$ to be a family of random variables. If the population size can explode to infinity at finite t, then $N(t)$ is not always a random variable for a given t.

One difficulty of studying the evolution of the continuous time Markov chain is the lack of one-step transition probability matrix. No matter how short a period of time is, we can always find a shorter period. This observation gives rise to the need of infinitesimal generator.

Definition 5.2

The transition probability $p_{ij}(s,t)$ is defined to be

$$p_{ij}(s,t) = P(X(t) = j|X(s) = i)$$

for $s < t$.

The chain is called time homogeneous if $p_{ij}(s,t) = p_{ij}(0, t-s)$ for all i, j, s, t and we write $p_{ij}(t-s)$ for $p_{ij}(s,t)$. $\diamondsuit$

We will assume the homogeneity for all continuous time Markov chain to be discussed unless otherwise specified. We also use notation $\mathbf{P}_t$ for the corresponding matrix. It turns out that the collection of all transition probability matrices form a stochastic semi-group. Forming a group means we can define an operation on this set such that the operation is closed and invertable. A semi-group may not have an inverse for each member in the same set.

Theorem 5.7

The family $\{\mathbf{P}_t : t \geq 0\}$ is a stochastic semi-group; that is, it satisfies the following conditions:

(a) $\mathbf{P}_0 = I$, the identity matrix;

(b) $\mathbf{P}_t$ is stochastic, that is $\mathbf{P}_t$ has non-negative entries and row sum 1;

(c) the Chapman-Kolmogorov equations: $\mathbf{P}_{s+t} = \mathbf{P}_s\mathbf{P}_t$ if $s, t, \geq 0$. ♢

PROOF: Obvious. ♢

Remark: a quick reference to mathematics definition reveals that only (c), and sometimes also (a) are properties of a semi-group.

For continuous time Markov chain, there is no t_0 such that P_{t_0} can be used to compute P_t for all t. This is different from the discrete time Markov chain. Is it possible to represent all $\mathbf{P}_t$ from a single entity?

The answer is positive if the semi-group satisfies some properties.

Definition 5.3

The semigroup $\{\mathbf{P}_t : t \geq 0\}$ is called standard if $\mathbf{P}_t \to I$ as $t \downarrow 0$, which is to say that $p_{ii}(t) \to 1$ and $p_{ij}(t) \to 0$. ♢

Being standard means that every entry of $\mathbf{P}_t$ is a continuous function of t, not only at $t = 0$, but also at any $t > 0$ by Chapman-Kolmogorov equations. It does not imply, though, that they are differentiable at $t = 0$.

Suppose that $X(t) = i$ at the the moment t. What might happy in the next short period $(t, t + h)$?

May be nothing will happen, the corresponding chance is $p_{ii}(h) + o(h)$.

It may switched into state j with corresponding chance $p_{ij}(h) + o(h)$.

The chance to have two or more transitions are assumed $o(h)$ here. I take it as an assumption, which means not all Markov chain has this property. The book mentioned that it can be proved. I am somewhat skeptical.

It turns out that when the semi-group is standard, the following claim is true:

$$p_{ij}(h) = g_{ij}h + o(h), \quad p_{ii}(h) = 1 + g_{ii}h + o(h) \tag{5.1}$$

where g_{ij}'s are a set of constants such that $g_{ij} > 0$ for $i \neq j$ and $-\infty \leq g_{ii} \leq 0$ for all i.

Unless additional retrictions are applied, some g_{ii} may take value $-\infty$. In which case I guess that the interpretation is

$$\frac{p_{ii}(h)-1}{h} \to -\infty.$$

When all the needed conditions are in place, we set up a matrix G for g_{ij} and call it **infinitesimal generator**.

Linking this expansion to the transition probability, we must have

$$\sum_j g_{ij} = 0$$

for all i. Due to the fact that some g_{ii} can be negative infinity, the above relationship does not apply to all Markov chains. We will discuss the conditions under which this result is guaranteed.

Once we admit the validity of (5.1), then we can easily obtain the forward and backward equations.

Forward equations $\mathbf{P}'_t = \mathbf{P}_t G$;

Backward equations $\mathbf{P}'_t = G\mathbf{P}_t$;

Example 5.2

Birth process. Write down the infinitesimal generator here. ♢

The validity of (5.1) does not guarantee the uniqueness of the solution to these two systems. It is known that the backward equations are satisfied when the semi-group is standard. The forward equations are proved when the semi-group is uniform which will be discussed a bit further.

When all entries of G are bounded by a common constant in absolute value, then two systems share a unique solution for $\mathbf{P}_t$. In fact, the solution can be written in the form

$$\mathbf{P}_t = \sum_{n=0}^{\infty} \frac{t^n}{n!} G^n.$$

We may also write it as

$$\mathbf{P}_t = \exp(tG).$$

Recall that for some Markov chain with standard semi-group transition probability matrices, the instantanuous rates g_{ii} could be negative infinity. In this case, state i is instantanuous. That is, the moment of the Markov chain entering state i is also the moment of leaving the state. Barring such possibilities by assuming g_{ij} have a common upper bound in absolute value, then waiting times for Markov chain leaving state i will have momoryless property. Therefore, we have the result as follows.

Theorem 5.8

Under some conditions, the random variable U_i is exponentially distributed with parameter $-g_{ii}$, where U_i is the waiting time for the Markov chain to leave state i.

Further, the probability that the chain jumps to state j from state i is $-g_{ij}/g_{ii}$. ♢

The result on the exponential distribution is the product of the Markov property. Since the Markov property implies the memoriless property, and the later implies the exponential distribution. The only hassle is the parameter $-g_{ii}$. As long as it is finite for all i, the proof is good enough.

The proof for the second part is less rigorous. Taking a small positive value h, we consider the probability under the assumption of $x < U < x + h$. The conditional probability converges to $-g_{ij}/g_{ii}$ when the boundedness conditions on g_{ii} are satisfied.

Example 5.3

A continuous Markov chain with finite state space is always standard and uniform. Thus, all the conclusions are valid. ♢

Example 5.4

A continuous Markov chain with only two possible states can have its forward and backward equations solved easily. ♢

5.3 Limiting probabilities

Our next issue is about limiting probabilities. Similar to discrete time Markov chain, we need to examine the issue of irreducibility. It turns out that for any state i and j, if

$$p_{ij}(t) > 0$$

for some $t > 0$, then $p_{ij}(t) > 0$ for all $t \geq 0$.

We say that a continuous time Markov chain is irreducible when $p_{ij}(t) > 0$ for all i, j and $t > 0$.

Definition 5.1

The vector π is a stationary distribution of the continuous Markov chain if $\pi_j \geq 0$, $\sum_j \pi_j = 1$, and $\pi = \pi \mathbf{P}_i$ for all $t \geq 0$. $\diamondsuit$

By Chapman-Kolmogorov equations, if $X(t)$ has distribution π, then $X_(t + s)$ also has distribution π for all $s \geq 0$.

If π is such a distribution, it must be unique and non-negative when the chain is irreducible. This is due to the fact that $\mathbf{P}_t$ is a transition probability matrix of some discrete times Markov chain which is irreducible. Hence, its stationary distribution is non-negative and unique.

Solving equations $\pi = \pi \mathbf{P}_t$ may not be convenient as $\mathbf{P}_t$'s are hard to specify. We hence seek help from the following theorem.

Theorem 5.9

If the Markov chain has uniform semi-group transition probability matrices, then $\pi = \pi \mathbf{P}_t$ for all t if and only if $\pi G = 0$. $\diamondsuit$

The proof is simple. When the Markov chain has uniform semi-group, then we have $\mathbf{P}_t = \exp(tG)$. Using the expansion for the exponential function results in the conclusion. Note that when $\pi G = 0$, it is easy to show that every term in the power expansion is zero. If $\pi \mathbf{P}_t = 0$ for all t, it implies the function is a zero function. An analytical function is a zero function if and only if all coefficients are zero. This implies that $\pi G = 0$.

The textbook does not specify the uniformity condition. The result is true for more general Markov chains such as the birth and death process.

Finally, we state the limiting theorem.

Theorem 5.10

Let X be irreducible with a standard semigroup $\{\mathbf{P}_t\}$ of transition probabiblities.

(a) If there exists a stationary distribution π, then it is unique and

$$p_{ij}(t) \to \pi_j$$

as $t \to \infty$ for all i and j.

(b) If there is no stationary distribution then $p_{ij}(t) \to 0$ as $t \to 0$ as $t \to \infty$ for all i and j.

♢

PROOF: For any $h > 0$, we may define $Y_n = X(nh)$. Then we have a discrete time Markov chain which is irreducible and ergodic. If it is non-null recurrent, then it has a unique stationary distribution $\pi^{(h)}$. In addition,

$$p_{ij}(nh) = P(Y_n = j | Y_0 = i) \to \pi_j^{(h)}.$$

If the chain is null recurrent or transient, we would have

$$p_{ij}(nh) = P(Y_n = j | Y_0 = i) \to 0$$

for all i, j.

For any two rational numbers h_1 and h_2, applying the above result implies that $\pi_j^{(h_1)} = \pi_j^{(h_2)}$.

Next, using continuity of $p_{ij}(t)$ to fill up the gap for all real numbers.

Remark: Unlike other result, we are told that the conclusion is true when the class of the transition matrix is a standard semi-group.

5.4 Birth-death processes and imbedding

Birth and death process is a more realistic model for population evolution. Suppose the random variable $X(t)$ represents the population size at time t.

(a) Its state space is $S = \{0, 1, \ldots\}$.

(b) Its infinitesimal transition probbilities are given by

$$P(X(t+h) = m+n | X(t) = n) = \begin{cases} \lambda_n h + o(h) & \text{if } m = 1, \\ \mu_n h + o(h) & \text{if } m = -1, \\ o(h) & \text{if } |m| > 1, \end{cases}$$

(c) the 'birthe rates' $\lambda_0, \ldots,$ and 'death rates' $\mu_0 = 0, \mu_1, \mu_2, \ldots$ satisfy $\lambda_n \geq 0$ and $\mu_n \geq 0$.

Due to memoryless property, the waiting time for the next transition has exponential distribution with rate $\lambda_n + \mu_n$ when $X(t) = n$. The probability that the next transition is a birth is given by $\lambda_n/(\lambda_n + \mu_n)$.

The infinitesimal generator G ha a nice form

$$G = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3 + \mu_3) & \lambda_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \end{pmatrix}.$$

The chain is uniform if and only if $\sup_n\{\lambda_n + \mu_n\} < \infty$. When $\lambda_0 = 0$, the chain is reducible. State 0 becomes absorbing. If $\lambda_n = 0$ for some $n > 0$ and $X(0) = m < n$, then the population size will be bounded by n. It seems the standard semi-group condition is satisfied as long as all rates are finite. Yet, we have to assume that the process is honest to maintain that it is a continuous time Markov chain.

The transition probabilities $p_{ij}(t)$ may be computed in principle, but may not be so useful. It has a stationary distribution under some simple conditions:

$$\pi_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \pi_0$$

with

$$\pi_0^{-1} = \sum_{n=0}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}.$$

The limit exists when $\pi_0 > 0$. Note that in this case, the condition of standard semi-group is satisfied. Hence, the conclusion for the existence of the limiting probabilities is solid.

Example 5.5 *Simple death with immigration.*

Consider the situation when the birth rates $\lambda_n = \lambda$ for all n, and $\mu_n = n\mu$ for $n = 1, 2, \ldots$. That is, there is a constant source of immigration, and each individual in the population has the same rate of die. Using the transition probability language, the book staes that

$$\begin{aligned} p_{ij}(h) &= P(X(t+h) = j | X(t) = i) \\ &= \begin{cases} P(j-i\text{arrivals, 0 deaths}) + o(h) & \text{if } j \geq i, \\ P(i-j\text{deaths, 0 arrivals}) + o(h) & \text{if } j < i \end{cases}. \end{aligned}$$

Since the chance to have two or more transitions in a short period of length h is $o(h)$, it reduces to

$$p_{i,i+1}(h) = \lambda h + o(h),$$
$$p_{i,i-1}(h) = (i\mu)h + o(h)$$

and $p_{ij}(h) = o(h)$ when $|i - j| \geq 2$.

It is very simple to work out its limiting probabilities.

Theorem 5.11

In the limit as $t \to \infty$, $X(t)$ is asymptotically Poisson distributed with parameter $\rho = \frac{\lambda}{\mu}$. That is,

$$P(X(t) = n) \to \frac{\rho^n}{n!} \exp(-\rho), \quad n = 0, 1, 2, \ldots.$$

The proof is very simple. Let us work out the distribution of $X(t)$ directly. Assume $X(0) = I$.

Let $p_j(t) = P(X(t) = j | X(0) = I)$. By the foward equations, we have

$$p_j'(t) = \lambda p_{j-1}(t) - (\lambda + j\mu)p_j(t) + \mu(j+1)p_{j+1}(t)$$

for $j \geq 1$ and $p_0'(t)a = \mu p_1(t)$.

Note that the probability generating function of $X(t)$ is given by $G(s,t) = E[s^{X(t)}]$, we have

$$\frac{\partial G}{\partial s} = \sum_{j=0}^{\infty} j s^{j-1} p_j(t)$$

and

$$\frac{\partial G}{\partial t} = \sum_{j=0}^{\infty} s^j p_j'(t).$$

Hence, by multiplying s^j on both sides of the forward system and sum up, we have

$$\begin{aligned} \frac{\partial G}{\partial t}(s,t) &= \lambda(s-1)G(s,t) - (\mu s - \mu)\frac{\partial G}{\partial s}(s,t) \\ &= (s-1)[\lambda G(s,t) - \mu\frac{\partial G}{\partial t}(s,t)]. \end{aligned}$$

It is easy to verify that

$$G(s,t) = \{1 + (s-1)\exp(-\mu t)\}^I \exp\{\rho(s-1)(1-e^{-\mu t})\}.$$

It is interesting to see that if $I = 0$, then the distribution of the $X(t)$ is Poisson. Otherwise, the distribution is a convolution of a binomial distribution and a Poisson distribution. Due to the uniqueness of the solution to the forward system, we do not have to look for other solutions anymore.

Example 5.6 *Simple birth-death*

When the birth and death rates are given by $\lambda_n = n\lambda$ and $\mu_n = n\mu$, we can work out the problem in exactly the same way.

It turns out that given $X(0) = I$, the generating function of $X(t)$ is given by

$$G(s,t) = \left[\frac{\mu(1-s) - (\mu - \lambda s)\exp\{-t(\lambda-\mu)\}}{\lambda(1-s) - (\mu - \lambda s)\exp\{-t(\lambda-\mu)}\right]^I.$$

The corresponding differential equation is given by

$$\frac{\partial G}{\partial t}(s,t) = (\lambda s - \mu)(s-1)\frac{\partial G}{\partial s}(s,t).$$

Does it look like a convolution of one binomial and one negative binomial distribution?

One can find the mean and variance of $X(t)$ from this expression.

$$E[X(t)] = I\exp\{(\lambda-\mu)t\}$$

$$Var(X(t)) = \frac{\lambda+\mu}{\lambda-\mu}\exp\{(\lambda-\mu)t\}[\exp\{(\lambda-\mu)t\}-1]I.$$

The limit for the expectation is either 0 or infinity depending on whether the ratio of λ/μ is larger than or smaller than 1.

The probability of extinction is given by the limit of $\eta(t) = P(X(t) = 0)$ which is $min(\rho^{-1}, 1)$. ♢

5.5 Embedding

If we ignore the length of the time between two consecutive transitions in a continuous time Markov chain, we obtain a discrete time Markov chain. In the example of linear birth and death, the waiting time for the next birth or death has exponential distribution with rate $n(\lambda+\mu)$ given $X(t) = n$. The probability that the transition is a birth is $(n\lambda)/[n(\lambda+\mu)] = \lambda/(\lambda+\mu)$.

Think of the transition as the movement of a particle from the integer n to the new integer $n+1$ or $n-1$. The particle is performing a simple random walk with probability $p = \lambda/(\lambda+\mu)$. The state 0 is an absorbing state. The probability that the random walk will be absorbed to state 0 is given by $\min(1, q/p)$.

5.6 Markov chain Monte Carlo

In statistical applications, we often need to compute the joint commulative distribution functions of several random variables. For example, we might be interested to know the distribution of $s_n^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)$ where $X_1, \ldots, X_n$ are independent and identically distributed random variables, $\bar{X}_n = n^{-1}\sum_i X_i$.

When X_1 has normal distribution, the distribution of s_n^2 is known to be chisquare with $n-1$ degrees of freedom. The density function is well known, and the value of $P(s_n^2 \le x)$ can be easily obtained via numerical integration. If X_1 has exponential distribution, the density function of s_n^2 is no longer available in a simple analytical form. Straightforward numerical integration becomes very difficult. If one can generate 10,000 sets of independent random variables $Y_1, Y_2, \ldots, Y_n$ such that they all have exponential distribution, then

we may compute 10,000 many s_n^2 values. If 12% of them are smaller than $x = 5$, then it can be shown that $P(s_n^2 \leq 5)$ can be approximated by 12%. The precision improves when we generate more and more sets of such random variables.

A more general problem to compute

$$\int g(\theta)\pi(\theta)d\theta, \quad \text{or} \quad \sum_{\theta} g(\theta)\pi(\theta)$$

over $\theta \in \Theta$, where $\pi(\theta)$ is a density function or probability mass function on Θ. If we can generate random variables $X_1, X_2, \ldots$ such that they all take values in Θ and have $\pi(\theta)$ as their density function or probability mass function, then the law of large number in the probability theory implies

$$n^{-1} \sum_i g(X_i) \to E_\pi\{g(X_1)\}.$$

Hence, the value of $n^{-1} \sum_i g(X_i)$ provides an approximation to $\int g(\theta)\pi(\theta)d\theta$, or $\sum_\theta g(\theta)\pi(\theta)$.

Before this idea becomes applicable, we need to overcome a technical difficulty. How do we generate random variables with a specified distribution very quickly? It is found that some seemly simple functions can produce very unpredictable outcomes. Thus, when this operation is applied repeatedly, the outcomes appear to be completely random. A so called pseudo random number generator for uniform distribution is hence constructed. Starting from this point, one can then generate random numbers from most commonly used distributions, whether they are discreate or continuous.

In some applications, especially for Baysian analysis, it is often necessary to generate random numbers from a not well defined density function. For example, we may only know that the corresponding density function $\pi(\theta)$ is proportional to some known function. In theory, one needs only rescale the function so that the total probability becomes 1. In reality, to find the scale itself is numerically impractical.

The beauty of the Markov Chain Monte Carlo (MCMC) is to construct a discrete time Markov chain so that its stationary distribution is the same as the given density function $\pi(\theta)$, without the need of completely specifying $\pi(\theta)$. How this is achieved is the topic of this section.

In the eyes of a computer, nothing is continuous. Hence we deal with discrete $\pi(\theta)$. Assume the probability function $\pi = (\pi_i : i \in \Theta)$ is our target. We construct a discrete time Markov chain with state space given by Θ, and its stationary distribution is given by π.

For a given π, we have many choices of such Markov chains. The one which is reversible at equilibrium may have some advantage. Thus, we try to find a Markov chain with its transition probabilities satisfying

$$\pi_k p_{kj} = \pi_j p_{jk}$$

for all $k, j \in \Theta$. The following steps will create such a discrete time Markov chain.

Assume we have $X_n = i$ already. We need to generate X_{n+1} according to some transition probability.

(1) First, we pick an arbitrary stochastic matrix $\mathbf{H} = (h_{ij} : i, j \in \Theta)$. This matrix is called the 'proposal matrix'. We generate a random number Y according to the distribution given by $(h_{ik} : k \in \Theta)$. That is, $P(Y = k) = h_{ik}$. Since h_{ik} are well defined, we consider it feasible.

(2) Select a matrix $\mathbf{A} = (a_{ij} : i, j \in \Theta)$ be a matrix with entries between 0 and 1. The a_{ij} are called 'acceptance probabilities'. We first generate a uniform [0, 1] random variable Z and define

$$X_{n+1} = X_n I(Z > a_{ij}) + Y I(Z < a_{ij}).$$

Repeat this two steps, we have created a sequence of random numbers $X_0, X_1, \ldots$ with an arbitrary starting value X_0. Our next task is to choose $\mathbf{H}$ and $\{\mathbf{A}\}$ such that the stationary distribution is what we are look for: π.

With the given $\mathbf{A}$, the transition probabilities are given by

$$p_{ij} = \begin{cases} h_{ij} a_{ij} & \text{if } i \neq j \\ 1 - \sum_{k:k \neq i} h_{ik} a_{ik} & \text{if } i = j \end{cases}$$

This is because that the transition to j occurs only if both $Y = j$ and $Z < a_{ij}$ when $X_n \neq j$.

It turns out that the balance equation is satisfied when we choose

$$a_{ij} = \min\{1, (\pi_j h_{ji})/(\pi_i h_{ij})\}.$$

This choice results in the algorithm called the Hastings algorithm. Note also that in this algorithm, we do not need the knowledge of π_i, but π_i/π_j for all i and j.

Let us try to verify this result. Consider two states $i \neq j$. We have $\pi_i p_{ij} = \pi_i h_{ij} a_{ij} = \min\{\pi_i h_{ij}, \pi_j h_{ji}\}$. Similarly, $\pi_j p_{ji} = \pi_j h_{ji} a_{ji} = \min\{\pi_i h_{ij}, \pi_j h_{ji}\}$. Hence, $\pi_i p_{ij} = \pi_j p_{ji}$.

When two states are the same, the balance equation is obviously satisfied.

When Θ is multi-dimensional, there is some advantage of trying to generate the random vectors one component a time. That is, if X_n is the current random vector, we make X_{n+1} equal X_n except of one of its component. The ultimate goal can be achieved by randomly select a component or by rotating the component.

Gibbs sampler, or heat bath algorithm Assume $\Theta = S^V$. That is, its dimension is V. The state space S is finite and so is the dimension V. Each state in Θ can be written as $i = (i_w : w = 1, 2, \ldots, V)$. Let

$$\Theta_{i,v} = \{j \in \Theta : j_w = i_w \text{ for } w \neq v\}.$$

That is, it is the set of all states which have the same components as state i other than the vth component. Assume $X_n = i$, we now choose a state from $\Theta_{i,v}$ for X_{n+1}. For this purpose, we select

$$h_{ij} = \frac{pi_j}{\sum_{k \in \Theta_{i,v}} \pi_k}, \quad j \in \Theta_{i,v}$$

in the first step of the Hastings algorithm. That is, the choice in $\Theta_{i,v}$ is based on the conditional distribution given the vth component.

The next step of the Hastings algorithm is the same as before. It turns out that $a_{ij} = 1$ for all $j \in \Theta_{i,v}$. That is, there is no need of the second step.

We need to determine the choice of v. We may rotate the components or randomly pick a component each time we generate the next random vector.

Metropolis algorithm If the matrix $\mathbf{H}$ is symmetric, then $a_{ij} = \min\{1, \pi_j/\pi_i\}$ and hence $p_{ij} = h_{ij} \min\{1, \pi_j/\pi_i\}$.

We find such a matrix by placing a uniform distribution on $\Theta_{i,v}$ as defined in the last Gibbs samples algorithm. That is

$$h_{ij} = [|\Theta_{i,v}| - 1]^{-1}$$

for $i \neq j$.

Finally, even the limiting distribution of X_n is π. The distribution of X_n may be far from π until n is hugh. It is also very hard to tell how large n has to be before the distribution of X_n well approximates the stationary distribution.

In applications, researchers often propose to make use of a burn out period M. That is, we throw away $X_1, X_2, \ldots, X_M$ for a large M and start using $Y_1 = X_{M+1}$, $Y_2 = X_{M+2}$ and so on to compute various characteristics of π. For example, we estimate $E_\pi g(\theta)$ by $n^{-1}\sum_{i=1}^{n} g(X_{M+i})$. It is hence very important to have some idea on how large this M must be before the random vectors (numbers) can be used.

Let P be a transition probability matrix of a finite irreducible Markov chain with period d. Then,

(a) $\lambda_1 = 1$ is an eigenvalue of P,

(b) the d complex roots of unity,

$$\lambda_1 = \omega^0,\ \lambda_2 = \omega^1, \ldots,\ \lambda_d = \omega^{d-1},$$

are eigenvalues of P.

(c) the remaining eigenvalues $\lambda_{d+1}, \ldots, \lambda_N$ satisfy $|\lambda_j| < 1$.

When all eigenvalues are distinct, then $P = B^{-1}\Lambda B$ for some matrix B and diagonal matrix with entries $\lambda_1, \ldots, \lambda_N$. Thus, this decomposition allows us to compute the n step transition probability matrix easily. That is, $P^n = B^{-1}\Lambda^n P$.

When not all eigenvalues are distinct, then we can still decompose P as $B^{-1}MB$. However, M cannot be made diagonal any more. The best we can is block diagnonal $diag(J_1, J_2, \ldots)$ such that

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & 0 & \cdots \\ 0 & \lambda_i & 1 & 0 & \cdots \\ 0 & 0 & \lambda_i & 1 & \cdots \\ 0 & 0 & 0 & \lambda_i & \cdots \\ \cdots & \cdots & \cdots & \cdots & \end{pmatrix}.$$

Fortunately, M^n still have a very simple form. Hence, once such a decomposition is found, the n-step transition probability matrix is easily obtained.

Remember that the limiting probability is closed linked to the n-step transition probability matrix. Suppose that the Markov chain in MCMC algorithm is aperiodic. Then, it can be shown that

$$p_{ij}^{(n)} = \pi_j + O(n^{m-1}|\lambda_2|^n)$$

where λ_2 is the eigenvalue with the second largest modulus and m is the multiplicity of this eigenvalue. Note that the transition probability converges to π_j at exponential rate when $m = 1$.

Theorem 5.12

Let X be an aperiod irreducible reversible Markov chain on the finite state space Θ, with transition probability matrix P and stationary distribution π. Then

$$\sum_k |p_{ik} - \pi_k| \leq |\Theta| \cdot |\lambda_2|^n \sup\{|\nu_r(i))| : r \in \Theta\}, \quad i \in \Theta, n \geq 1,$$

where $\nu_r(i)$ is the ith term of the rth right-eigenvector V_r of P. $\diamondsuit$

Chapter 6

General Stochastic Process in Continuous Time

The term stochastic process is generally referred to a collection of random variables denoted as $\{X(t), t \in \mathcal{T}\}$. Usually, these random variables are linked by some index generally referred to as time. If so, $\mathcal{T}$ is a set of time points. When $\mathcal{T}$ contains integer values only, we get a discrete time stochastic process; while when $\mathcal{T}$ is an interval of the real numbers, we have a continuous time stochastic process. Examples other than $\mathcal{T}$ being time includes the situation when $\mathcal{T}$ represents geological location.

6.1 Finite Dimensional Distribution

Let us make a simplifying assumption that $\mathcal{T} = [0, T]$ for some positive number T. Let us recall that a random variable is a measurable function on a probability space $(\Omega, \mathcal{F}, P)$. That is, for each $t \in T$, $S(t)$ is a map in the form

$$S(t): \quad \Omega \to R$$

where R is the set of all real numbers. That is, at any sample point $\omega \in \Omega$, $S(t)$ takes a real value: $S(\omega, t)$.

It becomes apparent that $S(\omega, t)$ is a bivariate function. One of its variables is time t, and the other variable is the sample point ω. Given t, we

get a random variable. At the same time, once we fix $\omega = \omega_0$, $S(\omega_0, t)$ is a function of t. This function is called a sample path of $S(t)$. We denote $s(t)$ as realized sample path of $S(t)$. This is analog to use x for a realized value of a random variable X.

We may also put it as follows: a random variable is a map from the sample space to the space of real numbers; a stochastic process is a map from the sample space to the space of real valued functions. Recall that a random variable is required to be a measurable map. Thus, there must be some kind of measurability requirement on stochastic process.

Compared to the space of real numbers, the space of real functions on $[0, T]$ is much more complex. The problem of setting up a suitable σ-algebra is beyond this course. The space of all continuous functions, denoted as $C[0, T]$, has a simpler structure, and a proper σ-field is available. In applications, however, stochastic processes with non-continuous sample paths are useful. The well known Poisson process, for example, does not have continuous sample path. A slight extension of $C[0, T]$ is to add functions which are right-continuous with left limits. We denote this space as $D[0, T]$. These functions are referred as **regular right-continuous function** or **càdàg** functions. It turns out that $D[0, T]$ is large enough for usual applications, yet is still simple enough to allow a useful σ-field.

To properly investigate the properties of a random variable X, we want to be able to compute $P(X \leq x)$ for any real number x. We refer to $F(x) = P(X \leq x)$ as the cumulative distribution function of X, or simply the distribution of X. If $S(t)$ is a stochastic process on $D[0, T]$, then for any $t_1, t_2 \in [0, T]$, we must have

$$P(S(t_1) \leq s_1, S(t_2) \leq s_2)$$

well defined. That means, the set $S(t_1) \leq s_1, S(t_2) \leq s_2$ must be a member of $\mathcal{F}$. It is easy to see that this requirement goes from two time points to any finite number of time points, and to any countable number of time points. It can be shown that if we can give a proper probability to all such sets (called cylinder sets), then the probability measure can be extended uniquely to the smallest σ-algebra that contains all cylinder sets.

As consequence of this discussion is the following theorem.

Theorem 6.1

A stochastic process taking values in $D[0,T]$ is uniquely determined by its finite dimensional distributions. ♢

6.2 Sample Path

Let us have a look of the following example.

Example 6.1

Let $X(t) = 0$ for all t, $0 \leq t \leq 1$, and τ is a uniformly distributed random variable on [0, 1]. Let $Y(t) = 0$ for $t \neq \tau$ and $Y(t) = 1$ if $t = \tau$.

Now we have two stochastic processes: $\{X(t) : t \in [0,1]\}$ and $\{Y(t) : t \in [0,1]\}$. It appears that they are very different. All sample paths of X is identically 0 function, and every sample path of $Y(t)$ has a jump point at $t = \tau$.

Yet we notice that for any fixed t,

$$P(Y(t) \neq 0) = P(\tau = t) = 0.$$

That is, $P(Y(t) = 0) = 1$. Hence, for any $t \in [0,1]$, $X(t), Y(t)$ have the same distribution. Further, for any set of $0 \leq t_1 < t_2 < \cdots < t_n \leq 1$, the joint distribution of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ is identically to that of $\{Y(t_1), Y(t_2), \ldots, Y(t_n)\}$. That is, $X(t)$ and $Y(t)$ have the same finite dimensional distributions. ♢

The moral of this example is: even if two stochastic processes have the same distribution, they can still have very different sample paths. This observation prompts the following definition.

Definition 6.1

Two stochastic processes are called a version (modification) of one another if

$$P\{X(t) = Y(t)\} = 1 \text{ for all } t, 0 < t < T.$$

♢

In the case when a number of versions exist for a given distribution, a stochastic process with the smoothest sample path will be the version for investigation.

Having $P\{X(t) = Y(t)\} = 1$ does not mean the event $X(t) \neq Y(t)$ is an empty set. Let this set be called N_t. While the probability of each is zero, the probability of its union over $t \in [0, T]$ can be non-zero. In the above example, it equals 1. When $P(\cup_{0 \leq t \leq T} N_t) = 0$, the two stochastic processes are practically the same. They are **indistinguishable**.

Given a distribution of a stochastic process, can we always find a version of it so that its sample paths are continuous or regular (with only jump discontinuities)?

Theorem 6.2

Let $S(t), 0 \leq t \leq T$ be a real valued stochastic process.

(a) A continuous version of $S(t)$ exists if we can find positive constants α, ϵ and C such that

$$E|S(t) - S(u)|^{\alpha} \leq C|t - u|^{1+\epsilon}$$

for all $0 \leq t, u \leq T$.

(b) A regular version of $S(t)$ exists if we can find positive constants $\alpha_1, \alpha_2, \epsilon$ and C such that

$$E\{|S(u) - S(v)|^{\alpha_1}|S(t) - S(v)|^{\alpha_2}\} \leq C|t - u|^{1+\epsilon}$$

for all $0 \leq u \leq v \leq t \leq T$. $\diamondsuit$

There is no need to memorize this theorem. What we should make out of it? Under some continuity condition on expectations, a regular enough version of a stochastic process exists. We hence have legitimate base to investigate stochastic processes with this property.

Finally, let us summarize the results we discussed in this section. The distribution of a stochastic process is determined by its finite dimensional distributions, whatever it means. Even if two stochastic processes have the same distribution, their sample paths may have very different properties. However, for each given stochastic process, there often exists a continuous

version, or regular version, whether the given one itself has continuous sample paths or not.

Ultimately, we focus on the most convenient version of the given stochastic process in most applications.

6.3 Gaussian Process

Let us make it super simple.

Definition 6.1

Let $\{X(t) : 0 \leq t \leq T\}$ be a stochastic process. $X(t)$ is a Gaussian process if all its finite dimensional distributions are multivariate normal. $\diamondsuit$

We are at ease when working with normal, multivariate normal random variables. Thus, we also feel more comfortable with Gaussian processes when we have to deal with continuous time stochastic process taking arbitrary real values. We will be obsessed with Gaussian processes.

Example 6.2 *A simple Gaussian Process*

Let X and Y be two independent standard normal random variables. Let

$$S(t) = \sin(t)X + \cos(t)Y.$$

It is seen that $S(t)$ is a Gaussian process with certain covariance structures. $\diamondsuit$

6.4 Stationary Processes

Since the random behavior of a general stochastic process is very complex, we often first settle down with very simple ones and then timidly move toward more complex ones. The simple random work is the first example of stochastic process. We then introduce discrete time Markov chain and so on. We still hope to stay in the familiar territory by moving forward a little: introducing the concept of stationary processes here.

Definition 6.1 *Strong Stationarity.*

Let $\{X(t) : 0 \leq t \leq T = \infty\}$ be a stochastic process. If for any choices of $0 \leq t_1 < t_2 \ldots \leq t_n$, with any n and positive constant s, the joint distribution of

$$X(t_1 + s), X(t_2 + s), \ldots, X(t_n + s)$$

does not depend on s, then we say that $\{X(t) : 0 \leq t \leq T = \infty\}$ is strongly stationary. $\diamondsuit$

Definition 6.2 *Weak Stationarity.*

Let $\{X(t) : 0 \leq t \leq T = \infty\}$ be a stochastic process. If for any choices of $0 \leq t_1 < t_2 \ldots \leq t_n$, with any n and positive constant s, the mean and covariance matrix of

$$X(t_1 + s), X(t_2 + s), \ldots, X(t_n + s)$$

do not depend on s, then we say that $\{X(t) : 0 \leq t \leq T = \infty\}$ is weakly stationary, or simply stationary. $\diamondsuit$

The above definition can be easily applied to the case when $T < \infty$, or to discrete time stochastic processes. The obstacles are merely notational.

Being strongly stationary does not necessary imply the weak stationary, because of the moment requirements of the latter.

Most results for stationary processes are rather involved. We only mention two theorems.

Theorem 6.3 *Strong and Weak Laws of Large Numbers (Ergodic Theorems).*

(a) Let $X = \{X_n : n = 1, 2, \ldots\}$ be a strongly stationary process such that $E|X_1| < \infty$. There exists a variable Y with the same mean as the X_n such that

$$n^{-1} \sum_{i=1}^{n} X_i \to Y$$

almost surely, and in mean.

(b) Let $X = \{X_n : n = 1, 2, \ldots\}$ be a weakly stationary process. There exists a variable Y with the same mean as the X_n such that

$$n^{-1} \sum_{i=1}^{n} X_i \to Y$$

in the second moment. ♢

If a continuous time stationary process can be discretised, the above two results can also be very useful.

6.5 Stopping Times and Martingales

We are often interested in knowing how the property of a stochastic process following the moment when some event has just occurred. This is a setting point of a renewal process. These moments are random variables with special properties.

Consider the example of simple random walk $\{X_n : n = 0, 1, \ldots\}$, the time of returning to zero is a renewal event. At any moment, we examine the value of X_n to determine whether the process has renewal itself at n with respect to this particular renewal event.

In general, we need to examine the values of $\{X_i : i = 0, 1, \ldots, n\}$ to determine the occurrence of the renewal or other events. The time for the occurrence of such events is therefore a function of $\{X_i : i = 0, 1, \ldots, n\}$. In measure theory language, if τ represents such moments, then the event

$$\tau \leq n$$

is $\mathcal{F}_n = \sigma(X_0, \ldots, X_n\}$ measurable.

Definition 6.1 *Stopping Time*

Let $\{X(t) : 0 \leq t \leq T = \infty\}$ be a stochastic process. A nonnegative random variable τ, which is allowed to take value ∞, is called a stopping time if for each t, the event

$$\{\tau \leq t\} \in \sigma(X_s, 0 \leq s \leq t\}.$$

♢

Example 6.3 *To be invented.*

Chapter 7

Brownian Motion or Wiener Process

It was botanist R. Brown who first described the irregular and random motion of a pollen particle suspected in fluid (1828). Hence, such motions are called Brownian motion. Interestingly, the particle theory in physics was not widely accepted until the Einstein used Brown motion to argue that such motions are caused by bombardments of molecules in 1905. The theory is far from complete without a mathematical model developed later by Wiener (1931). This stochastic model used for Brown motion is hence also called Wiener process. We will see that the mathematical model has many desirable properties, and a few that do not fit well with physical laws for Brownian motion. As early as 1931, Brown motion has been used in mathematical theory for stock prices. It is nowadays a fashion to use stochastic processes to study financial market.

Definition 7.2 *Defining properties of Brownian motion.*

Brownian motion $\{B(t)\}$ is a stochastic process with the following properties:

1. (Independent increment) The random variable $B(t) - B(s)$ is independent of $\{B(u) : 0 \leq u \leq s\}$ for any $s < t$.

2. (Stationary Gaussian increment) The distribution of $B(t) - B(s)$ is normal with mean 0 and variance $t - s$.

3. (Continuity) The sample paths of $B(t)$ are continuous.

$\diamondsuit$

We should compare this definition with that of Poisson process: the increment distribution is changed from Poisson to normal. The sample path requirement is needed so that we will work on a definitive version of the process.

Once the distribution of $B(0)$ is given, then all the finite dimensional distributions of $B(t)$ are completely determined, which then determines the distribution of $B(t)$ itself. This notion is again analog to that of a counting process with independent, stationary and Poisson increments.

In some cases, $B(0) = 0$ is part of the definition. We will more or less adopt this convention, but will continue to spell it out.

Example 7.1

Assume that $B(t)$ is a Brownian motion and $B(0) = 0$.

1. Find the marginal distributions of $B(1)$, $B(2)$ and $B(2) - B(1)$.
2. Find the joint distribution of $B(1)$ and $B(2)$, and compute $P(B(1) > 1, B(2) < 0)$.

$\diamondsuit$

Recall the multivariate normal distribution is completely determined by its mean vector and covariance matrix. Thus, the finite dimensional distributions of a Gaussian process is completely determined by the mean function and the correlation function. As a special Gaussian process, it is useful to calculate the mean and correlation function of the Brownian motion.

Example 7.2 *Compute the correlation function of the Brownian motion.*

$\diamondsuit$

Based this calculation, it is clear that the above Brownian motion is not stationary.

Example 7.3 *Distribution of* $\int_0^1 B(t)dt$ *given* $B(t) = 0$.

♢

Example 7.4

Let $\xi_0, \xi_1, \ldots$ be a sequence of independent standard normally distributed random variables. Define

$$S(t) = \frac{t}{\sqrt{\pi}}\xi_0 + \frac{2}{\sqrt{\pi}}\sum_{i=1}^{\infty}\frac{\sin(it)}{i}\xi_i.$$

Assume that we can work with infinite summation as if nothing is unusual. Then we can verify the covariance structure of $S(t)$ resembles that of the Brownian motion with $B(0) = 0$ over $[0, \pi]$. Thus, $S(t)$ is a Brownian motion and it provides a very concrete way of represent the Brownian motion. It can be seen that all sample paths of $S(t)$ are continuous, yet they are nowhere differentiable. ♢

Example 7.5 *Sample paths of the Brownian motion.*

Brownian motions have the following well known but very peculiar properties.

1. Almost all sample paths are continuous;

2. Almost every sample path is not monotone in any interval, no matter how small the interval is;

3. Almost every sample path is not differentiable at any point;

4. Almost every sample path has infinite variation on any interval, not matter how short it is;

5. Almost all sample paths have quadratic variation on $[0, t]$ equal to t, for any t.

♢

Definition 7.3

Let $s(x)$ be a function of real variable.

(a) The (total) variation of $s(x)$ over $[0, t]$ is

$$\lim \sum_{i=1}^{n} |s(x_i) - s(x_{i-1})|$$

where the limit is taken over any sequence of sets of $(x_1, \ldots, x_n)$ such that

$$0 = x_0 < x_1 < x_2 < \cdots < x_n = t$$

and $\max\{x_i - x_{i-1} : i = 1, \ldots, n\} \to 0$.

(b) The quadratic variation over the interval $[0, t]$ is

$$\lim \sum_{i=1}^{n} \{s(x_i) - s(x_{i-1})\}^2$$

where the limit is taken over any sequence of sets of $(x_1, \ldots, x_n)$ such that

$$0 = x_0 < x_1 < x_2 < \cdots < x_n = t$$

and $\max\{x_i - x_{i-1} : i = 1, \ldots, n\} \to 0$. ♢

If $s(x)$ is monotone over $[0, t]$, then the total variation is simply $|s(t) - s(0)|$. Having a non-zero quadratic variation implies that the function fluctuates up and down very very often. It is hard to think of a function with such property. One example is $s(x) = \sqrt{x}\sin(1/x)$ over $[0, 1]$.

Theorem 7.1

If $s(x)$ is continuous and of finite total variation, then its quadratic variation is zero. ♢

Property 1 is true by definition. Property 2 can be derived from Property 4. If a sample path is monotone over an interval, then its variation is finite, which contradicts Property 4. Property 3 fits well with Property 2; if a sample path is nowhere monotone, it is very odd to be differentiable. This point can be made rigorous.

Finally, it is seen that Property 4 follows from Property 5 because of these arguments. It is hence vital to show that Brownian motion has Property 5.

Let $0 = t_0 < t_1 < t_2 < \cdots < t_n = t$ be a partition of the interval $[0, 1]$, where each of t_i depends on n though not spelled out. Let

$$T_n = \sum_{i=1}^{n} \{B(t_{i-1}) - B(t_i)\}^2$$

be the variation based on this partition. We find

$$\begin{aligned} E\{T_n\} &= \sum_{i=1}^{n} E\{B(t_{i-1}) - B(t_i)\}^2 \\ &= \sum_{i=1}^{n} Var\{B(t_{i-1}) - B(t_i)\} \\ &= \sum_{i=1}^{n} (t_i - t_{i-1}) = t. \end{aligned}$$

That is, the expected sample path quadratic variation is t, regardless of how the interval is partitioned.

Property 5, however, mean that $\lim T_n = t$ almost surely along any sequence of parititions such that $\delta_n = \max |t_i - t_{i-1}| \to 0$. We give a partial proof when $\sum d_n < \infty$. In this case, we have

$$\begin{aligned} Var\{T_n\} &= \sum_{i=1}^{n} Var\{B(t_{i-1}) - B(t_i)\}^2 \\ &= \sum_{i=1}^{n} 3(t_i - t_{i-1})^2 \\ &\leq 3t \max |t_i - t_{i-1}| = 3\delta_n t. \end{aligned}$$

Therefore, $\sum Var\{T_n\} = \sum E\{T_n - E(T_n)\}^2 < \infty$. Using monotone convergence theorem in measure theory, it implies $E \sum \{T_n - E(T_n)\}^2 < \infty$. (Note the change of order between summation and expectation). The latter implies $T_n - E(T_n) \to 0$ almost surely, which is the same as $T_n \to t$ almost surely.

Despite the beauty of this result, this property also shows that the mathematical Brownian motion does not model the physical Brownian motion in microscope level. If Newton's laws hold, the acceleration cannot be instantaneous, and the sample paths of a diffusion particle should be smooth.

7.1 Existence of Brownian Motion

One technical consideration is whether there exist stochastic processes with the properties prescribed by the definition of Brownian motion. This question is usually answered by defining a sequence of stochastic processes such that its limit exists, and having the defining properties of Brownian motion. One rigorous proof of the existence is scratched as follows.

Let $\xi_i, i = 1, 2, \ldots$ be independent and normally distributed with mean 0 and variance 1. The existence of such sequence and the corresponding probability space is well discussed in probability theory and is assumed here. Let

$$Y_n = \sum_{i=1}^{n} \xi_i$$

and define

$$B_n(t) = \frac{1}{\sqrt{n}}\{Y_{[nt]} + (nt - [nt])\xi_{[nt]+1}\}$$

which is a smoothed partial sum of ξ's. It can be verified that the covariance structure of $B_n(t)$ approaches what is assumed for Brownian motion. All finite dimensional distributions of $B_n(t)$ are normal. With a dose of asymptotic theory which requires the verification of tightness, it is seen that $B_n(t)$ has a limit, and the limiting process has all the defining properties of a Brownian motion. Thus, the existence is verified. See Page 62 of Billingsley (1968).

One may replace $\xi_i, i = 1, 2, \ldots$ by independent and identically distributed random variables taking $\pm a_n$ values. We may define the partial sum Y_n in the same way, and define a similar stochastic process. By properly scaling down the time, and letting $a_n \to 0$, the limiting process will also have the defining properties of a Brownian motion. Thus, Brownian motion is regarded as a limit of simple random walk. Many of other properties of Brownian motion to be discussed have their simple random walk versions.

7.2 Martingales of Brownian Motion

The reason for Brownian motion being the centre of most textbooks and courses in stochastic processes can be attributed to the fact that it is a good

example of any important concepts in continuous time stochastic process. This section introduces its martingale aspects.

Recall that each random variable X generates a σ-field: $\sigma(X)$. This extends to a set of random variables. If $X(t)$ is a stochastic process on $[0, T]$, then $\{X(u) : 0 \leq u \leq t\}$ is a set of random variables and it generates a σ-field which is usually denoted as $\mathcal{F}_t$. As sets of sets, these σ-fields satisfy

$$\mathcal{F}_s \subset \mathcal{F}_t$$

for all $0 \leq s < t \leq T$.

We may put forward a sequence of σ-field $\mathcal{F}_t$ not associated with any stochastic processes. As long as $\{\mathcal{F}_t, t \geq 0\}$ has the increasing property, it is called a filtration.

Definition 7.1 *Martingale*

A stochastic process $\{X(t), t \geq 0$ is a martingale with respect to $\mathcal{F}_t$ if for any t it is integrable, $E|X(t)| \leq \infty$, and for any $s > 0$

$$E\{X(t+s)|\mathcal{F}_t\} = X(t).$$

$\diamondsuit$

With this definition, we have implied that $X(t)$ is $\mathcal{F}_t$ measurable. In general, $\mathcal{F}_t$ represents information about $X(t)$ available to an observer up to time t. If any decision is to be made at time t by this observer, the decision is a $\mathcal{F}_t$ measurable function. Unless otherwise specified, we take $\mathcal{F}_t = \sigma\{X(s) : s \leq t\}$ when a stochastic process $X(t)$ and a filtration $\mathcal{F}_t$ are subjects of some discussion.

One of the defining properties of Brownian motion can be reworded as the Brownian motion is a martingale.

We would like to mention two other martingales.

Theorem 7.2

Let $B(t)$ be a Brownian motion.

(a). $B^2(t) - t$ is a martingale;

(b). For any θ, $\exp\{\theta B(t) - \frac{1}{2}\theta^2 t\}$ is a martingale. $\diamondsuit$

Proofs will be provided in class.

7.3 Markov Property of Brownian Motion

Let $\{X(t), t \geq 0\}$ be a stochastic process and $\mathcal{F}_t$ is the corresponding filtration. The Markov property is naturally extended as follows.

Definition 7.1 *Markov Process.*

If for any $s, t > 0$,

$$P\{X(t+s) \leq y | \mathcal{F}_t\} = P\{X(t+s) \leq y | X_t\}$$

almost surely for all y, then $\{X(t), t \geq 0\}$ is a Markov process. $\diamondsuit$

Compared to the definition of Markov chain, this definition removes the requirement that the state space of $X(t)$ is countable.

Theorem 7.3 *A Brownian motion is a Markov Process.*

PROOF: Working out the conditional moment generating function will be sufficient. $\diamondsuit$

In applications, we often want to know the property following the occurrence of some event. A typical example is the behavior of a simple random walk after its return to 0. Since the returning time itself is random, we must be careful when claim that the simple random walk following that moment will have the same property as a simple random walk starting from 0.

If we recall, this claim turns out to be true. Regardless, we must be prepared to prove that this is indeed the case. As we mentioned before, Brownian motion in many respects generalizes the simple random walk. They share many similar properties. One of them is the strong Markov property. As a preparation, we re-introduce the concept of stopping times.

Definition 7.2

Let $\{X(t), t \geq 0\}$ be a stochastic process, and $\mathcal{F}_t = \sigma\{X(s) : s \leq t\}$. A random time T is called a stopping time for $\{X(t), t \geq 0\}$ if $\{T \leq t\} \in \mathcal{F}_t$ for all $t > 0$. $\diamondsuit$

If we stripe the jargon of σ-field, a stopping time T is a random quantity such that after we observed the stochastic process up to and include time t, we can tell whether $T \leq t$ or not.

Let T to be the time when a Brownian motion first exceeds value 1. Then T is a stopping time. By examining the values of $X(s)$ for all s smaller or equal t, the truthfulness of $T \leq t$ is a simple matter.

Let T be the time when the value of stock price will drop by 50% in the next unit of time. If T were a stopping time, we would make a lot of money. In this case, you really have to be able to see into the future. We do not need strong mathematics to find that T is not a stopping time.

Another preparation is: for each non-random time t, $\mathcal{F}_t$ is the σ-field generated by random variables $X(s)$ including all $s \leq t$. If T is a stopping time, we hope to be able to use it as if it is a non-random time. We define

$$\mathcal{F}_T = \{A : A \in \mathcal{F}, A \cap \{T \leq t\} \in \mathcal{F}_t \text{ for any } t\}.$$

Finally, let us return to the issue of strong Markov property.

Theorem 7.4 *Strong Markov Property.*

Let T be a stopping time associated with Brownian motion $B(t)$. Then for any $t > 0$,

$$P\{B(T+t) \leq y|\mathcal{F}_T\} = P\{B(T+t) \leq y|B(T)\}$$

for all y. ♢

We cannot afford to spend more time at proving this result. Let us tentatively accept it as a fact and see what will be the consequences.

Let us define $\hat{B}(t) = B(T+t) - B(T)$. When T is a constant, it can be easily verified that $\hat{B}(t)$ is also a Brownian motion. When T is a stopping time, it remains to be true due to the strong Markov property.

7.4 Exit Times and Hitting Times

Let $T(x) = \inf\{t > 0 : B(t) = x\}$ be a stopping time for Brownian motion $B(t)$. It is seen that $T(x)$ is the first moment when the Brownian motion hits target x.

Assume $B(0) = x$ and $a < x < b$. Define $T = \min\{T(a), T(b)\}$. Hence T is the time when the Brownian motion escapes the area formed by two horizontal lines. The question is: will it occur in finite time? If so, what is the average time it takes?

The first question is answered by computing $P(T < \infty)$, and the second question is question is answered by computing $E\{T\}$ under the assumption that the first answer is affirmative. One may link this problem with simple random walk and predict the outcome. The probability for the first passage of "1" for a symmetric simple random walk is 1. Since the normal distribution is symmetric, the simple random walk result seems to suggest that $P(T < \infty) = 1$. We do not have similar results for $E\{T\}$, but for $E\{T(a)\}$ or $E\{T(b)\}$ for simple random walk, and the later are likely finite from the same consideration. The real answers are given as follows.

Theorem 7.5

The hitting time T defined above have the properties:

(a) $P(T < \infty | B(0) = x) = 1$;

(b) $E\{T | B(0) = x\} < \infty$, under the assumption that $a < x < b$.

$\diamondsuit$

PROOF: It is seen that the event $\{T > 1\}$ implies $a < B(t) < b$ for all $t \in [0, 1]$. Hence,

$$\begin{aligned} P(T > 1 | B(0) = x) &\leq P\{a < B(1) < b | B(0) = x\} \\ &= \Phi(b - x) - \Phi(a - x). \end{aligned}$$

As long as a, b are finite,

$$\alpha = \sup_x \{\Phi(b - x) - \Phi(a - x)\} < 1.$$

Next, for any positive integer n,

$$\begin{aligned} p_n &= P\{T > n | B(0) = x\} \\ &= P\{T > n | T > n - 1, B(0) = x) p_{n-1} \end{aligned}$$

$$\begin{aligned}
&= P\{a < B(s) < b, \text{ for } n-1 \le s \le n | a < B(s) < b, \\
&\quad \text{for } 0 \le s \le n-1, B(0) = x\} p_{n-1} \\
&= P\{a < B(s) < b, \text{ for } n-1 \le s \le n | a < B(n-1) < b\} p_{n-1} \\
&= P\{a < B(s) < b, \text{ for } 1 \le s \le 1 | a < B(0) < b\} p_{n-1} \\
&\le \alpha p_{n-1}.
\end{aligned}$$

Applying this relationship repeatedly, we get $p_n \le \alpha^n$. Thus, we have

$$P(T < \infty | B(0) = x) = 1 - \lim_{n\to\infty} p_n = 1.$$

Further,

$$E(T|B(0) = x) \le \sum_{n=0}^{\infty} P\{T > n | B(0) = x\} < \infty.$$

♢

I worry slightly that I did not really make use of strong Markov property, but the Markov property itself.

The following results are obvious:

$$P\{T(b) < \infty | B(0) = a\} = 1; \quad P\{T(a) < \infty | B(0) = a\}.$$

7.5 Maximum and Minimum of Brownian Motion

Define, for a given Brownian motion,

$$M(t) = \max\{B(s) : 0 \le s \le t\} \text{ and } m(t) = \min\{B(s) : 0 \le s \le t\}.$$

The significance of these two quantities are obvious.

Again, the following result resembles a result of simple random walk, and I believe that a proof using some limiting approach is possible.

Theorem 7.6

For any $x > 0$,

$$P\{M(t) > m | B(0) = 0\} = 2P\{B(t) \le m | B(0) = 0\}.$$

PROOF: We will omit the condition that $B(0) = 0$. Also, we use $T(a)$ as the stopping time when $B(t)$ first hits a.

First, for any m, we have

$$\begin{aligned} P\{M(t) \geq m\} &= P\{M(t) \geq m, B(t) \geq m\} \\ &\quad + P\{M(t) \geq m, B(t) < m\}. \end{aligned}$$

Note that $B(T(m)) = m$ as the sample paths are continuous almost surely, plus $\{M(t) > m\}$ is equivalent to $T(m) \leq t$.

$$\begin{aligned} & P\{M(t) \geq m, B(t) < m\} \\ &= P\{T(m) < t, B(t) < B(T(m))\} \\ &= P\{B(t) < B(T(m)) | T(m) < t\} P\{T(m) < t\} \\ &= P\{B(t) > B(T(m)) | T(m) < t\} P\{T(m) < t\} \\ &\qquad\qquad \text{Strong Markov Property} \\ &\qquad\qquad \text{Reflection Principle} \\ &= P\{M(t) \geq m, B(t) \geq m\}. \end{aligned}$$

Applying this back, we get

$$P\{M(t) \geq m\} = 2P\{M(t) \geq m, B(t) \geq m\}.$$

Since $B(t) \geq m$ implies $M(t) \geq m$, the above result becomes

$$P\{M(t) \geq m\} = 2P\{B(t) \geq m\}.$$

This completes the proof. ♢

This result can be used to find the distribution function of $T(x)$ under the assumption that $B(0) = 0$.

Theorem 7.7

The density function of $T(x)$ is given by

$$f(t) = \frac{|x|}{\sqrt{2\pi t^3}} \exp\{-\frac{x^2}{2t}\} \quad t \geq 0.$$

PROOF: It is done by using the equivalence between $T(x) \leq t$ and $M(t) \geq x$. $\diamondsuit$

We call $T(x)$ the **first passage time**.

The reflection idea is formally stated as the next theorem.

Theorem 7.8

Let T be a stopping time. Define $\hat{B}(t) = B(t)$ for $t \leq T$ and $\hat{B}(t) = 2B(t) - B(t)$ for $t > T$. Then $\hat{B}(t)$ is also Brownian motion. $\diamondsuit$

The proof is not hard, and we should draw a picture to understand its meaning.

7.6 Zeros of Brownian Motion and Arcsine Law

Assume $B(0) = 0$. How long does it take for the motion to come back to 0? This time, the conclusion is very different from that of simple random walk.

Theorem 7.9

The probability that $B(t)$ has a least one zero in the time interval (a, b) is given by

$$\frac{2}{\pi} \arccos \sqrt{a/b}.$$

PROOF: Let $E(a, b)$ be the event that $B(t) = 0$ for at lease some $t \in (a, b)$.

$$P\{E(a, b)\} = E[P\{E(a, b)|B(a)\}]$$

For any $x > 0$,

$$P\{E(a, b)|B(a) = -x\} = P\{T(x) < b - a|B(0) = 0\} = P\{T(x) < b - a\}$$

by Markov property and so on.

Hence,

$$\begin{aligned} E[P\{E(a,b)|B(a)\}] &= \int_{-\infty}^{\infty} P\{E(a,b)|B(a)=x\}f_a(x)dx \\ &= 2\int_0^{\infty} P\{E(a,b)|B(a)=-x\}f_a(x)dx \\ &= 2\int_0^{\infty} P\{T(x)<b-a\}f_a(x)dx \end{aligned}$$

where $f_a(x)$ is the density function of $B(a)$ and known to be normal with mean 0 and variance a. The density function of $T(x)$ is given earlier. Substituting two expressions in and finishing the job of integration, we get the result. ♢

When $a \to 0$, so that $a/b \to 0$, this probability approaches 1. Hence, the probability that $B(t) = 0$ in any small neighborhood of $t = 0$ (not including 0) is 1 given $B(0) = 0$. This conclusion can be further strengthened. There are infinite number of zeros in any neighborhood of $t = 0$, no matter how small this neighborhood is. This result further illustrates that the sample paths of Brownian motion, though continuous, is nowhere smooth.

The famous Arcsine law now follows.

Theorem 7.10

The probability that Brownian motion has no zeros in the time interval (a, b) is given by $(2/\pi)\arcsin\sqrt{(a/b)}$. ♢

There is a similar result for simple random walk.

7.7 Diffusion Processes

To be continued.