Notes on the Wiener Process and Ito's Lemma

JOHN R. SMITH
email: jrsmith@ucdavis.edu

Physics Department
University of California, Davis
Davis, CA 95616-8677, USA

August 23, 2017

1 Introduction

A stochastic process $\{X(t), t \in T\}$ is a parameterized collection of random variables. That is, for each $t \in T, X(t)$ is a random variable. The set T is the *parameter space* or *index set* of the process and contains the allowed values of t. An index set T is said to be a *linear* index set if it has the property that the sum $t+h$, of any two members t and h of T, also belongs to T. If t is *time*, we can call the stochastic process a *time series*. Many results in stochastic processes are for time series – where observations are made over a period of time for particular random quantities, e.g. the observation of the whole time history of the temperature at a particular location every hour for a year.

$X(t)$ represents the *state* of the process at time t. When T is countable, then the stochastic process is said to be a *discrete-time* process. If T is an interval of the real line, the stochastic process is said to be a *continuous-time* process. Discrete-time processes can be represented as $\{X_n, n = 1, 2, ...\}$; while continuous-time processes can be represented as $\{X(t), t \geq 0\}$.

Statistical problems generally work with a sample chosen from a given set of elementary outcomes that can be selected from a larger space which is denoted by Ω. Ω is the universe of possible outcomes and is called the *sample space*. The elementary outcomes contained in Ω are labeled ω. Possible sets of outcomes are referred to as *events*. If A is an event, then A consists of a set of possible outcomes and is itself a subset contained in $\Omega : A \subset \Omega$. With time series, each elementary outcome represents an entire set of random variables, e.g., $\{X(t), t \in T\}$. We will use the notation $\{X(t, \omega), t \in T, \omega \in \Omega\}$, or the shorter version $\{X(t), t \in T\}$ where is is understood that $\{X(t), t \in T\}$ corresponds to a single value of ω.

In experimental practice, it is possible to vary the length of the observed time series by changing the time interval T, however it is usually impossible to observe more than one time series at a time. So we end up with a single outcome of the stochastic process and a single observation of the random

variable of interest at time t. We view the observed time series as one outcome in a much larger sample space (denoted as Ω which could have an infinity of possible outcomes). The particular observed time series is said to be a *realization* of the stochastic process. Here one outcome in the sample space is a single and entire sample function defined on the whole interval T.

The theory of stochastic processes can be viewed as the dynamical part of probability theory.

We will study probability and stochastic processes using various concepts such as *mean squared convergence* which is simpler than other kinds of convergence summarized in Section [14.3]. We will be concerned in these notes with the properties of a particular stochastic process: The Wiener Process or Brownian Motion. An important fact is that there are different ways to work with stochastic processes, each with varying degrees of difficulty and with different degrees of completeness. *Ordinary* calculus is too restrictive because it requires the results to hold for *every* possible sample function. For most cases of interest, it is not necessary for a result to hold for all sample functions – but to hold for *almost all* sample functions. Therefore a less restrictive approach than using ordinary calculus is to ignore those sample functions belonging sets in the sample space with probability zero (probability zero is the same as measure zero). If results are required to converge at the individual sample function level (one complete sample function at a time) with probability one (wp1), then the tools require some sophistication in the methods used such as Lebesgue measure theory.

Many results in stochastic processes are not required to hold wp1 at the sample function level – but only to hold *on average* for the most important cases which influence the statistical averages. This is the motivation for using *mean-squared calculus* – it discards the difficulties involved with dealing wp1 at the sample function level and works only with the *important* sample functions that affect the statistical averages – which covers most of the cases of interest in applications. In a nut shell, what mean-squared calculus has too offer is *simplicity*. We will occasionally work with results at the sample function level – such as the quadratic and total variation properties of the Brownian paths as well as the Kolmogorov continuity theorem. We will make comparisons occasionally between the mean-squared and the sample-function-level approaches.

2 Probability Spaces & Random Variables

Let us summarize some of the basic concepts in probability theory.

The sample space Ω consists of the set of all elementary outcomes of an experiment. The family of sets $\mathcal{F}$ (called a "σ-algebra") on Ω form a measurable space $(\Omega, \mathcal{F})$. The σ-algebra is a family of subsets of all the possible subsets of Ω such that a probability function can be defined on the members of the family (which is, in general, an impossible task for all the possible subsets of Ω).

The family, $\mathcal{F}$, consists of sets with the following properties:

(i) $\emptyset \in \mathcal{F}$
(ii) $F \in \mathcal{F} \implies \mathcal{F}^{\mathcal{C}} \in \mathcal{F}$, where $F^C = \Omega \backslash F$ is the complement of F in Ω
(iii) $A_1, A_2, ... \in \mathcal{F} \implies A := \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Point functions like $f : A \to B$ map the set A into the set B. The domain of f consists points in A and the range is the set $\{f(x)|x \in A\}$. We think of $f(x)$ as the *image* of the point x. Probability, on the other hand, is a *set function* whose domain consists of sets. In other words, probability is a function that assigns a number between 0 and 1 to a set A, i.e., the probability of A occurring is given by the function $P(A)$.

Probability is a *continuous set function*, $P(A)$, defined on the space $(\Omega, \mathcal{F})$ containing subsets $A \subset \Omega$. It is sometimes called a probability measure and is an example of a measure. The mathematical concept of measure is analogous to determining the size of set. For finite sets one could use a function that counted the number of entries in a set to determine the "size". A probability measure on a measurable space $(\Omega, \mathcal{F})$ is a function $P : \mathcal{F} \to [0, 1]$ such that

(Axiom 1) if $A \in \mathcal{F}$, then $0 \leq P(A) \leq 1$ (as defined above).

(Axiom 2) $P(\emptyset) = 0, P(\Omega) = 1$

(Axiom 3) *Countable Additivity:* if $A_1, A_2, ... \in \mathcal{F}$ and $\{A_i\}_{i=1}^{\infty}$ is disjoint (i.e., $A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i)$$

The triple $(\Omega, \mathcal{F}, P)$ is called a probabilty space.

Using these concepts we can develop the idea of a "random variable" in terms of a function of the elementary outcomes $\omega \in \Omega$. A scalar random variable $x(\cdot)$ is a real-valued point function which assigns a real scalar value to each point ω in the sample space Ω, denoted as $x(\omega) = x$, such that every set $A \subset \Omega$ of the form

$$A = \{\omega : x(\omega) \leq \xi\}$$

for any value ξ on the real line ($\xi \in \mathbb{R}$), is an element of the σ-algebra $\mathcal{F}$ (i.e., $A \in \mathcal{F}$) and therefore lies in the domain of the probability function. The name "random variable" is perhaps unfortunate in that it does not seem to imply the fact that we are talking about a function, as opposed to values the function can assume. In fact, $x(\cdot)$ is a function, or mapping, from Ω into $\mathbb{R}$.

A vector random variable or random vector $\mathbf{x}(\cdot)$ is just the generalization of the random variable concept to the multidimensional (vector) case: a real-valued point function which assigns a real vector to each point ω in Ω, denoted as $\mathbf{x}(\omega)$, such that every set A of the form

$$A = \{\omega : \mathbf{x}(\omega) \leq \xi\}$$

for any $\xi \in \mathbb{R}^n$, is an element of $\mathcal{F}$.

3 Probability as a Sequentially Continuous Set Function

3.1 Sequential Continuity

Definition In ordinary calculus a sequence $x_1, \ldots, x_n \ldots$ *converges to a limit* L iff for every $\epsilon > 0$ there is a positive integer N such that

$$|x_n - L| < \epsilon \text{ whenever } n > N. \text{ We also write this as } x_n \to L \text{ as } n \to \infty.$$

Definition An infinite sequence $\{x_n\}$ is called a *Cauchy sequence* iff for each $\epsilon > 0$, there is a positive integer N such that

$$|x_n - x_m| < \epsilon \quad \text{for all } m > N \text{ and } n > N. \tag{1}$$

Theorem 3.1 *(Cauchy convergence criterion). A necessary and sufficient condition for convergence of a sequence* $\{x_n\}$ *is that it be a Cauchy sequence.*

Proof See any book on real analysis such as [1]. ∎

The Cauchy convergence criterion provides a means of establishing the convergence of a sequence to a limit without knowing the limit itself.

The idea of a continuous function in ordinary analysis is given by the following

Definition Suppose that D is a subset of $\mathbb{R}$ and $f : D \to \mathbb{R}$. The function f is continuous at a iff (1) the point a is in in an open interval $I \subset D$, and (2) there is a positive number δ such that

$$|f(x) - f(a)| < \epsilon \quad \text{whenever } |x - a| < \delta.$$

The above definition can be extended to functions which act on subsets of a metric space S, e.g., $S = \mathbb{R}^n$. In that case continuity of a function f is defined as

Definition Let A be a subset of a metric space S, and suppose $f : A \to \mathbb{R}$. Let $p_0 \in A$. We say that f is *continuous with respect to* A *at* p_0 iff (1) $f(p_0)$ is defined, and (2) either p_0 is an isolated point of A or p_0 is a limit point of A and

$$f(p) \to f(p_0) \quad \text{as } p \to p_0, \quad p \in A.$$

We say that *f is continuous on A* iff f is continuous with respect to A at every point of A.

The above ideas and definitions are basic to real analysis and can be found in [1].

The concept of continuity can be extended to the probability function which is defined on the sets of the σ-algebra.

We begin by considering *increasing* and *decreasing* sequences of sets in the σ-algebra and show how to define the probability for limits of such sequences.

A sequence of events $\{A_n, n \geq 1\}$ is said to be an *increasing sequence* if

$$A_1 \subset A_2 \subset \ldots \subset A_n \subset A_{n+1} \subset \ldots$$

A sequence of events $\{A_n, n \geq 1\}$ is said to be a *decreasing sequence* if

$$A_1 \supset A_2 \supset \ldots \supset A_n \supset A_{n+1} \supset \ldots$$

If $\{A_n, n \geq 1\}$ is an increasing sequence of events, then it is possible to define an event representing the limit by $\lim_{n\to\infty} A_n$ as

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$$

Similarly if $\{A_n, n \geq 1\}$ is a decreasing sequence of events, we define $\lim_{n\to\infty} A_n$ by

$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$$

Proposition 3.2 *If $\{A_n, n \geq 1\}$ is either an increasing or decreasing sequence of events, then*

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

Suppose that $\{A_n, n \geq 1\}$ is an increasing sequence of events and define new events $C_n, n \geq 1$ by

$$\begin{aligned} C_1 &= A_1 \\ C_n &= A_n \cap \Big(\bigcup_{i=1}^{n-1} A_i\Big)^c = A_n \cap C_{n-1}^c \qquad n > 1 \end{aligned}$$

where C_i^c is the set complement of C_i and also $\bigcup_{i=1}^{n-1} C_i = C_{n-1}$, since C_i is an increasing sequence of events. So C_n consists of the points in A_n which are not in any of the previous $A_i, i < n$. Now the C_i are mutually exclusive sets and also

$$\bigcup_{i=1}^{\infty} C_i = \bigcup_{i=1}^{\infty} A_i \qquad \text{and} \qquad \bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} A_i \qquad n > 1$$

We have the following sequential continuity property: $P\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$.

$$\begin{aligned}
P(\lim_{n\to\infty} A_n) &= P(\bigcup_{i=1}^{\infty} A_i) && A_n \text{ is a nested increasing sequence} \\
&= P(\bigcup_{i=1}^{\infty} C_i) && [\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} C_i \quad \text{for all } n \text{ (including } n=\infty)] \\
&= \sum_{i=1}^{\infty} P(C_i) && \text{Axiom 3 (Countable Additivity)} \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} P(C_i) && \text{definition} \\
&= \lim_{n\to\infty} P(\bigcup_{i=1}^{n} C_i) && \text{Axiom 3 (Countable Additivity)} \\
&= \lim_{n\to\infty} P(\bigcup_{i=1}^{n} A_i) && [\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} C_i \quad \text{for all } n \text{ (including } n=\infty)] \\
&= \lim_{n\to\infty} P(A_n) && A_n \text{ is a nested increasing sequence}
\end{aligned}$$

for increasing sequences $\{A_n, n \geq 1\}$.

If $\{A_n, n \geq 1\}$ is a decreasing sequence, then $\{A_n^c, n \geq 1\}$ is an increasing sequence and we can use the previous results

$$P\Big(\bigcup_{i=1}^{\infty} A_i^c\Big) = \lim_{n\to\infty} P(A_n^c)$$

But since $\bigcup_{i=1}^{n} A_i^c = (\bigcap_{i=1}^{\infty})^c$ using DeMorgan's laws (see the Appendix [14.4]), we find

$$P\Big[\Big(\bigcap_{i=1}^{\infty} A_i\Big)^c\Big] = \lim_{n\to\infty} P(A_n^c)$$

Since $P(A^c) = 1 - P(A)$ we can write the previous relation as

$$1 - P\Big(\bigcap_{i=1}^{\infty} A_i\Big) = \lim_{n\to\infty} [1 - P(A_n)] = 1 - \lim_{n\to\infty} P(A_n)$$

or

$$P\Big(\bigcap_{i=1}^{\infty} A_i\Big) = \lim_{n\to\infty} P(A_n).$$

We can define limits for general sequences of events $\{A_n, n \geq 1\}$ which are not necessarily nest as in

the above examples. Let $A_1, A_2, \ldots$ be a sequence of arbitrary events. Construct the events

$$\begin{aligned} B_n &\equiv \bigcup_{i=n}^{\infty} A_i, \qquad B_n \text{ is a nested decreasing sequence of events, and} \\ C_n &\equiv \bigcap_{i=n}^{\infty} A_i, \qquad C_n \text{ is a nested increasing sequence of events.} \end{aligned}$$

Notice that

$$C_n \equiv \bigcap_{i=n}^{\infty} A_i \subset A_n \subset \bigcup_{i=n}^{\infty} A_i \equiv B_n$$

for all n. B_n and C_n are legitimate events because countable intersections and unions of events are always events belonging to the σ-algebra. Now B_n is nested decreasing and C_n is nested increasing and we can use the results above to describe their limits.
For example, consider the new events denoted by $\limsup_{n\to\infty} A_n$ and $\liminf_{n\to\infty} A_n$ as follows

$$\begin{aligned} B &= \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} B_n \\ C &= \liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} C_n \end{aligned}$$

$\limsup_{n\to\infty} A_n$ is the intersection of the events $\bigcup_{i=n}^{\infty} A_i$ and it follows that a point $x \in B$ if, for all n, it is contained in at least one of the events A_i for $i \geq n$. But this is equivalent to x being contained in an infinite number of the events $A_n, n \geq 1$. Therefore $\limsup_{n\to\infty} A_n$ consists of all x that are contained in an infinite number of the events $A_n, n \geq 1$, or that x occurs *infinitely often* (i.o.).

On the other hand a point $x \in C$ if, for some n, x is contained in $\bigcap_{i=n}^{\infty} A_i$. Therefore for some $n, x \in A_i$ for all $i \geq n$.. But this is equivalent to x being contained in all but a finite number of the events $A_n, n \geq 1$.

Definition If $\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n$ then we say that the $\lim_{n\to\infty} A_n$ exists and is equal to the common limit:

$$\lim_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n.$$

In this case we have:

Theorem 3.3 *For any sequence of events $\{A_n, n \leq 1\}$ for which $\lim_{n\to\infty} A_n$ exists,*

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

Proof Since $\left\{ \bigcup_{i=n}^{\infty} A_i, n \geq 1 \right\}$ is a decreasing sequence of events, it follows from Proposition [3.2] that

$$P\Big(\limsup_{n\to\infty} A_n \Big) = \lim_{n\to\infty} P\Big(\bigcup_{i=n}^{\infty} A_i \Big)$$

and, because $A_n \subset \bigcup_{1=n}^{\infty} A_i$ then $P(A_n) \leq P(\bigcup_{i=n}^{\infty} A_i)$ and therefore

$$\lim_{n\to\infty} P\Big(\bigcup_{i=n}^{\infty} A_i \Big) \geq \overline{\lim}_{n\to\infty} P(A_n)$$

(where for any sequence of real numbers $b_n, n \geq 1, \overline{\lim}_n b_n$ is defined to equal the largest limit point of the set $\{b_n, n \geq 1\}$. In addition, $\underline{\lim}_n b_n$ is defined to equal the smallest limit point. We say that $\lim_{n\to\infty} b_n$ exists if $\overline{\lim}_{n\to\infty} b_n = \underline{\lim}_n b_n$). Therefore

$$P\Big(\limsup_{n\to\infty} A_n \Big) \geq \overline{\lim}_n P(A_n)$$

Similarly, since $\left\{ \bigcap_{i=n}^{\infty} A_i, n \geq 1 \right\}$ is an increasing sequence of events, it follows from Proposition 3.2 that

$$\begin{aligned} P\Big(\liminf_{n\to\infty} A_n \Big) &= P\Big(\bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i \Big) \\ &= \lim_{n\to\infty} P\Big(\bigcap_{i=n}^{\infty} A_i \Big) \\ &\leq \underline{\lim}_n P(A_n) \end{aligned}$$

where the last inequality follows because $\bigcap_{i=n}^{\infty} A_i \subset A_n$. Therefore if $\lim_{n\to\infty} A_n$ exists, then we have

$$\overline{\lim}_n P(A_n) \leq P(\limsup_{n\to\infty} A_n) = P(\lim_{n\to\infty} A_n) = P(\liminf_{n\to\infty} A_n) \leq \underline{\lim}_n P(A_n)$$

which, since $\overline{\lim}_n P(A_n) \geq \underline{\lim}_{n\to\infty} P(A_n)$ gives

$$P\Big(\lim_{n\to\infty} A_n \Big) = \overline{\lim_{n\to\infty}} P(A_n) = \underline{\lim_{n\to\infty}} P(A_n) \qquad \blacksquare$$

3.2 Distribution Functions

In dealing with specific random variables which can take on continuous values we will find it useful to express the probability that a random variable takes on any value less than a specified number. We define the Cumulative Distribution Function (CDF) as follows

$$F_X(x) = P\{x(\omega) \leq x\}$$

Then the probability that $x < x(\omega) \leq x + \Delta x$ can be expressed in terms of the CDF as

$$P\{x < x(\omega) \leq x + \Delta x\} = F_X(x + \Delta x) - F_X(x) \approx \frac{dF_X(x)}{dx} \cdot \Delta x = f_X(x)\Delta x$$

where $f_X(x) = \frac{dF_X(x)}{dx}$ (assuming the derivative of $F_X(x)$ exists). In differential form we can write

$$dF_X(x) = f_X(x)\, dx. \tag{2}$$

$f_X(x)$ is called the Probability Density Function (PDF) and can be used to calculate the probability that x is the infinitesmal range

$$P\{x < x(\omega) \leq x + \Delta x\} \approx \frac{dF_X(x)}{dx}\Delta x = f_X(x)\Delta x.$$

This computation can be extended to any finite range ($x_1 < x \leq x_2$) accordingly by integration

$$P\{x_1 < x(\omega) \leq x_2\} = \int_{x_1}^{x_2} dF_X(x) = \int_{x_1}^{x_2} f_X(x)dx.$$

4 Gaussian Random Variables

The main random variable we will concern ourselves with is the Gaussian or normal random variable with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$. The Standard Normal Distribution has parameters $\mu = 0$ and $\sigma = 1$ and is denoted by $N(0, 1)$. If X is a normal random variable distributed according to $N(\mu, \sigma^2)$ we write $X \sim N(\mu, \sigma^2)$ The Gaussian random variable is continuous and can take on any value in the real number line. The probability distribution function for a $X \sim N(\mu, \sigma^2)$ random variable is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty$$

The probability for X to take on values in a subset $B \subset R$ is given by

$$P\{X \in B\} = \int_{x \in B} f_X(x)\, dx.$$

The probability density function is normalized to unity. Using the change-of-variable $u = (x - \mu)/\sigma$:

$$\int_{-\infty}^{\infty} f_X(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du = 1 \tag{3}$$

If $X \sim N(\mu, \sigma^2)$, then $Z = \alpha X + \beta$ will be distributed according to $Z \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$:

Proof

$$\begin{aligned} F_Z(a) &= P(Z \leq a) \\ &= P(\alpha X + \beta \leq a) \\ &= P\left(X \leq \frac{a-\beta}{\alpha}\right) \\ &= F_X\left(\frac{a-\beta}{\alpha}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\left(\frac{a-\beta}{\alpha}\right)} e^{-(x-\mu)^2/2\sigma^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{[z-(\alpha\mu+\beta)]^2}{2\alpha^2\sigma^2}}\, dz \end{aligned}$$

As shown above, $Z \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$. ∎

The CDF for the standard normal random variable is $\Phi(a)$ and is given by

$$\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-z^2/2}\, dz \tag{4}$$

5 Expectation and Variance

5.1 1-dimensional Random Variables

Expectation is the averaging operation carried out over the outcomes ω in Ω according to their respective probabilities.

Consider the example of a continuous random variable X with probability density function $f_X(x)$. Then the *Mean*, μ, is given by

$$\mu = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

in which case $E[X]$ is just the definition of the mean value of X. This concept can be extended to functions of the random variable X, i.e., g(X) and we find that

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

This result is intuitively reasonable but needs to be proven also from the definition of the averaging method over the random variable $Y = g(X)$.

The *Variance*, σ^2, is given by

$$\sigma^2 = \mathrm{Var}[X] = E[(X - E[X])(X - E[X])] = E[X^2] - (E[X])^2.$$

Since $\mu = E[X]$, this result can also be written as

$$\sigma^2 = E[X^2] - \mu^2.$$

If one has two random variables X and Y then we will distinguish the respective means by $\mu_X = E[X]$ and $\mu_Y = E[Y]$ and define the *Covariance* by Cov[Y,Y] which is given by

$$\mathrm{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[(X - \mu_X)(Y - \mu_Y)].$$

For a Gaussian random variable X, let us compute $E[X]$:

$$E[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma^2}\, dx$$

Making the substitution $u = (x - \mu)/\sigma, du = dx/\sigma, E[X]$ becomes

$$\begin{aligned} E[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma u + \mu) e^{-u^2/2}\, \sigma du \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma u + \mu) e^{-u^2/2}\, du \\ &= \sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u e^{-u^2/2}\, du + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du \end{aligned}$$

Now, the first term has an anti-derivative given by

$$\frac{d}{du}(-e^{-u^2/2}) = u e^{-u^2/2}$$

and therefore

$$\int_{-\infty}^{\infty} u e^{-u^2/2}\, du = -e^{-u^2/2}\Big|_{-\infty}^{\infty} = 0$$

Also using Eq. [3] the integral multiplying μ is equal to 1 and therefore $E[X] = \mu$.

The Moment Generating Function (m.g.f. see Section[14.7] for more details) for the $N(\mu, \sigma^2)$ Gaussian random variable is defined by

$$\psi(u) = E[e^{uX}]$$

Using the probability density function for a $N(\mu, \sigma^2)$ we have

$$\begin{aligned} \psi(u) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} e^{xu}\, dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2} + xu\right)}\, dx \end{aligned}$$

Using the substitution $y = x - \mu$ we have

$$\psi(u) = e^{\mu u} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{\left(-\frac{y^2}{2\sigma^2} + \mu y\right)} \, dy \tag{5}$$

From an integral table we find the Gaussian integral formula, viz.,

$$\int_{-\infty}^{\infty} e^{(-ay^2 + by)} \, \frac{dy}{2\pi} = \frac{1}{\sqrt{4\pi a}} \, e^{b^2/4a} \tag{6}$$

Using Eq. [6] in Eq. [5] we obtain the m.g.f for an $N(\mu, \sigma^2)$ random variable:

$$\psi(u) = \exp\left\{\sigma^2 u^2/2 + \mu u\right\} \tag{7}$$

Taking the derivative of $\psi(u)$ with respect to u we find

$$\psi'(u) = (\mu + u\sigma^2) \exp\{\sigma^2 u^2/2 + \mu u\}.$$

The relationship between the derivative of the m.g.f and the first moment of the respective random variable, $E[X]$ can be obtained by setting evaluating $\psi'(0)$ and we see

$$E[X] = \psi'(0) = \mu.$$

We can use the m.g.f. also to obtain the second moment as well and find

$$\psi''(u) = (\mu + u\sigma)^2 \exp\{\sigma^2 u^2/2 + \mu u\} + \sigma^2 \exp\{\sigma^2 u^2/2 + \mu u\}$$

and therefore $E[X^2] = \psi''(0) = \mu^2 + \sigma^2$.

To find the variance of the Gaussian random variable we have to compute the second moment about the mean or

$$\text{Var}[X] = E[(X - (E[X])^2] = E[X^2] - (E[X])^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2.$$

The m.g.f. can also be defined with respect to $(X - \mu)$.

$$\psi_{(X-\mu)}(u) = E[e^{u(X-\mu)}] = \exp\{\sigma^2 u^2/2\}$$

which is the same functional form as $\psi(u)$ and depends only on the parameter σ^2. This means that all the central moments, $E[(X - \mu)^n]$ are functions only of σ^2. Let $\psi^{(n)}(u) \equiv \frac{d^n}{du^n} \psi(u)$. Then we can show that

$$\psi^{(n+2)}(0) = (n-1)\sigma^2 \psi^{(n)}(0).$$

With the initial condition $\psi^{(1)}(0) = 0$, (first central moment), it immediately follows that $\psi^{(n)}(0) = 0$ for all odd $n \geq 1$. With the initial conditions $\psi^{(0)}(0) = 1$, we have

$$\psi^{(n)}(0) = 1 \cdot 3 \cdot 5 \ldots (n-1)\, \sigma^n = \{(n-1)!!\}\sigma^n \qquad \text{for all even } n \geq 2. \tag{8}$$

5.2 Multi-dimensional Random Variables

In dealing with more than one random variable, the concept of joint probability distribution arises. The continuous random variables $X_1, X_2, ..., X_n$ are said to be *jointly distributed* if they are defined on the same probability space. They may be characterized by their *joint distribution function* or *cummulative distribution function*:

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \text{Probability}\{x_1(\omega) \leq x_1, x_2(\omega) \leq x_2, \ldots, x_n(\omega) \leq x_n\},$$

where

$$\{x_1(\omega), ..., x_n(\omega) \leq x_n\} = \{x_1(\omega) \leq x_1\} \cap \{x_2(\omega) \leq x_2\} \cap ... \cap \{x_n(\omega)\}$$

or alternatively by their *joint density function*

$$F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_{X_1,...,X_n}(\xi_1, ..., \xi_n)\, d\xi_1 ... d\xi_n.$$

It follows that the joint density function is related to the joint distribution function by

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1, \partial x_2, ... \partial x_n} F_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n)$$

The joint density function can be used to define the probability distribution for a vector of random numbers. Consider such a vector $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

The matrix transpose of $\mathbf{X}$ is denoted by

$$\mathbf{X}^T = (X_1, X_2, \ldots, X_n)$$

Now for $j, k = 1, 2, \ldots, n$ let

$$m_j = E[X_j], \qquad K_{jk} = \text{Cov}[X_j, X_k].$$

Let the matrix of covariances be given by

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & \ldots K_{1n} \\ K_{21} & K_{22} & \ldots K_{2n} \\ \vdots & & \\ K_{n1} & K_{n2} & \ldots K_{nn} \end{bmatrix}$$

and let the inverse matrix be represented by

$$\mathbf{K}^{-1} = \begin{bmatrix} K^{11} & K^{12} & \dots K^{1n} \\ K^{21} & K^{22} & \dots K^{2n} \\ \vdots & & \\ K^{n1} & K^{n2} & \dots K^{nn} \end{bmatrix}$$

Now the n random variables $X_1, \dots X_n$ are jointly normally distributed if their joint probability density, for any real numbers $x_1, \dots x_n$, is given by

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|K|^{1/2}} \exp\Big\{ -\frac{1}{2} \sum_{j,k=1}^{n} (x_j - m_j) K^{jk} (x_k - m_k) \Big\}$$

where $|K|$ is the determinant of the $\mathbf{K}$ matrix (i.e., $|K| = \det \mathbf{K}$).

Let

$$\mathbf{m} = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Using matrix multiplication we can also express the joint probability distribution more succiently as

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|K|^{1/2}} \exp\Big\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T \mathbf{K}^{-1} (\mathbf{x} - \mathbf{m}) \Big\}. \tag{9}$$

The quantity

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{K}^{-1} (\mathbf{x} - \mathbf{y})}$$

is known in statistics as the "Mahalanobis distance" [7] and it measures the separation between $\mathbf{x}$ and $\mathbf{y}$. The argument of the exponential in Eq. [9] involves $D_M^2(\mathbf{x}, \mathbf{m})$.

The m.g.f. for a n-vector Gaussian random variable X defined by the above $\mathbf{m}$ and $\mathbf{K}$ matrices is given by (Theorem [14.5] in Section [14.7])

$$\psi_X(\mathbf{u}) = \exp\{\mathbf{u}^T \mathbf{m} + \frac{1}{2} \mathbf{u}^T \mathbf{K} \mathbf{u}\},$$

where $\mathbf{u}$ is an n-vector.

If a linear transformation is performed on a vector of Gaussian random variables, then the transformed random variable also is a multivariate Gaussian.

Theorem 5.1 *Let the n-vector random variable $Z \sim N(m_z, K_z)$. Let $W = CZ + a$ where C is a $q \times n$ constant matrix, a is a constant q-vector, and W is a random q-vector. Then $W \sim N(Cm_z + a, CK_zC^T)$.*

Proof Write the m.g.f. for W

$$\begin{aligned} \psi_W(\mathbf{u}) &= E[\exp[\mathbf{u}^T\mathbf{w}] = \exp(\mathbf{u}^T\mathbf{a})E[\exp(\mathbf{u}^T\mathbf{Cz})] \\ &= \exp[\mathbf{u}^T\mathbf{w}] = \exp(\mathbf{u}^T\mathbf{a})\psi_Z(\mathbf{C}^T\mathbf{z}) \end{aligned}$$

But Z is Gaussian, so that (see Theorem [14.5])

$$\psi_Z(\mathbf{u}) = \exp\{\mathbf{u}^T(\mathbf{Cm_z}) + \frac{1}{2}\mathbf{u}^T\mathbf{CK_zC}^T\mathbf{u}\}$$

Therefore

$$\psi_W(\mathbf{u}) = \exp\{\mathbf{u}^T(\mathbf{Cm_z} + \mathbf{a}) + \frac{1}{2}\mathbf{u}^T[\mathbf{CK_zC}^T]\mathbf{u}\} \qquad \blacksquare$$

5.3 Jensen's Inequality

Jensen's inequality generalizes the statement that the secant line of a convex function lies *above* the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function (for $t \in [0,1]$),

$$tf(x_1) + (1-t)f(x_2),$$

while the graph of the function is the convex function of the weighted means,

$$f\left(tx_1 + (1-t)x_2\right).$$

Thus, Jensen's inequality is

$$f\left(tx_1 + (1-t)x_2\right) \leq tf(x_1) + (1-t)f(x_2).$$

In the context of probability theory, it is generally stated in the following form: if X is a random variable and ϕ is a convex function, then

$$\varphi\left(\mathrm{E}[X]\right) \leq \mathrm{E}\left[\varphi(X)\right].$$

The difference between the two sides of the inequality, $\mathrm{E}\left[\varphi(X)\right] - \varphi\left(\mathrm{E}[X]\right)$, is called the Jensen gap.

A graphical demonstration of Jensen's inequality for convex functions is shown in Fig. 1

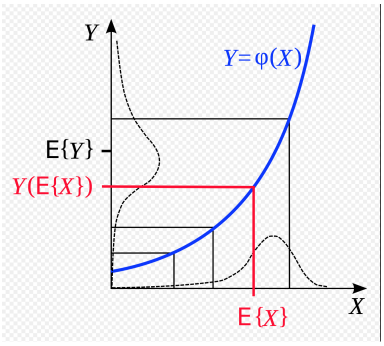

Figure 1: Jensen's Inequality.

6 Conditional Probability, Independence and Expectation

Consider two events A and B and ask: What is the probability of the event A given that the event B has already been observed? Since B has been observed, the available sample space Ω is reduced to the set of outcomes in B and hence we must consider the effects of a reduced set of possible outcomes. In effect B becomes the new sample space of possible outcomes. To compute the relative probability of observing A given that B has been observed we construct the conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

The left hand side is read as "the conditional probability of A given B" and the right hand side are the probabilities computed with respect to the original sample space Ω. In this way we can take into account that the sample space Ω has been reduced by observing an event in the set B. It is unimportant if $P(A|B)$ is undefined if $P(B) = 0$, since we would not consider finding $A \cap B$ in the first place (there being zero probability of having first observed B).

Conditional probability can be shown to satisfy the three axioms of probability given in Section [2] and hence has all the properties of a probability function.

Two events are said to be *independent* if the observation of the first event does not effect the probability of observing the second event. In other words

$$P(A|B) = P(A).$$

This is equivalent to

$$P(A \cap B) = P(A)P(B)$$

where, again, the probabilities $P(A)$ and $P(B)$ are defined with respect to the original sample space Ω.

If the joint probability of events $x \in A$ and $y \in B$ is given in terms of a probability density function

$$f_{X,Y}(x, y).$$

Then we can determine the separate or marginal probability functions for x with y unrestricted by

$$f_X(x) = \int f_{X,Y}(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x, y)\, dx$$

We will use the shorthand X and Y to stand for that random variables $x(\omega)$ and $y(\omega)$ respectively. Let us now find the density function for finding $Y \in A$ given that we have observed $Y \in B$. This will require a limiting process (since $P\{Y = y\} = 0$ for an isolated continuous random variable in B). We will choose an interval about y of the form $(y, y + \Delta y]$ and compute the probability

$$P\{y < Y \leq y + \Delta y\} = \int_y^{y+\Delta y} f_Y(y)dy \approx f_Y(y)\Delta y.$$

First let us compute the conditional probability that $X \leq x$ given B.

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

For $Y \approx y$, we take B as the set $B = \{y < Y \leq y + \Delta y\}$ with probability $P(B) = f_Y(y)\Delta y$ and we write the conditional probability distribution function $P\{X \leq x|B\}$ as

$$\begin{aligned} F_{X|y<Y\leq y+\Delta y} &= \frac{P\{X \leq x, y < Y \leq y + \Delta y\}}{f_Y(y)\Delta y} \\ &= \frac{\int_{-\infty}^{x} \int_y^{y+\Delta y} f_{X,Y}(u, v)dudv}{f_Y(y)\Delta y} \\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(u, y)du\Delta y}{f_Y(y)\Delta y} \\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(u, y)du}{f_Y(y)} \end{aligned}$$

Taking the limit as $\Delta y \to 0$ we obtain $\lim_{\Delta} Y \to 0 F_{X,|y<Y\leq y+\Delta y} = F_{X|Y}(x, y)$ and

$$F_{X|Y}(x|y) = \frac{\int_{-\infty}^{x} f_{X,Y}(u, y)du}{f_Y(y)}.$$

Differentiating with respect to x we obtain

$$f_{X|Y}(x|y) \triangleq \frac{\partial F_{X|Y}(x|y)}{\partial x} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Rewriting the above, we also have

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) \cdot f_Y(y) \quad \text{and, by symmetry} \quad = f_{Y|X}(y|x) \cdot f_X(x) \tag{10}$$

Given the conditional probability density function $f_{X|Y}(x|y)$ or $f_{Y|X}(x|y)$ we can show that it satisfies all the axioms of probability and can be interpreted as a probability in its own right. Therefore we can use $f_{Y|X}(x|y)$ to define the Conditional Expectation

Definition The conditional expectation of the random variable (vector) Y given the random variable (vector) X (given $x(\omega) = x$ is defined by

$$E[Y|X] \triangleq \int y f_{Y|X}(y|x) dy$$

If Y is a random vector, then

$$E[Y|X] = \big(E[Y_1|X], \ldots, E[Y_n|X]\big) \tag{11}$$

Notice that

$$\begin{aligned} f_{Y|X}(y|x) &\geq 0, \\ \int f_{Y|X}(y|x)\, dy &= 1, \\ f_Y(y) &= E[f_{Y|X}(y|x)] = \int f_{Y|X}(y|x) f_X(x)\, dx. \\ f_{Y|X}(y|x) &= f_Y(y) \quad \text{if } X \text{ and } Y \text{ are independent.} \end{aligned}$$

If one is dealing with random variables that are a mixture of discrete values and continuous values, it is useful to employ the Riemann-Stieltjes Integral which in the continuous case is based on the differential identity $dF_X(x) = f_X(x)dx$. Using this notation, the conditional expectation, $E[Y|X = x]$ is given by

$$E[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy \tag{12}$$

and if one takes the expectation of $E[Y|X = x]$ with respect to the probability function for X, then it is reasonable that we should find

$$E[Y] = \int_{-\infty}^{\infty} E[Y|X = x] dF_X(x) \tag{13}$$

It is also reasonable that the conditional expectation of $E[g(X)Y|X = x]$ should satisfy the equation

$$E[g(X)Y|X = x] = g(x)E[Y|X = x] \tag{14}$$

for any function $g(X)$ of a random variable X such that $E[g(X)Y] < \infty$. If we put Eq. [13] and Eq. [14] together we should expect

$$E[g(X)Y|X = x] = \int_{-\infty}^{\infty} g(x)E[Y|X = x]dF_X(x) \tag{15}$$

Now we will use the notation that $E[Y|X] = E[Y|X = x]$, so that $E[X|Y]$ denotes the function whose value when $X = x$ is defined to be equal to $E[Y|X = x]$. Using this notation we may rewrite Eq. [15] as

$$E[g(X)Y] = E[g(X)E[Y|X]]. \tag{16}$$

In the particular case of $g(X) = 1$ we find

$$E[Y] = E[E[Y|X]] \tag{17}$$

provided that $E[Y] < \infty$ and we see how the unconditional expectation is related to the conditional expectation. An important formula relates the unconditional variance to conditional expectations. If $E[Y^2] < \infty$, then

$$\mathrm{Var}[Y] = E\big[\mathrm{Var}[Y|X]\big] + \mathrm{Var}\big[E[Y|X]\big] \tag{18}$$

Or *the variance is equal to the mean of the conditional variance plus the variance of the conditional mean.* To prove Eq. [18] use Eq. [17] to write

$$\mathrm{Var}[Y] = E\big[(Y - E[Y])^2\big] = E\big[E[(Y - E[Y])^2|X]\big] \tag{19}$$

Now for any random variable Z and any constant a we have

$$E[(Z - a)^2] = E[(Z - E[Z])^2] + (E[Z] - a)^2. \tag{20}$$

To prove Eq. [20] expand both sides and compare terms. The LHS expands out to:

$$\begin{aligned} E[(Z - a)^2] &= E[Z^2 - 2aZ + a^2) \\ &= E[Z^2] - 2aE[Z] + a^2 \end{aligned}$$

The RHS expands out to

$$\begin{aligned} E[(Z - E[Z])^2] + (E[Z] - a)^2 &= E[Z^2 - 2ZE[Z] + (E[Z])^2] + (E[Z])^2 - 2aE[Z] + a^2 \\ &= E[Z^2] - 2(E[Z])^2 + (E[Z])^2 + (E[Z])^2 - 2aE[Z] + a^2 \\ &= E[Z^2] - 2aE[Z] + a^2 \end{aligned}$$

and we see that the LHS = RHS verifying Eq. [20]. Applying Eq. [20] to the conditional expectation $E\big[(Y - E[Y])^2|X\big]$ we obtain

$$E[(Y - E[Y])^2|X] = E[(Y - E[Y|X])^2|X] + (E[Y|X] - E[Y])^2 \tag{21}$$

Taking a final expectation of Eq. [21] with respect to X and noting that $E[Y] = E\big[E[Y|X]\big]$ results in

$$\begin{aligned}\mathrm{Var}[Y] &= E\big[E[(Y - E[Y])^2|X]\big] \\ &= E\big[E[(Y - E[Y|X])^2|X]\big] + E\big[(E[Y|X] - E[E[Y|X]])^2\big] \\ &= E\big[\mathrm{Var}[Y|X]\big] + \mathrm{Var}\big[E[Y|X]\big]\end{aligned}$$

7 Stochastic Processes

A stochastic process $\{X(t), t \in T\}$ is a parameterized collection of random variables. For every value of the parameter $t \in T, X(t)$ is a random variable. Often t is taken to be time and $X(t)$ is referred to as the *state* of the stochastic process at time t. T is the *index set* of the process and can consist of discrete values or can be in interval of the real line. When T is a countable set the stochastic process is said to be a *discrete-time process* or a *discrete parameter process* and denoted by $\{X_t\}_{t\in T}$. If T is an interval of the real line, then the stochastic process is said to be a *continuous-time process* or a *continuous parameter process* and denoted, for example by $\{X(t), t \geq 0\}$ where T is the halfline $[0, \infty)$. The *state space* of the process is defined to be the set of all possible values that the random variables $X(t)$ can take on. The random variables $X(t)$ can be scalars or vectors. If the values of $X(t)$ are discrete, then the stochastic process is said to have a *discrete state space*. If they are continuous, the process is said to have a *continuous state space*.

The random variables that make up $X(t)$ are defined on a probability space $(\Omega, \mathcal{F}, P)$ and assume values in R if the process is scalar-valued or $\mathbb{R}^n$ if the process is vector-valued. A continuous parameter space T is usually the halfline $[0, \infty)$ but it may also be an interval $[a, b]$, the non-negative integers in the case of a discrete parameter process or even subsets of $\mathbb{R}^n$ for $n \geq 1$.

Using the notations $X_t(\omega)$ or $X(t, \omega)\big)$ to represent the dependence of the stochastic process on t as well as the random variable ω, then for each $t \in T$ fixed we have a random variable

$$\omega \to X_t(\omega), \quad \omega \in \Omega.$$

On the other hand, fixing $\omega \in \Omega$ we can consider the function

$$t \to X_t(\omega), \quad t \in T$$

which is called the path of X_t. Once ω is selected, the entire path $X_t(\omega)$ is determined.

One can also consider the stochastic process as a function of two variables t and ω, i.e., $X(t, \omega)$ in place of $X_t(\omega)$. Then we can consider the stochastic process corresponding to the mapping

$$(t, \omega) \to X_t(\omega)$$

from $T \times \Omega \to \mathbb{R}^n$.

The concepts of mean and variance given above for a statistical sample, can be extended accordingly into functions for stochastic processes as follows

Mean: The mean function $\mu(t)$ is defined by

$$\mu(t) = E[X(t)]$$

Variance: The variance function $\sigma^2(t)$ is defined by

$$\sigma^2(t) = \mathrm{Var}[X(t)]$$

Autocorrelation: The autocorrelation function is given by

$$\Gamma_X(t_1, t_2) = E[X(t_1)X(t_2)]$$

Autocovariance: The variance function alone is not enough to specify the second moments of a sequence of random variables. We therefore include the autocovariance function (covariance kernel)

$$C_X(t_1, t_2) = E[(X(t_1) - \mu(t_1))(X(t_2) - \mu(t_2))]$$

We will confine ourselves to stochastic processes with

$$E[X^2(t)] < \infty$$

for all t. Such stochastic processes are said to have *finite average power* and are also known as *second-order processes*. The second moment of such a stochastic process is finite. This also implies by the Cauchy-Schwarz inequality that $E[|X(t)|]$ is finite (see Section [14.6]. One interesting conclusion about process with finite $|X(t)|$ is

$$P(|X| < \infty) = 1.$$

For if there were any finite non-zero probability that $|X| = \infty$, then $E[|X|] = \infty$. Therefore $P(|X| = \infty) = 0$.

Dropping the explicit dependence on ω (and using the notation $\{X(t)\}$) a stochastic process $X(t)$ for values of t given in steps as

$$0 \leq t_0 < t_1 < ... < t_k$$

can be displayed by values $X(t_0), X(t_1), ..., X(t_k)$. A stochastic process $\{X(t), t \in T\}$, whose index set T is linear, is said to be

(1) *strictly stationary of order k*, where k is a positive integer, if for any points $t_1, \ldots, t_k \in T$ and any h in T, the k-dimensional vectors

$$(X(t_1), \ldots, X(t_k)) \quad \text{and} \quad (X(t_1 + h), \ldots, X(t_k + h))$$

are identically distributed.

(2) *strictly stationary* if for any integer k it is strictly stationary of order k.

A continuous parameter stochastic process $\{X(t), 0 \leq t > \infty\}$ is said to have *independent increments* if $X(0) = 0$ and, for all choices of the indices $t_0 < t_1 < \ldots t_k$, the k random variables

$$X(t_1) - X(t_0), X(t_2) - X(t_1), ..., X(t_k) - X(t_{k-1})$$

are independent. If in addition to being *independent*, the increments have the property of *stationarity*, or that $X(t + h) - X(s + h)$ has the same distribution as $X(t) - X(s)$ for any t and s and for all h, then the stochastic process is said to have *stationary independent increments* (SII).

Another notion of stationarity is *covariance stationarity*: A stochastic process $\{X(t), t \in T\}$ is said to be covariance stationary if it has finite second moments, if its index set T is linear, and if its covariance kernel $C_X(s, t)$ is a function only of the absolute difference $|s - t|$, in the sense that there exists a function $R(v)$ such that for all s and t in T

$$C_X(s, t) = R(s - t).$$

$R(v)$ has the property that for every t and v in T

$$\text{Cov}[X(t), X(t + v)] = R(v).$$

8 Random Walk in One Dimension

Let $U = (U_1, U_2, \ldots)$ be a sequence of independent random variables, each taking the values -1 and 1 with probabilities $p \in [0, 1]$ and $q = 1 - p$ respectively. Let $V = (V_0, V_1, V_2 \ldots)$ be the partial sum process associated with U starting with the initial step counter value $V_0 = 0$:

$$V_0 = 0, \quad V_n = V_0 + \sum_{i=1}^{n} U_i, \quad n = 1, 2, \ldots$$

The sequence $V_0, V_1, V_2 \ldots$ is called a simple random walk in one dimension with parameter p.

We imagine a person walking or a particle moving on an axis, so that at each discrete time step of duration τ, the walker moves either one unit to the right with probability p or one unit to the left

with probability $1-p$, independently from step to step. V_n keeps count of the net number of steps in the positive direction from the origin.

The indicator function I_k for U_k is defined as follows:

$$I_k = \begin{cases} 1 & \text{if } U_k = \quad 1 \\ 0 & \text{if } U_k = -1 \end{cases}$$

In terms of U_k we can write the indicator function as

$$I_k = \frac{1}{2}(U_k + 1). \tag{22}$$

The expectation of the Indicator Function satisfies:

$$E[I_k] = \sum_{x_k \in \{I_k = 1\}} P(x_k) = P\{I_k = 1\} = p.$$

Also, $E[I_k^2] = p$ by similar reasoning. Hence

$$\text{Var}[I_k] = E[I_k^2] = (E[I_k])^2 = p - p^2 = p(1-p) = pq.$$

Using Eq. [22] we find that $U_k = 2I_k - 1$ and $U_k^2 = 4I_k^2 - 4I_k + 1$. Therefore the respective mean and variance of U_k is given by

$$\begin{aligned} E[U_k] &= 2p - 1 = p - q \\ \text{Var}[U_k] &= E[4I_k^2 - 4I_k + 1] - (2p-1)^2 \\ &= 1 - (2p-1)^2 = 1 - 4p^2 + 4p - 1 \\ &= 4p(1-p) = 4pq \end{aligned}$$

Since the $U_k, k = 1, 2, \ldots, I_n$ are a set of n independent random variables, we find that

$$\begin{aligned} E[V_n] &= n(p-q) \\ \text{Var}[V_n] &= 4npq \end{aligned}$$

For large values of n and moderate p we can invoke the Central Limit Theorem (see Section 14.11 and Theorem 14.11) with asymptotic mean and variance given by $\mu_n = n(p-q)$ and $\sigma_n^2 = 4npq$ respectively, to deduce the asymptotic limiting distribution of V_n for large n to be

$$P\left\{\frac{V_n - \mu_n}{\sigma_n} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-x^2/2\right)\, dx.$$

Therefore the asymptotic probability density function for V_n is given by the Gaussian:

$$f_{V_n}(v) \approx \frac{1}{\sqrt{8\pi npq}} \exp\left\{-\frac{(v - n[p-q])^2}{8npq}\right\} \tag{23}$$

The probability distribution for the random walker step count follows an approximate Gaussian probability distribution. Hence we anticipate that the same type of probability distribution will describe the position of a particle undergoing Brownian motion after n "steps" (where each step is a collision that is assumed to knock the particle either to the right or to the left by the same distance l). If the length of each step in the random walk is l then the position from the origin can be described in terms of $x = lv$. Using Eq. [23] we can relate the probability that a particular undergoing the above "random walk" will be located in an interval dx about x is

$$\begin{aligned} f_X(x)dx &= f_{V_n}(v)\,dv = f_{V_n}(x/l)\frac{dx}{l} \\ &\approx \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}\,dx, \end{aligned}$$

where $\mu = nl(p-q)$ and $\sigma = 2l\sqrt{npq}$. From this derivation we see that the probability density for a random walker which starts at the origin to be within an interval of width dx at x after n steps is given by

$$f_{X_n}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-(x-\mu)^2/2\sigma^2\Big\} \tag{24}$$

Eq. [24] is an important property for random walks and also demonstrates the connection between Brownian motion and the normal distribution.

Let us re-introduce the duration of a time step τ take a limit as the step size l and τ approach zero as $n \to \infty$ is such a way that

$$\frac{\sigma^2}{2\tau} = D, \quad nl(p-q) = x_0, \quad n\tau = t$$

where D is a constant, x_0 is the mean position and t is the total duration of n steps. Taking the limit we find

$$\lim_{n\to\infty} f_{X_n}(x) = \frac{1}{\sqrt{4\pi Dt}}\exp\Big\{-(x-x_0)^2/4Dt\Big\} \tag{25}$$

Eq. [25] was derived by Albert Einstein in his original description of the probability law for the position of a physical particle undergoing Brownian motion [8]. The constant D is known as the constant of diffusion and can be related to other physical constants in the case of Brownian Motion:

$$D = \frac{2RT}{Nf},$$

where R is the universal gas constant, T is the absolute temperature, N is the Avogadro number, and f the friction coefficient specific to the particles undergoing Brownian motion. In light of the connection with the Random Walk, we see that Einstein's result emerges in probability theory as a consequence of the Central Limit Theorem. We will use Eq. [25] as one of the fundamental properties of the mathematical abstraction of Brownian Motion known as the Wiener Process.

9 Brownian Motion – The Wiener Process

The mathematical idealization of Brownian motion is called the Wiener process and it is an example of a stochastic process with stationary independent increments.

A stochastic process $\{B(t), t \geq 0\}$ taking values in R is said to be a Brownian motion in 1-dimension (a Wiener process) if

(a) $\{B(t), t \geq 0\}$ has stationary independent increments,
(b) For every $t > 0, B(t)$ is normally distributed,
(c) For all $t > 0, E[B(t)] = 0$
(d) $B(0) = 0$

Obviously the above stochastic process can be extended to $\mathbb{R}^n$ by allowing $B(t)$ to be vector-valued with n components.

We use the notation $E[B(t)]$ to represent the expectation, or the mean, of the random number $B(t)$ averaged over the elementary outcomes ω in Ω.

We will also use both notation $B_t(\omega)$ in place of $B(t)$ to emphasize the dependence of the Brownian motion on the particular elementary outcome $\omega \in \Omega$ that gives rise to it.

(Without loss of generality the above Brownian motion is assumed to start from the origin 0. We could use any point $B(0) = x$ on the real number line as the starting point.)

Brownian motion can be considered as the integral with respect to time of "white noise" (more on this later). Because integration "improves" continuity properties representations of Brownian motion, $B(t)$ will have better continuity properties than "white noise". It appears to be reasonable to describe Brownian motion in terms of the observed values taken at discrete time intervals, say at the points $t_1, t_2, ..., t_n$ and hope to characterize it by finding a joint probability function for $B(t_1), B(t_2), ..., B(t_n)$.

Let us examine some of the consequences of discretizing the t (or time) parameter space. Let us break up the interval $[0, t)$ into $[0, s) \cup [s, t)$. This can be thought of as several points in order $0 \leq s \leq t$.

By (b) above, we see that

$$E[B(t) - B(s)] = 0$$

because $E[B(t) - B(s)] = E[B(t)] - E[B(s)] = 0 - 0 = 0$

However more characteristics can be obtained by using properties (a) through (d) above. We use (d) out of convenience – there is nothing sacred about setting the coordinate system up so that at $t = 0$ we have $B(0) = 0$. It is a convenient choice that simplifies our derivations.

Now consider the time interval $0 \leq s \leq t$. We can assign values t_1 and t_2 to the lengths of the respective separations as

$$\begin{aligned} s &= t_1 \\ t &= t_1 + t_2 \quad \text{and conversely} \\ t_1 &= s \\ t_2 &= t - s \end{aligned}$$

Notice that we can write

$$B(t) = B(s) + B(t) - B(s)$$

Now $E[B(s)] = 0$ and $E[B(t) - B(s)] = 0$ for $t \geq s \geq 0$ and using the property of stationary independent intervals we also have

$$E[B(s)(B(t) - B(s))] = E[B(s)] \cdot E[B(t) - B(s)] = 0$$

Since $B(s)$ and $B(t) - B(s)$ for $t \geq s \geq 0$ are independent random variables we can express the variance $\text{Var}[B(t)]$ as

$$\text{Var}[B(t)] = \text{Var}[B(s)] + \text{Var}[B(t) - B(s)]$$

Using the property of Stationary Independent Increments we see that

$$\text{Var}[B(t) - B(s)] = \text{Var}[B(t - s) - B(0)] = \text{Var}[B(t - s)] = \text{Var}[B(t_2)]$$

Using this result and noting that $s = t_1$ and $t = t_1 + t_2$ we can write

$$\text{Var}[B(t_1 + t_2)] = \text{Var}[B(t_1)] + \text{Var}[B(t_2)]$$

So that $f(t) = \text{Var}[B(t)]$ satisfies the property that

$$f(t_1 + t_2) = f(t_1) + f(t_2)$$

One can easily extend this by induction to

$$f(\sum_{i=1}^{n} t_i) = \sum_{i=1}^{n} f(t_i)$$

Consider a rational number $t = n/m$ for integers n and m. Then

$$f(t) = f(n/m) = \sum_{i=1}^{n} f(1/m) = n \cdot f(1/m)$$

Also, $f(1) = f(m/m) = m \cdot f(1/m)$ and hence

$$f(1/m) = (1/m) \cdot f(1).$$

Putting these facts together, we obtain

$$f(t) = f(n/m) = (n/m)f(1).$$

Now if we represent $f(1)$ by the constant c, then

$$f(t) = f(n/m) = (n/m) \cdot f(1) = c \cdot t.$$

Since $\mathrm{Var}[X] \geq 0$ for any random variable X, then $c > 0$ which we will denote by $c = \sigma^2$.

Using these facts we find that $\mathrm{Var}[B(t)] = \sigma^2 t$ and therefore

$$\mathrm{Var}[B(t)] = \mathrm{Var}[B(s)] + \mathrm{Var}[B(t) - B(s)]$$

can be written as

$$\mathrm{Var}[B(t) - B(s)] = \sigma^2(t - s)$$

when $t \geq s \geq 0$. The general case can be written as

$$\mathrm{Var}[B(t) - B(s)] = \sigma^2|t - s|. \tag{26}$$

Setting $s = 0$ we obtain

$$\mathrm{Var}[B(t)] = E[(B(t))^2] - (E[B(t))^2 = E[B(t)^2] = \sigma^2 t.$$

Another consequence of Stationary Independent Increments using $B(t) = B(s) + B(t) - B(s)$ is that

$$\begin{aligned} E[B(s)B(t)] &= E[B(s) \cdot (B(s) + B(t) - B(s))] \\ &= E[B^2(s)] + E[B(s) \cdot (B(t) - B(s)] \\ &= E[B^2(s)] = \sigma^2 s \end{aligned}$$

since $E[B(s) \cdot (B(t) - B(s))] = 0$ for $t \geq s \geq 0$. Since we assumed that $t \geq s \geq 0$ in the derivation we can summarize the general case as

$$E[(B(t)B(s)] = \sigma^2 \min(s, t) \tag{27}$$

The parameter σ^2 above characterizes the normal distribution referred to in part (b) above. If we take only one coordinate x of the Brownian path and consider the motion as a function of time, then taking $\sigma = 1$ normalizes the motion to fit the time scale. In this case we can develop a probability

distribution for $x(t)$. Let us take $x(t) = 0$ at $t = 0$ and consider the position of the particle seen at set of times $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$. The assuming the time scale is chosen so that $\sigma = 1$, then the probability that the particle lies between x_1 and $x_1 + dx_1$ at time $t_1, \ldots$, between x_n and $x_n + dx$ at time t_n is [14]

$$\frac{\exp\left[-\frac{x_1^2}{2t_1} - \frac{(x_2-x_1)^2}{2(t_2-t_1)} - \ldots - \frac{(x_n-x_{n-1})^2}{2(t_n-t_{n-1})}\right]}{\sqrt{|(2\pi)^n t_1(t_2-t_1)\ldots(t_n-t_{n-1})|}}\, dx_1 \ldots dx_n$$

The above probability measure was developed by Norbert Wiener in 1921 and represents an advance over the early work of Albert Einstein on Brownian motion [8].

In what follows we will examine further properties of Brownian Motion and Ito's Integral. It is occasionally useful to define two stochastic processes which are versions of each other:

Definition Suppose that $\{X(t,\omega)\}$ and $\{Y(t,\omega)\}$ are stochastic processes on $(\Omega, \mathcal{F}, P)$. Then we say that $\{X(t)\}$ is a version of $\{Y(t)\}$ if

$$P(\{\omega : X(t,\omega) = Y(t,\omega)\}) = 1 \quad \text{for all t.}$$

Thus if $X(\omega, t) \in \mathbb{R}^n$, then from the point of view that a stochastic process is a probability law on $(\mathbb{R}^n)^{[0,\infty)}$ two such processes are indistinguishable, but nevertheless their path properties may be different. Two stochastic processes that are different versions of each other can be thought of as belonging to the same equivalence class (see Section [14.2]).

The properties of Brownian motion will be discussed further in the next section, but in passing, please note the following famous theorem of Kolmogorov (see [4], [5] and [17]).

Theorem 9.1 *(Kolmogorov's continuity theorem) Suppose that the process $X = \{X(t), t \geq 0\}$ satisfies the following condition: For all $T > 0$ there exists positive constants α, β, D such that*

$$E[|X(t) - X(s)|^\alpha] \leq D \cdot |t-s|^{1+\beta}; 0 \leq s, t \leq T. \tag{28}$$

Then there exists a continuous version of $X(t)$.

Proof We follow van Zanten's proof[1]. First observe that by Chebyshev's inequality, Eq. [28] implies that the process $X(t)$ is continuous in probability. Without loss of generality, we assume that $t \in [0, T]$ with $T = 1$ and work with subintervals, Δ_n, made up of the dyadic rationals: $\Delta_n = k/2^n : k = 0, 1, \ldots, 2^n$ and let $D = \bigcup_{n=1}^{\infty} \Delta_n$. Then D is a countable set, and D is dense[2] in $[0, 1]$. Our next aim is to show

[1] See Harry van Zanten, [5], section [1.3]

[2] $\overline{A}$ is the closure of A, i.e., the set A together with all its limit point (see [1], page 366): Let A be a metric space. We say a set $A \subset S$ is *dense* in S iff $\overline{A} = S$

that with probability 1, the process $X(t)$ is uniformly continuous on D. Pick any $\gamma \in [0, \beta/\alpha]$. Using Chebyshev's inequality again, we see

$$P(|X(k/2^n) - X((k-1)/2^n)| \geq 2^{-\gamma n}) \leq \frac{K}{2^{n(1+\beta-\alpha\gamma)}}$$

for some positive constant K. It follows that

$$P(\max_{1\leq k\leq 2^n} |X(k/2^n) - X((k-1)/2^n)| \geq 2^{-\gamma n}) \leq \sum_{k=1}^{2^n} P(|X(k/2^n) - X((k-1)/2^n)| \geq 2^{-\gamma n}) \leq \frac{K}{2^{n(\beta-\alpha\gamma)}}$$

Hence, by the Borel-Cantelli lemma (see Section [14.12]), there almost surely exists an $N \in \mathbb{N}$ such that

$$\max_{1\leq k\leq 2^n} |X(k/2^n) - X((k-1)/2^n)| \leq 2^{-\gamma n} \tag{29}$$

for all $n \geq N$. Next, consider an arbitrary pair $s, t \in D$ such that $0 < t - s < 2^{-N}$. We aim to show that

$$|X(t) - X(s)| \leq K|t - s|^\gamma, \tag{30}$$

for some positive constant K. There exists an $n \geq N$ such that $2^{-(n+1)} \leq t - s < 2^{-n}$. We claim that if $s, t \in \Delta_m$ for $m \geq n + 1$, then

$$|X(t) - X(s)| \leq 2 \sum_{k=n+1}^{m} 2^{-\gamma k} \tag{31}$$

To see this we use induction. Suppose first that $s, t \in \Delta_{n+1}$. Then necessarily, $t = k/2^{n+1}$ and $s = (k-1)/2^{n+1}$ for some $k \in (1, \ldots, 2^{n+1}$. By Eq. [29], it follows that

$$|X(t) - X(s)| \leq 2^{-\gamma(n+1)}$$

which proves the claim for $m = n+1$. Now suppose that it is true for $m = n+1, \ldots, l$ and assume that $s, t \in \Delta_{l+1}$. Define the numbers $s', t' \in \Delta_l$ by $s' = \min\{u \in \Delta_l : u \geq s\}, t' = \max\{u \in \Delta_l : u \leq t\}$. Then by construction, $s \leq s' \leq t' \leq t$ and $s' - s \leq 2^{-(l+1)}, t' - t \leq 2^{-(l+1)}$. Hence, by the triangle inequality, Eq. [29] and the induction hypothesis,

$$\begin{aligned} |X(t) - X(s)| &\leq |X(s') - X(s)| + |X(t') - X(t)| + |X(t') - X(s')| \\ &\leq 2^{-\gamma(l+1)} + 2^{-\gamma(l+1)} + 2 \sum_{k=n+1}^{l} 2^{-\gamma k} = 2 \sum_{k=n+1}^{l+1} 2^{-\gamma k}, \end{aligned}$$

so the claim is true for $m = l + 1$ as well. The proof of Eq. [30] is now straightforward. Indeed, since $t, s \in \Delta_m$ for some large enough m, relation Eq. [31] implies that

$$|X(t) - X(s)| \leq 2 \sum_{k=n+1}^{\infty} = \frac{2}{1 - 2^{-\gamma}} 2^{-\gamma(n+1)} \leq \frac{2}{1 - 2^{-\gamma}} |t - s|^\gamma$$

Observe that Eq. [30] implies in particular that, almost surely, the process $X(t)$ is uniformly continuous on D. In other words, we have an event $O \in \Omega$ with $P(O) = 1$ such that for all $\omega \in O$, the sample path $X(t, \omega)$ is uniformly continuous on the countable, dense set D. Now we define a new stochastic process

$$Y(t) = \begin{cases} X(t) & \text{if } t \in D \\ \lim_{t_n \to t, t_n \in D} X(t_n, \omega) & \text{if } t \notin D. \end{cases}$$

The uniform continuity of $X(t)$ implies that $Y(t)$ is a well-defined, continuous stochastic process. Since $X(t)$ is continuous in probability (see the first part of the proof), $Y(t)$ is a version of $X(t)$. ∎

As an example, we will apply Kolmogorov's continuity theorem to n-dimensional Brownian motion: $B(t) = [B^{(1)}(t), B^{(2)}(t), \ldots, B^{(n)}(t)]$, where the components $B^{(j)}(t), 1 \leq j \leq n$ are independent 1-dimensional Brownian motions.

Example (*Brownian motion exists)*
Let $B(t)$ be a standard 1-dimensional Brownian motion on $\mathbb{R}$ with $B(0) = 0$ and $\sigma^2 = 1$. The moment generating function (see Section [14.7]) is given by

$$E[e^{uB(t)}] = \exp(\frac{1}{2}u^2 t) \quad \text{for all } u \in \mathbb{R}^n. \tag{32}$$

From this we can conclude that

$$E[e^{uB(t)}] = \sum_{k=0}^{\infty} \frac{u^k E[B^k(t)]}{k!} = \sum_{j=0}^{\infty} \frac{t^j u^{2j}}{2^j j!} \tag{33}$$

Hence

$$E[B^{2k+1}(t)] = 0 \quad \text{and} \quad E[B^{2k}(t)] = \frac{(2k)! t^k}{k! 2^k}$$

for $k = 0, 1, \ldots$. Note that all the odd powers of k give zero expectations.

We now extend the above argument to Brownian motion in $\mathbb{R}^n$:

$$\begin{aligned} E[|B(t) - B(s)|^4] &= \sum_{i=1}^{n} E[(B^i(t) - B^i(s))^4 + \sum_{i \neq j} E[(B^{(i)}(t) - B^{(i)}(s))^2 (B^{(j)}(t) - B^{(j)}(s))^2] \\ &= n \cdot \frac{4!}{4 \cdot 2!} \cdot (t-s)^2 + n(n-1)(t-s)^2 \\ &= n(n+2)(t-s)^2 \end{aligned}$$

Hence we see that n-dimensional standard Brownian motion satisfies the criterion for Kolmogorov's continuity theorem with $\alpha = 4, \beta = 1$ and $D = n(n+2)$ and therefore it has a continuous version. We will look at this property again from a different point of view based on mean-square calculus and an intuitive approach representing Brownian motion as the integral of "white noise". Kolmogorov's continuity theorem is a powerful method.

10 Mean-Square Calculus

Please see [3] and [13] for a general description of mean-square Calculus.

As mentioned in Section [2], mean-square (m.s.) calculus is used simplify many of the derivations but still keep the subject general enough to describe a large body of the interesting subjects in stochastic processes. The main cornerstones on which mean-square calculus stands are the ideas of *mean-square limit, mean-square continuity, mean-square differentiability, and mean-square integrability.* The concept of mean-square convergence appears in many other areas such as the inner product spaces (Hilbert space) of Quantum Mechanics.

Earlier we defined the correlation function for a stochastic process by $\Gamma(t_1, t_2) = E[X(t_1)X(t_2)]$ The correlation function can also be thought of as a kind of inner product

$$\langle X(t_1), Y(t_2)\rangle = E[X(t_1)X(t_2)]$$

The inner product exists as a real number for every $X(t_1)$ and $Y(t_2)$ which have finite average power and satisfies the following rules:

1.) $\langle X(t_1), X(t_1)\rangle \geq 0$ and $\langle X(t_1), X(t_2)\rangle = 0$ iff $X(t) = 0$ almost surely (i.e., $P[X(t) = 0] = 1$),
2.) $\langle X(t_1), Y(t_2)\rangle = \langle Y(t_2), X(t_1)\rangle$,
3.) $\langle cX(t_1), Y(t_2)\rangle = c\langle X(t_1), Y(t_2)\rangle$, where $c \in \mathbb{R}$.

The equivalence of $\langle X(t), X(t)\rangle = 0$ and $X(t) = 0$ identically is a general requirement of an inner product, but this is not true in the above case since we only have $X(t) = 0$ almost surely (wp1). There are ways around this subtle point. For example we can use $\langle , \rangle = 0$ to define an equivalence relation (see Section [14.2]) between $X(t)$ and $Y(t)$ and say that XRY if $\langle X(t)-Y(t), X(t)-Y(t)\rangle = 0$ and then take the continuous version from this equivalence class as the representative version (assuming it exists). With this understanding we define the *vector norm* of a vector-valued process similarly:

$$\|X(t)\| = \sqrt{\langle X(t), X(t)\rangle}$$

This vector norm satisfies:

1.) $\|X(t)| \geq 0$ and $\|X(t)\| = 0$ iff $X(t) = 0$ almost surely (i.e., $P[X(t) = 0] = 1$),
2.) $\|cX(t)\| = c\|X(t)\|$, where $c \in \mathbb{R}$.
3.) $\|X(t) + Y(t)\| \leq \|X(t)\| + \|Y(t)\|$ (the triangle inequality).

If $E[X(t)] = 0$, then $\|X(t)\|$ is the standard deviation of $X(t)$.

We assume that the stochastic processes of interest have finite average power

$$E[X^2(t)] < \infty$$

for all t. We also assume, without loss of generality, that all finite-power processes have zero mean. Since the assumption of $E[X^2(t)] < \infty$ implies that $E[X] < \infty$ by the Cauchy-Schwarz inequality we can transform variables to

$$Y(t) = X(t) - E[X]$$

that have zero mean.

10.1 Mean-Square Limit

The various kinds of convergences used are summarized in Section [14.3]. Fundamental to the subject of mean-square limits is the Cauchy convergence theorem:

Theorem 10.1 *Let $X(t)$ be a real-valued, finite power stochastic process. The mean-square limit*

$$Y(t) = \operatorname*{l.i.m}_{t' \to t} X(t') = \operatorname*{l.i.m}_{\epsilon \to 0} X(t+\epsilon)$$

exists as a unique (in m.s. sense) stochastic process iff

$$\lim_{t_1, t_2 \to t} E\Big[[X(t_1) - X(t_2)]^2\Big] = \lim_{\epsilon_1, \epsilon_2 \to 0} E\Big[[X(t+\epsilon_1) - X(t+\epsilon_2)]^2\Big] = 0$$

Definition The mean-square limit $\operatorname*{l.i.m}_{n \to \infty} X(t+t_n)$ exists as a unique stochastic process iff the double limit

$$\operatorname*{l.i.m}_{n,m \to \infty} [X(t+t_n) - X(t+t_m)] = 0$$

Because we are working with zero-mean processes with finite average power the correlation function is finite by the Cauchy-Schwarz inequality:

$$|\Gamma(t_1, t_2)| = |E[X(t_1)X(t_2)]| \leq \sqrt{E[X^2(t)]E[X^2(t_2)]} < \infty$$

10.2 Mean Square Continuity

Definition A stochastic process $X(t)$ is *mean-square (m.s.) continuous* at time t if

$$X(t) = \operatorname*{l.i.m}_{t' \to t}[X(t') - X(t)] \equiv \operatorname*{l.i.m}_{\epsilon \to 0}[X(t+\epsilon) - X(t)] = 0$$

Another way to express this limit is

$$\lim_{t' \to t} \|X(t') - X(t)\|^2 = \lim_{t' \to t} E\left[(X(t') - X(t))^2\right] = \lim_{\epsilon \to 0} E\left[(X(t+\epsilon) - X(t))^2\right] = 0$$

Mean-square continuity does not imply continuity at the individual sample function level.
The condition for mean-square continuity is given by the following theorem:

Theorem 10.2 *At time t, a stochastic process $X(t)$ is mean-square continuous iff the correlation function $\Gamma(t_1, t_2)$ is continuous at $t = t_1 = t_2$.*

Proof Continuity of $\Gamma(t_1, t_2)$ is sufficient for m.s. continuity of $X(t)$:

$$\begin{aligned} E[(X(t') - X(t))^2] &= E[X^2(t')] - E[X(t')X(t)] - E[X(t)X(t')] + E[X^2(t)] \\ &= \Gamma(t', t') - \Gamma(t', t) - \Gamma(t, t') + \Gamma(t, t) \end{aligned}$$

Therefore if $\Gamma(t_1, t_2)$ is continuous at $t = t_1 = t_2$, then the RHS of the above equation is equal to zero as $t' \to t$ and $X(t)$ is mean-square continuous at t.

Now let us assume that $X(t)$ is mean-square continuous at t. Form the difference $\Gamma(t_1, t_2) - \Gamma(t, t)$:

$$\begin{aligned} \Gamma(t_1, t_2) - \Gamma(t, t) &= E[X(t_1)X(t_2)] - E[X(t)X(t)] \\ &= E[(X(t_1) - X(t))(X(t_2) - X(t))] + E[(X(t_1) - X(t))X(t)] \\ &= E[X(t)(X(t_2) - X(t))] \end{aligned}$$

and therefore

$$\begin{aligned} |\Gamma(t_1, t_2) - \Gamma(t, t)| &\leq |E[(X(t_1) - X(t))(X(t_2) - X(t))]| + |E[(X(t_1) - X(t))X(t)]| \\ &+ |E[X(t)(X(t_2) - X(t))]| \end{aligned}$$

Apply the Cauchy-Schwarz inequality to each term on the RHS of the above to obtain

$$\begin{aligned} |\Gamma(t_1, t_2) - \Gamma(t, t)| &\leq \sqrt{E[(X(t_1) - X(t))^2 E(X(t_2) - X(t))^2]} + \sqrt{E[(X(t_1) - X(t))^2 E[X^2(t)]} \\ &+ \sqrt{E[X^2(t)]E(X(t_2) - X(t))^2]} \end{aligned}$$

Since $X(t)$ is m.s. continuous at t, the RHS of the above approaches zero as $t_1, t_2 \to t$ as follows

$$\lim_{t1, t2 \to t} \Gamma(t_1, t_2) = \Gamma(t, t)$$

and $\Gamma(t_1, t_2)$ is continuous at $t = t_1 = t_2$. ∎

10.3 Mean-Square Differentiation

We introduce the following operator on functions of one variable to simplify the expressions which follow:

$$\Delta_\epsilon f(x) = f(x + \epsilon) - f(x).$$

Because the operations which follow involve functions of two variables we define two delta-operators

$$\begin{aligned}\Delta_\epsilon^{(1)} f(t,s) &= f(t+\epsilon,s) - f(t,s)\\ \Delta_\epsilon^{(2)} f(t,s) &= f(t,s+\epsilon) - f(t,s)\end{aligned}$$

A stochastic process $X(t)$ has a mean-square (m.s.) derivative denoted by $\dot{X}(t)$ if there exists a finite power stochastic process

$$\dot{X}(t) = \operatorname*{l.i.m}_{\epsilon\to 0}\left[\frac{X(t+\epsilon)-X(t)}{\epsilon}\right] = \operatorname*{l.i.m}_{\epsilon\to 0}\left[\frac{\Delta_\epsilon X(t)}{\epsilon}\right]$$

These equations can also be expressed as

$$\lim_{\epsilon\to 0} E\left[\left(\frac{X(t+\epsilon)-X(t)}{\epsilon} - \dot{X}(t)\right)^2\right] = \lim_{\epsilon\to 0} E\left[\left(\frac{\Delta_\epsilon X(t)}{\epsilon} - \dot{X}(t)\right)^2\right] = 0$$

According to the Cauchy Convergence Theorem, as $\epsilon \to 0$, the quantity $\Delta_\epsilon X(t)/\epsilon$ has a m.s. limit iff

$$\operatorname*{l.i.m}_{\epsilon_1,\epsilon_2\to 0}\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1} - \frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right) = 0.$$

This condition can be written as

$$\begin{aligned}\lim_{\epsilon_1,\epsilon_2\to 0} E\left[\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1} - \frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right)^2\right] &= \lim_{\epsilon_1,\epsilon_2\to 0} E\left[\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1}\right)^2 - 2\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1}\right)\left(\frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right) + \left(\frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right)^2\right]\\ &= 0.\end{aligned}$$

There are two terms in the above which are equal to

$$\begin{aligned}\lim_{\epsilon\to 0} E\left[\left(\frac{\Delta_\epsilon X(t)}{\epsilon}\right)^2\right] &= \lim_{\epsilon\to 0} E\left[\left(\frac{X(t+\epsilon)-X(t)}{\epsilon}\right)^2\right]\\ &= \lim_{\epsilon\to 0}\frac{\Gamma(t+\epsilon,t+\epsilon) - \Gamma(t+\epsilon,t) - \Gamma(t,t+\epsilon) + \Gamma(t,t)}{\epsilon^2}\end{aligned}$$

and a cross term that is equal to

$$\begin{aligned}\lim_{\epsilon_1,\epsilon_2\to 0} E\left[\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1}\right)\left(\frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right)\right] &= \lim_{\epsilon_1,\epsilon_2\to 0} E\left[\left(\frac{X(t+\epsilon_1)-X(t)}{\epsilon_1}\right)\left(\frac{X(t+\epsilon_2)-X(t)}{\epsilon_2}\right)\right]\\ &= \lim_{\epsilon_1,\epsilon_2\to 0}\frac{\Gamma(t+\epsilon_1,t+\epsilon_2) - \Gamma(t+\epsilon_1,t) - \Gamma(t,t+\epsilon_2) + \Gamma(t,t)}{\epsilon_1\epsilon_2}\end{aligned}$$

Substituting these results back into the expression for the Cauchy series above we obtain

$$\begin{aligned}\textbf{(A)}\quad \lim_{\epsilon_1,\epsilon_2\to 0} E\left[\left(\frac{\Delta_{\epsilon_1} X(t)}{\epsilon_1} - \frac{\Delta_{\epsilon_2} X(t)}{\epsilon_2}\right)^2\right] &= 2\lim_{\epsilon\to 0}\frac{\Gamma(t+\epsilon,t+\epsilon) - \Gamma(t+\epsilon,t) - \Gamma(t,t+\epsilon) + \Gamma(t,t)}{\epsilon^2}\\ &\quad -2\lim_{\epsilon_1,\epsilon_2\to 0}\frac{\Gamma(t+\epsilon_1,t+\epsilon_2) - \Gamma(t+\epsilon_1,t) - \Gamma(t,t+\epsilon_2) + \Gamma(t,t)}{\epsilon_1\epsilon_2}.\end{aligned}$$

Putting all these conditions together leads us to the

Theorem 10.3 *A finite-power stochastic process $X(t)$ is mean-square differentiable at t iff, the double limit*

$$\lim_{\epsilon_1,\epsilon_2\to 0} \frac{\Delta^{(1)}_{\epsilon_1}\Delta^{(2)}_{\epsilon_2}\Gamma(t,t)}{\epsilon_1\epsilon_2} = \lim_{\epsilon_1,\epsilon_2\to 0} \frac{\Gamma(t+\epsilon_1,t+\epsilon_2)-\Gamma(t+\epsilon_1,t)-\Gamma(t,t+\epsilon_2)+\Gamma(t,t)}{\epsilon_1\epsilon_2}$$

exists and is finite.

Proof Sufficient Condition: if $\frac{\partial\Gamma(t_1,t_2)}{\partial t_1}$, $\frac{\partial\Gamma(t_1,t_2)}{\partial t_2}$, and $\frac{\partial^2\Gamma(t_1,t_2)}{\partial t_1\partial t_2}$ exist in a neighborhood of $(t_1,t_2)=(t,t)$ and $\frac{\partial^2\Gamma(t_1,t_2)}{\partial t_1\partial t_2}$ is continuous at $(t_1,t_2)=(t,t)$, then the above limit exist by the Calculus, and the process $X(t)$ will be differentiable at t.

If the above sufficient condition does not hold, then $X(t)$ is m.s. differentiability implies that the above mean-square expectation **(A)** is zero, independently of how ϵ_1 and ϵ_2 approach zero which means that the right-hand-side of the expression in the above exists and is finite. On the other hand if the right-hand-side of the expression in theorem [10.3] exists and has the value R, then this must hold regardless of how ϵ_1 and ϵ_2 approach zero. That means the limit of the first term on the RHS of the expectation **(A)** is 2R and therefore the two terms cancel out exactly and $X(t)$ is m.s. differentiable. ∎

Assuming that the sufficient condition in the 1st part of the above proof holds, the correlation function for the m.s. derivative $\dot{X}(t)$ is

$$\Gamma_{\dot{X}}(t_1,t_2) = E[\dot{X}(t_1)\dot{X}(t_2)] = \frac{\partial^2\Gamma(t_1,t_2)}{\partial t_1\partial t_2}.$$

Example Brownian Motion. We saw that the correlation function for Brownian motion is given by $\Gamma(t,s)=E[B(t)B(s)]=\sigma^2\min(s,t)$.

Now

$$\min(s,t) = \begin{cases} t, & t<s \\ s, & t>s \end{cases}$$

so that

$$\frac{\partial}{\partial s}\min(s,t) = \begin{array}{ll} 0, & t<s \\ 1, & t>s \end{array}$$

which is the Heavyside unit step function (in t) and its derivative with respect to t is the Dirac delta function: $\delta(t-s)$.

The correlation function for the m.s. derivative of Brownian motion is therefore

$$\frac{\partial^2\Gamma(t,s)}{\partial t\partial s} = \sigma^2\delta(t-s).$$

Hence the second derivative does not exist in the sense of ordinary calculus function and therefore Brownian motion is not m.s. differentiable (we will see more on this in Section [11]). However, in the sense of *generalized functions* we see that the second derivative of the correlation function is the delta-function which is the correlation function for the "white noise" stochastic process. Hence in the sense of generalized functions, Brownian motion is the integral of white noise, even though it is not m.s. differentiable. △

10.4 Mean-Square Riemann Integration

The last subject we will explore in mean-square calculus is mean-square integration. Integrals of stochastic process appear in many applications – for example a slowly varying signal might be corrupted by high-frequency noise. In another application a more realistic model for physical Brownian motion (which is a finite length path) is given by the Langevin equation[20],[21] which requires integrating "white noise" in the presence of frictional effects such as viscosity. Integrals of stochastic processes will therefore enter in the construction of *Filters* for *Smoothing* as well as *Prediction*.

As in ordinary calculus we will partition the finite interval $[a, b]$ into subdivisions labeled by points $t_k, k = 0, 1, \ldots, n$ such that

$$a = t_0 < t_1 < \ldots < t_n = b.$$

We will also label the increments in t as

$$\Delta t_i = t_i - t_{i-1}, \quad 1 \leq i \leq n$$

We will label such a partition by P_n which is made up of the $n + 1$ points $t_k, k = 0, 1, \ldots n$. Also, although for simplicity we can take all the increments to be of equal size, we allow for the general case and define the upper bound on the increment (or mesh) size by

$$\Delta_n = \max_k \Delta t_k$$

As the partitions are refined into more and more points, Δ_n will decrease in a well behaved manner. Now, as in ordinary calculus we can form a Riemann sum for a finite-power stochastic process $X(t)$ as

$$\sum_{k=1}^{n} X(t_k^*)\Delta t_k, \quad \text{where} \quad t_k^* \in (t_{k-1}, t_k].$$

From this starting point we define the mean-square Riemann integral on the interval $[a, b]$ (known as the Wiener integral) as

$$\int_a^b X(t)dt \equiv \operatorname*{l.i.m}_{\Delta_n \to 0} \sum_{k=1}^{n} X(t_k^*)\Delta t_k. \tag{34}$$

If everything is well behaved, then as $\Delta_n \to 0$, then $n \to \infty$ and the Riemann sum converges in mean-square to the mean-square Riemann integral. We state the requirement for convergence to occur:

Theorem 10.4 *The Wiener integral, Eq. [34], exists, iff the ordinary double integral*

$$\int_a^b \int_a^b \Gamma(s,t) ds dt \tag{35}$$

exists and is finite.

Proof We will apply the Cauchy Convergence Criteria to prove the above result. Let P_n and P_m denote two distinct partitions of the $[a,b]$ interval. Let us take them to be given by

$$P_n : \left\{ \begin{array}{rl} a & = t_0 < t_1 < \ldots < t_n = b \\ \Delta t_i & = t_i - t_{i-1} \\ \Delta_n & = \max_i \Delta t_i \end{array} \right.$$

$$P_m : \left\{ \begin{array}{rl} a & = s_0 < s_1 < \ldots < s_n = b \\ \Delta s_i & = s_i - s_{i-1} \\ \Delta_m & = \max_i \Delta s_i \end{array} \right.$$

According to the Cauchy Convergence Criteria, the Wiener integral (Eq. [34]) exists iff,

$$\lim_{\Delta_n, \Delta_m \to 0} E\left[\left(\sum_{k=1}^{n} X(t_k^*) \Delta t_k - \sum_{j=1}^{m} X(s_j^*) \Delta s_j \right)^2 \right] = 0 \tag{36}$$

Now expand the square in Eq. [36] and take the expectations to obtain

$$\sum_{k=1}^{n}\sum_{i=1}^{n} \Gamma(t_k^*, t_i^*) \Delta_{t_k} \Delta_{t_i} - 2\sum_{k=1}^{n}\sum_{j=1}^{m} \Gamma(t_k^*, s_i^*) \Delta_{t_k} \Delta_{s_j} + \sum_{j=1}^{m}\sum_{i=1}^{m} \Gamma(s_k^*, s_i^*) \Delta_{s_k} \Delta_{s_i} = 0 \tag{37}$$

As Δ_n and Δ_m approach zero, Eq. [36] is true iff

$$\lim_{\Delta_n, \Delta_m \to 0} \sum_{k=1}^{n}\sum_{j=1}^{m} \Gamma(t_k^*, s_i^*) \Delta_{t_k} \Delta_{s_j} = \int_a^b \int_a^b \Gamma(t,s) dt ds \tag{38}$$

The cross-term in Eq. [37] must converge independent of the paths that n and m take as $n, m \to \infty$. If this happens the first and third sums on LHS of Eq. [37] converge to the same double integral, and Eq. [36] holds true. ∎

We end this section with an example based on Brownian motion.

Example Let $X(t)$ be the Brownian motion process, and consider the m.s. integral

$$Y(t) = \int_0^t X(s)ds \tag{39}$$

Now, for the Brownian motion process $\Gamma(u,v) = \sigma^2 \min(u,v)$ and this can be integrated to obtain

$$\begin{aligned} \int_0^t \int_0^t \Gamma(u,v)dudv &= \int_0^t \int_0^t \sigma^2 \min(u,v)dudv \\ &= \sigma^2 \int_0^t [\int_0^v udu + \int_v^t vdu]\, dv \\ &= \sigma^2 t^3/3 \end{aligned}$$

Therefore, Brownian motion is m.s. Riemann integrable. △

11 Properties of Brownian Motion Paths

Most of the ideas in this section were explained to me by Howard Weiner of UC Davis [6].

There are several important properties of Brownian motion:

1) Brownian motions have finite total Quadratic Variation.
2) Brownian motions have infinite length almost surely (a.s.)
3) Brownian motions are non-differentiable
4) Brownian motions are continuous functions t or have continuous versions.

In what follows in this section let B(t) be the standard Brownian Motion (B.M.) with mean zero and variance parameter $\sigma^2 = 1$.

11.1 Brownian motions have finite Quadratic Variation

The exposition in this section works with one complete Brownian motion chosen from the sample space at a time, or at the sample function level.

Define the Quadratic Variation of B.M. as follows (on $[0,1]$):
Let

$$V_n \equiv \sum_{l=1}^{2^n} \left[B(\frac{l}{2^n}) - B(\frac{l-1}{2^n}) \right]^2,$$

where we again sum over the 2^n dyadic intervals of length $1/2^n$.

By stationary independent increments, the random variables

$$\Big[B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\Big]$$

are independent identically distributed (IID) $\sim N(0, 1/2^n)$. Therefore $E[V_n] = 1$.

From Eq.[8], $X \sim N(0,\sigma^2) \implies E[X^4] = 3\sigma^4$ so that

$$\mathrm{Var}[X^2] = E[X^4] - (E[X^2])^2 - 3\sigma^4 - \sigma^4 = 2\sigma^4.$$

Then

$$\mathrm{Var}[V_n] = E[(V_n - 1)^2] = \left(2(\frac{1}{2^n})^2\right) \cdot 2^n = \frac{1}{2^{n-1}}$$

Therefore

$$\lim_{n\to\infty} E[(V_n - 1)^2] \to 0 \quad \text{and} \quad \lim_{n\to\infty} V_n = 1 \quad \text{in q.m.}$$

Then define the Quadratic Variation of BM on $[0,1]$ by

$$V \equiv \lim_{n\to\infty} V_n.$$

Therefore $V \to 1$ on $[0,1]$ in q.m.

To go a little further[3] define a corresponding set of random variables X_l by

$$X_l = [B(t_{l+1}) - B(t_l)]^2 - [t_{l+1} - t_l]$$

and use the partition, π_n of $[0,1]$ given by $t_l = \frac{l}{2^n}, l = 0, 1, 2, \ldots 2^n, \Delta t_n = 2^{-n}$.
Then

$$E\Big[\left(\sum_{i=0}^{n-1} [B(t_{l+1}) - B(t_l)]^2 - (t_{l+1} - t_l)\right)^2\Big] = E\Big[\left(\sum_{i=0}^{n-1} [B(t_{l+1}) - B(t_l)]^2 - t\right)^2\Big]$$

Since the X_l's are independent with zero means we can write the above as

$$E\Big[\left(\sum_{i=0}^{n-1} \Big[\big(B(t_{l+1}) - B(t_l)\big)^2\Big] - t\right)^2\Big] = E\left[\Big(\sum_{i=0}^{n-1} X_i\Big)^2\right] = \sum_{i=0}^{n-1} E[X_i^2]$$

Now since the 4th moment of a normally distributed random variable with variance σ^2 is $3\sigma^4$, we have

$$\begin{aligned} E[X_i^2] &= E[(B(t_{i+1}) - B(t_i))^4] - 2(t_{i+1} - t_i)E[(B(t_{i+1}) - B(t_i))^2] + (t_{i+1} - t_i)^2 \\ &= 3(t_{i+1} - t_i)^2 - 2(t_{i+1} - t_i)^2 + (t_{i+1} - t_i)^2 \\ &= 2(t_{i+1} - t_i)^2 \end{aligned}$$

[3]Following the discussion in https://idontgetoutmuch.wordpress.com/2012/03/17/the-quadratic-variation-of-brownian-motion

Therefore

$$\begin{aligned}\sum_{i=0}^{n-1} E[X_i^2] &= 2\sum_{i=0}^{n-1}(t_{i+1}-t_i)^2 \\ &\leq 2(\Delta t_n)\sum(t_{i+1}-t_i) \\ &= 2t\Delta t_n\end{aligned}$$

and so

$$E\left[\Big(\sum_{i=0}^{n-1} X_i\Big)^2\right] \leq 2t\Delta t_n$$

and by Chebyshev's inequality (see appendix)

$$P\left(\Big|\sum_{i=0}^{n-1} X_i\Big| > \epsilon\right) \leq \frac{E\left[\left(\sum_{i=0}^{n-1} X_i\right)^2\right]}{\epsilon^2} \leq \frac{2t\Delta t_n}{\epsilon^2}$$

Now choose a sequence π_n of the 2^n dyadic divisions of the interval $[0,1]$ with $n = 1, 2, \ldots$ and note that $\sum_{i=0}^{\infty} \frac{1}{2^n} = 2$. Then

$$\sum_{n=1}^{\infty} P\left(\Big|\sum_{t_i\in\pi_n} [B(t_{i+1})-B(t_i)]^2 - (t_{i+1}-t_i)\Big| > \epsilon\right) \leq \frac{2t\sum_{n=1}^{\infty}\Delta t_n}{\epsilon^2} < \infty$$

By the first Borel-Cantelli lemma (see appendix) there can only be a finite number of members of the π_n sequence that are such that

$$\Big|\sum_{t_i\in\pi_n} [B(t_{i+1})-B(t_i)]^2 - (t_{i+1}-t_i)\Big| > \epsilon$$

Therefore we can conclude that there must exist an $N \geq 1$ such that for all $n > N$ we must have

$$\Big|\sum_{t_i\in\pi_n} [B(t_{i+1})-B(t_i)]^2 - (t_{i+1}-t_i)\Big| < \epsilon$$

Hence we conclude that

$$\sum_{t_i\in\pi_n} [B(t_{i+1})-B(t_i)]^2 - (t_{i+1}-t_i) \to 0$$

wp1 (a.s.) Therefore we have shown that the quadratic variation $V \to 1$ on $[0,1]$ in q.m. and a.s.

11.2 Brownian motions have infinite length

This section also works at the sample function level.

Now define the Total Variation of B.M. on $[0,1]$ by

$$T = \lim_{n\to\infty} T_n \equiv \lim_{n\to\infty} \sum_{l=1}^{2^n} \left|B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\right|.$$

By the triangle inequality, since the (n+1)-th mesh cuts the n-th in half, T_n is monotonically increasing and converges to T, i.e., $T_1 < T_2 < \ldots < T_n < T_{n+1} < \ldots$ and

$$\lim_{n\to\infty} T_n \to T$$

which we designate with the $\uparrow$ symbol (see Section [14.1]), viz.,

$$T_n \leq T_{n+1} \uparrow T$$

So that T is well-defined a.s.and is a lower bound for the path length, since $\{B(t)\}$ is continuous a.s. Now

$$V_n = \sum_{l=1}^{2^n} \left[B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\right]^2$$

so that

$$V_n \leq \left[\max_{1\leq l\leq 2^n} \left|B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\right|\right] \cdot \left(\sum_{l=1}^{2^n} \left|B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\right|\right)$$

or

$$V_n \leq M_n \cdot T_n$$

where $M_n = \left[\max_{1\leq l\leq 2^n} \left|B(\frac{l}{2^n}) - B(\frac{l-1}{2^n})\right|\right]$. Then

$$T \geq T_n \geq \frac{V_n}{M_n}.$$

As $n \to \infty, M_n \to 0$ (a.s.) by B.M. path continuity. Since $V_n \to 1$ (a.s.) it follows that $T = \infty$ (a.s.)

11.3 On the Non-Differentiability of B.M. paths

Define, for $\{Z_i\}$ IID $N(0,1)$, $\{$ $r_i(x), 0 \leq x \leq 1$ orthonormal (o.n.) functions$\}$

$$W_n(t) = \sum_{i=1}^{n} Z_i \int_0^t r_i(x)\, dx,$$

Then $W_n(t) \to B(t)$ in quadratic mean (q.m.). If $B(t)$ had differentiable paths, one would suppose that the derivatives

$$\frac{d}{dt} W_n(t) = \sum_{i=1}^{n} Z_i r_i(t)$$

would be a Cauchy sequence in q.m., that is that

$$E[\left(\frac{d}{dt}W_n(t)\frac{d}{dt}W_n(t)\right)^2] \to 0$$

as $n \to \infty$. But

$$\mathrm{Var}[\frac{d}{dt}W_n(t)] = \sum_{i=1}^{n} r_i^2(t) \uparrow \infty,$$

so that

$$\{\frac{d}{dt}W_n(t)\}$$

cannot converge in q.m.

11.4 Brownian motions are "continuous"

Define a process $K(t)$ to be a Normal Process with 0 mean $E[K(t)] = 0$ and with covariance function given by

$$\begin{aligned} \mathrm{Cov}[K(s), K(t)] &= \sigma^2\delta(|t-s|) \\ \mathrm{Var}[K(t)] &= \infty \end{aligned}$$

where $\delta(x)$ is the Dirac-delta function.

Claim

$$W(t) = \int_0^t K(v)dv,$$

is B.M. with parameter σ^2.

Proof:

$$E[W(t)] = \int_0^t E[K(v)]\,dv = 0$$

and for $s < t$

$$\begin{aligned}
\text{Cov}[W(s), W(t)] &= E[\int_0^s K(u)\,du \int_0^t K(v)\,dv] \\
&= \int_0^s \int_0^t E[K(u)K(v)]dudv \\
&= \sigma^2 \int_0^s \int_0^t \delta(|u-v|)dudv \\
&= \sigma^2 \int_0^s \left[\int_0^s \delta(|u-v|) + \int_s^t \delta(|u-v|)\right] \\
&= \sigma^2 \int_0^s \int_0^s \delta(|u-v|) = \sigma^2 \int_0^s dv = \sigma^2 s
\end{aligned}$$

So we see that $W(t)$ has the same properties as B.M. and hence B.M. can be written as an "integral"

$$B(t) = \int_0^t K(v)\,dv.$$

Since integration "improves" continuity properties we can conclude that "$B(t)$ is continuous".

However, as shown above $\frac{d}{dt}B(t) = K(t)$ "does not exist." because B.M. is non-differentiable.

Note: The stochastic process $K(v)$ used in the above derivation is known as "white noise".

12 Ito's Isometry and Ito's Lemma

In this section we will investigate the existence, in a certain sense, of

$$\text{“}\int_0^t f(s,\omega)dB(s,\omega)''$$

where $B(t,\omega)$ is a 1-dimensional Brownian motion starting at the origin, for a class of functions $f : [0,\infty] \times \Omega \to R$.

Suppose $0 \le S \le T$ and $f(t,\omega)$ is given, We want to define

$$\int_S^T f(s,\omega)dB(s,\omega).$$

To do so we will develop the integral for a simple class of functions f and then extend to more general functions by some approximation method. Let us first assume that f has the form

$$\phi(t,\omega) = \sum_{j\geq 0} e_j(\omega)\cdot \chi_{(j\cdot 2^{-n},(j+1)2^{-n}]}(t),$$

where $\chi_{[t_j,t_{j+1})}(t)$ is the Indicator Function which is equal to 1 if $t\in[t_j,t_{j+1})$ and 0 if t is outside that interval and n is a natural number. For such functions it is reasonable to define

$$\int_S^T f(s,\omega)dB_s(\omega) = \sum_{j\geq 0}[B(t_{j+1})-B(t_j)](\omega),$$

where

$$t_k = t_k^{(n)} = \left\{ \begin{array}{lll} k\cdot 2^{-n} & \text{if} & S\leq k\cdot 2^{-n}\leq T \\ S & \text{if} & k\cdot 2^{-n} < S \\ T & \text{if} & k\cdot 2^{-n} > T \end{array} \right\}$$

Without any further assumptions on the functions $e_j(\omega)$ the above definition leads to difficulties as the next example shows.

Example Choose

$$\begin{aligned} \phi_1(t,\omega) &= \sum_{j\geq 0} B(j\cdot 2^{-n},\omega)\cdot \chi_{(j\cdot 2^{-n},(j+1)\cdot 2^{-n}]}(t) \\ \phi_2(t,\omega) &= \sum_{j\geq 0} B((j+1)\cdot 2^{-n},\omega)\cdot \chi_{(j\cdot 2^{-n},(j+1)\cdot 2^{-n}]}(t) \end{aligned}$$

Then

$$E[\int_S^T \phi_1(t,\omega)dB(t,\omega)] = \sum_{j\geq 0} E[B(t_j)(B(t_{j+1})-B(t_j))] = 0,$$

since $B(t,\omega)$ has independent increments. But

$$\begin{aligned} E[\int_S^T \phi_2(t,\omega)dB(t,\omega)] &= \sum_{j\geq 0} E[B(t_{j+1})(B(t_{j+1})-B(t_j))] \\ &= \sum_{j\geq 0} E[(B(t_{j+1})-B(t_j))^2] = T \end{aligned}$$

So, in spite of both ϕ_1 and ϕ_2 appearing to be reasonable appoximations to $f(t,\omega)=B(t,\omega)$, their integrals are not close to each other at all no matter how large n is taken to be.

It is natural to approximate a function $f(t,\omega)$ by

$$\sum_j f(t_j^*,\omega)\cdot \chi_{(t_j,t_{j+1}]}(t)$$

where the points t_j^* belong to the intervals $(t_j, t_{j+1}]$, and then define $\int_S^T f(t,\omega)dB(t,\omega)$ as the limit of $\sum_j f(t_j^*,\omega)[B(t_{j+1}) - B(t_j)](\omega)$ as $n \to \infty$. However as the example above shows that – unlike the Riemann-Stieiltjes integral – it does make a difference here what points t_j^* we choose.

There are two standard choices that have turned out to be the most useful:

1) $t_j^* = t_j$ (the left end point), which leads to the *Ito integral* which we will denote by

$$\int_S^T f(t,\omega)dB(t,\omega),$$

2) $t_j^* = (t_j + t_{j+1})/2$ (the mid point), which leads to the *Stratonovich integral* denoted by

$$\int_S^T f(t,\omega) \circ dB(t,\omega),$$

We will work with square-integrable functions whose mean-square is assumed to be finite, i.e.,

$$E[\int_S^T f(t,\omega)^2\, dt] < \infty.$$

We will also work with "elementary functions" of the form

$$\phi(t,\omega) = \sum_j e_j(\omega) \cdot \chi_{[t_j,t_{j+1})}(t)$$

where $e_j(\omega)$ are a set of n real numbers indexed by $j = 1, 2, ..., n$. $\phi(t,\omega)$ is a "simple function" and can used to approximate an arbitrary measurable function on (T, Ω).

Ito's Isometry follows. An *Isometry* is a transformation that preserves the norm – in this case the L_2-norm with respect to square-integrable functions. Another famous isometry is the Plancherel-Parseval theorem (see Eq. [58]).

Lemma 12.1 *The Ito isometry: If $\phi(t,\omega)$ is bounded and elementary then*

$$E\Big[\Big(\int_S^T \phi(t,\omega)\, dB_t(\omega)\Big)^2\Big] = E\Big[\int_S^T \big(\phi(t,\omega)\big)^2\, dt\Big].$$

Proof Put $\Delta B_j = B(t_{j+1}) - B(t_j)$. Then

$$E[e_i e_j \Delta B_i \Delta B_j] = \begin{cases} 0 & \text{if} \quad i \neq j \\ E[e_j^2] & \text{if} \quad i = j \end{cases}$$

using the fact that $e_i e_j \Delta B_i$ and ΔB_j are independent if $i < j$. Therefore

$$\begin{aligned} E[(\int_S^T \phi\, dB)^2] &= \sum_{i,j} E[e_i e_j \Delta B_i \Delta B_j] = \sum_j E[e_j^2] \cdot (t_{j+1} - t_j) \\ &= E[\int_S^T \phi^2 dt]. \quad \blacksquare \end{aligned}$$

It is possible to extend the application of the Ito isometry to functions, f which can be represented by the limit as $n \to \infty$ of a sequence of elementary functions ϕ_n in the sense of mean squares, i.e.,

$$E[\int_S^T (f - \phi_n)^2 dt] \to 0 \text{ as } n \to \infty.$$

In this case we would define the Ito integral $\mathcal{I}$ as

$$\mathcal{I}[f](\omega) := \lim_{n\to\infty} \int_S^T \phi_n(t,\omega)\, dB(t,\omega).$$

In this case the Ito isometry becomes

$$E\Big[\Big(\int_S^T f(t,\omega)\, dB_t(\omega)\Big)^2\Big] = E\Big[\int_S^T f(t,\omega)^2\, dt\Big].$$

Let us know look at Ito's lemma:

Lemma 12.2 *Ito's lemma*

$$\int_0^t B(s,\omega)\, dB(s,\omega) = \frac{1}{2}B^2(t,\omega) - \frac{1}{2}t.$$

Before going into the proof the Ito's lemma consider the values of the Brownian motion (a continuous function) at discrete times $t \in \{t_1, t_2, ..., t_n\}$ or the set $\{B_1, B_2, ..., B_n\}$.

Let $\Delta B_j = B_j - B_{j-1}$ and let $\Delta B_j^2 = B_j^2 - B_{j-1}^2$. Then

$$(\Delta B_j)^2 = B_j^2 + B_{j-1}^2 - 2B_{j-1}B_j$$

and

$$\begin{aligned}(\Delta B_j)^2 + 2B_j\Delta B_j &= B_j^2 + B_{j-1}^2 - 2B_{j-1}B_j \\ &= B_j^2 + B_{j-1}^2 - 2B_{j-1}B_j + 2B_{j-1}B_j - 2B_{j-1}^2 \\ &= B_j^2 - B_{j-1}^2 \\ &= \Delta B_j^2\end{aligned}$$

or

$$\Delta B_j^2 = 2B_{j-1}\Delta B_j + (\Delta B_j)^2$$

Or equivalently,

$$B_{j-1}\Delta B_j - \frac{1}{2}\Delta B_j^2 + \frac{1}{2}(\Delta B_j)^2 = 0$$

Taking "infinitesmal limits" we can write

$$\Delta B^2 = 2\, B\, \Delta B + (\Delta B)^2$$

Now Ito's lemma can be shown to hold "on average" by taking the expectation of the above result. Because Brownian motions have properties of normal distributions, then this gives

$$\begin{aligned} E[\Delta B^2] &= E[2\,B\,\Delta B + (\Delta B)^2] \\ &= 2B\,\Delta B + \sigma^2 \Delta t \end{aligned}$$

where σ^2 is the standard deviation associated with the Brownian motion. This is Ito's lemma. In Integral Form is can be written as

$$\int B\,dB = \frac{1}{2}B^2 - \frac{1}{2}\sigma^2 t$$

Let us know prove Ito's lemma in detail

Proof Put $\phi_n(s,\omega) = \sum B_j(\omega) \cdot \chi_{[t_j,t_{j+1})}(s)$, where $B_j = B(t_j,\omega)$ and $B_s = B(s,\omega)$. Then

$$\begin{aligned} E[\int_0^t (\phi_n - B(s,\omega))^2 ds] &= E[\sum_j \int_{t_j}^{t_{j+1}} (B_j - B_s)^2\,ds] \\ = \sum_j \int_{t_j}^{t_{j+1}} (s - t_j) ds &= \sum_j \frac{1}{2}(t_{j+1} - t_j)^2 \to 0 \text{ as } \Delta t \to 0. \end{aligned}$$

Therefore

$$\int_0^t B_s\,dB_s = \lim_{\Delta t_j \to 0} \int_0^t \phi_n\,dB_s.$$

Now

$$\begin{aligned} & E[(\frac{1}{2}B_t^2 - \frac{1}{2}t - \sum_j B_j \Delta B_j)^2] \\ = \; & E[\frac{1}{4}B_t^4 + \frac{1}{4}t^2 + (\sum_j B_j \Delta B_j)^2 - B_t^2 \cdot \sum_j B_j \Delta B_j - \frac{1}{2}tB_t^2 + t \cdot \sum_j B_j \Delta B_j] \\ = \; & \frac{3}{4}t^2 + \frac{1}{4}t^2 + E[(\sum B_j \Delta B_j)^2] - E[B_t^2 \cdot \sum B_j \Delta B_j] - \frac{1}{2}t^2 + 0. \end{aligned}$$

Also,

$$E[(\sum B_j \Delta B_j)^2] = \sum_{i,j} E[B_i B_j \Delta B_i \Delta B_j] = \sum_i E[B_i^2 (\Delta B_i)^2] \sum_i t_i \Delta t_i$$

and

$$
\begin{aligned}
& E[B_t^2 \cdot \sum B_j \Delta B_j] = \sum_j E[B_t^2 B_j \Delta B_j] = \sum_j E[\{(B_t - B_{j+1}) + \Delta B_j + B_j\}^2 \cdot B_j \cdot \Delta B_j] \\
= & \sum_j E[(B_t - B_{j+1})^2 B_j \Delta B_j] + \sum_j E[B_j \cdot (\Delta B_j)^3] + \sum_j E[B_j^3 \cdot \Delta B_j] \\
+ & 2 \sum E[(B_t - B_{j+1}) B_j \Delta B_j^2] + 2 \cdot \sum_j E[(B_t - B_{j+1}) B_j^2 \Delta B_j + 2 \sum_j E[B_j^2 \Delta B_j^2] \\
= & 0 + 0 + 0 + 0 + 0 + 2 \sum_j t_j \cdot \Delta t_j
\end{aligned}
$$

Hence we have Ito's lemma:

$$
E[(\frac{1}{2} B_t^2 - \frac{1}{2} t - \sum_j B_j \Delta B_j)^2] = \frac{1}{2} t^2 - \sum_j t_j \Delta t_j \to 0 \text{ as } \Delta t_j \to 0 \qquad \blacksquare
$$

This is the end of the demonstration of Ito's lemma.

As is well-known, Langevin's description of Brownian motion introduces a physical modification to take into account viscosity (friction) [20], [21]. The result is that physical Brownian motion has finite total variation and the particles undergoing Brownian motion do not travel at infinite speed.

It would be interesting to explore the effects of special relativity which require that the speed of any particle is less than the speed of light (relativistic Brownian Motion). Also, in Quantum Field Theory – the commutators of physical observables vanish outside the light-cone, however the propagators associated with quantum mechanical wave-functions have exponentially decaying tails outside the light cone. We talk about time-like and space-like particles to distinguish particles which have time-like or space-like momentum 4-vectors, or equivalently are propagating inside or outside the light cone. The paradox is that even though it is possible for a propagator associated with a physical particle not to vanish outside the light cone, nevertheless there is a strict requirement of microscopic causality that the commutators associated with physical observables must vanish identically outside the light cone. A study of relativistic Brownian motion may shed some light on this subject.

13 Martingales

13.1 Martingales

The standard reference on martingales is [19]. A stochastic process $\{X(t), t \geq 0\}$ with finite means is said to be a *continuous parameter martingale* if for any set of times

$$
t_1 < t_2 < \ldots < t_n < t_{n+1}, \tag{40}
$$

$$E[X(t_{n+1})|X(t_1),\ldots,X(t_n)] = X(t_n). \tag{41}$$

Therefore the conditional expectation of $X(t_{n+1})$, given the history of values $X(t_1),\ldots,X(t_n)$ is equal to the most recently observed value $X(t_n)$. Consider a stochastic process $\{X(t), t \geq 0\}$ with independent increments and finite means together with the above set of times in Eq. [40]. Note that

$$X(t_{n+1}) = X(t_n) + [X(t_{n+1}) - X(t_n)].$$

Because $X(t)$ has independent increments, then $[X(t_{n+1}) - X(t_n)]$ is independent of any elements in the set $\{X(t_1),\ldots,X(t_n)\}$ because those can be considered boundary points of the respective intervals $[X(t_i) - X(0)]$ for each i (e.g., assuming $X(0) = 0$). Therefore

$$E[X(t_{n+1}) - X(t_n)|X(t_1),\ldots,X(t_n)] = E[X(t_{n+1}) - X(t_n)] = m(t_{n+1}) - m(t_n) \tag{42}$$

where $m(t) = E[X(t)] < \infty$ by assumption. Putting these facts together we see that

$$E[X(t_{n+1})|X(t_1),\ldots,X(t_n)] = X(t_n) + m(t_{n+1}) - m(t_n) \tag{43}$$

If $m(t_{n+1}) = m(t_n)$, then

$$E[X(t_{n+1})|X(t_1),\ldots,X(t_n)] = X(t_n) \tag{44}$$

and the stochastic process is a martingale.

A stochastic process $\{X_n, n = 1, 2\ldots,\}$ with finite means is said to be a *discrete parameter martingale* if for any integer n

$$E[X_{n+1}|X_1,\ldots,X_n] = X_n. \tag{45}$$

Let us consider some example martingales:

Example *Sums of independent, zero mean, random variables* Consider the consecutive sums of independent random variables with zero mean

$$S_n = \sum_{i=1}^{n} X(t_i), \qquad E[X(t_i)] = 0 \tag{46}$$

Then S_n is a martingale because $S_{n+1} = X_{n+1} + S_n$ and therefore

$$\begin{aligned} E[S_{n+1}|S_1,\ldots,S_n] &= E[S_{n+1}|X(t_1),\ldots,X(t_n)] = E[X(t_{n+1}) + S_n|X(t_1),\ldots,X(t_n)] \\ &= E[X(t_{n+1})|X(t_1),\ldots,X(t_n)] + E[S_n|X(t_1),\ldots,X(t_n)] \\ &= E[X(t_{n+1})] + E[S_n|X(t_1),\ldots,X(t_n)] = 0 + S_n \\ &= S_n \end{aligned}$$

since S_n is fixed if $X(t_1),\ldots,X(t_{n+1})$ are given. △

Example $Y_n = (S_n^2 - \sigma^2 n)$ *is a martingale.* Consider again the sums of independent, zero mean random variables with $E[X_i^2] = \sigma^2$ and square the sums and form the difference $S_n^2 - \sigma^2 n$. Then take the conditional expectation to obtain

$$\begin{aligned} E[(S_n^2 - \sigma^2 n)|X_1, \ldots, X_{n-1}] &= E[(S_{n-1} + X_n)^2 - \sigma^2 n | X_1, \ldots X_{n-1}] \\ &= E[S_{n-1}^2 | X_1, \ldots X_{n-1}] + E[X_n^2 | X_1, \ldots X_{n-1}] + 2E[X_n]E[S_{n-1}|X_1, \ldots, X_{n-1}] - \sigma^2 n \\ &= S_{n-1}^2 + E[X_n^2] - \sigma^2 n \\ &= S_{n-1}^2 + \sigma^2 - \sigma^2 n \\ &= S_{n-1}^2 - \sigma^2 (n-1) \end{aligned}$$

Since Y_i is determined completely by $X_1, \ldots, X_i, \quad i = 1, 2, \ldots, n-1$, we have by the above that

$$E[Y_n | Y_1, \ldots Y_{n-1}] = Y_{n-1}$$

and therefore $Y_n = (S_n^2 - \sigma^2 n)$ is a martingale. △

One can also conclude that the unconditional expectations will form a chain such that

$$E[S_n^2 - \sigma^2 n] = E[S_{n-1}^2 - \sigma^2 (n-1)] = \ldots = E[S^1 - \sigma^2] = E[X_1^2] - 1 = 0$$

and therefore that

$$E[S_n^2] = \sigma^2 n. \tag{47}$$

Example *Brownian motion is a martingale* For $s < t$ we have

$$\begin{aligned} E[B(t)|B(s)] &= E[B(t) - B(s) + B(s)|B(s)] \\ &= E[B(t) - B(s)|B(s)] + E[B(s)|B(s)] \\ &= 0 + B(s) \quad \text{by independent increments} \\ &= B(s) \end{aligned}$$

Therefore Brownian motion satisfies the martingale condition $E[B(t)|B(s)] = B(s)$ for $s < t$. △

Let us compute the conditional expectation for exponentiated Brownian motion. Assuming $s < t$

$$\begin{aligned} E\big[\exp[\theta B(t)]|B(s)\big] &= E[\exp\big(\theta[B(t) - B(s)] + \theta B(s)\big)|B(s)] \\ &= \exp[\theta B(s)] E[\exp\big(\theta[B(t) - B(s)]\big)] \end{aligned}$$

by independent increments. Since $B(t) - B(s)$ is normally distributed with mean 0 and variance $\sigma^2(t-s)$ we can make use of the m.g.f. for the normal distribution to compute

$$E\big[\exp\big(\theta[B(t) - B(s)]\big)\big] = \exp[\theta^2 \sigma^2 (t-s)/2] \tag{48}$$

and arrive at

$$E\big[e^{\theta B(t)}|B(s)\big] = e^{\theta B(s) + \frac{1}{2}\theta^2 \sigma^2 (t-s)} \tag{49}$$

So, there is an extra term on the RHS of Eq. [49] which means that $\exp[\theta B(t)]$ is not a martingale. However, as this can be rectified as shown in the following.

Example *Exponentiated Brownian Motion Martingale* Define a stochastic process

$$X(t,\theta) = e^{\theta B(t) - \frac{1}{2}\theta^2\sigma^2 t}$$

In this case $X(t,\theta)$ will be a martingale because

$$\begin{aligned} E[X(t,\theta)|X(s,\theta)] &= E[X(t,\theta)|B(s),\sigma^2(t-s)] \\ &= e^{\theta B(s) + \frac{1}{2}\theta^2\sigma^2(t-s) - \frac{1}{2}\theta^2\sigma^2 t} \\ &= e^{\theta B(s) - \frac{1}{2}\theta^2\sigma^2 s} \\ &= X(s,\theta) \end{aligned}$$

△

The above martingale has the interesting property that it can be used to *generate* more martingales by differentiating $X(t,\theta)$ with respect to θ and setting $\theta = 0$. In other words, there is a whole series of martingales which are related to Brownian motion by using $X(t,\theta) = e^{\theta B(t) - \frac{1}{2}\theta^2\sigma^2 t}$ as a generating function.

Suppose we are observing a 1-dimensional Brownian motion, $B(t)$ that begins at $B(0) = 0$. Given two real numbers a and b we ask what is the probability, p, that the Brownian motion hits b before it hits $-a$? (b is to the right of 0 and $-a$ is to the left). Let T be the time to hit b or $-a$ whichever is first. T is an example of a *stopping time*. One needs to show that T is finite – in other words, it must be shown that the Brownian motion process does not stay between $-a$ and b forever. We assume that this is true and that the Brownian motion eventually hits $-a$ or b. The probability p is given by

$$p = P[\text{hit } +b \text{ before } -a] = P[B(T) = b] \tag{50}$$

Solution: $B(t)$ is a martingale. Therefore $E[B(t)|B(s)] = B(s)$ in general and for $t = T$ we have

$$E[B(T)|B(0)] = B(0) = 0$$

Now since $B(T) = b$ with probability p, then $B(T) = -a$ with probability $1 - p$ and so

$$E[B(T)|B(0)] = bp + (-a)(1-p) = B(0) = 0 \tag{51}$$

or

$$\begin{aligned} 0 &= bp - a + ap \\ a &= p(a+b) \\ p &= \frac{a}{a+b} \end{aligned}$$

Therefore the probability that the Brownian motion will first hit b is $a/(a+b)$.

Let us further seek to find the expected time $E[T]$ when the Brownian motion will first encounter either b or $-a$. Now Brownian motion satisfies $E[B^2(t)] = \sigma^2 t$ and therefore $E[B^2(t)|t = T] = \sigma^2 T$ and so

$$E\big[E[B^2(t)|t = T]\big] = \sigma^2 E[T] \tag{52}$$

Now, based on the probabilities we found above, we have

$$B^2(T) = \begin{cases} b^2 & \text{probability} \quad p \\ (-a)^2 & \text{probability} \quad 1-p \end{cases} \tag{53}$$

Putting everything together we find

$$\begin{aligned} \sigma^2 E[T] &= E[B^2(T)] \\ &= b^2 p + a^2(1-p) \quad \text{and we have that } p = a/(a+b) \\ &= b^2 \frac{a}{a+b} + a^2 \frac{b}{a+b} \\ &= ab\frac{a+b}{a+b} \\ &= ab \end{aligned}$$

In summary, the probability that B(t) hits b before $-a$ is $p = a/(a+b)$ and the expected time to hit is $E[T] = ab/\sigma^2$.

13.2 Sub-and-super martingales

A *submartingale* is a stochastic process that satisfies the following inequality

$$E[X_k|X_1, \ldots, X_{k-1}] \geq X_{k-1} \tag{54}$$

A *supermartingale* is a stochastic process that satisfies the following inequality

$$E[X_k|X_1, \ldots, X_{k-1}] \leq X_{k-1} \tag{55}$$

Example Consider again the squares of sums S_k^2 above.

$$\begin{aligned} E[S_k^2|X_1, \ldots, X_{k-1}] &= E[(X_k + S_{k-1})^2|X_1, \ldots, X_{k-1}] \\ &= E[X_k^2|X_1, \ldots, X_{k-1}] + E[S_{k-1}^2|X_1, \ldots, X_{k-1}] + 2E[X_k]E[S_{k-1}|X_1, \ldots, X_{k-1}] \\ &\geq E[S_{k-1}^2|X_1, \ldots, X_{k-1}] = S_{k-1}^2 \end{aligned}$$

because $E[X_k] = 0$.Therefore S_k^2 is a submartingale because $E[S_k^2|X_1, \ldots, X_{k-1}] \geq S_{k-1}^2$. $\triangle$

14 Appendices

14.1 Abbreviations and Notation

$\triangleq$ is defined as
IID Independent Identically Distributed
SII Stationary Independent Increments
iff if and only if
a.s. almost surely
i.o. infinitely often
m.s. mean square (same as quadratic mean)
o.n. orthonormal
q.m. quadratic mean (same as mean square)
wp1 with probability 1
w.r.t with respect to
$\{x_n\}$ is a list or sequence of numbers $x_n, n = 1, 2, \ldots$
$x_n \to x$ means $\lim_{n\to\infty} x_n = x$
$\uparrow$ is used to denote monotonically increasing sequence $\{x_n\}$, which converges from below:

$$x_n \uparrow x \implies x_1 < x_2 < \ldots \text{ and } \lim_{n\to\infty} x_n = x.$$

For sets $A_n \uparrow A$ means $A_1 \subset A_2 \subset \ldots$ and $A = \bigcup_{n=1}^{\infty} A_n$.

$\downarrow$ is used to denote a monotonically decreasing series which converges from above:

$$x_n \downarrow x \implies x_1 > x_2 > \ldots \text{ and } \lim_{n\to\infty} x_n = x.$$

For sets $A_n \downarrow A$ means $A_1 \supset A_2 \supset \ldots$ and $A = \bigcap_{n=1}^{\infty} A_n$.

14.2 Equivalence Relations

An important technique for dividing a set into mutually exclusive subsets is the method of equivalence relations [2].

Definition Suppose in a set A we have a relation R defined between pairs of elements x and y, xRy indicating that x and y stand in the relation R. Suppose that R has the following three properties:

(1) R is reflexive: i.e., xRx for all $x \in A$.
(2) R is symmetric: i.e., $xRy \implies yRx$.
(3) R is transitive: i.e., xRy and $yRz \implies xRz$. Then R is an *Equivalence Relation.*

Theorem 14.1 *An equivalence relation R divides A into mutually exclusive sets so that every element of A is in one and only one subset, and so that two elements are in the same subset iff they stand in the the relation R to one another.*

Proof Given any element x in A consider all the elements y such that xRy. These elements form a subset of A which we call A_x. Now two elements are in the same subset A_x, iff they stand in the relation R with one another. For suppose yRz and $y \in A_x$. Then xRy and, because R is transitive, xRz. Then $z \in A_x$. Also, if y and z are both in A_x, then xRy and xRz. By the symmetric property yRx. Together with xRz we have yRx and xRz and therefore yRz by transitivity.

For each element $x \in A$ we have a subset A_x. These subsets will not all be distinct. We next show that any two such subsets are either mutually exclusive of identical. Suppose both A_x and A_y both have an element z in common. Then xRz and yRz,. By the symmetric property we have zRy and therefore xRy by transitivity. Now pick any element w of A_x, xRw. Because xRy we also have yRx and therefore yRw also by transitivity. Therefore $w \in A_y$, or $A_x \subset A_y$. By the same reasoning, exchanging A_x and A_y and going through the same line of reasoning, we must also have $A_y \subset A_x$. So A_x and A_y are identical. Therefore we have a list of mutually exclusive subsets $A_{x_1}, A_{x_2}, \ldots$. Also, for any element $z \in A$ we have zRz by the reflexive property, z must be in one of the subsets, i.e., A_z. ∎

14.3 Various Modes of Convergence

Given random variables $Z, Z_1, Z_2 \ldots$ we say that

i) the sequence $\{Z_n\}$ converges to Z *surely* if, given any $\epsilon > 0$ there exists an integer $N > 0$ such that

$$|Z_n - Z| < \epsilon \quad \text{for all } n \geq N$$

This mode of convergence is also denoted by $Z_n \to Z$.

ii) the sequence $\{Z_n\}$ converges to Z *with probability one* (wp1) iff for every $\epsilon > 0$

$$\lim_{N \to \infty} P\Big[\Big(\max_{n \geq N} |Z_n - Z|\Big) > \epsilon\Big] = 0.$$

Such a sequence is said to converge *almost surely* (a.s.) which is also denoted by

$$P[\lim_{n \to \infty} Z_n = Z] = 1.$$

iii) the sequence $\{Z_n\}$ converges to Z *in probability* if for every $\epsilon > 0$

$$\lim_{n \to \infty} P\big[|Z_n - Z| > \epsilon\big] = 0,$$

iv) the sequence $\{Z_n\}$ converges to Z in *mean square* (m.s.) or in *quadratic mean* (q.m.), if each random variable Z_n has a finite mean square and if

$$\lim_{n\to\infty} E\big[(Z_n - Z)^2\big] = 0.$$

v) a sequence of random variables can also converge *in distribution* if their probability distribution functions converge to a limiting function

$$\lim_{n\to\infty} f_{Z_n}(z) = f_Z(z)$$

everywhere except on a set of measure zero.

vi) In addition to the above convergence types a sequence $\{R(v), v = 0, 1, \ldots\}$ is said to converge to 0 in Cesàro mean as $v \to \infty$ if

$$\lim_{t\to\infty} \frac{1}{t} \sum_{v=0}^{t-1} R(v) = 0.$$

14.4 DeMorgan's Laws

The following useful relationship between the three basic operations of forming unions, intersections, and complements of sets is known as *DeMorgan's laws*

$$\Big(\bigcup_{i=1}^{n} A_i\Big)^c = \bigcap_{i=1}^{n} A_i^c$$
$$\Big(\bigcap_{i=1}^{n} A_i\Big)^c = \bigcup_{i=1}^{n} A_i^c$$

This is a standard result in set theory. The usual method of proof is the following: Suppose that $x \in (\bigcup_{i=1}^{n} A_i)^c$ then x is not contained in $\bigcup_{i=1}^{n} A_i$, which means it is not contained in any of the events $A_i, i = 1, 2, \ldots n$, implying that x is contained in A_i^c for all $i = 1, 2, \ldots n$ and therefore contained in $\bigcap_{i=1}^{n} A_i^c$. Going in the other direction, suppose $x \in \bigcap_{i=1}^{n} A_i^c$. Then x is contained in A_i^c for all $i = 1, 2, \ldots n$, which means that x is not contained in A_i for any $i = 1, 2, \ldots n$, implying that x is not contained in $\bigcup_{i=1}^{n} A_i$, which means $x \in (\bigcup_{i=1}^{n} A_i)^c$.

To prove the second DeMorgan's laws, apply the first to A_i^c and use the fact that $A_i = (A_i^c)^c$ for any set. We obtain

$$\begin{aligned}\Big(\bigcup_{i=1}^{n} A_i^c\Big)^c &= \bigcap_{i=1}^{n} (A_i^c)^c \\ &= \bigcap_{i=1}^{n} A_i\end{aligned}$$

By taking the complement of both sides we find the second of DeMorgan's laws:

$$\bigcup_{i=1}^{n} A_i^c = \Big(\bigcap_{i=1}^{n} A_i\Big)^c$$

14.5 Fourier Theorems

For n-vectors x and k we define the dot product of these vectors by

$$< x \cdot k >= \sum_{i=1}^{n} x_i k_i$$

Here x is an element of $\mathbb{R}^n$ and k is in the "Fourier transform space" or the "momentum space" corresponding to $\mathbb{R}^n$ which is also n-dimensional, but is a different space (i.e., the k is an element of the dual-space of $\mathbb{R}^n$). Let $f(x)$ be a subset of integrable and square-integrable functions on $\mathbb{R}^n$, $f(x) \subset L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$. The Fourier transform of $f(x)$ is denoted by $\hat{f}(k)$ and is given in symmetric form by

$$\hat{f}(k) = \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} f(x) e^{-i<x\cdot k>}\, d^n k \tag{56}$$

The inverse Fourier transform is given also in symmetric form by

$$f(x) = \frac{1}{\sqrt{(2\pi)^n}} \int_{\mathbb{R}^n} \hat{f}(k) e^{i<x\cdot k>}\, d^n x \tag{57}$$

One of the main theorems of Fourier Analysis is the Plancherel-Parseval theorem which states that the Fourier transform map is an *Isometry* with respect to the L^2 norm.

Theorem 14.2 *Plancherel-Parseval theorem:*

$$\int_{-\infty}^{\infty} f(x)\overline{g(x)}\, d^n x = \int_{-\infty}^{\infty} \hat{f}(k)\overline{\hat{g}(k)}\, d^n k, \tag{58}$$

where $\overline{f}$ is the complex-conjugate of f

As can be seen above the functions for which the Fourier integral is applied are square-integrable functions (the are in the function space L^2). The norm in this space is defined by

$$\|f\| = \int_{-\infty}^{\infty} f(x)\overline{f(x)}\, d^n x \tag{59}$$

and we assume that $\|f\| < \infty$ implicitly in using Fourier integrals. Fourier series deals with periodic functions. However, no meaning can be attached to the integral

$$f_X(\omega) = \int_{-\infty}^{\infty} e^{-it\omega} X(t) dt \tag{60}$$

for many stochastic processes, such as covariance stationary processes, since their sample functions are non-periodic and are undamped (do not fall off at infinity in such a manner as to be in the space of L^2 functions). It is possible to define a generalized harmonic analysis of stochastic processes (a method of assigning to each frequency ω a measure of its contribution to the "content" of the process) as shown in [16].

14.6 Cauchy-Schwarz Inequality

Here we include the proof of the Cauchy-Schwarz inequality for sequences and, by extension, to integrals:

(Cauchy-Schwarz inequality). Let $x = (x_1, x_2, ..., x_N)$ and $y = (y_1, y_2, ..., y_N)$ be elements of $\mathbb{R}^N$. Then

$$\left| \sum_{i=1}^{N} x_i y_i \right|^2 \leq \sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2$$

This inequality has many different proofs. We present the following proof. If every x_i is zero, then the equality sign in the above expression holds. So let us assume that there is at least one x_i not zero. We form the function

$$f(\lambda) = \sum_{i=1}^{N} (y_i - \lambda x_i)^2 \geq 0$$

which is nonnegative for all real values of λ. We set

$$A = \sum_{i=1}^{N} x_i^2 \quad B = \sum_{i=1}^{N} x_i y_i \quad C = \sum_{i=1}^{N} y_i^2$$

Then we can expand $f(\lambda)$ as

$$f(\lambda) = A\lambda^2 - 2B\lambda + C \geq 0, \quad \text{for} \quad A > 0.$$

Since the only real root for $f(\lambda)$ occurs when $y_i = \lambda x_i$, then unless this is the case $f(\lambda)$ can only have complex roots. Examining the solutions to the quadratic equation we have

$$\lambda = \frac{2B \pm \sqrt{4B^2 - 4AC}}{2A}$$

or

$$\lambda = \frac{B \pm \sqrt{B^2 - AC}}{A}.$$

$f(\lambda)$ has complex roots if $B^2 < AC$ or

$$\left| \sum_{i=1}^{N} x_i y_i \right|^2 \leq \sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2$$

Let us define the norm of an n-vector.

Definition The vector dot-product of two n vectors u and v is defined as

$$u \cdot v = \sum_{i=1}^{n} u_i v_i.$$

With the above definition $u \cdot u$ is the square of the magnitude of u. The magnitude of u is also known as the norm of u and is denoted by $\|u\|$.

Definition The norm of an n-vector u is given in terms of the vector dot-product by

$$\|u\| = \sqrt{u \cdot u}$$

If u is complex then let $\overline{u}$ stand for the complex-conjugate and define the norm of u as

$$\|u\|^2 = \sum_{i=1}^{n} u_i \overline{u_i} = u \cdot \overline{u}$$

A useful shorthand for $u \cdot u = \|u\|^2$ is u^2, where the squared vector is obtained by dotting the vector u with itself.

With the above definition of norm we can also write the Cauchy-Schwarz inequality as

$$\|x \cdot y\| \leq \|x\| \cdot \|y\|$$

The Cauchy-Schwarz inequality can be used to prove the Triangle Inequality for n-vectors.

$$\|u + v\| \leq \|u\| + \|v\|$$

Proof Take u and v to be n-vectors. Then the squared norm of $u + v$ is

$$(u + v)^2 = u^2 + 2u \cdot v + v^2$$

and by the Cauchy-Schwarz in equality $\|u \cdot v\| \leq \|u\|\|v\|$. Therefore

$$(u + v)^2 \leq u^2 + 2\|u\|\|v\| + v^2 = (\|u\| + \|v\|)^2$$

and finally,

$$\|u + v\| \leq \|u\| + \|v\|$$

The Cauchy-Schwarz inequality can be extended to the inner product space of square-integrable complex-valued functions:

$$\left| \int_{\mathbb{R}^N} f(x)\overline{g(x)} \right|^2 \leq \int_{\mathbb{R}^N} |f(x)|^2 \, dx \cdot \int_{\mathbb{R}^N} |g(x)|^2 dx$$

where $\overline{g(x)}$ is the complex conjugate of $g(x)$ and $|f(x)|^2 = f(x)\overline{f(x)}$.

We can also write the above inequality using the concept of expectation for random variables.

Cauchy-Schwarz Inequality for random variables. For any two random variables X and Y we can write the Cauchy-Schwarz inequality as

$$(E[XY])^2 \leq E[X^2]E[Y^2],$$

where equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in R$.

Theorem 14.3 *If $E[X^2] < \infty$, then $E[|X|] < \infty$*

Proof Taking the square root in the Cauchy-Schwarz inequality and setting $y \equiv 1$, we have

$$E[|X|] < (E[|X^2|])^{\frac{1}{2}} < \infty. \quad \blacksquare$$

It can also be shown (we skip the proof) that

Theorem 14.4 *If $E[|X|^n] < \infty$, then $E[|X|^m] < \infty$ for all $1 \leq m \leq n$.*

Hence if the n-th moment of X is finite, then all the lower moments of X will also be finite.

14.7 Generating Functions

The Moment Generating Function (m.g.f) $\psi_X(t)$ of the random variable X is defined for all real values of t by

$$\begin{aligned} \psi_X(t) &= E[e^{tX}] \\ &= \begin{cases} \sum_x e^{tx} p(x) & \text{if } X \text{ is discrete with mass function } p(x) \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & \text{if } X \text{ is continuous with density } f_X(x) \end{cases} \end{aligned}$$

$\psi_X(t)$ is called the m.g.f. because all the moments of X can be obtained by differentiating $\psi_X(t)$ a sufficient number of times and setting $t = 0$ afterwards. For example

$$\begin{aligned} \psi'_X(t) &= \frac{d}{dt} E[e^{tX}] \\ &= E[\frac{d}{dt}(e^{tX})] \\ &= E[Xe^{tX}] \end{aligned}$$

Therefore $E[X] = \psi'(0)$. Similarly all the higher moments can be generated by successive differentiation.

Consider the m.g.f. for the standard Gaussian random variable $Z \in N(0,1)$:

$$\begin{aligned}
\psi_Z(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx}\, e^{-x^2/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x^2-2tx)/2}\, dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}+\frac{t^2}{2}}\, dx \\
&= e^{t^2/2}\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}\, dx \\
&= e^{t^2/2}\, \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\, dy \quad \text{substitution: } y = x - t \\
&= e^{t^2/2}
\end{aligned}$$

Given that Z is $N(0,1)$ we know that $X = \mu + \sigma Z$ will be normally distributed according to $N(\mu, \sigma^2)$. Therefore the m.g.f. for X can be obtained using the above as follows:

$$\begin{aligned}
\psi_X(t) &= E[e^{tX}] \\
&= E[e^{t(\mu+\sigma Z)}] \\
&= e^{\mu t} E[e^{t\sigma Z}] \\
&= e^{\mu t} \psi_Z(\sigma t) \\
&= e^{\mu t} e^{(\sigma t)^2/2} \\
&= \exp\left\{\sigma^2 t^2/2 + \mu t\right\}
\end{aligned}$$

Closely related to the m.g.f. is the Characteristic Function $\phi(u)$ of X which is defined, for any real number u, by

$$\phi_X(u) = E[e^{iuX}]. \tag{61}$$

It is possible that a random variable does not have a finite mean or variance or m.g.f. But a random variable always possesses a characteristic function.

Theorem 14.5 *Consider a set of jointly normally distributed random variables $X_1, \ldots, X_n$ with means m_j and covariances K_{jk} given. The m.g.f. for the set is given by*

$$\psi_{X_1,\ldots,X_n}(u_1, \ldots, u_n) = \exp\Big\{ \sum_{j=1}^{n} u_j m_j + \frac{1}{2} \sum_{j,k=1}^{n} u_j K_{jk} u_k \Big\} \tag{62}$$

Proof By definition the m.g.f for the jointly normal random variables is given by

$$\begin{aligned} E[e^{u_1X_1+\ldots+u_nX_n}] &= \int f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, e^{u_1x_1+\ldots+u_nx_n}\, dx_1\ldots dx_n \text{ in matrix notation:} \\ &= \int \frac{1}{(2\pi)^{n/2}}\frac{1}{|K|^{1/2}} \exp\Big\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})+\mathbf{u}^T\cdot\mathbf{x}\Big\}\, dx_1\ldots dx_n. \end{aligned}$$

Let us examine the exponent in the above equation. Using the change of variables $\mathbf{y}=\mathbf{x}-\mathbf{m}$ we have

$$\begin{aligned} -\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})+\mathbf{u}^T\cdot\mathbf{x} &= -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}+\mathbf{u}^T\cdot(\mathbf{y}+\mathbf{m}) \\ &= \Big\{-\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}+\mathbf{u}^T\cdot\mathbf{y}\Big\}+\mathbf{u}^{\mathbf{T}}\cdot\mathbf{m} \end{aligned}$$

The last term on the RHS is independent of the variable of integration and can be moved outside the integral. The method of completing the square can be used to search for a $\mathbf{z}$ and λ such that

$$\Big\{-\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}+\mathbf{u}^T\cdot\mathbf{y}\Big\} = -\frac{1}{2}(\mathbf{y}-\mathbf{z})^T\mathbf{K}^{-1}(\mathbf{y}-\mathbf{z})+\lambda \tag{63}$$

Now we assume that $\mathbf{K}^{-1}$ (and also the inverse matrix $\mathbf{K}$) is symmetric and expand the RHS of Eq. [63] to obtain:

$$\begin{aligned} -\frac{1}{2}(\mathbf{y}-\mathbf{z})^T\mathbf{K}^{-1}(\mathbf{y}-\mathbf{z})+\lambda &= -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}-\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{z}-\frac{1}{2}\mathbf{z}^T\mathbf{K}^{-1}\mathbf{y}-\frac{1}{2}\mathbf{z}^T\mathbf{K}^{-1}\mathbf{z}+\lambda \\ &= -\frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\mathbf{y}-\mathbf{z}^T\mathbf{K}^{-1}\mathbf{y}-\frac{1}{2}\mathbf{z}^T\mathbf{K}^{-1}\mathbf{z}+\lambda \text{ by symmetry of } \mathbf{K} \end{aligned}$$

By comparing the LHS of Eq. [63] to the RHS of the above we find that

$$\begin{aligned} \mathbf{z}^T &= \mathbf{u}^{\mathbf{T}}\mathbf{K} \\ \lambda &= \frac{1}{2}\mathbf{z}^T\mathbf{K}^{-1}\mathbf{z}=\frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{K}^{-1}\mathbf{K}^T\mathbf{u}=\frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u} \text{ also by symmetry of } \mathbf{K} \end{aligned}$$

Inserting these results into the expression for $E[e^{u_1X_1+\ldots+u_nX_n}]$, we find

$$\begin{aligned} E[e^{u_1X_1+\ldots+u_nX_n}] &= \int f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\, e^{u_1x_1+\ldots+u_nx_n}\, dx_1\ldots dx_n \text{ in matrix notation:} \\ &= \int \frac{1}{(2\pi)^{n/2}}\frac{1}{|K|^{1/2}} \exp\Big\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})+\mathbf{u}^T\cdot\mathbf{x}\Big\}\, dx_1\ldots dx_n. \\ &= \exp\{\mathbf{u}^T\cdot\mathbf{m}+\frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u}\}\frac{1}{(2\pi)^{n/2}}\frac{1}{|K|^{1/2}}\int \exp\Big\{-\frac{1}{2}(\mathbf{y}-\mathbf{z})^T\mathbf{K}^{-1}(\mathbf{y}-\mathbf{z})\Big\}\, dy_1\ldots dy_n. \\ &= \exp\{\mathbf{u}^T\cdot\mathbf{m}+\frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u}\}, \end{aligned}$$

where

$$1 = \frac{1}{(2\pi)^{n/2}} \frac{1}{|K|^{1/2}} \int \exp\{ -\frac{1}{2}(\mathbf{y}-\mathbf{z})^T \mathbf{K}^{-1}(\mathbf{y}-\mathbf{z})\}\, dy_1 \ldots dy_n$$

for any given vector $\mathbf{z}$. Putting all the above together we find

$$\begin{aligned} \psi_{X_1,\ldots,X_n}(u_1,\ldots,u_n) &= E[e^{u_1X_1+\ldots+u_nX_n}] \\ &= \exp\{\mathbf{u}^T \cdot \mathbf{m} + \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u}\} \\ &= \exp\{ \sum_{j=1}^{n} u_j m_j + \frac{1}{2} \sum_{j,k=1}^{n} u_j K_{jk} u_k\} \qquad \blacksquare \end{aligned}$$

Note that the *inverse of the covariance matrix*, $\mathbf{K}^{-1}$, appears explicitly in the joint probability function

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|K|^{1/2}} \exp\{ -\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\}$$

whereas the covariance matrix $\mathbf{K}$ appears in the m.g.f.

$$\psi_{X_1,\ldots,X_n}(u_1,\ldots,u_n) = \exp\{\mathbf{u}^T\mathbf{m} + \frac{1}{2}\mathbf{u}^T\mathbf{K}\mathbf{u}\}.$$

14.8 Probabilistic Inequalities

Proposition 14.6 *Markov's inequality: If X is a random variable that takes only non-negative values, then for any value $a > 0$*

$$P\{x \geq a\} \leq \frac{E[X]}{a}$$

Proof We give a proof for the case where X is continuous with density $f(x)$.

$$\begin{aligned} E[X] &= \int_0^\infty x f(x)\, dx \\ &= \int_0^a x f(x)\, dx + \int_a^\infty x f(x)\, dx \\ &\geq \int_a^\infty x f(x)\, dx \\ &\geq \int_a^\infty a f(x)\, dx \\ &= a \int_a^\infty f(x)\, dx \\ &= aP\{X \geq a\} \qquad \blacksquare \end{aligned}$$

Proposition 14.7 *Chebyshev's inequality: If X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$*

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

Proof Since $(X - \mu)^2$ is a non-negative random variable, we can apply Markov's in equality (with $a = k^2$) to obtain

$$P\big\{(X - \mu)^2 \geq k^2\big\} \leq \frac{E[(X - \mu)^2]}{k^2}$$

But since $(X - \mu)^2 \geq k^2$ if and only if $|X - \mu| \geq k$, the above equation is equivalent to

$$P\big\{|X - \mu| \geq k\big\} \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2} \qquad \blacksquare$$

Proposition 14.8 *Kolmogorov's inequality: Let $X_1, X_2, \ldots X_n$ be n independent random variables such that $E[X_i] = 0$ and $Var[X_i] = \sigma_i^2 < \infty, i = 1, 2, \ldots n$. Then, for all $a > 0$,*

$$P\Big\{\max_{i=1,\ldots n} |X_1 + \ldots X_i| > a\Big\} \leq \sum_{i=1}^{n} \frac{\sigma_i^2}{a^2}$$

Proof Consider all the sums $S_l = \sum_{i=1}^{l} X_i, 1 \leq l \leq n$. Then $E[S_l] = 0$ because $E[X_i] = 0$. Define an event T as follows

$$T = \begin{cases} l & \text{if } l \text{ is the first time } |S_l| > a, 1 \leq l \leq n \\ \infty & \text{if } |S_l| \leq a, \text{ for all } 1 \leq l \leq n \end{cases}$$

Notice that the sets of events $\{T = i\}, i = 1, 2, \ldots n$ are mutually exclusive since $\{T = l\} \cap \{T = k\} = \emptyset$ for $1 \leq l \neq k \leq n$. Also, note that $\{T < \infty\} = \bigcup_{l=1}^{n} \{T = l\}$.

The variance of S_n can be written as

$$\begin{aligned} E[S_n^2] &= E[(S_k + (S_n - S_k))^2] \\ &\geq E[S_n^2; T < \infty] = \sum_{k=1}^{n} E[S_n^2; T = k] \\ &= E\Big[S_k^2 + (S_n - S_k)^2 + 2S_k(S_n - S_k); T = k\Big] \end{aligned}$$

$$E[S_n^2] \geq \sum_{k=1}^{n} E[S_k^2; T = k] + \sum_{k=1}^{n} E[(S_n - S_k)^2); T = k] + 2\sum_{k=1}^{n} E[(S_n - S_k)(S_k; T = k)]$$

Now

$$\sum_{k=1}^{n} E[(S_n - S_k)^2; T = k] \geq 0$$

and

$$\sum_{k=1}^{n} E[(S_n - S_k)(S_k; T = k)] = \sum_{k=1}^{n} E[(S_n - S_k)]E[(S_k; T = k)] = 0$$

by independence because $E[(S_n - S_k)] = 0$. Therefore we have found

$$E[S_n^2] \geq \sum_{k=1}^{n} E[S_k^2; T = k] \geq \sum_{k=1}^{n} (a^2) P\{T = k\}$$

or

$$E[S_n^2] \geq a^2 \sum_{k=1}^{n} P\{T = k\} = a^2 P\{T < \infty\}$$

which is the same as

$$P\{T < \infty\} \leq \frac{E[S_n^2]}{a^2} = \sum_{i=1}^{n} \frac{\sigma_i^2}{a^2} \qquad \blacksquare$$

Kolmogorov's inequality can be regarded as a generalization of Chebyshev's inequality. However, Kolmogorov's inequality is much stronger than Chebyshev's inequality. Take $X_1, \ldots, X_2$ to be independent zero-mean random variables with variance $\mathrm{Var}[X_i] = \sigma_i^2$, then Chebyshev's inequality gives

$$P\{|X_1 + \ldots + X_n| > a\} \leq \sum_{i=1}^{n} \frac{\sigma_i^2}{a^2}$$

whereas Kolmogorov's inequality gives the same bound for the probability of a larger set, namely

$$\bigcup_{i=1}^{n} \{|X_1 + \ldots + X_i| > a\}$$

14.9 Kronecker's Lemma

Proposition 14.9 *(Kronecker's Lemma) If $a_1, a_2, \ldots$ are real numbers such that $\sum_{i=1}^{\infty} a_i/i < \infty$, then*

$$\lim_{n \to \infty} \sum_{i=1}^{n} \frac{a_i}{n} = 0.$$

14.10 Strong Law of Large Numbers for Independent Random Variables

Theorem 14.10 *Let $X_1, X_2, \ldots$ be independent zero-mean random variables with variance $Var[X_i] = \sigma_i^2 < \infty$. If $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$, then with probability 1,*

$$\frac{X_1 + \ldots + X_n}{n} \to 0, \quad \text{as } n \to \infty$$

Proof For any n and any $a > 0$, it follows from Kolmogorov's inequality that

$$P\Big\{\max_{j=1,\ldots n}\left|\sum_{i=1}^{j}\frac{X_i}{i}\right| > a\Big\} \leq \frac{\sum_{i=1}^{n}\mathrm{Var}[X_i/i]}{a^2} \leq \frac{\sum_{i=1}^{\infty}\sigma_i^2/i^2}{a^2} \tag{64}$$

Define

$$E_n = \Big\{\max_{j=1,\ldots n}\left|\sum_{i=1}^{j}\frac{X_i}{i}\right| > a\Big\}$$

Then, because the E_n is an increasing sequence of events, it follows from the continuity property of probability that

$$\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n) = P(\bigcup_{1}^{\infty} E_n) = P\Big\{\max_{j\geq 1}\left|\sum_{i=1}^{j}\frac{X_i}{i}\right| > a\Big\}$$

Hence, Eq. [64] gives us

$$P\Big\{\max_{j\geq 1}\left|\sum_{i=1}^{j}\frac{X_i}{i}\right| > a\Big\} \leq \frac{\sum_{i=1}^{\infty}\sigma_i^2/i^2}{a^2} \tag{65}$$

or

$$P\Big\{\max_{j\geq 1}\left|\sum_{i=1}^{j}\frac{X_i}{i}\right| \leq a\Big\} \geq 1 - \frac{\sum_{i=1}^{\infty}\sigma_i^2/i^2}{a^2} \tag{66}$$

Because $\max_{j\geq 1}\left|\sum_i = 1^j X_i/i\right| \leq a$ implies that $\sum_{i=1}^{\infty} X_i/i \leq a \leq infty$ we find

$$P\Big\{\sum_{i=1}^{\infty}\frac{X_i}{i} < \infty\Big\} \geq 1 - \frac{\sum_{i=1}^{\infty}\sigma_i^2/i^2}{a^2} \tag{67}$$

By taking $a \to \infty$ we find

$$P\Big\{\sum_{i=1}^{\infty}\frac{X_i}{i} < \infty\Big\} = 1 \tag{68}$$

and by Kronecker's lemma

$$\lim_{n\to\infty} P\Big\{\sum_{i=1}^{n}\frac{X_i}{i} = 0\Big\} = 1 \qquad \blacksquare \tag{69}$$

If the random variables are assumed not only to be independent but also identically distributed with mean μ and finite variance σ^2, then as $\sum i = 1^{\infty}\sigma^2/i^2 < \infty$ by assumption, it also follows that wp1,

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{(X_i - \mu)}{n} = 0$$

of that

$$\lim_{n\to\infty}\sum_{i=1}^{n}\frac{X_i}{n} = \mu.$$

14.11 Central Limit Theorem

We will follow the method in Ross[4].

Theorem 14.11 *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. The distribution of*

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. For $-\infty < a < \infty$,

$$P\left\{\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-x^2/2\right)\, dx.$$

We will use the following lemma, which is stated without proof:

Lemma 14.12 *Let $Z_1, Z_2, \ldots$ be a sequence of random variables having distribution functions F_{Z_n} and m.f.g $M_{Z_n}(t), n \geq 1$, and let Z be a random variable having distribution function F_Z and m.f.g $M_Z(t)$. If $M_{Z_n}(t) \to M_Z(t)$ for all t, then $F_{Z_n}(z) \to F_Z(z)$ for all z at which $F_Z(z)$ is continuous.*

In the case where $F_Z(z)$ is the standard normal random variable, then $M_Z(t) = e^{t^2/2}$. It follows from Lemma 14.12 that if $M_{Z_n}(t) \to e^{t^2/2}$ as $n \to \infty$, then $F_{Z_n}(z) \to \Phi(z)$ as $n \to \infty$ ($\Phi(z)$ is the CDF for the standard normal distribution defined in Eq. [4]).

Proof (Proof of Theorem 14.11) Let us first assume that $\mu = 0$ and $\sigma = 1$. We also assume that the m.g.f. of the $X_i, M(t)$, exists and is finite. The m.f.g for $X_i/\sqrt{n}$ is given by

$$E\left[\exp\left\{\frac{tX_i}{\sqrt{n}}\right\}\right] = M\Big(\frac{t}{\sqrt{n}}\Big).$$

Then the m.g.f. of $\displaystyle\sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}$ is given by $\left[M\Big(\frac{t}{\sqrt{n}}\Big)\right]^n$. Now let $L(t) = \log M(t)$ and notice that $M(0) = 1$ and

$$\begin{aligned}
L(0) &= 0 \\
L'(0) &= \frac{M'(0)}{M(0)} \\
&= \mu = 0 \\
L''(0) &= \frac{M(0)M''(0) - [M'(0)]^2}{[M(0)]^2} \\
&= E[X^2] = 1
\end{aligned}$$

[4]Sheldon Ross, "A First Course in Probability", 9th Edition, Pearson Education, Inc., pp. 370-371 (2014).

In order to prove Theorem 14.11 we must show that $[M(t/sqrtn) \to e^{t^2/2}$ as $n \to \infty$, or that $nL(t/\sqrt{n}) \to t^2/2$ as $n \to \infty$. In order to show this, notice that

$$\begin{aligned}
\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}} \quad \text{by L'Hôpital's rule} \\
&= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})t}{-2n^{-1/2}} \\
&= \lim_{n\to\infty} \frac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}} \quad \text{also by L'Hôpital's rule} \\
&= \lim_{n\to\infty} \frac{L''(\frac{t}{\sqrt{n}})t^2}{2} \\
&= \frac{t^2}{2}
\end{aligned}$$

Therefore the Central Limit Theorem is proven in the case $\mu = 0$ and $\sigma = 1$. The result can be extended to the more general case of independent identically distributed random variables with non-zero mean and arbitrary variance by transforming X_i to the standard normal variable $X_i^* = (X_i - \mu)/\sigma$ and applying the result obtained above. Since $E[X_i^*] = 0, \text{Var}[X_i^*] = 1$ we have directly

$$\frac{X_1^* + X_2^* + \ldots + X_n^*}{\sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \tag{70}$$

Since the theorem was proven for the expression on the LHS of Eq. [70], it will also hold for the RHS since these expressions are equivalent. ∎

14.12 Borel-Cantelli Lemma

Lemma 14.13 *Borel-Cantelli*

Let $A_1, A_2, \ldots$ be random events in a probability space Ω.

1) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n, i.o.) = 0$;

2) If $A_1, A_2, \ldots$ are independent, and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(A_n, i.o.) = 1$

Proof (Borel-Cantelli 1) Let B_k be the event $\bigcup_{i=k}^{\infty} A_i$ for $k = 1, 2, \ldots$, If x is in the event $\{A_i, \text{i.o.}\}$, then $x \in B_k$ for all k, therefore $x \in \bigcup_{k=1}^{\infty} B_k$.

Conversely, if $x \in B_k$ for all k, then we can show that x is in $\{A_i, \text{i.o.}\}$. This is true because $x \in B_1 = \bigcup_{i=1}^{\infty} A_i$ means that $x \in A_{j_1}$ for some j_1. However, $x \in B_{j_1+1}$ implies that $x \in A_{j_1}$ for some j_2 that

is strictly larger than j_1. Therefore we can produce an infinite sequence of integers $j_1 < j_2 < j_3 < \ldots$ such that $x \in A_{j_i}$ for all i. Let E be the event $\{x : x \in A_i, \text{i.o.}\}$. By the definition of $\limsup_{n\to\infty} A_n$ we have

$$E = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i$$

From $E \subset B_k$ for all k, it follows that $P(E) \leq P(B_k)$ for all k. By the property of union bound, we have that $P(B_k) \leq \sum_{i=k}^{\infty} P(A_i)$. By hypothesis since $\sum_{i=1}^{\infty} P(A_i)$ is finite and hence $P(B_k) \to 0$ as $k \to \infty$. Therefore $P(E) = 0$. ∎

Proof (Borel-Cantelli 2) Let E denote the set of samples that are in A_i infinitely often. We have to show that complement of E (denoted by E^c) has probability zero.

Taking the complement of E we find using the definition in the proof of Borel-Cantelli 1) and DeMorgan's laws

$$E^c = \bigcup_{k=1}^{\infty} \bigcap_{i=k}^{\infty} A_i^c$$

But for each k, assuming that the A_i's are independent,

$$\begin{aligned} P(\bigcup_{i=k}^{\infty} A_i^c) &= \prod_{i=k}^{\infty} P(A_i^c) \\ &= \prod_{i=k}^{\infty} (1 - P(A_i)) \end{aligned}$$

The inequality $1 - a \leq e^{-a}$ and the assumption that the sum of $P(A_i)$ diverges together imply that

$$P(\bigcup_{i=k}^{\infty} A_i^c) \leq \exp\Big(-\sum_{i=k}^{\infty} P(A_i)\Big) = 0$$

Therefore E^c is a union of countable number of events, each of them has probability zero. So $P(E^c) = 0$. ∎

14.13 Partitioned Matrices

Consider the matrix given by

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A and D are square matrices not necessarily of the same size and B and C are not necessarily square matrices. Assume A and D have inverses A^{-1} and D^{-1} respectively. Multiply the top row by CA^{-1} and subtract that result from the bottom row. The resulting matrix is

$$M' = \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix}$$

Now this type of transformation can be thought of as a particular type of elementary row operation that leaves the determinant of M unchanged: $\det M = \det M'$. Using the notation $|M|$ as a shorthand for $\det M$ we can conclude

$$|M| = |A| \cdot |D - CA^{-1}B|.$$

By doing a similar row operation by multiplying the second row by BD^{-1} and subtracting it from the top row we obtain

$$M'' = \begin{bmatrix} A - BD^{-1}C & 0 \\ C & D \end{bmatrix}$$

So we also have that $\det M = \det M''$ and hence

$$|M| = |D| \cdot |A - BD^{-1}C|.$$

The above identities are handy in dealing with some of the applications of stochastic processes to the problem of filtering.

Let us develop a couple of interesting matrix identities which occur in filters. We will develop these identities by performing the inversion procedure on the matrix M using the following work matrix to keep track of the inversion steps:

$$\left[\begin{array}{cc|cc} A & B & I & 0 \\ C & D & 0 & I \end{array}\right],$$

First subtract CA^{-1} times the first row from the second row and multiply the first row by A^{-1}. The work matrix becomes:

$$\left[\begin{array}{cc|cc} I & A^{-1}B & A^{-1} & 0 \\ 0 & D - CA^{-1}B & -CA^{-1} & 1 \end{array}\right],$$

Next subtract $A^{-1}B[D - CA^{-1}B]^{-1}$ times the second row from the first row and multiply the second row by $[D - CA^{-1}B]^{-1}$. The work matrix now becomes:

$$\left[\begin{array}{cc|cc} I & 0 & A^{-1} - A^{-1}B[D - CA^{-1}B]^{-1}CA^{-1} & -A^{-1}B[D - CA^{-1}B]^{-1} \\ 0 & I & -[D - CA^{-1}B]^{-1}CA^{-1} & [D - CA^{-1}B]^{-1} \end{array}\right],$$

Now, in order to discover the interesting identities let use an alternative ordering of row operations. In this second case we first subtract BD^{-1} times the second row from the first row and multiply the second row by D^{-1}. The work matrix becomes:

$$\left[\begin{array}{cc|cc} A - BD^{-1}C & 0 & I & -BD^{-1} \\ D^{-1}C & I & 0 & D^{-1} \end{array}\right],$$

Next subtract $D^{-1}C[A-BD^{-1}C]^{-1}$ times the first row from the second row and multiply the first row by $[A-BD^{-1}C]^{-1}$. The work matrix now becomes:

$$\left[\begin{array}{cc|cc} I & 0 & [A-BD^{-1}C]^{-1} & -[A-BD^{-1}C]^{-1}BD^{-1} \\ 0 & I & -D^{-1}C[A-BD^{-1}C]^{-1} & D^{-1}-D^{-1}C[A-BD^{-1}C]^{-1}BD-1 \end{array}\right],$$

Since these are two ways of obtaining the inverse of M, the matrix elements must be equal term-by-term. Hence we obtain some interesting identities. For example: using Partitioned Matrices we have found the Sherman-Morrison-Woodbury formula

$$[A-BD^{-1}C]^{-1} = A^{-1} - A^{-1}B[D-CA^{-1}B]^{-1}CA^{-1}. \tag{71}$$

Let us apply the ideas of joint random variables, conditional probabillity and expectation to a random $(n+m)$-vector $Z^T = [X^T, Y^T]$, where X is an n-vector and Y is an m-vector and assume that Z is a multivariate Gaussian $Z \sim N(m_Z, P_{ZZ})$ (with $m_Z = E[Z]$ is a vector of means and $P_{ZZ} = E[(z-m_Z)(z-m_Z)^T]$ is the covariance matrix). This also assumes that X and Y are jointly random Gaussian variables. Then Z has a probability density function given by

$$f_Z(z) = \frac{1}{\sqrt{(2\pi)^{n+m}|P_{ZZ}|}} \exp\big\{-\frac{1}{2}(z-m_z)^T P_{ZZ}^{-1}(z-m_z)\big\}. \tag{72}$$

Now we can use the partition matrix approach given in Section [14.13] to write the covariance matrix P_{ZZ} as

$$P_{ZZ} = E[(z-m_Z)(z-m_Z)^T] = \begin{bmatrix} E[(X-m_X)(X-m_X)^T] & E[(X-m_X)(Y-m_Y)^T] \\ E[(Y-m_Y)(X-m_Z)^T] & E[(Y-m_Y)(Y-m_Y)^Y] \end{bmatrix} = \begin{bmatrix} P_{XX} & P_{XY} \\ P_{YX} & P_{YY} \end{bmatrix}$$

Then we have

Theorem 14.14 *If the random vectors X and Y are distributed according to Eq. [72] then X and/or Y is also a multivariate Gaussian (with appropriate parameters).*

Proof Let show the theorem holds for Y. Accordingly let use consider a transformation of variables to $W^T = [W_1^T, W_2^T]$ such that

$$W_1 = X - P_{XY}P_{YY}^{-1}Y, \quad W_2 = Y.$$

Observe that the Jacobian of the above transformation is 1.

From Theorem [5.1] W is normally distributed. Let us compute $E[(W_1 - E[W_1])(W_2 - E[W_2])]$,

$$\begin{aligned} E[W_1] &= m_X - P_{XY}P_{YY}^{-1}m_Y \\ W_1 - E[W_1] &= (X-m_X) - P_{XY}P_{YY}^{-1}(Y-m_Y) \\ E[W_2] &= m_Y \\ W_2 - E[W_2] &= (Y-m_Y) \end{aligned}$$

and so

$$\begin{aligned} E[(W_1 - E[W_1])(W_2 - E[W_2])^T] &= E[\big((X - m_X) - P_{XY}P_{YY}^{-1}(Y - m_Y)\big)\big(Y - m_Y\big)^T] \\ &= E[(X - m_X)(Y - m_Y)^T] - E[P_{XY}P_{YY}^{-1}(Y - m_Y)(Y - m_Y)^T] \\ &= P_{XY} - P_{XY}P_{YY}^{-1}P_{YY} \\ &= P_{XY} - P_{XY} = 0 \end{aligned}$$

and by symmetry $E[(W_2 - E[W_2])(W_1 - E[W_1])^T] = 0$.

Also,

$$\begin{aligned} E[(W_1 - E[W_1])(W_1 - E[W_1])^T] &= E[\big((X - m_X) - P_{XY}P_{YY}^{-1}(Y - m_Y)\big)\big((X - m_X) - P_{XY}P_{YY}^{-1}(Y - m_Y)^T\big)] \\ &= E[(X - m_X)(X - m_X)^T] - E[P_{XY}P_{YY}^{-1}(Y - m_Y)(Y - m_Y)^T P_{YY} - 1^T P_{XY}^T] \\ &= P_{XX} - P_{XY}P_{YY}^{-1}P_{YX} \end{aligned}$$

Therefore

$$P_{ZZ} = E[(z - m_Z)(z - m_Z)^T] = \begin{bmatrix} P_{XX} - P_{XY}P_{YY}^{-1}P_{XY} & 0 \\ 0 & P_{YY} \end{bmatrix} \quad \blacksquare$$

Therefore we can say that X and Y are marginally Gaussian (with appropriate parameters).

It is clear that if two vectors of random variables X and Y are independent, then they are uncorrelated. An important property of vectors of Gaussian random variables is the converse.

Theorem 14.15 *If two jointly distributed Gaussian random vectors are uncorrelated then they are also independent.*

Proof Let X and Y are random vectors of dimension n and m respectively, that are jointly Gaussian distributed and uncorrelated Define Z to be a random vector of dimension $(n+m)$ composed of the components of X and Y as above:

$$Z^T = [X^T, Y^T], \quad Z = \left[\frac{X}{Y}\right]$$

Then Z has a mean given by

$$m_Z = E[Z] = \left[\frac{E[X]}{E[Y]}\right] = \left[\frac{m_X}{m_Y}\right]$$

Because X and Y are uncorrelated, then we have

$$E[XY^T] = m_X m_Y^T, \quad E[YX^T] = m_Y m_X^T$$

and the covariance P_{ZZ} becomes
The second moment of Z is given by

$$E[ZZ^T] = \left[\begin{array}{c|c} E[XX^T] & E[XY^T] \\ \hline E[YX^T] & E[YY^T] \end{array}\right], \quad P_{ZZ} = E[ZZ^T] - m_Z m_Z^T$$

Therefore the covariance P_{ZZ} becomes block diagonal:

$$P_{ZZ} = E[ZZ^T] - m_Z m_Z^T = \left[\begin{array}{c|c} E[XX^T] & m_X m_Y^T \\ \hline m_Y m_X^T & E[YY^T] \end{array}\right] - \left[\begin{array}{c|c} m_X m_X^T & m_X m_Y^T \\ \hline m_Y m_X^T & m_Y m_Y^T \end{array}\right] = \left[\begin{array}{c|c} P_{XX} & 0 \\ \hline 0 & P_{YY} \end{array}\right]$$

Therefore the determinant of P_{ZZ} is given by the product

$$|P_{ZZ}| = |P_{XX}||P_{YY}|$$

Let ξ be the vector composed of n x-values and m y-values

$$\xi = \left[\frac{x}{y}\right]$$

Given the block diagonal structure of P_{ZZ} we can form the quadratic form $[\xi - m_Z]^T P_{ZZ}^{-1}[\xi - m_Z]$ which can be expanded using the vector representation of ξ in terms of x and y values

$$\begin{aligned} [\xi - \mathbf{m_Z}]^T \mathbf{P_{ZZ}^{-1}} [\xi - \mathbf{m_Z}] &= \left[\frac{x - m_X}{y - m_Y}\right]^T \left[\begin{array}{c|c} P_{XX}^{-1} & 0 \\ \hline 0 & P_{YY}^{-1} \end{array}\right] \left[\frac{x - m_X}{y - m_Y}\right] \\ &= [x - m_X]^T P_{XX}^{-1}[x - m_X] + [y - m_Y]^T P_{YY}^{-1}[y - m_Y] \end{aligned}$$

Therefore the joint probability distribution can be written as

$$\begin{aligned} f_Z(\xi) &= \frac{1}{(2\pi)^{(n+m)/2}} \frac{1}{|P_{ZZ}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\xi - \mathbf{m_Z})^T \mathbf{P_{ZZ}^{-1}} (\xi - \mathbf{m_Z}) \Big\} \\ &= \left[\frac{1}{(2\pi)^{n/2}} \frac{1}{|P_{XX}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x} - \mathbf{m_X})^T \mathbf{P_{XX}^{-1}} (\mathbf{x} - \mathbf{m_X}) \Big\}\right] \\ &\times \left[\frac{1}{(2\pi)^{m/2}} \frac{1}{|P_{YY}|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{y} - \mathbf{m_Y})^T \mathbf{P_{YY}^{-1}} (\mathbf{y} - \mathbf{m_Y}) \Big\}\right] \end{aligned}$$

And so if the Gaussian random vectors X and Y are uncorrelated, then their joint probability density factors into

$$f_Z(\xi) = f_X(x) \times f_Y(y) \tag{73}$$

and therefore X and Y are also independent random variables. ∎

By extension, if $X_1, X_2, \ldots, X_n$ are normally distributed random variables and are pair-wise uncorrelated, then they are independent.

Theorem 14.16 *Let X and Y be jointly Gaussian distributed as in Eq. [72]. Then the conditional probability density of X given Y is normal with mean*

$$m_X + P_{XY} P_{YY}^{-1} (Y - m_Y) \tag{74}$$

and covariance matrix

$$P_{XX} - P_{XY} P_{YY}^{-1} P_{YX}. \tag{75}$$

Proof We have from Theorem [14.14] that W_1 and W_2 are Gaussian and independent. Therefore, their joint density, which is also Gaussian, is the product of their individual probability distributions:

$$\begin{aligned} W_1 &\sim N(m_X - P_{XY}P_{YY}^{-1}m_Y, \quad P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}) \\ W_2 &\sim N(m_Y, P_{YY}) \end{aligned}$$

Therefore the combined joint probability distribution for W can be written as

$$\begin{aligned} f_{W_1,W_2}(w_1,w_2) &= \frac{1}{\sqrt{2\pi)^n|P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}|}} \\ &\exp\Big\{-\frac{1}{2}(w_1 - m_X + P_{XY}P_{YY}^{-1}m_Y)^T(P_{XX} - P_{XY}P_{YY}^{-1}P_{YX})^{-1} \\ &(w_1 - m_X + P_{XY}P_{YY}^{-1}m_Y)\Big\} \\ &\frac{1}{\sqrt{2\pi)^n|P_{YY}|}}\exp\Big\{-\frac{1}{2}(w_2 - m_Y)^T P_{YY}^{-1}(w_2 - m_Y)\Big\} \end{aligned}$$

Since the Jacobian of the transformation between X, Y and W_1, W_2 is equal to 1, we can recover the density $f_{X,Y}(x,y)$ by eliminating W_1 and W_2 in terms of X and Y. Doing do, we obtain

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{\sqrt{2\pi)^n|P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}|}} \\ &\exp\Big\{-\frac{1}{2}[x - m_X - P_{XY}P_{YY}^{-1}(y - m_Y)]^T(P_{XX} - P_{XY}P_{YY}^{-1}P_{YX})^{-1} \\ &[x - m_X - P_{XY}P_{YY}^{-1}(y - m_Y)]\Big\} \\ &\frac{1}{\sqrt{2\pi)^n|P_{YY}|}}\exp\Big\{-\frac{1}{2}(y - m_Y)^T P_{YY}^{-1}(y - m_Y)\Big\} \end{aligned}$$

Now the right hand side of the above is the marginal probability function of Y. So by comparison with Eq. [10] we see that the conditional density function of X given Y is

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{1}{\sqrt{2\pi)^n|P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}|}} \\ &\exp\Big\{-\tfrac{1}{2}[x - m_X - P_{XY}P_{YY}^{-1}(y - m_Y)]^T[P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}]^{-1} \\ &[x - m_X - P_{XY}P_{YY}^{-1}(y - m_Y)]\Big\} \end{aligned} \tag{76}$$

Eq. [76] is important for applications in Filtering theory.

15 Acknowledgements

These notes grew out of discussions I have had with Professor Howard Weiner of the University of California, Davis. Many of the methods are taken directly from my notes of Professor Weiner's explanations.

References

[1] M.H. Protter and C.B. Morrey, "A First Course in Real Analysis", Springer-Verlag, Inc., New York (1977).

[2] F. M. Hall, "An Introduction to Abstract Algebra", Volume 2, Cambridge University Press, (1969).

[3] John Stensby, "EE603 Class Notes", http://www.ece.uah.edu/courses/ee385/, University of Alabama, Huntsville (2016).

[4] Fabrice Baudoin,"Stochastic Processes and Brownian Motion",http://www.math.purdue.edu/~fbaudoin/MA539.pdf, Purdue University, Indiana (2010).

[5] Harry van Zanten,"An Introduction to Stochastic Processes in Continuous Time", http://www.math.vu.nl/sto/onderwijs/sp/sp_2007.pdf, University of Leiden (2007).

[6] Howard Weiner, UC Davis Mathematics Department, private communication (2007-2017).

[7] Mahalanobis, Prasanta Chandra (1936)."On the generalised distance in statistics", Proceedings of the National Institute of Sciences of India. 2 (1): 49-55.

[8] Albert Einstein, "Investigations on the Theory of the Brownian Movement", Dover Publications, Inc.(1956).

[9] Sheldon Ross, "A First Course in Probability", First Edition, Macmillan Publishers Co, Inc., New York, NY (1976).

[10] Sheldon Ross, "Introduction to Probability Models", Sixth Edition, Academic Press, San Diego, CA (1997).

[11] Emanuel Parzen, "Stochastic Processes", Holden-Day, Inc., San Francisco, CA (1962).

[12] Jyotiprasad Medhi, "Stochastic Processes", Wiley Eastern Limited, New Delhi (1983).

[13] Andrew H. Jazwinski,"Stochastic Processes and Filtering Theory", Dover Publications, Inc., Mineola, NY (2007).

[14] Norbert Wiener, "Cybernetics: or Control and Communication in the Animal and the Machine", The MIT Press, Cambridge, MA (1948).

[15] Norbert Wiener,"Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications", First Paperback Edition, The MIT Press, Cambridge, MA (1964).

[16] Norbert Wiener, "Generalized Harmonic Analysis", Acta Mathematica, V. 55, p. 117 (1930).

[17] Bernt Øskendal, "Stochastic Differential Equations, An Introduction with Applications", Third Edition, Springer-Verlag, Berlin (1992).

[18] Bernt Øskendal, "Stochastic Differential Equations, An Introduction with Applications", Sixth Edition, Springer-Verlag, Berlin (2005).

[19] J. L. Doob, "Stochastic Processes", John Wiley & Sons, Inc., New York, (1953).

[20] Lennart Sjogren. "Brownian Motion: Langevin Equation", Chapter 6, http://physics.gu.se/~frtbm/joomla/media/mydocs/LennartSjogren/kap6.pdf

[21] "Statistical Mechanics", Physics 127b http://www.cmp.caltech.edu/~mcc/Ph127/b/Lecture16.pdf