1

Stochastic Processes

A Brief Review

1.1 Introduction

In this chapter, we introduce the basic mathematical tools we will use. We assume the reader has a good understanding of probability spaces and random variables. For more details we refer to [67, 70]. This chapter is not meant to be a replacement for a book. To get the fundamentals please consult [70, 117]. In this chapter, we are reviewing fundamental notions for the rest of the book.

So, *what is a stochastic process?* When asked this question, *R.A. Fisher* famously replied, "What is a stochastic process? Oh, it's just one darn thing after another." We hope to elaborate on Fisher's reply in this introduction.

We start the study of stochastic processes by presenting some commonly assumed properties and characteristics. Generally, these characteristics simplify analysis of stochastic processes. However, a stochastic process with these properties will have simplified dynamics, and the resulting models may not be complex enough to model real-life behavior. In Section 1.6 of this chapter, we introduce the simplest stochastic processes: the coin toss process (also known as the Bernoulli process) which produces the simple random walk.

We start with the definition of a stochastic process.

Definition 1.1.1 Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, a stochastic process is any collection $\{X(t) : t \in \mathcal{I}\}$ of random variables defined on this probability space, where $\mathcal{I}$ is an index set. The notations X_t and $X(t)$ are used interchangeably to denote the value of the stochastic process at index value t.

Specifically, for any fixed t the resulting X_t is just a random variable. However, what makes this index set $\mathcal{I}$ special is that it confers the collection of random variables a certain structure. This will be explained next.

Quantitative Finance, First Edition. Maria C. Mariani and Ionut Florescu.

1.2 General Characteristics of Stochastic Processes

1.2.1 The Index Set $\mathcal{I}$

The set $\mathcal{I}$ indexes and determines the type of stochastic process. This set can be quite general but here are some examples:

- If $\mathcal{I} = \{0, 1, 2 \ldots\}$ or equivalent, we obtain the so-called discrete-time stochastic processes. We shall often write the process as $\{X_n\}_{n\in\mathbb{N}}$ in this case.
- If $\mathcal{I} = [0, \infty)$, we obtain the continuous-time stochastic processes. We shall write the process as $\{X_t\}_{t\geq 0}$ in this case. Most of the time t represents time.
- The index set can be multidimensional. For example, with $\mathcal{I} = \mathbb{Z} \times \mathbb{Z}$, we may be describing a discrete random field where at any combination $(x, y) \in \mathcal{I}$ we have a value $X(x, y)$ which may represent some node weights in a two-dimensional graph. If $\mathcal{I} = [0, 1] \times [0, 1]$ we may be describing the structure of some surface where, for instance, $X(x, y)$ could be the value of some electrical field intensity at position (x, y).

1.2.2 The State Space $\mathcal{S}$

The state space is the domain space of all the random variables X_t. Since we are discussing about random variables and random vectors, then necessarily $\mathcal{S} \subseteq \mathbb{R}$ or $\mathbb{R}^n$. Again, we have several important examples:

- If $\mathcal{S} \subseteq \mathbb{Z}$, then the process is integer valued or a process with discrete state space.
- If $\mathcal{S} = \mathbb{R}$, then X_t is a real-valued process or a process with a continuous state space.
- If $\mathcal{S} = \mathbb{R}^k$, then X_t is a k-dimensional vector process.

The state space $\mathcal{S}$ can be more general (for example, an abstract Lie algebra), in which case the definitions work very similarly except that for each t we have X_t measurable functions.

We recall that a real-valued function f defined on Ω is called measurable with respect to a sigma algebra $\mathcal{F}$ in that space if the inverse image of set B, defined as $f^{-1}(B) \equiv \{\omega \in E : f(\omega) \in B)\}$ is a set in sigma algebra $\mathcal{F}$, for all Borel sets B of $\mathbb{R}$.

A sigma algebra $\mathcal{F}$ is a collection of sets F of Ω satisfying the following conditions:

1) $\emptyset \in \mathcal{F}$.
2) If $F \in \mathcal{F}$ then its complement $F^c \in \mathcal{F}$.
3) If $F_1, F_2, \ldots$ is a countable collection of sets in $\mathcal{F}$ then their union $\cup_{n=1}^{\infty} F_n \in \mathcal{F}$

Suppose we have a random variable X defined on a space Ω. The sigma algebra generated by X is the smallest sigma algebra in Ω that contains all the pre images of sets in $\mathbb{R}$ through X. That is,

$$\sigma(X) = \sigma\left(\{X^{-1}(B) \mid \text{ for all } B\text{Borel sets in } \mathbb{R}\}\right)$$

This abstract concept is necessary to make sure that we may calculate any probability related to the random variable X.

1.2.3 Adaptiveness, Filtration, and Standard Filtration

In the special case when the index set $\mathcal{I}$ possesses a total order relationship,[1] we can discuss about the information contained in the process $X(t)$ at some moment $t \in \mathcal{I}$. To quantify this information we generalize the notion of sigma algebras by introducing a sequence of sigma algebras: the filtration.

Definition 1.2.1 (Filtration). A probability space $(\Omega, \mathscr{F}, \mathbf{P})$ is a filtered probability space if and only if there exists a sequence of sigma algebras $\{\mathscr{F}_t\}_{t\in\mathcal{I}}$ included in $\mathscr{F}$ such that $\mathscr{F}$ is an increasing collection i.e.:

$$\mathscr{F}_s \subseteq \mathscr{F}_t, \quad \forall s \leq t, \quad s, t \in \mathcal{I}.$$

The filtration is called *complete* if its first element contains all the null sets of $\mathscr{F}$. If, for example, 0 is the first element of the index set (the usual situation) then $\forall N \in \mathscr{F}$, with $\mathbf{P}(N) = 0 \Rightarrow N \in \mathscr{F}_0$. This particular notion of a complete filtration is not satisfied and may lead to all sorts of contracdictions and counterexamples. To avoid any such case we shall assume that any filtration defined in this book is complete and all filtered probability spaces are complete.

In the particular case of continuous time (i.e. $I = [0, \infty)$), it makes sense to discuss about what happens with the filtration when two consecutive times get close to one another. For some specific time $t \in I$ we define the left and right sigma algebras:

$$\mathscr{F}_{t+} = \bigcap_{u>t} \mathscr{F}_u = \lim_{u \downarrow t} \mathscr{F}_u,$$

$$\mathscr{F}_{t-} = \sigma\left(\bigcup_{u>t} \mathscr{F}_u\right).$$

The countable intersection of sigma algebras is always a sigma algebra [67], but a union of sigma algebras is not necessarily a sigma algebra. This is why we modified the definition of $\mathscr{F}_{t-}$ slightly. The notation used $\sigma(\mathscr{C})$ represents the smallest sigma algebra that contains the collection of sets $\mathscr{C}$.

1 i.e. for any two elements $x, y \in \mathcal{I}$, either $x \leq y$ or $y \leq x$.

Definition 1.2.2 (Right and Left Continuous Filtrations). A filtration $\{F_t\}_{t\in\mathcal{I}}$ is right continuous if and only if $\mathcal{F}_t = \mathcal{F}_{t+}$ for all t, and the filtration is left continuous if and only if $\mathcal{F}_t = \mathcal{F}_{t-}$ for all t.

In general we shall assume throughout (if applicable) that any filtration is right continuous.

Definition 1.2.3 (Adapted Stochastic Process). A stochastic process $\{X_t\}_{t\in I}$ defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{P}, \{\mathcal{F}_t\}_{t\in\mathcal{I}})$ is called adapted if and only if X_t is $\mathcal{F}_t$-measurable for any $t \in \mathcal{I}$.

This is an important concept since in general, $\mathcal{F}_t$ quantifies the flow of information available at any moment t. By requiring that the process be adapted, we ensure that we can calculate probabilities related to X_t based solely on the information available at time t. Furthermore, since the filtration by definition is increasing, this also says that we can calculate the probabilities at any later moment in time as well.

On the other hand, due to the same increasing property of a filtration, it may not be possible to calculate probabilities related to X_t based only on the information available in $\mathcal{F}_s$ for a moment s earlier than t (i.e. $s < t$). This is the reason why the conditional expectation is a crucial concept for stochastic processes. Recall that $\mathbf{E}[X_t|F_s]$ is $\mathcal{F}_s$-measurable. Suppose we are sitting at time s and trying to calculate probabilities related to the random variable X_t at some time t in the future. Even though we may not calculate the probabilities related to X_t directly (nobody can since X_t will be in the future), we can still calculate its distribution according to its best guess based on the current information. That is precisely $\mathbf{E}[X_t|\mathcal{F}_s]$.

Definition 1.2.4 (Standard Filtration). In some cases, we are only given a standard probability space (without a separate filtration defined on the space). This typically corresponds to the case where we assume that all the information available at time t comes from the stochastic process X_t itself. No external sources of information are available. In this case, we will be using the standard filtration generated by the process $\{X_t\}_{t\in\mathcal{I}}$ itself. Let

$$\mathcal{F}_t = \sigma(\{X_s : s \le t, s \in \mathcal{I}\}),$$

denote the sigma algebra generated by the random variables up to time t. The collection of sigma algebras $\{\mathcal{F}_t\}_t$ is increasing and obviously the process $\{X_t\}_t$ is adapted with respect to it.

Notation In the case when the filtration is not specified, we will always construct the standard filtration and denote it with $\{F_t\}_t$.

In the special case when $\mathcal{I} = \mathbb{N}$, the set of natural numbers, and the filtration is generated by the process, we will sometimes substitute the notation $X_1, X_2, \dots, X_n$ instead of $\mathcal{F}_n$. For example we may write

$$\mathbf{E}[X_T^2|\mathcal{F}_n] = \mathbf{E}[X_T^2|X_n, \dots, X_1]$$

1.2.4 Pathwise Realizations

Suppose a stochastic process X_t is defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Recall that by definition for every $t \in \mathcal{I}$ fixed, X_t is a random variable. On the other hand, for every fixed $\omega \in \Omega$ we shall find a particular realization for any time t's, this outcome is typically denoted $X_t(\omega)$. Therefore, for each ω we can find a collection of numbers representing the realization of the stochastic process. That is a path. This realization may be thought of as the function $t \mapsto X_t(\omega)$.

This pathwise idea means that we can map each ω into a function from $\mathcal{I}$ into $\mathbb{R}$. Therefore, the process X_t may be identified as a subset of all the functions from $\mathcal{I}$ into $\mathbb{R}$.

In Figure 1.1 we plot three different paths each corresponding to a different realization ω_i, $i \in \{1, 2, 3\}$. Due to this pathwise representation, calculating probabilities related to stochastic processes is equivalent with calculating

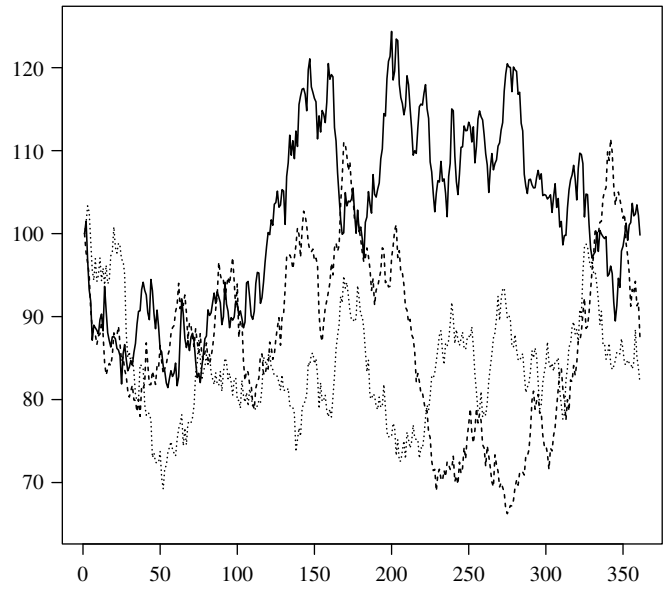

Figure 1.1 An example of three paths corresponding to three ω's for a certain stochastic process.

the distribution of these paths in subsets of the two-dimensional space. For example, the probability

$$\mathbf{P}(\max_{t\in[0,1]} X_t \leq 1 \text{and } \min_{t\in[0,1]} X_t \geq 0)$$

is the probability of the paths being in the unit square. However, such a calculation is impossible when the state space is infinite or when the index set is uncountable infinite such as the real numbers. To deal with this problem we need to introduce the concept of finite dimensional distribution.

1.2.5 The Finite Dimensional Distribution of Stochastic Processes

As we have seen, a stochastic process $\{X_t : t \geq 0\}_{t\in\mathcal{I}}$ is a parametrized collection of random variables defined on a probability space and taking values in $\mathbb{R}$. Thus, we have to ask: what quantities characterize a random variable? The answer is obviously its distribution. However, here we are working with a lot of variables. Depending on the number of elements in the index set $\mathcal{I}$, the stochastic process may have a finite or infinite number of components. In either case we will be concerned with the joint distribution of a finite sample taken from the process. This is due to practical consideration and the fact that in general we cannot study the joint distribution of a continuum of random variables. The processes that have a continuum structure on the set $\mathcal{I}$ serve as subject for a more advanced topic in stochastic differential equations (SDE). However, even in that more advanced situation, the finite distribution of the process still is the primary object of study.

Next, we clarify what we mean by finite dimensional distribution. Let $\{X_t\}_{t\in\mathcal{I}}$ be a stochastic process. For any $n \geq 1$ and for any subset $\{t_1, t_2, \ldots, t_n\}$ of $\mathcal{I}$, we denote with $F_{X_{t_1},X_{t_2},\ldots,X_{t_n}}$ the joint distribution function of the variables $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$. The statistical properties of the process X_t are completely described by the family of distribution functions $F_{X_{t_1},X_{t_2},\ldots,X_{t_n}}$ indexed by the n and the t_i's. This is a famous result due to Kolmogorov in the 1930's. Please refer to [119] and [159] for more details.

If we can describe these finite dimensional joint distributions for all n and t's, we completely characterize the stochastic process. Unfortunately, in general this is a complicated task. However, there are some properties of the stochastic processes that make this calculation task much easier. Figure 1.1 has three different paths. It is clear that every time the paths are produced they are different but the paths may have common characteristics. In the example plotted, the paths tend to keep coming back to the starting value, and they seem to have large oscillations when the process has large values and small oscillations when the process is close to 0. These features, if they exist, will help us calculate probabilities related to the distribution of the stochastic processes. Next we discuss the most important types of such features.

1.2.6 Independent Components

This is one of the most desirable properties of stochastic processes, however, no reasonable real life process has this property. For any collection $\{t_1, t_2, \dots, t_n\}$ of elements in $\mathcal{I}$, the corresponding random variables $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ are independent. Therefore, the joint distribution $F_{X_{t_1}, X_{t_2}, \dots, X_{t_n}}$ is just the product of the marginal distributions $F_{X_{t_i}}$. Thus, it is very easy to calculate probabilities using such a process. However, every new component being random implies no structure. In fact, this is the defining characteristic of a noise process.

Definition 1.2.5 (White Noise Process). A stochastic process $\{X_t\}_{t\in\mathcal{I}}$ is called a white noise process if it has independent components. That is, for any collection $\{t_1, t_2, \dots, t_n\}$ of index elements, the corresponding random variables $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ are independent. Additionally, for any time t, the random variables X_t have the same distribution $F(x)$, with the expected value $\mathbf{E}[X_t] = 0$.

The process is called a Gaussian white noise process if it is a white noise process, and in addition the common distribution of the stochastic process X_t is a normal with mean 0.

Please note that independent components do not require the distribution to be the same for all variables. In practical applications, modeling a signal often means eliminating trends until eventually reaching this noise process. At that time the process does not expose any more trends since only "noise" remains. Typically the modeling process is complete at that point.

1.2.7 Stationary Process

A stochastic process X_t is said to be *strictly stationary* if the joint distribution function of the vectors

$$(X_{t_1}, X_{t_2}, \dots, X_{t_n}) \quad \text{and} \quad (X_{t_1+h}, X_{t_2+h}, \dots, X_{t_n+h})$$

are the same for all $h > 0$ and all arbitrary selection of index points $\{t_1, t_2, \dots, t_n\}$ in $\mathcal{I}$. In particular, the distribution of X_t is the same for all t. Note that this property simplifies the calculation of the joint distribution function. The condition implies that the process is in equilibrium. The process will behave the same regardless of the particular time at which we examine it.

A stochastic process X_t is said to be *weak stationary* or *covariance stationary* if X_t has finite second moments for any t and if the covariance function $\mathrm{Cov}(X_t, X_{t+h})$ depends only on h for all $t \in \mathcal{I}$. Note that this is a weaker version than the notion of strict stationarity. A strictly stationary process with finite second moments (so that covariance exists) is going to be automatically covariance stationary. The reverse is not true. Indeed, examples of processes which are covariance stationary but are not strictly stationary include autoregressive

conditionally heteroscedastic (ARCH) processes. ARCH processes are knows as discrete-time stochastic variance processes.

The notion of weak stationarity was developed because of the practical way in which we observe stochastic processes. While strict stationarity is a very desirable concept, it is not possible to test it with real data. To prove strict stationarity means we need to test all joint distributions. In real life the samples we gather are finite so this is not possible. Instead, we can test the stationarity of the covariance matrix which only involves bivariate distributions.

Many phenomena can be described by stationary processes. Furthermore, many classes of processes eventually become stationary if observed for a long time. The white noise process is a trivial example of a strictly stationary process.

However, some of the most common processes encountered in practice – the Poisson process and the Brownian motion – are not stationary. However, they have stationary and independent **increments**. We define this concept next.

1.2.8 Stationary and Independent Increments

In order to discuss the increments for stochastic processes, we need to assume that the set $\mathcal{I}$ has a total order, that is, for any two elements s and t in $\mathcal{I}$ we have either $s \leq t$ or $t \leq s$. As a clarifying point a two-dimensional index set, for example, $\mathcal{I} = [0,1] \times [0,1]$ does not have this property.

A stochastic process X_t is said to have *independent increments* if the random variables

$$X_{t_2} - X_{t_1},\ X_{t_3} - X_{t_2},\ \ldots,\ X_{t_n} - X_{t_{n-1}}$$

are independent for any n and any choice of the sequence $\{t_1, t_2, \ldots, t_n\}$ in $\mathcal{I}$ with $t_1 < t_2 < \cdots < t_n$.

A stochastic process X_t is said to have *stationary increments* if the distribution of the random variable $X_{t+h} - X_t$ depends only on the length $h > 0$ of the increment and not on the time t.

Notice that this is not the same as stationarity of the process itself. In fact, with the exception of the constant process, there exists no process with stationary and independent increments *which is also* stationary. This is proven in the next proposition.

Proposition 1.2.1 If a process $\{X_t, t \in [0, \infty)\}$ has stationary and independent increments then,

$$\mathbf{E}[X_t] = m_0 + m_1 t$$
$$\mathrm{Var}[X_t - X_0] = \mathrm{Var}[X_1 - X_0]t,$$

where $m_0 = \mathbf{E}[X_0]$, and $m_1 = \mathbf{E}[X_1] - m_0$.

Proof. We present the proof for the variance, and the result for the mean is entirely similar (see [119]). Let $f(t) = \text{Var}[X_t - X_0]$. Then for any t, s we have

$$\begin{aligned} f(t+s) &= \text{Var}[X_{t+s} - X_0] = \text{Var}[X_{t+s} - X_s + X_s - X_0] \\ &= \text{Var}[X_{t+s} - X_s] + \text{Var}[X_s - X_0] \text{ (because the increments are independent)} \\ &= \text{Var}[X_t - X_0] + \text{Var}[X_s - X_0] \text{ (because of stationary increments)} \\ &= f(t) + f(s) \end{aligned}$$

that is, the function f is additive (the above equation is also called Cauchy's functional equation). If we assume that the function f obeys some regularity conditions,[2] then the only solution is $f(t) = f(1)t$ and the result stated in the proposition holds. □

1.3 Variation and Quadratic Variation of Stochastic Processes

The notion of the variation of a stochastic process is originated from deterministic equivalents. We recall these deterministic equivalents.

Definition 1.3.1 (Variation for Deterministic Functions). Let $f : [0, \infty) \to \mathbb{R}$ be a deterministic function. Let $\pi_n = (0 = t_0 < t_1 < \dots t_n = t)$ be a partition of the interval $[0, t]$ with n subintervals. Let $\| \pi_n \| = \max_i (t_{i+1} - t_i)$ be the length of the largest subinterval in the partition. We define the first order variation as

$$FV_t(f) = \lim_{\|\pi_n\| \to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|.$$

We define the quadratic variation as

$$[f, f]_t = \lim_{\|\pi_n\| \to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|^2.$$

In general, the d-order variation is defined as

$$\lim_{\|\pi_n\| \to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|^d.$$

Next, we remark why we have not used a notation to the higher order variations (that is for orders three or more).

2 These regularity conditions are either (i) f is continuous, (ii) f is monotone, and (iii) f is bounded on compact intervals. In particular the third condition is satisfied by any process with finite second moments. The linearity of the function under condition (i) was first proven by [34].

Lemma 1.3.1 The first order variation at point t of a differentiable function $f(t)$ with continuous derivative is the length of the curve from 0 to t, that is,

$$FV_t(f) = \int_0^t |f'(s)|ds$$

Proof. This lemma is easy to prove using the mean value theorem. Recall that for any differentiable function f with continuous derivative ($f \in \mathscr{C}^1([0,\infty))$), the mean value theorem states that

$$f(t_{i+1}) - f(t_i) = f'(t_i^*)(t_{i+1} - t_i),$$

where t_i^* is some point between t_i and t_{i+1}. Hence, we obtain

$$\begin{aligned} FV_t(f) &= \lim_{\|\pi_n\|\to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)| \\ &= \lim_{\|\pi_n\|\to 0} \sum_{i=0}^{n-1} |f'(t_i^*)|(t_{i+1} - t_i) \\ &= \int_0^t |f'(s)|ds, \end{aligned}$$

recognizing that the last sum is just a Darboux sum which converges to the integral. □

Lemma 1.3.2 For a deterministic function f which is differentiable with continuous first order derivative, all d-order variations with $d \geq 2$ are zero. We denote $\mathscr{C}^1([0,\infty)$ the collection of all functions with first derivative continuous.

Proof. This lemma is the reason why we do not need to discuss about higher order variations for deterministic function, since they are all 0. To prove the lemma, we look at the formula for the quadratic variation. All higher d-orders ($d > 2$) use the same reasoning. We have

$$\begin{aligned} [f,f]_t &= \lim_{\|\pi_n\|\to 0} \sum_{i=0}^{n-1} |f(t_{i+1}) - f(t_i)|^2 \\ &= \lim_{\|\pi_n\|\to 0} \sum_{i=0}^{n-1} |f'(t_i^*)|^2 (t_{i+1} - t_i)^2 \\ &\leq \lim_{\|\pi_n\|\to 0} \| \pi \| \sum_{i=0}^{n-1} |f'(t_i^*)|^2 (t_{i+1} - t_i) \\ &= \lim_{\|\pi_n\|\to 0} \| \pi \| \lim_{\|\pi_n\|\to 0} \sum_{i=0}^{n-1} |f'(t_i^*)|^2 (t_{i+1} - t_i). \end{aligned} \tag{1.1}$$

The second term in 1.1 is the integral $\int_0^t |f'(s)|^2 ds$. This integral is bounded, since the function has a continuous first order derivative and furthermore the first term converges to 0. Therefore, the product goes to 0. □

We note that the only way the product at the end of the above proof does not equal 0 is when the integral is infinite. However, as we know the integral of any derivable function of finite intervals is finite. Therefore, it must be that the functions with finite quadratic variation on $[0, t]$ have to be non-derivable. In fact, for any point t we may repeat this argument for an arbitrary interval $[t - \Delta t, t + \Delta t]$; thus we can easily conclude that the functions with finite quadratic variation are not derivable at any point in $\mathbb{R}$.

The notion of a function which is continuous but not derivable at any point is very strange. However, it is this strange behavior that is the defining characteristic for stochastic processes.

Definition 1.3.2 (Quadratic Variation for Stochastic Processes). Let X_t be a stochastic process on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with filtration $\{\mathcal{F}_t\}_t$. Let $\pi_n = (0 = t_0 < t_1 < \dots t_n = t)$ be a partition of the interval $[0, t]$. We define the quadratic variation process

$$[X, X]_t = \lim_{\|\pi_n\| \to 0} \sum_{i=0}^{n-1} |X_{t_{i+1}} - X_{t_i}|^2,$$

where the limit of the sum is defined in probability.

The quadratic variation process is a stochastic process. The quadratic variation may be calculated explicitly only for some classes of stochastic processes. The stochastic processes used in finance have finite second order variation. The third and higher order variations are all zero while the first order is infinite. This is the fundamental reason why the quadratic variation has such a big role for stochastic processes used in finance.

1.4 Other More Specific Properties

- **Point Processes.** These are special processes that count rare events. They are very useful in practice due to their frequent occurrence. For example, consider the process that gives at any time t the number of buses passing by a particular point on a certain street, starting from an initial time $t = 0$. This is a typical rare event ("rare" here does not refer to the frequency of the event, rather to the fact that there are gaps between event occurrence). Or, consider the process that counts the number of defects in a certain area of material (say 1 cm^2). Two particular cases of such a process (and the most important) are the Poisson and jump diffusion processes which will be studied in Chapter 11.

- **Markov Processes.** In general terms this is a process with the property that at time s and given the process value X_s, the future values of the process (X_t with $t > s$) only depend on this X_s and not any of the earlier X_r with $r < s$. Or equivalently the behavior of the process at any future time when its present state is exactly known is not modified by additional knowledge about its past. The study of Markov processes constitutes a big part of this book. The finite distribution of such a process has a much simplified structure. Using conditional distributions, for a fixed sequence of times $t_1 < t_2 < \cdots < t_n$ we may write

$$\begin{aligned} F_{X_{t_1},X_{t_2},\ldots,X_{t_n}} &= F_{X_{t_n}|X_{t_{n-1}},\ldots,X_{t_1}} F_{X_{t_{n-1}}|X_{t_{n-2}},\ldots,X_{t_1}} \cdots F_{X_{t_2}|X_{t_1}} F_{X_{t_1}} \\ &= F_{X_{t_n}|X_{t_{n-1}}} F_{X_{t_{n-1}}|X_{t_{n-2}}} \cdots F_{X_{t_2}|X_{t_1}} F_{X_{t_1}} \\ &= F_{X_{t_1}} \prod_{i=2}^{n} F_{X_{t_i}|X_{t_{i-1}}} \end{aligned}$$

  which is a much simpler structure. In particular it means that we only need to describe one-step transitions.
- **Martingales.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A martingale sequence of length n is a set of variables $X_1, X_2, \ldots, X_n$ and corresponding sigma algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ that satisfy the following relations:
  1) Each X_i is an integrable random variable adapted to the corresponding sigma algebra $\mathcal{F}_i$.
  2) The $\mathcal{F}_i$'s form a filtration.
  3) For every $i \in [1, 2, \ldots, n-1]$, we have

$$X_i = \mathbf{E}[X_{i+1}|\mathcal{F}_i].$$

  This process has the property that the expected value of the future given the information we have today is going to be equal to the known value of the process today. These are some of the oldest processes studied in the history of probability due to their tight connection with gambling. In fact in French (the origin of the name is attributed to Paul Lévy) a martingale means a winning strategy (winning formula).

Examples of martingales include the standard Brownian motion, Brownian motion with drift, Wald's martingale and several others.

In the next section, we present some examples of stochastic processes.

1.5 Examples of Stochastic Processes

1.5.1 The Bernoulli Process (Simple Random Walk)

We will start the study of stochastic processes with a very simple process – tosses of a (not necessarily fair) coin. Historically, this is the first stochastic process ever studied.

Table 1.1 Sample outcome.

Y_i	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**
N_i	0	0	1	1	1	2	2	2	2	2	3	4	5

Let $Y_1, Y_2, \ldots$ be independent and identically distributed (iid) Bernoulli random variables with parameter p, i.e.

$$Y_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

To simplify the analogy, let $Y_i = 1$ when a head appears and a tail is obtained at the i-th toss if $Y_i = 0$. Let

$$N_k = \sum_{i=1}^{k} Y_i,$$

be the number of heads up to the k-th toss, which we know is distributed as a Binomial (k, p) random variable ($N_k \sim$ Binom(k, p)). An example of the above analogy is presented in Table 1.1

A sample outcome may look like the following:

Let S_n be the time at which the n-th head (success) occurred. Then mathematically,

$$S_n = \inf\{k : N_k = n\}$$

Let $X_n = S_n - S_{n-1}$ be the number of tosses to get the n-th head starting from the $(n-1)$-th head. We will present some known results about these processes.

Proposition 1.5.1

1) "Waiting times" $X_1, X_2 \ldots$ are independent and identically distributed "trials" ~Geometric(p) random variables.
2) The time at which the n-th head occurs is negative binomial, i.e. $S_n \sim$ negative binomial(n, p).
3) Given $N_k = n$ the distribution of $(S_1, \ldots, S_n)$ is the same as the distribution of a random sample of n numbers chosen without replacement from $\{1, 2, \ldots, k\}$.
4) Given $S_n = k$ the distribution of $(S_1, \ldots, S_{n-1})$ is the same as the distribution of a random sample of $n-1$ numbers chosen without replacement from $\{1, 2, \ldots, k-1\}$.
5) We have as sets:

$$\{S_n > k\} = \{N_k < n\}.$$

6) Central limit theorems (CLT):

$$\frac{N_k - \exp[N_k]}{\sqrt{\mathrm{Var}[N_k]}} = \frac{N_k - kp}{\sqrt{kp(1-p)}} \xrightarrow{D} N(0,1).$$

7)

$$\frac{S_n - \exp[S_n]}{\sqrt{\mathrm{Var}[S_n]}} = \frac{S_n - n/p}{\sqrt{n(1-p)/p}} \xrightarrow{D} N(0,1).$$

8) As $p \downarrow 0$

$$\frac{X_1}{\exp[X_1]} = \frac{X_1}{1/p} \xrightarrow{D} \text{Exponential}(\lambda = 1).$$

9) As $p \downarrow 0$

$$\mathbf{P}\left\{N_{[\frac{t}{p}]} = j\right\} \to \frac{t^j}{j!}e^{-t}.$$

We will prove several of these properties. The rest are assigned as exercises.

For 1) and 2) The distributional assertions are easy to prove; we may just use the definition of geometric (p) and negative binomial random variables. We need only to show that the X_i's are independent. See problem 1.

Proof for 3). We take $n = 4$ and $k = 100$ and prove this part only for this particular case. The general proof is identical to problem 2. A typical outcome of a Bernoulli process looks like as follows:

$$\omega : 00100101000101110000100$$

In the calculation of probability we have $1 \le s_1 < s_2 < s_3 < s_4 \le 100$. Using the definition of the conditional probability we can write:

$$\begin{aligned}
&\mathbf{P}(S_1 = s_1 \dots S_4 = s_4 | N_{100} = 4) \\
&= \frac{\mathbf{P}(S_1 = s_1 \dots S_4 = s_4 \text{and } N_{100} = 4)}{\mathbf{P}(N_{100} = 4)} \\
&= \frac{\mathbf{P}\left(\overbrace{00\dots0}^{s_1-1}\overbrace{100\dots0}^{s_2-s_1-1}\overbrace{100\dots0}^{s_3-s_2-1}\overbrace{100\dots0}^{s_4-s_3-1}\overbrace{100\dots0}^{100-s_4}\right)}{\binom{100}{4}p^4(1-p)^{96}} \\
&= \frac{(1-p)^{s_1-1}p(1-p)^{s_2-s_1-1}p(1-p)^{s_3-s_2-1}p(1-p)^{s_4-s_3-1}p(1-p)^{100-s_4}}{\binom{100}{4}p^4(1-p)^{96}} \\
&= \frac{(1-p)^{96}p^4}{\binom{100}{4}p^4(1-p)^{96}} = \frac{1}{\binom{100}{4}}.
\end{aligned}$$

The result is significant since it means that if we only know that there have been 4 heads by the 100-th toss, then any 4 tosses among these 100 are equally likely to contain the heads. □

Proof for 8).

$$\mathbf{P}\left(\frac{X_1}{1/p} > t\right) = \mathbf{P}\left(X_1 > \frac{t}{p}\right) = \mathbf{P}\left(X_1 > \left[\frac{t}{p}\right]\right)$$
$$= (1-p)^{\left[\frac{t}{p}\right]} = \left[(1-p)^{-\frac{1}{p}}\right]^{-p\left[\frac{t}{p}\right]} \to e^{-t},$$

since

$$\lim_{p\to 0} -p\left[\frac{t}{p}\right] = \lim_{p\to 0} -p\left(\frac{t}{p} + \left[\frac{t}{p}\right] - \frac{t}{p}\right)$$
$$= -t + \lim_{p\to 0} \underbrace{p\left(\frac{t}{p} - \left[\frac{t}{p}\right]\right)}_{\in[0,1]} = -t$$

Therefore $\mathbf{P}\left(\frac{X_1}{1/p} \le t\right) \to 1 - e^{-t}$ and the proof is complete. □

The next example of a stochastic process is the Brownian motion.

1.5.2 The Brownian Motion (Wiener Process)

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A Brownian motion is a stochastic process Z_t with the following properties:

1) $Z_0 = 0$.
2) With probability 1, the function $t \to Z_t$ is continuous in t.
3) The process Z_t has stationary and independent increments.
4) The increment $Z_{t+s} - Z_s$ has a $N(0, t)$ distribution.

The Brownian motion, also known as the Wiener process, may be obtained as a limit of a random walk. Assuming a random walk with probability $\frac{1}{2}$, we will have two variables: the time $t \in [0, T]$ and the position $x \in [-X, X]$.

For each $j = 1, 2, ..., n$, consider

$$u_{k,j} = P\left(x_k = j\right),$$

where k represents the time and j the position.

For $P(B) \neq 0$ we can define the conditional probability of the event A given that the event B occurs as

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

To the position j in time $k+1$ we can arrive only from the position $j-1$, or j at time k, so we have:

$$u_{k+1,j} = \frac{1}{2}\left(u_{k,j} + u_{k,j-1}\right). \tag{1.2}$$

We can rewrite (6.9) as

$$u_{k+1,j} = \frac{1}{2}\left[\left(u_{k,j+1} - u_{k,j}\right) - \left(u_{k,j} - u_{k,j-1}\right)\right] + u_{k,j}$$

or

$$u_{k+1,j} - u_{k,j} = \frac{1}{2}\left[\left(u_{k,j+1} - u_{k,j}\right) - \left(u_{k,j} - u_{k,j-1}\right)\right].$$

Using the notation,

$$u_{k,j} = u\left(t_k, x_j\right),$$

we obtain

$$u\left(t_{k+1}, x_j\right) - u\left(t_k, x_j\right) = \frac{1}{2}\left[\left(u\left(t_k, x_{j+1}\right) - u\left(t_k, x_j\right)\right) - \left(u\left(t_k, x_j\right) - u\left(t_k, x_{j-1}\right)\right)\right]. \tag{1.3}$$

Now let Δt and Δx be such that $X = n\Delta x,\ T = n\Delta t$ and then $\frac{X}{\Delta x} = \frac{T}{\Delta t}$. Then multiplying (6.10) by $\frac{1}{\Delta t}$ we obtain

$$\begin{aligned}&\frac{1}{\Delta t}\left(u\left(t_{k+1}, x_j\right) - u\left(t_k, x_j\right)\right)\\ &\quad = \frac{1}{2\Delta t}\left[\left(u\left(t_k, x_{j+1}\right) - u\left(t_k, x_j\right)\right) - \left(u\left(t_k, x_j\right) - u\left(t_k, x_{j-1}\right)\right)\right].\end{aligned} \tag{1.4}$$

If we take $\Delta t \to 0$ the first term in (6.11) converges to $\partial_t(u)$. For the second term, if we assume that

$$\Delta t \approx (\Delta x)^2$$

taking into account that $\Delta t = \frac{T\Delta x}{X}$,

$$\frac{1}{2\Delta t} = \frac{X}{2T\Delta x}$$

we can conclude that the second term converges to $\partial_{xx}(u)$. So, from the random walks we get a discrete version of the heat equation

$$\partial_t(u) = \frac{1}{2}\partial_{xx}(u)$$

As an example, consider the random walk with step $\Delta x = \sqrt{\Delta t}$, that is,

$$\begin{cases} x_{j+1} = x_j \pm \sqrt{\Delta t} \\ x_0 = 0 \end{cases}$$

We claim that the expected value after n steps is zero. The stochastic variables $x_{j+1} - x_j = \pm\sqrt{\Delta t}$ are independents and so

$$E\left(x_n\right) = VE\left(\sum_{j=0}^{n-1}\left(x_{j+1} - x_j\right)\right) = \sum_{j=0}^{n-1} VE\left(x_{j+1} - x_j\right) = 0$$

and

$$\mathrm{Var}\left(x_n\right) = \sum_{j=0}^{n-1}\mathrm{Var}\left(x_{j+1} - x_j\right) = \sum_{j=0}^{n-1}\Delta t = n\Delta t = T.$$

If we interpolate the points $\{x_j\}_{1\le j\le n-1}$ we obtain

$$x(j) = \frac{x_{j+1} - x_j}{\Delta t}\left(t - t_j\right) + x_j \tag{1.5}$$

for $t_j \le t \le t_{j+1}$. Equation (6.12) is a *Markovian process*, because:

1) $\forall a > 0$, $\{x\left(t_k + a\right) - x\left(t_k\right)\}$ is independent of the history $\{x(s) : s \le t_k\}$
2) $E\left(x\left(t_k + a\right) - x\left(t_k\right)\right) = 0$
   $\mathrm{Var}\left(x\left(t_k + a\right) - x\left(t_k\right)\right) = a$

Remark 1.5.1 If $\Delta t \ll 1$ then $x \approx N\left(0, \sqrt{a}\right)$ (i.e. normal distributed with mean *0* and variance a) and then,

$$P\left(x(t+a) - x(t) \ge y\right) \approx \frac{1}{\sqrt{2\pi}}\int_y^{\infty} e^{-\frac{t^2}{2a}}dt.$$

This is due to the CLT that guarantees that if N is large enough, the distribution can be approximated by Gaussian.

The Brownian motion will be discussed in details later in the book.

In the next sections, we recall some known results that would be used throughout this book.

1.6 Borel–Cantelli Lemmas

In probability theory, the Borel—Cantelli lemmas are statements about sequences of events. The lemmas state that, under certain conditions, an event will have probability either zero or one. We formally state the lemmas as follows: Let $\{A_n : n \ne 1\}$ be a sequence of events in a probability space. Then the event $A(i.o.) = \{A_n$ occurs for infinitely many $n\}$ is given by

$$A(i.o.) = \cap_{k=1}^{\infty}\cup_{n=k}^{\infty} A_n,$$

where *i.o.* stands for "infinitely often."

Lemma 1.6.1 If $\{A_n\}$ is a sequence of events and $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\{A_n \quad i.o\}) = 0$.

Lemma 1.6.2 Suppose that $\{A_n\}$ is a sequence of independent events and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\{A_n \quad i.o\}) = 1$.

The problems at the end of this chapter involves applications of the Borel–Cantelli lemmas to the Bernoulli process. Please refer to [67]. Section 1.4 for more details of the lemmas.

1.7 Central Limit Theorem

The CLT is the second fundamental theorem in probability which states that if S_n is the sum of n mutually independent random variables, then the distribution function of S_n is well approximated by a certain type of continuous function known as a normal density function, which is given by the formula

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.6}$$

where μ and σ are the mean and standard deviation, respectively. The CLT gives information on what happens when we have the sum of a large number of independent random variables each of which contributes a small amount to the total.

1.8 Stochastic Differential Equation

The theory of differential equations is the origin of classical calculus and motivated the creation of differential and integral calculus. A differential equation is an equation involving an unknown function and its derivative. Typically, a differential equation is a functional relationship

$$f(t, x(t), x'(t), x''(t), \ldots) = 0, \quad 0 \leq t \leq T \tag{1.7}$$

involving the time t, an unknown function $x(t)$, and its derivative. The solution of 1.7 is to find a function $x(t)$ which satisfies Equation 1.7.

Now consider the deterministic differential equation:

$$dx(t) = a(t, x(t))dt, \quad x(0) = x_0. \tag{1.8}$$

The easiest way to introduce randomness in this equation is to randomize the initial condition. The solution $x(t)$ then becomes a stochastic process $(X_t, t \in [0, T])$ defined as

$$dX_t = a(t, X_t)dt, \quad X_0(\omega) = Y(\omega). \tag{1.9}$$

Equation 1.9 is called a random differential equation. Random differential equations can be considered as deterministic equations with a perturbed initial condition. Note this is not a full SDE.

For a complete introduction and study of stochastic differential equations we refer to [159]. For our purpose, we introduce a simplified definition.

Definition 1.8.1 An SDE is defined as a deterministic differential equation which is perturbed by random noise.

Note this is very different from a random differential equation because the randomness is now included in the dynamics of the equation. Examples of SDE's include:

1) $dX_t = adt + bdB_t$.
2) $dX(t) = -\lambda X(t)dt + dB_t$.

and many, many others. dB_t is a notation for the derivative of the Brownian motion – a process that does not exist also called the white noise process.

A comprehensive discussion of SDE's will be presented in Chapter 9 of this book.

1.9 Stochastic Integral

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $\{\mathcal{F}_t\}$ a filtration on this space. Define for some fixed $S \leq T$ a class of functions $\nu = \nu(S, T)$:

$$f(t, \omega) : [0, \infty) \times \Omega \to \mathbb{R}$$

such that:

1) $(t, \omega) \mapsto f(t, \omega)$ is a $\mathcal{B} \times \mathcal{F}$ measurable function, where $\mathcal{B} = \mathcal{B}[S, T]$ is the Borel sigma algebra on that interval.
2) $\omega \mapsto f(t, \omega)$ is $\mathcal{F}_t$-adapted for all t.
3) $\mathbf{E}[\int_S^T f^2(t, \omega)dt] < \infty$

Then for every such $f \in \nu$ we can construct

$$\int_S^T f_t dB_t = \int_S^T f(t, \omega)dB_t(\omega),$$

where B_t is a standard Brownian motion with respect to the same filtration $\{\mathcal{F}_t\}$. This quantity is called a stochastic integral with respect to the Brownian motion B_t. We note that the stochastic integral is a random quantity.

1.9.1 Properties of the Stochastic Integral

- Linearity:

$$\int_S^T (af_t + bg_t)dB_t = a\int_S^T f_t dB_t + b\int_S^T g_t dB_t, \quad a.s.$$

- $$\int_S^T f_t dB_t = \int_S^U f_t dB_t + \int_U^T f_t dB_t, \quad a.s., \forall S < U < T$$

- $$\mathbf{E}\left[\int_S^T f_t dB_t\right] = 0$$

- Itô Isometry:

$$\mathbf{E}\left[\left(\int_S^T f_t dB_t\right)^2\right] = \mathbf{E}\left[\int_S^T f_t^2 dt\right]$$

- If $f \in \nu(0,T)$ for all T then $M_t(\omega) = \int_0^t f(s,\omega)dB_s(\omega)$ is a martingale with respect to $\mathcal{F}_t$ and

$$\mathbf{P}\left(\sup_{0\le t\le T}|M_t| \ge \lambda\right) \le \frac{1}{\lambda^2}\mathbf{E}\left[\int_0^T f_t^2 dt\right], \lambda, T > 0 \text{(Doob's inequality)}$$

A detailed discussion on the construction of stochastic integrals will be presented in Chapter 9 of this book.

1.10 Maximization and Parameter Calibration of Stochastic Processes

There are primarily two methods to estimate parameters for a stochastic process in finance. The difference is in the data that is available.

The first method uses a sample of observations of the underlying process. It recovers the parameters of the stochastic process under the objective probability measure P. The second method uses the particular data specific to finance. The input is derivative data (such as options futures) and it estimates the parameters under the equivalent martingale measure Q. The first method is appropriate if we want to identify features of the underlying process or we want to construct financial instruments based on it. For example, in portfolio optimization, assessing the risk of a portfolio are problems that need to use parameters estimated under the probability P. On the other hand estimating parameters under Q is needed if we want to price other derivatives which are not regularly traded on the market.

Method 1: Given a history of a stochastic process X_t: $x_0, x_1, \dots, x_n$ estimate the parameters of the process. To clarify we assume the process follows a stochastic differential equation:

$$dX_t = f(X_t, \theta)dt + g(X_t, \theta)dW_t, t \geq 0, X_0 = x0, \tag{1.10}$$

Here, the functions f, g are given and the problem is to estimate the vector of parameters θ. For example, to consider the Black–Scholes model, we take $f(x, \theta) = ux$, $g(x, \theta) = \sigma x$ and $\theta = (u, \sigma)$. For the Cox–Ingersoll–Ross (CIR) model, we take $f(x, \theta) = k(\overline{x} - x)$, $g(x, \theta) = \sigma\sqrt{x}$ and $\theta = (k, \overline{x}, \sigma)$. Almost all stochastic models used in the financial markets may be written in this form.

The classical approach is to write the likelihood function and maximize it to produce the maximum likelihood estimator (MLE). If $f(x_1, \dots, x_n|\theta)$ is the density of the probability

$$P(X_{t_1} \leq x_1, \dots, X_{t_n} \leq x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} f(x_1, \dots, x_n|\theta)dx_1 \dots dx_n$$

then we maximize the likelihood function:

$$L(\theta) = f(x_1, \dots, x_n|\theta),$$

for the observed $x_1, x_2, \dots, x_n$, as a function of θ.

Typically this distribution is hard to find. We can however write

$$f(x_0, x_1, \dots, x_n|\theta) = f(x_n|x_{n-1}, \dots, x_0, \theta)f(x_{n-1}|x_{n-2}, \dots, x_0) \dots f(x_0|\theta).$$

When X_t is a Markov process (any solution to the SDE 1.10 is) we can reduce the density:

$$f(x_0, \dots, x_n|\theta) = f(x_n|x_{n-1}, \theta)f(x_{n-1}|x_{n-2}, \theta) \dots f(x_0|\theta).$$

This can be calculated for diffusions of the type in Eq. (1.10) using the Kolmogorov backward equation (Fokker–Planck equation). Specifically, if f and g do not depend on time (i.e. $f(x, t, \theta) = f(x, \theta)$) as in the specifications of the Eq. (1.10), the Markov process is homogeneous and the transition density reduces to $p(t - s, x, y)$, which is the density of

$$X_t = y|X_s = x.$$

This density satisfies the equation

$$\frac{\partial p}{\partial t} = f(x)\frac{\partial p}{\partial x} + \frac{1}{2}g(x)^2\frac{\partial^2 p}{\partial x^2},$$

where $p(t, x, y)$ is the transition density function to be calculated. If we can somehow calculate or estimate p, we can express the maximum likelihood function as

$$L(\theta) = \Pi_{i=1}^n p_\theta(\Delta t, x_{i-1}, x_i)p_\theta(x_0)$$

where $p_\theta(x_0)$ is the initial distribution at time $t = 0$, $x_0, x_1, \dots, x_n$ are the observations at times $t_i = i\Delta t$ and θ is the vector of parameters. Since this function is hard to optimize we typically maximize a log transformation:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p_\theta(\Delta t, x_{i-1}, x_i) + \log p_\theta(x_0)$$

since the logarithm is an increasing function and it will not change the location of the maximum. This function is called the *score function*.

The major issue with this approach is that the function $p(t, x, y)$ may only be calculated exactly in very few cases. We present a way to deal with this issue next.

1.10.1 Approximation of the Likelihood Function (Pseudo Maximum Likelihood Estimation)

We adapt the method in [3] for the probability density function of V_T. The main idea is to transform the variance process into an equivalent process, which has a density function close to the normal density. Then carry out the series expansion around that normal density. Finally, we invert the transformation to obtain the density expansion for the original process.

In this instance, we replace p_θ with a density h_θ that has a simple functional form and same moments as p_θ.

Euler Method We discretize 1.10 using Euler scheme as follows:

$$X_{t+\Delta t} - X_t = f(X_t, \theta)\Delta t + g(X_t, \theta)\Delta W_t \tag{1.11}$$

Conditioned by $\mathcal{F}_t$ this is Gaussian so we use a transition density (approximate). Therefore $X_{t+\Delta} - X_t | \mathcal{F}_t$ is approximately normal with mean $f(X_t, \theta)\Delta t$ and variance $g(X_t, \theta)^2 \Delta t$ which implies that

$$h(\Delta t, x, y) = \frac{1}{\sqrt{2\pi g^2(x, \theta)\Delta t}} e^{\frac{-(y-x-f(x,\theta)\Delta t)^2}{2\Delta t g^2(x,\theta)}}$$

is a good approximation. Thus the approximate log likelihood is

$$\ln(\theta)(x_1, \dots, x_n) = \frac{-1}{2}\left[\sum_{i=1}^{n} \frac{(x_i - x_{i-1} - f(x_{i-1}, \theta)\Delta t)^2}{2\Delta t g^2(x_{i-1}, \theta)} + \sum_{i=1}^{n} \log(2\pi g^2(x_{i-1}, \theta)\Delta t).\right]$$

This approximate may be maximized easily.

1.10.2 Ozaki Method

The second approach we present is the Ozaki method, and it works for homogeneous stochastic differential equations. Given the following SDE,

$$dX_t = f(X_t, \theta)dt + \sigma dW_t$$

one can show that

$$X_{t+\Delta t}|X_t = x \sim N(E_x, V_x);$$

where

$$E_x = x + \frac{f(x)}{\partial f/\partial x}\left(e^{\frac{\partial f}{\partial x}\Delta t} - 1\right)$$

$$V_x = \sigma^2 \frac{e^{2K_x}\Delta t - 1}{2K_x}$$

and

$$K_x = \frac{1}{\Delta t}\log\left(1 + \frac{f(x)}{x\partial f/\partial x}\left(e^{\frac{\partial f}{\partial x}\Delta t} - 1\right)\right)$$

Also note that the general SDE may be transformed to a constant σ SDE using the Lamperti transform. For example, consider the Vasicek model: $f(x) = \theta_1(\theta_2 - x)$, $g(x) = \theta_3$. Take $\theta_1 = 3, \theta_2 = 0.7$ and $\theta_3 = 0.5$. Use pmle = "Ozaki" option in SimDiff.Proc in R.

1.10.3 Shoji-Ozaki Method

The Shoji–Ozaki method is an extension of the method to Ozaki. It is a more general case where the drift is allowed to depend on the time variable t, and also the diffusion coefficient can be varied. See [170] for more details of this methodology.

1.10.4 Kessler Method

Kessler [120] proposed to use a higher order It $\hat{o}$-Taylor expansion to approximate the mean and variance in a conditional Gaussian density. For example, consider the Hull White model:

$$dX_t = a(t)(b(t) - X_t)dt + \sigma(t)dW_t$$

For this example, take $a(t) = \theta_1 t$, $b(t) = \theta_2\sqrt{t}$, and $\sigma(t) = \theta_3 t$ where $\theta_1 = 2, \theta_2 = 0,7, \theta_3 = 0.8$, and $\Delta t = 0.001$. We refer to [170] for more details. An implementation of all these methods may be found in the Sim.DiffProc package in R [26].

Now we briefly discuss the second method of estimating parameters for a stochastic process by using option data. For this method, we obtain the option data $C_1, \ldots, C_n$ and assume a model (CIR, ...). In this model, we calculate a formula for the option price $C(K, T, \theta)$, i.e.

$$\min_\theta \sum_{i=1}^{n} (C(K, T, \theta) - C_i)^2$$

to obtain θ. Please refer to [83] for more details.

We now present some numerical approximation methods that are used to evaluate well defined integrals.

1.11 Quadrature Methods

Quadrature methods allow one to evaluate numerically an integral which is well defined (i.e. has a finite value).

To motivate the need of such methods suppose the price process follows a general dynamics of the form

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t, \quad X_0 = x.$$

The general formula for the price of a European-type derivative at time t is

$$F(t, X_t) = \mathbb{E}\left[e^{-\int_t^T r(s)ds} F(T, X_T)|F_t\right]$$

where $F(T, X_T)$ is the terminal payoff of the option at maturity time $t = T$. Suppose we can find the transition probability density for the diffusion process $X_t : f(\Delta, x, t, y)$. That is the density of the probability of going from $X_0 = x$ at time Δ to $X_t = y$ at time t. For the geometric Brownian motion, that is, the process $\frac{dS_t}{S_t} = rdt + \sigma dWt$, we can calculate this transition explicitly:

$$f(\Delta, x, t, y) = \frac{1}{y\sqrt{2\pi\sigma^2(t-\Delta)}} e^{\frac{-1}{2}\left(\frac{\ln y - \ln x - (r - \frac{\sigma^2}{2})(t-\Delta)}{\sigma\sqrt{t-\Delta}}\right)^2}.$$

In general if we know this function $f(\Delta, x, t, y)$ and we use the notation $P(t, T) = e^{-\int_t^T r(s)ds}$, we may rewrite the expectation as

$$F(t, X_t) = F(t, x) = \int_0^\infty P(t, T)F(T, y)f(\Delta, x, t, y)dy.$$

To now compute the option price we need to calculate this integral. This is usually hard and we typically do it in a numerical way. This is where the quadrature methods become useful.

The Problem Suppose we have a real function f. Consider

$$I(f) = \int_A f(x)dx$$

where $A \subset \mathbb{R}$ is some subset of the real axis.

Definition 1.11.1 A quadrature rule of order n is an expression of the form

$$I_n(f) = \sum_{i=1}^n w_i^n f(x_i^n)$$

where w_i^n are called weights and x_i^n are called abscissa or quadrature nodes such that $I_n(f) \to I(f)$ as $n \to \infty$.

The basic rules of constructing a quadrature approximation I_n are:

1) Approximate f by some interpolating function (polynomials) of order n such that $P_n(x_i^n) = f(x_i^n)$ for all i.
2) Integrate P_n and return $I(P_n)$ as an approximation to $I(f)$.

1.11.1 Rectangle Rule: ($n = 1$) (Darboux Sums)

For the rectangle rule, the main idea is to approximate f with a piece-wise constant function. Suppose we integrate on an interval $[a, b]$: take

$$\begin{aligned} x_0 &= a, \\ x_1 &= x_0 + \left(\frac{b-a}{n}\right) \\ x_i &= x_0 + \frac{b-a}{n} i \\ x_n &= b. \end{aligned}$$

Note that we are using equidistant points but we may use any points in the interval as long as the maximum interval $\max_i |x_i - x_{i-1}|$ is going to 0 with n.

Next we calculate $f(x_i)$ and we set

$$I_n(f) = \sum_{i=0}^{n-1} hf(x_i) \text{ where } h = \frac{b-a}{n}.$$

Here is a pseudocode that accomplishes the task:

- Input a, b, n, f
- Set $h = \frac{b-a}{n}$
- Set sum $= 0$
- For $i = 0$ to $n - 1$
  sum $=$ sum $+ h * f(a + ih)$
  $i \to i + 1$
  Return sum

If $a = \infty$ or $b = \infty$ the algorithm needs to be suitably modified to avoid calculating $f(a)$ or $f(b)$.

Since

$$I(f) = \lim_{||\pi \to 0||} \sum_{i=1}^{n} |x_i - x_{i-1}| f(\xi_i),$$

where $\pi = x_0, x_1, \ldots, x_n$, $||\pi|| = \min_i |x_i - x_{i-1}|$ where ξ_i is any number in the interval $[x_{i-1}, x_i]$, then in fact the algorithm works not only with the endpoints but also with any point in the interval $\xi_i \in [x_{i-1}, x_i]$. In particular the next rule uses the midpoint in the interval.

1.11.2 Midpoint Rule

The midpoint rule is the same as the rectangular rule but in this case, $I_n(f)$ is defined as

$$I_n(f) = \sum_{i=1}^{n} hf\left(\frac{x_i + x_{i-1}}{2}\right)$$

where $h = \frac{b-a}{n}$.

1.11.3 Trapezoid Rule

The idea is to use both endpoints of the interval to create a trapezoid. This is a linear approximation with a polynomial degree 1. Specifically, $I_n(f)$ is defined as

$$I_n(f) = \sum_{i=1}^{n} h\left(\frac{f(x_i) + f(x_{i-1})}{2}\right) = h\left(\frac{1}{2}f(x_0) + f(x_1) + \ldots + \frac{1}{2}f(x_n)\right).$$

The error of approximation is often quoted as $O(h^2)$ but this is an upper bound. Depending on the function the actual error may be quite a bit better.

1.11.4 Simpson's Rule

For this quadrature method, we approximate the function using a quadratic polynomial. We use two intervals to construct a second order polynomial so that $f(x_{i-1}) = P(x_{i-1}), f(x_i) = P(x_i)$ and $f(x_{i+1}) = P(x_{i+1})$.

1.11.4.1 Lagrange Interpolating Polynomial

Given $(x_1, y_1), \ldots, (x_k, y_k)$ all different, let

$$l_j(x) = \prod_{\substack{i=1 \\ i \neq j}}^{n} \frac{(x - x_j)}{(x_j - x_i)} = \frac{(x - x_1)}{(x_j - x_1)} \frac{(x - x_{j-1})}{(x_j - x_{j-1)}} \frac{(x - x_{j+1})}{(x_j - x_{j+1})} \cdots \frac{(x - x_n)}{(x_j - x_n)}$$

Then the polynomial $L(x) = \sum_{j=1}^{k} y_j l_j(x)$ passes through each of the points $(x_1, y_1), \ldots, (x_n, y_n)$, i.e. $L(x_j) = y_j$.

Going back and translating to the Simpson rule notation the quadrinomial interpolating polynomial is

$$\begin{aligned} P(x) = f(x_{i-1}) &\frac{(x - x_i)}{(x_{i-1} - x_i)} \frac{(x - x_{i+1})}{(x_{i-1} - x_{i+1})} + f(x_i) \frac{(x - x_{i-1})}{(x_i - x_{i-1})} \frac{(x - x_{i+1})}{(x_i - x_{i+1})} \\ &+ f(x_{i+1}) \frac{(x - x_{i-1})}{(x_{i+1} - x_{i-1})} \frac{(x - x_i)}{(x_{i+1} - x_i)}, \end{aligned}$$

So now we approximate $\int_{x_i}^{x_{i+1}} f(x)dx$ with $\int_{x_i}^{x_{i+1}} P(x)dx$. After integrating, we obtain assuming equidistant points $x_i - x_{i-1} = x_{i+1} - x_i$:

$$\int_{x_i}^{x_{i+1}} f(x)dx \simeq \frac{x_{i+1} - x_{i-1}}{6} \left(f(x_{i-1}) + 4f(x_i) + f(x_{i+1})\right).$$

In this exposition we used two consecutive intervals to calculate the approximating polynomial. However, we do not need to do this if we simply use the midpoint in each interval as the third point. That is, if we take the points x_i, $\frac{X_i + x_{i+1}}{2}$, and x_{i+1} to replace the points x_{i-1}, x_i, and x_{i+1} in the previous expression and we consider $x_{i+1} - x_{i-1} = h$ we obtain

$$I(f) \simeq \frac{h}{6} \sum_{i=1}^{n} \left(f(x_{i-1}) + 4f\left(\frac{x_i + x_{i-1}}{2}\right) + f(x_{i+1})\right).$$

There are many other quadrature rule variants; we stop here from the exposition and refer the reader for example to [78].

1.12 Problems

1. Prove that the random variables X_i's in Proposition 1.5.0 are in fact independent.
2. Give a general proof of parts 3) and 4) in Proposition 1.5.0 for any $n, k \in \mathbb{N}$.
3. Show that the equality of sets in part 5) of Proposition 1.5.0 holds by double inclusion.
4. Prove parts 6) and 7) of Proposition 1.5.0 by applying the CLT.
5. Prove part 9) of Proposition 1.5.0.
6. These exercises are due to [54]. Consider an infinite Bernoulli process with $p = 0.5$, that is, an infinite sequence of random variables $\{Y_i, i \in \mathbb{Z}\}$ with $\mathbf{P}(Y_i = 0) = \mathbf{P}(Y_i = 1) = 0.5$, for all $i \in \mathbb{Z}$. We would like to study the length of the maximum sequence of 1's. Let

   $$l_m = \max\{i \geq 1 : X_{m-i+1} = \cdots = X_m = 1\},$$

   be the length of the run of 1's up to the m-th toss and including it. Obviously, l_m will be 0 if the m-th toss is a tail. We are interested in the asymptotic behavior of the longest run from 1 to n for large n.
   That is we are interested in the behavior of L_n where

   $$\begin{aligned} L_n &= \max_{m \in \{1,\ldots,n\}} l_m \\ &= \max\{i \geq 1 : X_{m-i+1} = \cdots = X_m = 1, \text{ for some } m \in \{1, \ldots, n\}\} \end{aligned}$$

   (a) Explain why $\mathbf{P}(l_m = i) = 2^{-(i+1)}$, for $i = 0, 1, 2, \ldots$ and for any m.
   (b) Apply the first Borel–Cantelli lemma to the events

   $$A_n = \{l_n > (1 + \varepsilon)\log_2 n\}.$$

Conclude that for each $\varepsilon > 0$, with probability one, $l_n \leq (1+\varepsilon)\log_2 n$ for all n large enough.

Taking a countable sequence $\varepsilon_k \downarrow 0$ deduce that

$$\underset{n\to\infty}{limsup} \frac{L_n}{\log_2 n} \leq 1, \quad \text{a.s.}$$

(c) Fixing $\varepsilon > 0$ and letting $A_n = \{L_n < k_n\}$ for $k_n = (1-\varepsilon)\log_2 n$. Explain why

$$A_n \subseteq \bigcap_{i=1}^{m_n} B_i^c,$$

where $m_n = [n/k_n]$ (integer part) and $B_i = \{X_{(i-1)k_n+1} = \ldots = X_{ik_n} = 1\}$ are independent events.

Deduce that $\mathbf{P}(A_n) \leq \mathbf{P}(B_i^c)^{m_n} \leq \exp(-n^\epsilon/(2\log_2 n))$, for all n large enough.

(d) Apply the first Borel–Cantelli for the events A_n defined in problem 6(c), followed by $\varepsilon \downarrow 0$, to conclude that

$$\underset{n\to\infty}{liminf} \frac{L_n}{\log_2 n} \geq 1 \quad \text{a.s.}$$

(e) Combine problems 6(b) and 6(d) together to conclude that

$$\frac{L_n}{\log_2 n} \to 1 \quad \text{a.s.}$$

Therefore the length of the maximum sequence of heads is approximately equal to $\log_2 n$ when the number of tosses n is large enough.

7. Let Z be a Brownian motion defined in $[0, T]$. Given a partition $\mathscr{P}$ such that $0 = t_0 < t_1 < \ldots < t_n = T$, we define

$$V_{\mathscr{P}}(Z) = \sum_{j=0}^{n-1} (Z(t_{j+1}) - Z(t_j))^2$$

and the quadratic variation of Z as the limit (when it exists)

$$VC(Z) = \lim_{|\mathscr{P}|\to 0} V_{\mathscr{P}}(Z)$$

Prove that:

(a) $E[(Z(t_{j+1}) - Z(t_j))^2] = t_{j+1} - t_j$. Conclude that $E(V_{\mathscr{P}}(Z)) = T$.

(b) $\text{Var}[(Z(t_{j+1}) - Z(t_j))^2] = 2(t_{j+1} - t_j)^2$, and then

$$\text{Var}(V_{\mathscr{P}}(Z)) = \sum_{j=0}^{n-1} 2(t_{j+1} - t_j)^2 \to 0 \qquad \text{for} |\mathscr{P}| \to 0.$$

(c) Tchebycheff inequality: if X is an stochastic variable with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$, then for all $\varepsilon > 0$ we have that

$$P(|X - \mu| \geq \varepsilon) \leq \left(\frac{\sigma}{\varepsilon}\right)^2$$

Deduce that $VC(Z) = T$ with probability 1, i.e.:

$$P(|VC(Z) - T| \geq \varepsilon) = 0$$

for all $\varepsilon > 0$. Conclude that with probability 1, Z is not differentiable in any interval $[t, t + a]$.

8. Suppose that the price of an asset follows a Brownian motion :

$$dS = \mu S dt + \sigma S dz.$$

(a) What is the stochastic process for S^n?
(b) What is the expected value for S^n?

9. The Hull White model is

$$dX_t = a(t)(b(t) - X_t)dt + \sigma(t)dW_t.$$

In this problem take $a(t) = \theta_1 t$, $b(t) = \theta_2\sqrt{t}$ and $\sigma(t) = \theta_3 t$ where $\theta_1 = 2, \theta_2 = 0,7, \theta_3 = 0.8$, and $\Delta t = 0.001$.

(a) Generate a single path of the process by choosing a $\Delta t = 0.001$ from $t = 0$ to $t = 1$.
(b) Use the Sim.DiffProc package in R to estimate the parameters θ_1, θ_2, θ_3. Use the Euler method and compare with the known values of the parameters.
(c) Repeat part (b) with all the methods available in the package. Write a conclusion based on the results obtained.

10. The variance process in the Heston model satisfy a CIR process:

$$dV_t = \kappa(\overline{V} - V_t) + \sigma\sqrt{V_t}dW_t.$$

Use Ito to calculate the dynamics of the volatility process $U_t = \sqrt{V_t}$. Under which conditions on the parameters $\kappa, \overline{V}$, and σ the process becomes an Ørnstein–Uhlenbeck process, i.e. of the form

$$dU_t = \gamma X_t dt + \delta dW_t$$

for some parameters δ and γ?