Probabilistic Tools for Analysts

Jonathan Mui[*]

latest update: 6 April 2023

Abstract

This is a brief set of notes written for graduate students who have a strong foundation in analysis (real and complex analysis, measure theory and functional analysis), but who have had little exposure to stochastic analysis. It covers some of the most fundamental tools in this field — martingales, stopping times, Brownian motion, and Itô calculus — together with the basic theory of (ordinary) stochastic differential equations. Hopefully these notes will serve as a convenient starting point for further study of the subject itself, and for the application of probabilistic methods to other areas of analysis (partial differential equations, harmonic analysis, geometric analysis, dynamical systems, . . .). Much of the material is based on the lectures for the honours/masters level courses STAT4528 and MATH4512 at the University of Sydney in 2021, taught by Dr. Anna Aksamit and Prof. Ben Goldys.

Contents

[*] University of Sydney, jonathan.mui@sydney.edu.au

1 Foundations

1.1 Measures and algebras

While modern probability is built on the foundation of measure theory, it is certainly not merely a branch of measure theory. In the wise words of Terry Tao,

> At a purely formal level, one could call probability theory the study of measure spaces with total measure one, but that would be like calling number theory the study of strings of digits which terminate.[1]

We will therefore focus on measure theoretic concepts which have a distinct 'probabilistic' flavour (we refrain from trying to define this term rigorously), and which are particularly useful for applications. As an analogy[2], in differential geometry it is essential to emphasise concepts and constructions which are independent of a chosen coordinate system — such objects are considered to be 'truly' geometric in nature.

Definition 1.1. Let Ω be a non-empty set. A **σ-algebra** on Ω is a family $\mathcal{F}$ of subsets of Ω satisfying

(i) $\Omega \in \mathcal{F}$;

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;

(iii) If $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

The pair $(\Omega, \mathcal{F})$ is called a **measurable space**.

If, instead of (iii), the family $\mathcal{F}$ is only closed under finite unions, then $\mathcal{F}$ is called an **algebra** on Ω. Note that in the probability literature, (σ-)algebras are often called (σ-)**fields** instead.

In the probabilistic interpretation, a σ-algebra $\mathcal{F}$ should contain all possible information about a random process. Thus Ω is called the **sample space**, and elements of $\mathcal{F}$ are called **events**.

If $\{\mathcal{F}_i\}_{i\in I}$ is a family of σ-algebras on Ω, then it is very easy to check that the intersection $\mathcal{F} := \bigcap_{i\in I} \mathcal{F}_i$ is also a σ-algebra on Ω. (However, the union of σ-algebras in general does not form a σ-algebra — we leave it as an exercise to find a simple counterexample).

Definition 1.2. If $\mathcal{G}$ is a family of subsets of Ω, we define $\sigma(\mathcal{G})$ to be the smallest σ-algebra containing $\mathcal{G}$. Equivalently $\sigma(\mathcal{G})$ is the intersection of all σ-algebras on Ω containing $\mathcal{G}$. We say that $\sigma(\mathcal{G})$ is the σ-algebra **generated by** $\mathcal{G}$.

Example 1.1. If E is a topological space, then the **Borel σ-algebra** $\mathscr{B}(E)$ is the σ-algebra generated by the open subsets (equivalently, the closed subsets) of E.

Traditionally, when taking a first course in measure theory, we begin by defining measures on σ-algebras. Let us recall:

Definition 1.3. Let $(\Omega, \mathcal{F})$ be a measurable space. A map $\mu : \mathcal{F} \to [0, \infty]$ is called a **measure** if

(i) $\mu(\emptyset) = 0$;

(ii) If $A_k \in \mathcal{F}, k \in \mathbb{N}$ are mutually disjoint (i.e. $A_k \cap A_j = \emptyset$ if $k \neq j$), then

$$\mu\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k).$$

[1] Source: p. 2 of *Topics in Random Matrix Theory*.

[2] Also shamelessly taken from Tao's book on random matrix theory.

The triple $(\Omega, \mathcal{F}, \mu)$ is called a **measure space**. If $\mu(\Omega) < \infty$, μ is said to be a **finite measure**, and if $\mu(\Omega) = 1$, μ is a **probability measure**. In that case, we write $\mathbf{P} := \mu$. Finally, if $(\Omega, \mathcal{F}, \mu)$ is called a σ-finite measure space, and μ is a σ-finite measure, if Ω can be written as a countable union of elements in $\mathcal{F}$ with finite μ-measure.

Remark 1.2. If a certain statement A holds with probability 1, we say that A holds **almost surely** (or $\mathbf{P}$-almost surely, if we need to specify which probability measure), which is abbreviated 'a.s'.

We recall the following basic continuity properties.

Proposition 1.3 (Continuity properties of measures)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

(i) *Let $(A_n)_{n=1}^{\infty}$ be an increasing sequence of sets in $\mathcal{F}$, i.e. $A_n \subseteq A_{n+1}$ for all $n \in \mathbb{N}$. Define $A := \bigcup_{n=1}^{\infty} A_n$. Then $\lim_{n\to\infty} \mu(A_n) = \mu(A)$ and the convergence is monotone.*

(ii) *Let $(A_n)_{n=1}^{\infty}$ be a decreasing sequence of sets in $\mathcal{F}$, i.e $A_n \supseteq A_{n+1}$ for all $n \in \mathbb{N}$, and suppose $\mu(A_{n_0}) < \infty$ for some n_0. Define $A := \bigcap_{n=1}^{\infty} A_n$. Then $\lim_{n\to\infty} \mu(A_n) = \mu(A)$ and the convergence is monotone.*

Proof. (i): Firstly, we rewrite A as a countable disjoint union. Define $A_0 = \emptyset$ and $D_n = A_n \setminus A_{n-1}$. Then it is clear that the D_n are disjoint, and $A = \bigcup_{n=1}^{\infty} D_n$. Since $A_n = \bigcup_{k=1}^{n} D_n$, by the finite additivity of measures, we have $\sum_{k=1}^{n} D_k = \mu(A_n)$. Hence

$$\mu\left(\bigcup_{n=1}^{\infty} D_n\right) = \sum_{n=1}^{\infty} \mu(D_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(D_k) = \lim_{n\to\infty} \mu(A_n)$$

using countable additivity.

(ii): Since we are interested in $n \to \infty$, we may assume $\mu(A_1) < \infty$ without loss of generality. Note that the sequence $(A_1 \setminus A_n)_{n=1}^{\infty}$ is increasing and $\bigcup_{n=1}^{\infty}(A_1 \setminus A_n) = A_1 \setminus A$. By part (i) we deduce

$$\begin{aligned}\lim_{n\to\infty} \mu(A_n) &= \lim_{n\to\infty} (\mu(A_1) - \mu(A_1 \setminus A_n)) \\ &= \mu(A_1) - \mu(A_1 \setminus A) = \mu(A_1) - (\mu(A_1) - \mu(A)) = \mu(A).\end{aligned}$$ □

Exercise 1.1. (i) Give an example to show that the finiteness assumption in Proposition 1.3(ii) cannot be omitted.

(ii) Show that for finite measure spaces, the defining axioms for a measure are equivalent to the following conditions:

(a) $\mu(\emptyset) = 0$;

(b) If $A, B \in \mathcal{F}$ are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$.

(c) For every sequence of decreasing sets $(A_n)_{n=1}^{\infty} \subset \mathcal{F}$, it holds that $\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \mu(A)$, where $A := \bigcap_{n\in\mathbb{N}} A_n$.

One problem is that the σ-algebras we often encounter in applications are extremely large families of sets (try to imagine the family of all Borel sets of $\mathbb{R}$, for example!), and it is impractical to prove that certain statements are true for *all* events. A more 'probabilistic' approach is to find a 'convenient' family $\mathcal{G}$ of subsets — usually much smaller than a σ-algebra — of the sample space Ω which generates a sufficiently rich σ-algebra. The following exercise illustrates a simple but important example of this way of thinking.

Exercise 1.2. Let $\Omega = \mathbb{R}$, and consider the family of subsets

$$\mathcal{G} := \{(-\infty, r] \subset \mathbb{R} : r \in \mathbb{Q}\}.$$

Show that $\sigma(\mathcal{G}) = \mathscr{B}(\mathbb{R})$.

Thus, in practice, the goal is firstly to work with measures on a smaller family $\mathcal{G}$, then extend to the larger σ-algebra. There are many technical tools in probability theory developed for this very purpose.

Definition 1.4. Let Ω be a non-empty set. A family $\mathcal{K}$ of subsets of Ω is called a **π-system** if $\emptyset \in \mathcal{K}$, and $\mathcal{K}$ is closed under intersections.

A family $\mathcal{L}$ of subsets of Ω is called a **λ-system** if

(λ0) $\emptyset \in \mathcal{L}$;

(λ1) $A \in \mathcal{L} \implies A^c \in \mathcal{L}$;

(λ2) If $A_k \in \mathcal{L}, k \in \mathbb{N}$, are mutually disjoint, then $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{L}$.

Clearly a σ-algebra is a λ-system. Notice that a λ-system is 'almost' a σ-algebra, the key difference is in condition (λ2).

Exercise 1.3. (i) Give an example of a λ-system which is not a σ-algebra.

(ii) Show that a family $\mathcal{F}$ of subsets of Ω is a σ-algebra if and only if $\mathcal{F}$ is both a π-system and a λ-system.

Theorem 1.4 (Dynkin $\pi\lambda$ theorem)**.** *Let Ω be a non-empty subset, and suppose $\mathcal{K}$ is a π-system on Ω. If $\mathcal{L}$ is the smallest λ-system containing $\mathcal{K}$, then*

$$\sigma(\mathcal{K}) = \mathcal{L}.$$

Proof. The idea is to show that $\mathcal{L}$ is actually a π-system as well. Then by Exercise 1.3, $\mathcal{L}$ is a σ-algebra. However, by the assumption, it is then the smallest σ-algebra containing $\mathcal{K}$, and hence the theorem is proved.

Clearly $\emptyset \in \mathcal{K} \subseteq \mathcal{L}$. Fix an arbitrary $B \in \mathcal{L}$, and define

$$L_B := \{A \in \mathcal{L} : A \cap B \in \mathcal{L}\} \subseteq \mathcal{L}.$$

We will show that $L_B = \mathcal{L}$ for all $B \in \mathcal{L}$. Observe that if $A \in L_B$, then $A^c \cap B = (A \cap B)^c \cap B \in \mathcal{L}$. This shows that $A^c \in L_B$, so L_B satisfies condition (λ1). Now defining L_A analogously (for any $A \in \mathcal{L}$), we find that L_A satisfies (λ2). Indeed, if $B_k \in L_A, k \in \mathbb{N}$ are mutually disjoint, then $A \cap B_k \in \mathcal{L}$ are mutually disjoint, thus

$$(\bigcup_{k\in\mathbb{N}} B_k) \cap A = \bigcup_{k\in\mathbb{N}} (B_k \cap A) \in \mathcal{L}$$

because $\mathcal{L}$ is a λ-system. If we take $A \in \mathcal{K}$, it follows that $\mathcal{K} \subseteq L_A$ and hence $L_A = \mathcal{L}$ for all $A \in \mathcal{K}$, by the minimality condition of $\mathcal{L}$. In particular, we have proved that

$$A \in \mathcal{K}, B \in \mathcal{L} \implies A \cap B \in \mathcal{L}.$$

This shows that $\mathcal{K} \subseteq L_B$. However, by arguments used above, L_B satisfies (λ2) as well, so L_B is a λ-system. By minimality of $\mathcal{L}$, we have that $L_B = \mathcal{L}$ for all $B \in \mathcal{L}$, which is the desired conclusion. □

One of the most common applications of Dynkin's theorem is to show that if two probability measures defined on the same σ-algebra agree on 'sufficiently many' events, then the measures coincide. Let us prove a more general statement.

Theorem 1.5 (Uniqueness of measures). *Let $\mathcal{K}$ be a π-system on Ω, and let μ, ν be σ-finite measures defined on $\sigma(\mathcal{K})$. Assume that $\mu(B) = \nu(B)$ for all $B \in \mathcal{K}$. If there exists an increasing sequence of sets $(A_n)_{n\in\mathbb{N}} \subset \mathcal{K}$ such that $\bigcup_{n\in\mathbb{N}} A_n = \Omega$, and $\mu(A_n), \nu(A_n) < \infty$ for each $n \in \mathbb{N}$, then $\mu(B) = \nu(B)$ for all $B \in \sigma(\mathcal{K})$.*

Proof. For each $n \in \mathbb{N}$, we define

$$\mathcal{L}_n := \{B \in \sigma(\mathcal{K}) : \mu(A_n \cap B) = \nu(A_n \cap B)\}.$$

Since μ, ν agree on the π-system $\mathcal{K}$, it follows immediately that $\mathcal{K} \subseteq \mathcal{L}_n$ for all $n \in \mathbb{N}$. We need to show that each $\mathcal{L}_n$ is a λ-system.

It is trivial that $\emptyset \in \mathcal{L}_n$. If $B \in \mathcal{L}_n$, then

$$\mu(B^c \cap A) = \mu(A_n) - \mu(A_n \cap B) = \nu(A_n) - \nu(A_n \cap B) = \nu(B^c \cap A),$$

and hence $B^c \in \mathcal{L}_n$. Finally, let $(B_k)_{k\in\mathbb{N}} \subset \mathcal{L}_n$ be a disjoint sequence of sets. Then

$$\begin{aligned}\mu\left(A_n \cap (\bigcup_{k=1}^{\infty} B_k)\right) &= \sum_{k=1}^{\infty} \mu(A_n \cap B_k) \\ &= \sum_{k=1}^{\infty} \nu(A_n \cap B_k) = \nu\left(A_n \cap (\bigcup_{k=1}^{\infty} B_k)\right)\end{aligned}$$

which shows that $\bigcup_{k\in\mathbb{N}} B_k \in \mathcal{L}_n$. Hence each $\mathcal{L}_n$ is a λ-system.

By Dynkin's theorem (1.4), we obtain $\sigma(\mathcal{K}) \subseteq \mathcal{L}_n$ for every $n \in \mathbb{N}$. Hence $\mu(A_n \cap B) = \nu(A_n \cap B)$ for every $B \in \sigma(\mathcal{K})$, for all $n \in \mathbb{N}$. The proof is completed by taking $n \to \infty$. □

Corollary 1.6. *Let $(\Omega, \mathcal{F})$ be a measurable space, and suppose $\mathcal{K}$ is a π-system on Ω. If μ, ν are probability measures that agree on $\mathcal{K}$, then μ and ν agree on $\sigma(\mathcal{K})$.*

Proof. Define $\mathcal{L}$ to be the collection of all sets $B \subseteq \Omega$ such that $\mu(B) = \nu(B)$. Then $\mathcal{L}$ is a λ-system containing $\mathcal{K}$. The conclusion follows immediately from Theorem 1.5 by taking the trivial sequence $A_n = \Omega$ for all $n \in \mathbb{N}$, or alternatively it follows directly from Dynkin's theorem as well. □

In some situations, it is convenient to work with a family of subsets that is 'almost' a σ-algebra.

Definition 1.5. Let Ω be a non-empty set. A family $\mathcal{F}_0$ of subsets of Ω is called an **algebra** if

(i) $\Omega \in \mathcal{F}_0$;

(ii) If $A \in \mathcal{F}_0$, then $A^c \in \mathcal{F}_0$;

(iii) If $A_1, \ldots, A_n \in \mathcal{F}_0$, then $\bigcup_{k=1}^{n} A_k \in \mathcal{F}_0$.

A set function $\mu : \mathcal{F}_0 \to [0, \infty]$ is a **measure** on an algebra $\mathcal{F}_0$ if $\mu(\emptyset) = 0$, and if $(A_n)_{n=1}^{\infty}$ is a disjoint family of sets in $\mathcal{F}_0$ such that $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}_0$ (note that this is an assumption!), then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

As we can plainly see, the key difference is that an algebra is only closed under finite unions. It is also clear that an algebra is a π-system.

So far, we have started with a measure defined on a σ-algebra. However, since σ-algebras are typically very large families of sets, it is often more convenient in practice to define a measure on an algebra $\mathcal{F}_0$. We may then ask if such a measure can be extended to a σ-algebra containing $\mathcal{F}_0$. An affirmative answer is given by a fundamental theorem due to Carathéodory.

To include: Caratheodory extension theorem; Product measure

1.2 Random variables and independence

Definition 1.6. Let $(\Omega, \mathcal{F})$ and $(E, \mathcal{G})$ be measurable spaces. An E-valued **random variable** on Ω is a **measurable** function $X : \Omega \to E$. Measurability means that

$$X^{-1}(A) \in \mathcal{F} \text{ for all } A \in \mathcal{G}. \tag{1.1}$$

We often say that X is $\mathcal{F}$-measurable, especially in cases when we work with multiple σ-algebras on the same sample space. The codomain of the mapping $X : \Omega \to E$ is often called the **state space**, especially when studying stochastic differential equations and dynamical systems. In elementary probability, by far the most common choice of the state space $(E, \mathcal{G})$ is $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. In fact, the name 'random variable' is often reserved for measurable functions $X : \Omega \to \mathbb{R}^n$. For more advanced topics in analysis, E is typically a complete, separable metric space (often called a **Polish space**), and $\mathcal{G} = \mathscr{B}(E)$ is the σ-algebra of Borel subsets of E.

Exercise 1.4. Let $(\Omega, \mathcal{F}), (E, \mathcal{G})$ be measurable spaces, and $X : \Omega \to E$ a random variable. Show that

$$X^{-1}(\mathcal{G}) := \{X^{-1}(A) : A \in \mathcal{G}\}$$

is a σ-algebra on Ω.

The σ-algebra constructed in Exercise 1.4 is the smallest σ-algebra with respect to which X is measurable. We have a special notation reserved for this construction.

Definition 1.7. Let $(\Omega, \mathcal{F}), (E, \mathcal{G})$ be measurable spaces, and $X : \Omega \to E$ a random variable. We define

$$\sigma(X) := X^{-1}(\mathcal{G})$$

and call it the σ-algebra **generated by** X.

More generally, if $\{X_\alpha\}_{\alpha \in I}$ is a family of E-valued random variables on Ω, the σ-algebra generated by the family $\{X_\alpha\}_{\alpha \in I}$ is defined as

$$\sigma(X_\alpha : \alpha \in I) := \sigma(\{\sigma(X_\alpha)\}_{\alpha \in I}).$$

The reader who is thinking probabilistically should already suspect that in order to prove that a given function is a random variable, one does not need to check condition (1.1) for *all* $A \in \mathcal{G}$.

Proposition 1.7. *Let $(\Omega, \mathcal{F}), (E, \mathcal{G})$ be measurable spaces, and suppose that $\mathcal{G}_0$ is a family of subsets of E such that $\mathcal{G} = \sigma(\mathcal{G}_0)$. Then $X : \Omega \to E$ is a random variable if and only if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{G}_0$.*

Proof. Assume that $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{G}_0$. Consider the family of subsets

$$\mathcal{G}' := \{A \in E : X^{-1}(A) \in \mathcal{F}\}.$$

It is easy to verify that $\mathcal{G}'$ is a σ-algebra on E. By assumption, we have $\mathcal{G}_0 \subseteq \mathcal{G}'$. It follows that

$$\mathcal{G} = \sigma(\mathcal{G}_0) \subseteq \sigma(\mathcal{G}') = \mathcal{G}',$$

but this implies that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{G}$. □

Using the above proposition, we can obtain the following result which is often used as a definition in basic probability courses. We leave the proof as a simple exercise.

Corollary 1.8. *A map $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is a random variable if and only if*

$$(X \leq r) := X^{-1}((-\infty, r]) \in \mathcal{F} \text{ for all } r \in \mathbb{R}.$$

More generally, if X is an E-valued random variable, where $(E, \mathcal{G})$ is a measurable space, we can write

$$(X \in A) := \{\omega \in \Omega : X(\omega) \in A\} = X^{-1}(A) \quad \text{for all } A \in \mathcal{G}.$$

We have the following essential concept.

Definition 1.8. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and $(E, \mathcal{G})$ a measurable space. If $X : \Omega \to E$ is a random variable, then the **law** or **distribution** of X is denoted by $\mathbf{P}_X$ or $X_{\#}\mathbf{P}$, and is defined to be the measure

$$X_{\#}\mathbf{P}(A) := \mathbf{P}(X \in A) = \mathbf{P}(\{\omega \in \Omega : X(\omega) \in A\}), \quad \forall\, A \in \mathcal{G}.$$

It is also called the **pushforward** of $\mathbf{P}$ by X.

If $(E, \mathcal{G}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, then the function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(t) := \mathbf{P}(X \leq t)$$

is called the **distribution function** of X (or often, simply the **distribution** of X).

Another uniquely probabilistic concept is that of *independence* of events and random variables. Recall the following definition from elementary probability: if A, B are events with $\mathbf{P}(A) > 0$, then the *conditional probability* of B given A, denoted by $\mathbf{P}(B|A)$, is defined as

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

We say that the events A, B are independent if $\mathbf{P}(B|A) = \mathbf{P}(B)$. In other words, knowledge that A occurred tells us nothing more about the probability of B. From the above formula, we see that this is equivalent to

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

The above identity is then taken to be the definition of independence for a pair of events. Although we will introduce more abstract notions of independence, it is nonetheless useful to keep the elementary perspective in mind.

Definition 1.9. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A finite collection of events $\{A_i\}_{i=1}^n \subset \mathcal{F}$ is **independent** if for any subset of distinct indices $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$, it holds that

$$\mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbf{P}(A_{i_j}).$$

An infinite (possibly uncountable) collection of events $\{A_i\}_{i \in I} \subset \mathcal{F}$ is independent if every finite sub-collection $\{A_1, \ldots, A_n\}$ is independent in the sense defined above.

The following example shows that it is necessary to consider all finite sub-collections of the sets $A_1, \dots, A_n$.

Example 1.9. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, and consider the events $A_1 = \{1, 2, 3, 4\}$, $A_2 = A_3 = \{4, 5, 6\}$. Clearly these events are not independent, but nevertheless

$$\mathbf{P}(A_1 \cap A_2 \cap A_3) = \mathbf{P}(A_1)\mathbf{P}(A_2)\mathbf{P}(A_3) = \frac{1}{6}.$$

Let us extend the notion of independence to σ-algebras and random variables.

Definition 1.10. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A family $\{\mathcal{F}_i\}_{i \in I}$ of sub-σ-algebras of $\mathcal{F}$ is called **independent** if for every finite index set $J \subseteq I$ and every collection of events $\{A_j\}_{j \in J}$ with $A_j \in \mathcal{F}_j$, it holds that

$$\mathbf{P}\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}(A_j). \tag{1.2}$$

A family of $\{X_i\}_{i \in I}$ of random variables on Ω is independent if the corresponding family of σ-algebras $\{\sigma(X_i)\}_{i \in I}$ is independent in the sense defined above.

Exercise 1.5. Let $X_1, \dots, X_n$ be independent random variables on $(\Omega, \mathcal{F}, \mathbf{P})$. Find the distribution functions of the random variables $Y := \min_{1 \le k \le n} X_k$ and $Z := \max_{1 \le k \le n} X_k$.

Once again, we would like criteria for independence that involves smaller families of events. Unsurprisingly, the following result can be derived from Dynkin's theorem.

Proposition 1.10. *Let $\{\mathcal{K}_i\}_{i \in I}$ be a collection of π-systems on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and denote $\mathcal{F}_i := \sigma(\mathcal{K}_i)$. The σ-algebras $\{\mathcal{F}_i\}_{i \in I}$ are independent if for every finite index set $J \subseteq I$ and for any collection of events $A_j \in \mathcal{K}_j, j \in J$, identity (1.2) holds.*

Proof. Let $i_1, i_2 \in I$ and fix an arbitrary $A_2 \in \mathcal{K}_{i_2}$. Define

$$\mathcal{G} := \{A_1 \in \mathcal{F}_{i_1} : \mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2)\}.$$

We leave it as an exercise to check that $\mathcal{G}$ satisfies the conditions of Dynkin's theorem (1.4). Consequently $\mathcal{G} = \sigma(\mathcal{K}_{i_1}) = \mathcal{F}_{i_1}$. By an analogous argument for $\mathcal{K}_{i_2}$ (fixing an arbitrary $A_1 \in \mathcal{K}_{i_1}$), we find

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2) \quad \forall A_1 \in \mathcal{K}_{i_1}, \forall A_2 \in \mathcal{K}_{i_2}.$$

For an arbitrary finite index set $\{i_1, \dots, i_n\} \subseteq I$, we simply proceed by induction. □

The previous proposition yields the following extremely useful corollary.

Proposition 1.11. *The random variables $X_1, \dots, X_n : \Omega \to \mathbb{R}$ are independent if and only if*

$$\mathbf{P}\left(\bigcap_{k=1}^{n} (X_k \le x_k)\right) = \prod_{k=1}^{n} \mathbf{P}(X_k \le x_k) \tag{1.3}$$

for all choices of $x_k \in \mathbb{R}$.

We conclude this section by recalling some basic notions associated with $\mathbb{R}^n$-valued random variables.

Definition 1.11. Let $X : \Omega \to \mathbb{R}^n$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The **expectation** of X is the vector defined by

$$\mathbf{E}(X) := \int_\Omega X \, d\mathbf{P} = \int_\Omega X(\omega)\mathbf{P}(d\omega)$$

where the integral is understood in the Lebesgue sense. The random variable X is called **integrable** if $\mathbf{E}|X| < \infty$.

If $m_j := \mathbf{E}(X_j)$, then the **covariance matrix** Q of X is defined by

$$Q_{jk} := \mathrm{Cov}(X_j, X_k) := \mathbf{E}[(X_j - m_j)(X_k - m_k)] \quad (j, k = 1, \ldots, n).$$

The **variance** of $X : \Omega \to \mathbb{R}$ is

$$\mathrm{Var}(X) := \mathbf{E}(X - \mathbf{E}(X))^2 = \mathbf{E}(X^2) - \mathbf{E}(X)^2.$$

If X, Y are both real-valued random variables, their covariance is defined by

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

Exercise 1.6 (Important!)**.** Let $X, Y : \Omega \to \mathbb{R}$ be random variables.

(i) Show that if X, Y are independent, then

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y).$$

(ii) Is the converse statement true?

[*Hints*: (i) Verify the identity for indicator functions first. Then recall that non-negative random variables can be approximated from below by simple random variables, and finally, decompose into positive and negative parts to obtain the general result. (ii) No.]

We briefly recall the definitions of the Lebesgue spaces. If $X, Y : \Omega \to \mathbb{R}$ are random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, we can define an equivalence relation

$$X \sim Y \overset{\text{def.}}{\iff} \mathbf{P}(X = Y) = 1.$$

For $1 \leq p < \infty$, the space $L^p(\Omega, \mathcal{F}, \mathbf{P})$, or simply $L^p(\Omega)$ if we suppress the dependence on the measure and σ-algebra, is the vector space of (equivalence classes of) random variables such that

$$\|X\|_p := \left(\int_\Omega |X|^p \, d\mathbf{P} \right)^{1/p} = (\mathbf{E}|X|^p)^{1/p} < \infty.$$

If $p = \infty$, then the space $L^\infty(\Omega, \mathcal{F}, \mathbf{P})$ consists of all *essentially bounded* random variables, i.e. random variables such that

$$\|X\|_\infty := \text{ess sup}_{\omega \in \Omega} |X(\omega)| := \inf\{c \geq 0 : \mathbf{P}(|X(\omega)| \leq c) = 1\} < \infty.$$

For each $1 \leq p \leq \infty$, the functional $\| \cdot \|_p$ defines a norm on $L^p(\Omega)$. It is well-known that each $L^p(\Omega)$ is a Banach space with respect to L^p norm.

Proposition 1.12. *Let $X : \Omega \to \mathbb{R}$ be a random variable. For any non-negative Borel function $f : \mathbb{R} \to [0, \infty)$, it holds that*

$$\mathbf{E}[f(X)] = \int_\mathbb{R} f(x)\mathbf{P}_X(dx). \tag{1.4}$$

Moreover, $f \circ X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ if and only if $f \in L^1(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbf{P}_X)$. In this case, formula (1.4) holds as well.

Proof. We use the standard approximation argument. First assume that $f = \mathbf{1}_A$, where $A \in \mathscr{B}(\mathbb{R})$ (recall that $\mathbf{1}_A$ is the indicator function of an event A). Using the definition of the law $\mathbf{P}_X$, we have

$$\mathbf{E}(\mathbf{1}_A \circ X) = \int_\Omega \mathbf{1}_{(X \in A)}\, d\mathbf{P} = \mathbf{P}(X \in A) = \int_{\mathbb{R}} \mathbf{1}_A(x) \mathbf{P}_X(dx).$$

For a non-negative Borel function f, we approximate from below by simple functions and use the monotone convergence theorem to obtain (1.4). For a general Borel function, we can decompose into positive and negative parts, $f = f^+ - f^-$. Then $f^+ \circ X - f^- \circ X$ is the decomposition of $f \circ X$ into positive and negative parts, since $(f^+ \circ X)(\omega) > 0$ implies that $(f^- \circ X)(\omega) = 0$, and vice versa. The previous argument can then be applied separately to $f^+ \circ X$ and $f^- \circ X$.

The integrals in the formula are finite if and only if $f \circ X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ and $f \in L^1(\mathbb{R}, \mathscr{B}(\mathbb{R}), \mathbf{P}_X)$. □

1.3 Conditional expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and consider an event $A \in \mathcal{F}$ with $\mathbf{P}(A) > 0$. Given a random variable $X : \Omega \to \mathbb{R}$ with finite expectation, we may recall the following elementary formula for the *conditional expectation* of X given A:

$$\mathbf{E}(X|A) = \frac{1}{\mathbf{P}(A)} \int_A X\, d\mathbf{P}. \tag{1.5}$$

Thus $\mathbf{E}(X|A)$ is exactly the average of X over the set A. The intuition is that if we have prior knowledge about the random process modelled by X — e.g. if we know that a certain event A has happened — then we only need to average over that event.

Next, we would like to make sense of the expression $\mathbf{E}(X|Y)$, where Y is now another random variable. This should represent our 'best guess' of the values of the random variable X given the values of Y. In fact, this is a fundamental question of statistics: Y should be viewed as a random sample (or the outcome of an experiment), and we wish to infer information about the phenomenon modelled by X based on the limited information given by Y.

Let us consider the following simple but instructive example.

Example 1.13. Suppose the sample space Ω decomposes as a finite disjoint union $\Omega = \bigcup_{i=1}^n A_i$ of events $A_i \in \mathcal{F}$ such that $\mathbf{P}(A_i) > 0$ for all $i = 1, \ldots, n$, and let Y be the following random variable:

$$Y = \sum_{i=1}^n a_i \mathbf{1}_{A_i}, \qquad a_i \in \mathbb{R}.$$

Without loss of generality we can assume that the numbers a_i are distinct. Since Y is constant on each A_i, if we know the value of $Y(\omega)$, then we know which of the events $A_1, \ldots, A_n$ contains that sample point ω. Given *only* this information, our best guess of the values of X should therefore be the average over the corresponding events. Namely, we define the *random variable*

$$\mathbf{E}(X|Y)(\omega) := \frac{1}{\mathbf{P}(A_i)} \int_{A_i} X\, d\mathbf{P} \quad \text{for } \omega \in A_i \quad (i = 1 \ldots, n). \tag{1.6}$$

This can be rewritten more succinctly as

$$\mathbf{E}(X|Y) = \sum_{i=1}^n \mathbf{E}(X|A_i) \mathbf{1}_{A_i}.$$

The random variable (1.6) has the following properties (which the reader can verify easily):

(i) $\mathbf{E}(X|Y)$ is measurable with respect to $\mathcal{G} := \sigma(A_i : 1 \leq i \leq n)$, the σ-algebra generated by the sets A_i; and

(ii) $\int_B \mathbf{E}(X|Y)\, d\mathbf{P} = \int_B X\, d\mathbf{P}$ for all $B \in \mathcal{G}$. In particular, $\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}(X)$.

From the definition (1.6), one observes that the precise values of Y are not crucial, and it is rather the σ-algebra generated by Y that matters. This is the key insight behind the measure-theoretic treatment of conditional expectation.

Definition 1.12. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. Let $X : \Omega \to \mathbb{R}$ be an integrable random variable. A random variable $Z : \Omega \to \mathbb{R}$ is called a **conditional expectation** given $\mathcal{G}$ if

(i) Z is $\mathcal{G}$-measurable;

(ii) $\int_B Z\, d\mathbf{P} = \int_B X\, d\mathbf{P}$ for all $B \in \mathcal{G}$.

Theorem 1.14. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}|X| < \infty$. Then there exists a $\mathbf{P}$-almost surely unique conditional expectation, denoted by $\mathbf{E}(X|\mathcal{G})$.*

Proof. The proof is a straightforward application of the Radon-Nikodym theorem.

Existence: We first assume $X \geq 0$ almost surely. Consider the probability measure $\mathbf{P}$ restricted to $(\Omega, \mathcal{G})$, and define $\nu(A) := \mathbf{E}(X\mathbf{1}_A)$ for all $A \in \mathcal{G}$, which is clearly a finite measure on $(\Omega, \mathcal{G})$. If $\mathbf{P}(A) = 0$, then $\nu(A) = 0$, so ν is absolutely continuous with respect to $\mathbf{P}$. Hence, by the Radon-Nikodym theorem, there exists a $\mathcal{G}$-measurable function $g \geq 0$ such that

$$\nu(A) = \int_A X\, d\mathbf{P} = \int_A g\, d\mathbf{P} \qquad \forall\, A \in \mathcal{G}.$$

This shows that g is a conditional expectation. For a general X, we consider X^+ and X^- (the positive and negative parts) as usual, and the Radon-Nikodym theorem yields respectively a g^+ and g^-. Then $g = g^+ - g^-$ is the required conditional expectation.

Uniqueness: this is actually part of the Radon-Nikodym theorem as well, but we present the argument separately to highlight a useful technique. Consider conditional expectations Z_1, Z_2. Then both random variables are $\mathcal{G}$-measurable, so

$$B := (Z_1 > Z_2) := \{\omega \in \Omega : Z_1(\omega) > Z_2(\omega)\} \in \mathcal{G}.$$

By definition of conditional expectation, we have

$$\mathbf{E}(Z_1\mathbf{1}_B) = \mathbf{E}(X\mathbf{1}_B) = \mathbf{E}(Z_2\mathbf{1}_B).$$

Then $\mathbf{E}((Z_1 - Z_2)\mathbf{1}_B) = 0$, and since $Z_1 - Z_2 > 0$ on B, we conclude that B has measure 0. The same argument applies for the $\mathcal{G}$-measurable set $B' := (Z_2 > Z_1)$. Thus $Z_1 = Z_2$ holds $\mathbf{P}$-almost surely. □

Exercise 1.7. Suppose we toss a fair coin twice. The sample space for this experiment is $\Omega = \{HH, HT, TH, TT\}$, where H and T denote 'heads' and 'tails' respectively. Let A be the event that heads occurs first, and let X be the random variable that records the number of heads.

(i) Determine $\mathcal{G} := \sigma(\{A\})$.

(ii) Compute $\mathbf{E}(X|\mathcal{G})$ explicitly. [*Solution*: $\mathbf{E}(X|\mathcal{G}) = \frac{3}{2}\mathbf{1}_A + \frac{1}{2}\mathbf{1}_{A^c}$.]

We present some essential properties of the conditional expectation that will be frequently used.

Proposition 1.15. *Let the conditions of Definition 1.12 hold. In the following assertions, all pointwise equalities and inequalities are understood to hold* $\mathbf{P}$*-almost surely.*

(i) *If* X *is* $\mathcal{G}$*-measurable, then* $\mathbf{E}(X|\mathcal{G}) = X$.

(ii) Linearity*: If* X, Y *are integrable, then* $\mathbf{E}(aX + bY|\mathcal{G}) = a\mathbf{E}(X|\mathcal{G}) + b\mathbf{E}(Y|\mathcal{G})$.

(iii) Positivity preserving*: If* $X \geq 0$*, then* $\mathbf{E}(X|\mathcal{G}) \geq 0$.

(iv) Monotonicity*: If* $X_1 \leq X_2$ *and both are integrable, then* $\mathbf{E}(X_1|\mathcal{G}) \leq \mathbf{E}(X_2|\mathcal{G})$.

(v) $\mathbf{E}(X) = \mathbf{E}[\mathbf{E}(X|\mathcal{G})]$.

(vi) 'Tower property'*: If* $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$*, then*

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

(vii) *If* $\mathcal{G}$ *and* $\sigma(X)$ *are independent, then* $\mathbf{E}(X|\mathcal{G}) = \mathbf{E}(X)$.

(viii) *If* $0 \leq X_n \uparrow X$ *as* $n \to \infty$ *with* $\mathbf{E}(X) < \infty$*, then* $\mathbf{E}(X_n|\mathcal{G}) \uparrow \mathbf{E}(X|\mathcal{G})$ *as* $n \to \infty$.

Proof. Assertion (i) is obvious by the uniqueness of conditional expectation. Property (iii) follows from the Radon-Nikodym theorem, and (iv) clearly follows by combining (ii) and (iii). Identity (v) follows by taking $B = \Omega$ in the definition of conditional expectation, and (viii) is a consequence of the monotone convergence theorem — we leave it as an exercise to fill in the details.

(ii): Define $Z := aX + bY$, and let $A \in \mathcal{G}$ be arbitrary. Using the definition of conditional expectation and linearity of the Lebesgue integral, we compute

$$\begin{aligned}\int_A Z\,d\mathbf{P} = \int_A (aX + bY)\,d\mathbf{P} &= a\int_A X\,d\mathbf{P} + b\int_A Y\,d\mathbf{P} \\ &= a\int_A \mathbf{E}(X|\mathcal{G})\,d\mathbf{P} + b\int_A \mathbf{E}(Y|\mathcal{G})\,d\mathbf{P} \\ &= \int_A [a\mathbf{E}(X|\mathcal{G}) + b\mathbf{E}(Y|\mathcal{G})]\,d\mathbf{P}.\end{aligned}$$

Hence

$$\int_A \mathbf{E}(Z|\mathcal{G})\,d\mathbf{P} = \int_A Z\,d\mathbf{P} = \int_A [a\mathbf{E}(X|\mathcal{G}) + b\mathbf{E}(Y|\mathcal{G})]\,d\mathbf{P}.$$

Since the above holds for all $A \in \mathcal{G}$, by the uniqueness of conditional expectations, we conclude $\mathbf{E}(Z|\mathcal{G}) = a\mathbf{E}(X|\mathcal{G}) + b\mathbf{E}(Y|\mathcal{G})$.

(vi): From the definition of conditional expectation, $\mathbf{E}(X|\mathcal{G}_1)$ is $\mathcal{G}_1$-measurable. If $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\mathbf{E}(X|\mathcal{G}_1)$ is $\mathcal{G}_2$-measurable as well, and thus $\mathbf{E}(X|\mathcal{G}_1) = \mathbf{E}[\mathbf{E}(X|\mathcal{G}_1)|\mathcal{G}_2]$ by (i). Now take an arbitrary $A \in \mathcal{G}_1$. We have

$$\int_A \mathbf{E}(X|\mathcal{G}_1)\,d\mathbf{P} = \int_A X\,d\mathbf{P} = \int_A \mathbf{E}(X|\mathcal{G}_2)\,d\mathbf{P}$$

since $A \in \mathcal{G}_2$ as well. This proves that $\mathbf{E}(X|\mathcal{G}_1) = \mathbf{E}[\mathbf{E}(X|\mathcal{G}_2)|\mathcal{G}_1]$, by the uniqueness of conditional expectation.

(vii): If $\mathcal{G}$ and $\sigma(X)$ are independent, then for all $A \in \mathcal{G}$, we have

$$\mathbf{E}[\mathbf{E}(X)\mathbf{1}_A] = \mathbf{E}(X)\mathbf{E}(\mathbf{1}_A) \underset{\text{indep.}}{=} \mathbf{E}(X\mathbf{1}_A) = \mathbf{E}[\mathbf{E}(X|\mathcal{G})\mathbf{1}_A].$$

Hence, by the uniqueness of conditional expectation, we have $\mathbf{E}(X|\mathcal{G}) = \mathbf{E}(X)$. □

Exercise 1.8. Let the conditions of Definition 1.12 hold. In this exercise, all random variables are real-valued.

(i) Consider the Banach space $E = L^1(\Omega, \mathcal{F}, \mathbf{P})$ of random variables such that $\mathbf{E}|X| < \infty$. Define an operator T by

$$T(X) := \mathbf{E}(X|\mathcal{G}), \quad \forall\, X \in E.$$

Show that T is a bounded projection on E (i.e. T is a bounded linear operator on E such that $T^2 = T$). How can the tower property be interpreted in terms of properties of projection operators?

(ii) Prove that if $X \in L^\infty(\Omega, \mathcal{F}, \mathbf{P})$ (i.e. X is almost surely bounded), then

$$\|\mathbf{E}(X|\mathcal{G})\|_\infty \le \|X\|_\infty.$$

(iii) (*Geometric significance of conditional expectation*) Consider the Hilbert space $H = L^2(\Omega, \mathcal{F}, \mathbf{P})$. Check that the subspace $V := L^2(\Omega, \mathcal{G}, \mathbf{P})$ of $\mathcal{G}$-measurable random variables in H is a closed subspace of H. Then prove that

$$\mathbf{E}(X|\mathcal{G}) = \min_{Z \in V} \|X - Z\|_H,$$

and deduce that $\mathbf{E}(X|\mathcal{G})$ is the orthogonal projection of X onto V.

The next result is also a fundamental property of conditional expectation, and it is informally called 'taking out what is known'.

Proposition 1.16. *Suppose $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ and $Y \in L^\infty(\Omega, \mathcal{G}, \mathbf{P})$. Then*

$$\mathbf{E}(XY|\mathcal{G}) = Y\mathbf{E}(X|\mathcal{G}). \tag{1.7}$$

Proof. It is clear that $Y\mathbf{E}(X|\mathcal{G})$ is a $\mathcal{G}$-measurable random variable. We first prove that $\mathbf{E}(XY) = \mathbf{E}(Y\mathbf{E}(X|\mathcal{G}))$ for all $Y \in L^\infty(\Omega, \mathcal{G}, \mathbf{P})$. Indeed, recall from the proof of Theorem 1.14 that $\mathbf{E}(X|\mathcal{G})$ is the density of the measure $\nu(A) := \int_A X\, d\mathbf{P}$ with respect to $\mathbf{P}$. Hence

$$\mathbf{E}(Y\mathbf{E}(X|\mathcal{G})) = \int_\Omega Y\mathbf{E}(X|\mathcal{G})\, d\mathbf{P} = \int_\Omega Y\, d\nu = \int_\Omega XY\, d\mathbf{P} = \mathbf{E}(XY)$$

as claimed.

Now let $A \in \mathcal{G}$ be arbitrary. Then $Y\mathbf{1}_A$ is a bounded, $\mathcal{G}$-measurable random variable, and thus we use the above arguments with Y replaced by $Y\mathbf{1}_A$ to deduce

$$\mathbf{E}[\mathbf{E}(XY|\mathcal{G})\mathbf{1}_A] = \mathbf{E}(XY\mathbf{1}_A) = \mathbf{E}[Y\mathbf{1}_A\mathbf{E}(X|\mathcal{G})].$$

Since this holds for all $A \in \mathcal{G}$, we conclude that $\mathbf{E}(XY|\mathcal{G}) = Y\mathbf{E}(X|\mathcal{G})$. □

Remark 1.17. (i): In the above proof, we first had to show that $\mathbf{E}(XY) = \mathbf{E}(Y\mathbf{E}(X|\mathcal{G}))$ for all $Y \in L^\infty(\Omega, \mathcal{G}, \mathbf{P})$. In fact, this yields an equivalent definition of conditional expectation. More precisely: Z is a $\mathcal{G}$-measurable random variable such that $\mathbf{E}(XY) = \mathbf{E}(ZY)$ for all $Y \in L^\infty(\Omega, \mathcal{G}, \mathbf{P})$ if and only if $Z = \mathbf{E}(X|\mathcal{G})$. The 'if' direction is proved above; the converse follows by taking $Y = \mathbf{1}_A$ in the equation $\mathbf{E}(XY) = \mathbf{E}(ZY)$ and letting A vary over all events in $\mathcal{G}$.

(ii): Proposition 1.16 also holds if $X \in L^p(\Omega, \mathcal{F}, \mathbf{P})$ and $Y \in L^{p'}(\Omega, \mathcal{G}, \mathbf{P})$, where $\frac{1}{p} + \frac{1}{p'} = 1$.

We present Jensen's inequality for conditional expectation.

Theorem 1.18. *Let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbf{E}|X| < \infty$. If $\Phi : \mathbb{R} \to \mathbb{R}$ is a convex function such that $\mathbf{E}|\Phi(X)| < \infty$, then*

$$\Phi(\mathbf{E}(X|\mathcal{G})) \leq \mathbf{E}(\Phi(X)|\mathcal{G}).$$

Proof. Since Φ is convex, for all $x_0 \in \mathbb{R}$ there exists $m_0 \in \mathbb{R}$ such that

$$\Phi(x) - \Phi(x_0) \geq m_0(x - x_0), \qquad \forall\, x \in \mathbb{R}.$$

(That is, Φ lies above its tangent line at x_0). Let $x = X(\omega)$ and $x_0 = \mathbf{E}(X|\mathcal{G})(\omega)$. Then

$$\Phi(X(\omega)) - \Phi(\mathbf{E}(X|\mathcal{G})(\omega)) \geq m_0(X(\omega) - \mathbf{E}(X|\mathcal{G})(\omega)) \quad \forall\, \omega \in \Omega.$$

We apply conditional expectations to both sides to obtain

$$\mathbf{E}(\Phi(X)|\mathcal{G}) - \mathbf{E}[\Phi(\mathbf{E}(X|\mathcal{G}))|\mathcal{G}] \geq m_0 \mathbf{E}\big[X - \mathbf{E}(X|\mathcal{G})|\mathcal{G}\big] = 0.$$

Since $\Phi(\mathbf{E}(X|\mathcal{G}))$ is $\mathcal{G}$-measurable, we conclude

$$\mathbf{E}(\Phi(X)|\mathcal{G}) \geq \mathbf{E}[\Phi(\mathbf{E}(X|\mathcal{G}))|\mathcal{G}] = \Phi(\mathbf{E}(X|\mathcal{G}))$$

as claimed. □

Example 1.19. Commonly used convex functions include $x \mapsto |x|$ and $x \mapsto x^p$ for $x \geq 0$ and $p \geq 1$. In financial maths, the convex function $x \mapsto (x - a)^+$ for a fixed $a \in \mathbb{R}$ is fundamental (look up 'call options').

Exercise 1.9. Prove the *conditional Fatou lemma*: if $(X_n)_{n\in\mathbb{N}}$ is a sequence of non-negative random variables on Ω, then

$$\mathbf{E}(\liminf_{n\to\infty} X_n|\mathcal{G}) \leq \liminf_{n\to\infty} \mathbf{E}(X_n|\mathcal{G}) \tag{1.8}$$

for any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$.

We motivated the abstract definition of conditional expectation with an elementary interpretation of the quantity $\mathbf{E}(X|Y)$. This can now be defined in the measure-theoretic framework.

Definition 1.13. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and suppose $X, Y : \Omega \to \mathbb{R}$ are random variables such that $\mathbf{E}|X| < \infty$. We define

$$\mathbf{E}(X|Y) := \mathbf{E}(X|\sigma(Y)). \tag{1.9}$$

Since $\mathbf{E}(X|Y)$ is $\sigma(Y)$-measurable, intuitively speaking $\mathbf{E}(X|Y)$ is 'made up of information about Y', so we expect that $\mathbf{E}(X|Y) = g(Y)$ for some function $g : \mathbb{R} \to \mathbb{R}$. This intuition turns out to be correct, and is a direct consequence of the *Doob-Dynkin lemma*.

Theorem 1.20 (Doob, Dynkin). *Let $(\Omega, \mathcal{F})$ be a measurable space, and suppose $X : \Omega \to \mathbb{R}$ is a random variable. Then for any $\sigma(X)$-measurable random variable $Y : \Omega \to \mathbb{R}$, there exists a Borel function $g : \mathbb{R} \to \mathbb{R}$ such that $Y = g(X)$.*

Proof. We use again the standard approximation argument. If $Y = \mathbf{1}_A$ for some $A \in \mathcal{F}$. Since Y is $\sigma(X)$-measurable by assumption, we have necessarily $A \in \sigma(X)$, and thus $A = (X \in B)$ for some Borel set $B \subseteq \mathbb{R}$. Then $Y = \mathbf{1}_{(X\in B)} = \mathbf{1}_B \circ X$, so $g(x) := \mathbf{1}_B(x)$ is the required Borel function. Note that $g \geq 0$.

Let Y be a simple, $\sigma(X)$-measurable random variable, so $Y = \sum_{k=1}^n \alpha_k \mathbf{1}_{A_k}$ where $A_k \in \sigma(X)$ and $\alpha_k \in \mathbb{R}$. By the previous paragraph, there exist Borel functions $g_k : \mathbb{R} \to [0, \infty)$

such that $\mathbf{1}_{A_k} = g_k(X)$. Thus $Y = \sum_{k=1}^n \alpha_k g_k(X)$, and $g := \sum_{k=1}^n \alpha_k g_k$ is a non-negative Borel function such that $Y = g(X)$.

If $Y \geq 0$ is $\sigma(X)$-measurable, there exists a sequence $(Y_n)_{n\in\mathbb{N}}$ of non-negative, simple, $\sigma(X)$-measurable random variables such that $Y_n \uparrow Y$. From the previous step, there is a corresponding sequence of non-negative Borel functions $(g_n)_{n\in\mathbb{N}}$ such that $Y_n = g_n(X)$ for all $n \in \mathbb{N}$, and thus $Y(\omega) = \lim_{n\to\infty} g_n(X(\omega))$ for all $\omega \in \Omega$. We define $g : \mathbb{R} \to [0,\infty)$ by

$$g(x) := \begin{cases} \limsup\limits_{n\to\infty} g_n(x) & \text{if the limsup is finite} \\ 0 & \text{otherwise.} \end{cases}$$

Recall that the pointwise limit superior[3] of measurable functions is measurable, hence g is a non-negative Borel function. Then clearly we have $g(X) = \limsup_n g_n(X) = \lim_n g_n(X) = Y$.

Finally, for general sign-changing Y, we apply the above arguments to Y^+ and Y^- separately, yielding non-negative Borel functions g_1 and g_2 such that $g_1(X) = Y^+$ and $g_2(X) = Y^-$. We then conclude $Y = Y^+ - Y^- = g_1(X) - g_2(X)$, so $g := g_1 - g_2$ is the required Borel function. □

In many applications, we observe the outcome of some experiment described by a random variable Y, and we want to obtain the expectation of another random variable X given this information. This leads to an expression of the form $\mathbf{E}(X|Y = y)$. Although $(Y = y)$ is an event, in most cases of interest (e.g. the law of Y has a continuous density with respect to Lebesgue measure), it has probability 0, so $\mathbf{E}(X|Y = y)$ *a priori* does not appear to make sense. However, Theorem 1.20 now gives a way to define such an expression.

Definition 1.14. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space, and suppose $X, Y : \Omega \to \mathbb{R}$ are random variables with $\mathbf{E}|X| < \infty$. We define

$$\mathbf{E}(X|Y = y) := g(y) \tag{1.10}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a Borel function such that $g(Y) = \mathbf{E}(X|Y)$.

Exercise 1.10. Consider the setting of Example 1.13, where $Y = \sum_{k=1}^n a_k \mathbf{1}_{A_k}$. Verify that the quantity $\mathbf{E}(X|Y = y)$ when interpreted in the elementary way (1.6) is consistent with Definition 1.14.

The following example shows how to compute conditional expectations in practice.

Example 1.21 (Conditional densities). Let $X, Y : \Omega \to \mathbb{R}$ be random variables. Suppose that the law of the random vector $(X, Y) \in \mathbb{R}^2$ has a *joint density* with respect to the Lebesgue measure on $\mathbb{R}^2$, i.e. there exists an integrable function $f_{XY} : \mathbb{R}^2 \to [0,\infty)$ such that

$$\mathbf{P}((X,Y) \in A) = \int_A f_{XY}(x,y)\,dxdy \qquad \forall\, A \in \mathscr{B}(\mathbb{R}^2).$$

In particular, this means that

$$\mathbf{E}[g(X,Y)] = \int_\Omega g(X,Y)\,d\mathbf{P} = \int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\,dxdy$$

[3] Strictly speaking, we need to define $G := \limsup_n g_n$ first, which takes values in the extended halfline $[0,\infty]$, and then we obtain g as a modification. We will be slightly casual about this technical point.

for all suitable Borel functions $g : \mathbb{R}^2 \to \mathbb{R}$ (the precise conditions are analogous to Proposition 1.12). The individual densities of the laws of X and Y are recovered by the Fubini-Tonelli theorem. In the case of Y, we have

$$\mathbf{P}(Y \in B) = \mathbf{P}((X, Y) \in \mathbb{R} \times B) = \int_{\mathbb{R}} \int_B f_{XY}(x, y)\, dxdy = \int_B \int_{\mathbb{R}} f_{XY}(x, y)\, dx\, dy$$

for all Borel sets $B \subseteq \mathbb{R}$. A similar calculation holds for X, and thus we can identify the densities of $X_{\#}\mathbf{P}$ and $Y_{\#}\mathbf{P}$ respectively as

$$f_X(x) := \int_{\mathbb{R}} f_{XY}(x, y)\, dy, \qquad f_Y(y) := \int_{\mathbb{R}} f_{XY}(x, y)\, dx.$$

For every $y \in \mathbb{R}$ such that $f_Y(y) \neq 0$, we can define the **conditional density** of X given $Y = y$:

$$f_{Y=y}(x) := \frac{f_{XY}(x, y)}{f_Y(y)}.$$

Then the conditional expectation $\mathbf{E}(X|Y = y)$ can be computed by integration against the conditional density:

$$\mathbf{E}(X|Y = y) = \int_{\mathbb{R}} x f_{Y=y}(x)\, dx = \frac{1}{f_Y(y)} \int_{\mathbb{R}} x f_{XY}(x, y)\, dx. \tag{1.11}$$

Exercise 1.11. Verify that the function $g(y) := \mathbf{E}(X|Y = y)$ defined in (1.11) does in fact yield the conditional expectation of X given Y. More precisely, show that

$$\int_A g(Y)\, d\mathbf{P} = \int_A X\, d\mathbf{P} \qquad \forall\, A \in \sigma(Y)$$

and thus $g(Y) = \mathbf{E}(X|Y)$.

1.4 Supplement I: Gaussian random variables

In this section, we record some essential facts about Gaussian, or normal random variables. Recall that an $n \times n$ matrix Q is called positive definite if $\langle Qx, x\rangle \geq 0$ for all $x \in \mathbb{R}^n$.

Definition 1.15. A random variable $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ is **normal** or **Gaussian** with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ if the law of X has density

$$p_X(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad (x \in \mathbb{R}) \tag{1.12}$$

with respect to Lebesgue measure on $\mathbb{R}$. We write $X \sim N(\mu, \sigma^2)$. By convention, $X \sim N(\mu, 0)$ means that $X = \mu$ almost surely, corresponding to the Dirac measure at μ.

More generally, a random vector $X : \Omega \to \mathbb{R}^d, X = (X_1, \ldots, X_d)$, is called **multinormal** with mean $m = (m_1, \ldots, m_d) \in \mathbb{R}^d$ and covariance matrix $Q \in \mathbb{R}^{d\times d}$ if the law of X has density

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d (\det Q)}} \exp\left(-\frac{1}{2}\left\langle x - m, Q^{-1}(x - m)\right\rangle\right) \qquad (x \in \mathbb{R}^d) \tag{1.13}$$

with respect to Lebesgue measure on $\mathbb{R}^d$, where Q is a symmetric, non-singular, positive definite matrix (and thus $\det Q > 0$). The entries of Q are given by

$$q_{jk} = \mathbf{E}[(X_j - m_j)(X_k - m_k)], \qquad j, k = 1, \ldots, d.$$

Remark 1.22. A random variable $Z : \Omega \to \mathbb{R}$ is called **standard normal** if $Z \sim N(0,1)$. It is easy to check that if $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma}$ is standard normal.

The following result is easily established using the explicit form of the density (1.12) and integration by parts.

Proposition 1.23. *If $X : \Omega \to \mathbb{R}$ is Gaussian with mean 0 and variance $\sigma^2 > 0$, then*

$$\mathbf{E}(X^{2k}) = \frac{(2k)!}{2^k k!}\sigma^{2k} \quad \textit{and} \quad \mathbf{E}(X^{2k+1}) = 0 \quad (k \in \mathbb{N}). \tag{1.14}$$

The above formulas show that all central moments of an $N(0,\sigma^2)$ distribution are completely determined by the mean and variance.

The **characteristic function** of a random variable $X : \Omega \to \mathbb{R}^d$ is defined to be

$$\phi_X(\lambda) := \mathbf{E}\exp(i\lambda \cdot X) \qquad (\lambda \in \mathbb{R}^d). \tag{1.15}$$

It is essentially the Fourier transform (up to some constants). If $X \sim N(\mu, \sigma^2)$, it is straightforward to compute

$$\phi_X(\lambda) = \exp\left(i\mu\lambda - \frac{\sigma^2}{2}\lambda^2\right) \qquad (\lambda \in \mathbb{R}). \tag{1.16}$$

In $\mathbb{R}^d$, if $X \sim N(m, Q)$, then

$$\phi_X(\lambda) = \exp\left(i\langle \lambda, m\rangle - \frac{1}{2}\langle Q\lambda, \lambda\rangle\right) \qquad (\lambda \in \mathbb{R}^d). \tag{1.17}$$

It is a well-known fact in analysis that the Fourier transform maps $L^2(\Omega, \mathcal{F}, \mathbf{P})$ onto itself one-to-one (i.e. it is an automorphism). Thus the characteristic function uniquely determines the distribution of L^2 random variables. Many results about Gaussian random variables can be proved using characteristic functions.

Proposition 1.24. *If $X_1, \ldots, X_n : \Omega \to \mathbb{R}$ are independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$, then for any constants $a_1, \ldots, a_n \in \mathbb{R}$, it holds that*

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right).$$

Proof. This is a direct calculation using characteristic functions, and is left as an exercise. □

Theorem 1.25. *Let $X_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, n$, be random variables. The random vector $X = (X_1, \ldots, X_n)$ is normally distributed if and only if $\langle \lambda, X\rangle = \sum_{i=1}^n \lambda_i X_i$ is normally distributed for all vectors $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$.*

Proof. This is also a calculation using characteristic functions, see e.g. [Øks03, Theorem A.5] for the details. □

Two random variables X, Y are said to be **uncorrelated** if $\mathrm{Cov}(X,Y) = 0$. It is trivial that independent random variables are uncorrelated, while the converse is false in general. However, in the special case of Gaussian random variables, we have a partial converse.

Proposition 1.26. *Let (X, Y) be jointly normally distributed. Then X, Y are independent if and only if they are uncorrelated.*

Proof. Assume that (X, Y) is jointly normal and that $\mathrm{Cov}(X, Y) = 0$. Since characteristic functions uniquely determine the distribution of L^2 random variables, it suffices to prove

$$\phi_{(X,Y)}(\lambda) = \phi_X(\lambda_1)\phi_Y(\lambda_2)$$

for all $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$. Since $\mathrm{Cov}(X, Y) = 0$, the covariance matrix for the random vector (X, Y) is simply $\mathrm{diag}(q_X, q_Y)$, where $q_X = \mathrm{var}(X), q_Y = \mathrm{var}(Y)$. We write $m_X = \mathbf{E}(X)$ and $m_Y = \mathbf{E}(Y)$. Then by formula (1.17), we obtain

$$\begin{aligned}\phi_{(X,Y)}(\lambda) &= \exp\left(i(\lambda_1 m_X + \lambda_2 m_Y) - \frac{1}{2}(\lambda_1^2 q_X + \lambda_2^2 q_Y)\right)\\ &= e^{i\lambda_1 m_X - \frac{1}{2}\lambda_1^2 q_X} e^{i\lambda_2 m_Y - \frac{1}{2}\lambda_2^2 q_Y} = \phi_X(\lambda_1)\phi_Y(\lambda_2)\end{aligned}$$

for all $\lambda \in \mathbb{R}^2$, as required. □

Exercise 1.12 (Important!)**.** Let $X_n : \Omega \to \mathbb{R}^d$, $n \in \mathbb{N}$, be a sequence of normally distributed random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

(i) Establish the inequality

$$|e^{i\langle u,x\rangle} - e^{i\langle u,y\rangle}| \leq |u||x - y|$$

for all vectors $u, x, y \in \mathbb{R}^d$.

(ii) Prove that if $X_n \to X$ in $L^2(\Omega, \mathcal{F}, \mathbf{P})$, then X is normally distributed.

2 Stochastic processes

2.1 Basic notions

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. At the most basic level, a **stochastic process** is simply any collection of random variables $\{X_\alpha\}_{\alpha \in I}$ (where I is some indexing set) defined on $(\Omega, \mathcal{F}, \mathbf{P})$. However, such a general definition is not useful, and we would like to have some more structure. For practical applications, a stochastic process should capture the evolution of some random phenomenon, and it is reasonable to equip our probability space with some way of tracking the 'information' about the process available to us at a given time. We thus introduce the following fundamental definition.

Definition 2.1. Let $(\Omega, \mathcal{F})$ be a measurable space, and $(\mathbb{T}, \preceq)$ a totally ordered set. A **filtration** is a family of σ-algebras $(\mathcal{F}_t)_{t \in \mathbb{T}}$ such that

(i) $\mathcal{F}_t \subseteq \mathcal{F}$ for all $t \in \mathbb{T}$;

(ii) $\mathcal{F}_s \subseteq \mathcal{F}_t$ if $s \preceq t$.

Thus a filtration is an *increasing* family of σ-algebras. In most applications, we consider $\mathbb{T}$ to be $\mathbb{N}$ (this is the case for 'discrete' stochastic processes), $[0, \infty)$ or a compact interval $[0, T]$, all equipped with the usual ordering $\leq$ inherited from $\mathbb{R}$.

If a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ admits a filtration $(\mathcal{F}_t)_{t \geq 0}$, we call it a **filtered probability space**. We can now introduce a very general yet useful definition of a stochastic process.

Definition 2.2. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbf{P})$ be a filtered probability space, and let $(E, \mathcal{G})$ be a measurable space. A **stochastic process** $(X_t)_{t \geq 0}$ with values in E is a family of random variables $X_t : \Omega \to E$. The process is said to be **$\mathcal{F}_t$-adapted**, or simply **adapted**, if X_t is $\mathcal{F}_t$-measurable for every $t \geq 0$. (In particular, X_t is $\mathcal{F}_s$-measurable for all $0 \leq s \leq t$).

The concept of adaptedness heuristically means that we do not need to 'see into the future' in order to know everything about the random process at time t.

Remark 2.1 (On notation)**.** We follow the usual convention of denoting the 'time' parameter with a subscript. However, we will sometimes write $X(t)$ instead of X_t if there is possible confusion with subscripts for partial derivatives.

We have introduced stochastic processes as families of random variables indexed by a time variable t, i.e. for each $t \geq 0$ we obtain a random variable $X(t, \cdot) : \Omega \to E$. However, it is equally valid and very useful to view a stochastic process as a collection of *paths* in the target space E. From this viewpoint, for each sample point $\omega \in \Omega$, we obtain a function

$$[0, \infty) \ni t \mapsto X(t, \omega) \in E.$$

These maps are then called the **sample paths** or **trajectories** of the process. A large part of stochastic analysis involves understanding fine properties of sample paths of Brownian motion.

We introduce two filtrations that are ubiquitous in this subject.

Definition 2.3. Let $(X_t)_{t \geq 0}$ be a stochastic process. The **natural filtration** of the process is defined to be

$$\mathcal{F}_t^X := \sigma(X_s : 0 \leq s \leq t).$$

Any stochastic process is automatically adapted to its natural filtration.

Now let $(\mathcal{F}_t)_{t \geq 0}$ be any filtration. For each $t \geq 0$, we define

$$\mathcal{F}_{t+} := \bigcap_{u > t} \mathcal{F}_u.$$

This is sometimes called the σ-algebra of **events immediately after** t.

The idea of $\mathcal{F}_{t+}$ is that it captures the information available to us if we could 'see infinitesimally into the future'. This sounds quite strange at the moment, but it turns out to be a rather important technicality in the study of stochastic processes with *continuous* sample paths. Clearly $\mathcal{F}_t \subseteq \mathcal{F}_{t+}$, but it may be surprising that in general, $\mathcal{F}_{t+} \neq \mathcal{F}_t$. The following example is rather artificial but instructive.

Example 2.2. Consider the sample space $\Omega = \{0,1\}$ with σ-algebra $\mathcal{F} = \{\emptyset, \Omega, \{0\}, \{1\}\}$. Define a stochastic process by $X_t = 0$ for all $t \in [0,1)$. At $t = 1$, a fair coin is tossed, and we define $X_t(\omega) = 1 - t$ if $\omega = 0$, corresponding to tails, and $X_t(\omega) = t - 1$ if $\omega = 1$, corresponding to heads. Let $\mathcal{F}_t$ be the natural filtration. It is easy to check that for $t \in [0,1]$, we have $\mathcal{F}_t = \mathcal{F}_1 = \{\emptyset, \Omega\}$. However for $t > 1$, $\mathcal{F}_t = \mathcal{F}$. Hence $\mathcal{F}_1 \neq \mathcal{F}_t$ for all $t > 1$, which implies $\mathcal{F}_{t+} \neq \mathcal{F}$.

Filtrations in which $\mathcal{F}_t$ does coincide with $\mathcal{F}_{t+}$ play an important role in stochastic analysis.

Definition 2.4. We say a filtration $(\mathcal{F}_t)_{t\geq 0}$ is **right-continuous** if $\mathcal{F}_t = \mathcal{F}_{t+}$ for all $t \geq 0$.

Given a filtration on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, we say that $(\mathcal{F}_t)_{t\geq 0}$ is **complete** if each $\mathcal{F}_t$ contains all sets of probability 0. If $(\mathcal{F}_t)_{t\geq 0}$ on $(\Omega, \mathcal{F}, \mathbf{P})$ is right-continuous and complete, we say that the filtration satisfies the **usual conditions**.

add some technical notes about completeness of filtration?

2.2 Stopping times

We continue with the natural interpretation that t represents the flow of time, and the filtration $(\mathcal{F}_t)_{t\geq 0}$ keeps track of the information available to us about the random process $X(t)$ at time t. Quite often we will be interested in the *first time* a phenomenon occurs: the first time a stock price rises above (or falls below) a certain value, the first time a solution to a stochastic differential equation exits an interval $J \subseteq \mathbb{R}$, and so on. This time will, of course, be a random variable $\tau : \Omega \to [0, \infty]$. It is intuitively clear that the event $(\tau \leq t)$ — whether the phenomenon has occurred before time t — should be part of the available information. This discussion motivates the definition of a *stopping time*.

Definition 2.5. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ be a filtered probability space. A random variable $\tau : \Omega \to [0, \infty]$ is called a **stopping time** if $(\tau \leq t) \in \mathcal{F}_t$ for all $t \geq 0$. Moreover, τ is called an **optional time** if $(\tau < t) \in \mathcal{F}_t$ for all $t \geq 0$.

Remark 2.3. (i): Note that the definition of stopping times depends on the filtration!

(ii): For stochastic processes in discrete time $(X_n)_{n\in\mathbb{N}}$, stopping times are $\mathbb{N}$-valued random variables. In that case, it suffices to check that the event $(\tau = n) \in \mathcal{F}_n$ for each $n \in \mathbb{N}$.

Proposition 2.4. *Every stopping time is an optional time. If the filtration $(\mathcal{F}_t)_{t\geq 0}$ is right-continuous, then every optional time is a stopping time.*

Proof. Assume that τ is a stopping time. Then

$$(\tau < t) = \bigcup_{n\in\mathbb{N}} (\tau \leq t - \tfrac{1}{n}) \in \mathcal{F}_t$$

since $(\tau \leq t - \frac{1}{n}) \in \mathcal{F}_{t-1/n} \subseteq \mathcal{F}_t$ for all $n \in \mathbb{N}$.

Now assume that τ is an optional time, and that the filtration is right-continuous. Then

$$(\tau \leq t) = \bigcap_{n\in\mathbb{N}} (\tau < t + \tfrac{1}{n}) \in \bigcap_{n\in\mathbb{N}} \mathcal{F}_{t+1/n} = \mathcal{F}_{t+} = \mathcal{F}_t,$$

hence τ is a stopping time. □

The following facts are often used in calculations involving stopping times.

Proposition 2.5. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with a filtration $(\mathcal{F}_t)_{t\geq 0}$.*

(i) *Fix $t_0 \geq 0$. If $\tau = t_0$ almost surely, then τ is a stopping time.*

(ii) *If τ_1, τ_2 are stopping times, then $\tau_1 \wedge \tau_2$, $\tau_1 \vee \tau_2$, and $\tau_1 + \tau_2$ are stopping times.*

(iii) *If $(\tau_n)_{n\geq 1}$ is an increasing sequence of stopping times, i.e. $\tau_n \leq \tau_{n+1}$ almost surely, then $\tau := \lim_{n\to\infty} \tau_n$ is a stopping time. The same conclusion holds if τ_n is a decreasing sequence, $\tau_n \downarrow \tau$, and the filtration is right-continuous.*

Proof. (i): This is trivial.

(ii): Observe that $(\tau_1 \wedge \tau_2 \leq t) = (\tau_1 \leq t) \cap (\tau_2 \leq t) \in \mathcal{F}_t$, and similarly $(\tau_1 \vee \tau_2 \leq t) = (\tau_1 \leq t) \cup (\tau_2 \leq t) \in \mathcal{F}_t$.

To show that $\tau_1 + \tau_2$ is a stopping time, consider the decomposition

$$\begin{aligned}(\tau_1 + \tau_2 > t) = (\tau_1 = 0, \tau_2 > t) \cup (0 < \tau_1 < t, \tau_1 + \tau_2 > t) \cup (\tau_1 > t, \tau_2 = 0) \cup (\tau_1 \geq t, \tau_2 > 0)\\ =: A_1 \cup A_2 \cup A_3 \cup A_4.\end{aligned}$$

The events A_1 and A_3 clearly belong to $\mathcal{F}_t$. The event A_3 can be rewritten as

$$A_3 = (\tau_1 \geq t) \cap (\tau_2 > 0) = (\tau_1 < t)^c \cap (\tau_2 \leq 0)^c \in \mathcal{F}_t$$

where we have used Proposition 2.4 to deduce that $(\tau_1 < t) \in \mathcal{F}_t$ (and thus its complement also belongs to $\mathcal{F}_t$). Finally, the event A_2 can be rewritten as

$$A_2 = \bigcup_{q\in\mathbb{Q}\cap(0,t)} (q < \tau_1 < t) \cap (t - q < \tau_2),$$

and we leave it to the reader to verify carefully that $A_2 \in \mathcal{F}_t$. Hence $(\tau_1 + \tau_2 > t) \in \mathcal{F}_t$, which implies that the complementary event $(\tau_1 + \tau_2 \leq t)$ also belongs to $\mathcal{F}_t$.

(iii): Fix $t > 0$, and consider the event $(\tau \leq t)$. If the sequence (τ_n) is increasing, we have

$$(\lim_{n\to\infty} \tau_n \leq t) = \bigcap_{n\geq 1} (\tau_n \leq t) \in \mathcal{F}_t$$

and hence τ is a stopping time. If the filtration is right-continuous and $\tau_n \downarrow \tau$, then

$$(\tau < t) = \bigcup_{n\in\mathbb{N}} (\tau_n < t) \in \mathcal{F}_t,$$

see the first calculation of Proposition 2.4. This shows that τ is an *optional* time. As the filtration is right-continuous, we conclude from Proposition 2.4 that τ is a stopping time. □

Now we will formalise the notion of 'sampling at a random time'.

Definition 2.6. Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and let τ be an $\mathcal{F}_t$-stopping time. We define the **stopped σ-algebra**[4] $\mathcal{F}_\tau$ by

$$A \in \mathcal{F}_\tau \overset{\text{def.}}{\iff} A \cap (\tau \leq t) \in \mathcal{F}_t \quad \text{for all } t \geq 0.$$

[4] This is my own term. There does not seem to be a consensus for the name of this construction. In [KS91], one finds the accurate but highly inconvenient name 'σ-field of events determined prior to the stopping time τ.'

The definition is reasonable on intuitive grounds; however, we ought to show that $\mathcal{F}_\tau$ really is a σ-algebra. This is left as a simple exercise for the reader. Note that τ itself is $\mathcal{F}_\tau$-measurable. Indeed, for any $s, t \geq 0$, we have

$$(\tau \leq s) \cap (\tau \leq t) = (\tau \leq s \wedge t) \in \mathcal{F}_{s \wedge t} \subseteq \mathcal{F}_t,$$

which proves the claim.

Exercise 2.1 (Approximating a stopping time from above)**.** Let T be a $\mathcal{F}_t$-stopping time defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$.

(i) If S is an $\mathcal{F}_T$-measurable random variable with values in $[0, \infty]$ such that $T \leq S$ a.s., then S is an $\mathcal{F}_t$-stopping time.

(ii) Show that the random variables

$$T_n := \sum_{k=0}^{\infty} \frac{k+1}{2^n} \mathbf{1}_{(k2^{-n} < T \leq (k+1)2^{-n})} + \infty \cdot \mathbf{1}_{(T=\infty)}, \quad n = 1, 2, 3, \ldots$$

are $\mathcal{F}_t$-stopping times such that $T_n \downarrow T$ as $n \to \infty$.

2.3 A taste of martingale theory

Martingales are one of the fundamental tools of modern analysis and probability theory, and originated in the study of gambling strategies. For the moment, we will present some elementary facts, followed by some key results of an analytic nature. We will have much more to say about martingales when we introduce stochastic integrals.

Definition 2.7. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ be a filtered probability space. A stochastic process $(M_t)_{t\geq 0}$ is a **martingale** if:

(M1) $(M_t)_{t\geq 0}$ is $\mathcal{F}_t$-adapted;

(M2) $\mathbf{E}|M_t| < \infty$ for each $t \geq 0$;

(M3) ('Fair game') $\mathbf{E}(M_t|\mathcal{F}_s) = M_s$ for all $0 \leq s \leq t$.

If $(M_t)_{t\geq 0}$ is a stochastic process such that (M1), (M2) hold, then

(M3$^+$) if $\mathbf{E}(M_t|\mathcal{F}_s) \geq M_s$ for all $0 \leq s \leq t$, then $(M_t)_{t\geq 0}$ is a **submartingale** ('winning game').

(M3$^-$) if $\mathbf{E}(M_t|\mathcal{F}_s) \leq M_s$ for all $0 \leq s \leq t$, then $(M_t)_{t\geq 0}$ is a **supermartingale** ('losing game').

To emphasise the dependence on the filtration, we can say that M_t is an $\mathcal{F}_t$-martingale (e.g. it could be that there is another filtration $(\mathcal{G})_{t\geq 0}$ on the same probability space, but M_t need not be adapted to $\mathcal{G}_t$).

Of course, the above definitions can be reformulated for discrete time stochastic processes $(M_n)_{n\in\mathbb{N}}$. Obviously, M_t is a supermartingale if and only if $-M_t$ is a submartingale, and M_t is a martingale if and only if M_t is both a sub- and supermartingale. These terms actually derive from classical potential theory, in the study of harmonic, sub- and superharmonic functions.

Proposition 2.6. *Let $(M_t, \mathcal{F}_t)$ be a martingale. The following assertions hold.*

(i) $\mathbf{E}(M_t) = \mathbf{E}(M_0)$ *for all* $t > 0$.

(ii) *If $(N_t, \mathcal{F}_t)$ is another martingale (note that N_t is adapted to the same filtration!), then $aM_t + bN_t$ is also a martingale with respect to $\mathcal{F}_t$.*

(iii) *If $f : \mathbb{R} \to \mathbb{R}$ is a convex (resp. concave) function such that $\mathbf{E}|f(X)| < \infty$, then $(f(M_t), \mathcal{F}_t)$ is a submartingale (resp. supermartingale).*

We leave the easy proofs as exercises: (i) follows from (M3) and the tower property of conditional expectation, (ii) is obvious, and (iii) is proved using the conditional Jensen inequality.

Exercise 2.2. Let $X : \Omega \to \mathbb{R}$ be a fixed, integrable random variable on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$. Show that $M_t := \mathbf{E}(X|\mathcal{F}_t)$ defines an $\mathcal{F}_t$-martingale.

Martingales have a wide range of applications in mathematical analysis, probability, economics and finance, well beyond their initial purpose for the study of gambling strategies. We recommend [Wil91, Chapter 15] as a tasting menu. From the stochastic analysis perspective, a particularly important feature of martingales is that they 'co-operate' well with stopping times. The following theorem of Doob is fundamental.

Theorem 2.7 (Optional sampling theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with right-continuous filtration $(\mathcal{F}_t)_{t\geq 0}$, and suppose $(M_t)_{t\geq 0}$ is an $\mathcal{F}_t$-(sub)martingale whose sample paths are almost surely right-continuous. If S, T are stopping times such that $S \leq T < \infty$ almost surely, then*

$$\mathbf{E}(M_T|\mathcal{F}_S)\,(\geq) = M_S \quad \textit{almost surely}. \tag{2.1}$$

An easy but important corollary is the following:

Corollary 2.8. *Let $(\mathcal{F}_t)_{t\geq 0}$ be a right-continuous filtration, and suppose $(M_t)_{t\geq 0}$ is a continuous martingale adapted to $\mathcal{F}_t$. For any stopping time τ, the* stopped process *$(M_t^\tau)_{t\geq 0}$, defined by*

$$M_t^\tau := M_{t\wedge\tau},$$

is an $\mathcal{F}_t$-martingale.

Proof. Let $s < t$. Since $t \wedge \tau$ is a stopping time for all $t \geq 0$, it follows that

$$\mathbf{E}(M_{t\wedge\tau}|\mathcal{F}_{s\wedge\tau}) = M_{s\wedge\tau}$$

by Theorem 2.7. □

To include: Doob maximal and L^p inequalities

2.4 A technical digression

We consider the general problem of constructing an appropriate probability space $(\Omega, \mathcal{F}, \mathbf{P})$, that should, heuristically speaking, contain 'all the information' about a stochastic process. What does it take, then, to 'know' a stochastic process? Recall that we can view a stochastic process $(X(t))_{t\geq 0}$ as a collection of *paths* or *trajectories*. Let us consider just the case where our process takes values in $\mathbb{R}$. Our candidate sample space is

$$\Omega = \mathbb{R}^{[0,\infty)} := \{\text{all functions } f : [0,\infty) \to \mathbb{R}\}.$$

Fix $n \in \mathbb{N}$ and moments in time $0 \leq t_1 < t_2 < \ldots < t_n$, and corresponding intervals $I_1, I_2, \ldots, I_n$ (the type of interval does not matter). If we 'know' the process, then we should be able to determine the *finite dimensional* distribution

$$\mathbf{P}(X(t_1) \in I_1, X(t_2) \in I_2, \ldots, X(t_n) \in I_n).$$

This is a very practical constraint, since in real-world applications, we can only ever perform finitely many observations of a random process. Thus we require that all sets of the form $\prod_{i=1}^{n}(X(t_j) \in I_j)$ are measurable, for all $n \in \mathbb{N}$ and choices of intervals $I_1, \ldots, I_n$. For each finite ordered collection $[t]^n := (t_1, \ldots, t_n)$ of non-negative real numbers and corresponding intervals $I_1, \ldots, I_n$, define the **cylinder set**

$$C_{[t]^n}(I_1 \times I_2 \times \cdots \times I_n) := \{f : [0, \infty) \to \mathbb{R} \mid f(t_k) \in I_k, 1 \leq k \leq n\}. \tag{2.2}$$

Then we consider the σ-algebra

$$\sigma(I_1 \times I_2 \times \cdots \times I_n : I_k \subseteq \mathbb{R} \text{ intervals}),$$

but this is exactly $\mathscr{B}(\mathbb{R}^n)$, the Borel σ-algebra on $\mathbb{R}^n$ (this is essentially the higher-dimensional analogue of Exercise 1.2).

Let $\mathcal{T}$ denote the set of all finite ordered collections $[t]$ of distinct, non-negative real numbers. Suppose that for each $[t] \in \mathcal{T}$, we have a probability measure $Q_{[t]}$ on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. For every $n \in \mathbb{N}$ and every $[t]^n = (t_1, \ldots, t_n) \in \mathcal{T}$, we define

$$\mathbf{P}(C_{[t]^n}(I_1 \times I_2 \times \cdots \times I_n)) := Q_{[t]^n}(I_1 \times I_2 \times \cdots \times I_n)$$

for each cylinder set. The measure $Q_{[t]^n}$ can then be extended to all Borel sets $B \subseteq \mathbb{R}^n$. The collection of probability measures $(Q_{[t]})_{[t]\in\mathcal{T}}$ is called a **family of finite-dimensional distributions**.

Definition 2.8. A family $(Q_{[t]})_{[t]\in\mathcal{T}}$ of finite-dimensional distributions is called **consistent** if the following conditions hold:

(i) if $[\pi(t)] := (t_{\pi(1)}, \ldots, t_{\pi(n)})$ is a permutation of $[t]$, where $\pi \in \mathrm{Sym}(n)$ denotes a permutation on n letters, then for any $B_i \in \mathscr{B}(\mathbb{R}), i = 1, \ldots, n$, we have

$$Q_{[t]}(B_1 \times B_2 \times \cdots \times B_n) = Q_{[\pi(t)]}(B_{\pi(1)} \times B_{\pi(2)} \times \cdots \times B_{\pi(n)}).$$

(ii) if $[t] = (t_1, \ldots, t_n)$ with $n \geq 2$ and $[s] = (t_1, \ldots, t_{n-1})$ and $A \in \mathscr{B}(\mathbb{R}^{n-1})$, then

$$Q_{[t]}(A \times \mathbb{R}) = Q_{[s]}(A).$$

Let $\mathcal{F} := \mathscr{C}(\mathbb{R}^{[0,\infty)})$ denote the σ-algebra generated by all cylinder sets. Of course, if we are magically given a probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$, then we can define a family of finite-dimensional distributions by

$$Q_{[t]}(A) := \mathbf{P}(\{\omega \in \Omega : (\omega(t_1), \ldots, \omega(t_n)) \in A\})$$

for all $A \in \mathscr{B}(\mathbb{R}^n)$ and $[t] \in \mathcal{T}$. One can check that this definition yields a consistent family. We are interested in the converse: we would like to prescribe all finite-dimensional distributions for a stochastic process, and construct an appropriate probability space. That the consistency condition is sufficient is the non-trivial claim of the following result.

Theorem 2.9 (Daniell-Kolmogorov extension theorem). *Let $(Q_{[t]})_{[t]\in\mathcal{T}}$ be a consistent family of finite-dimensional distributions. Then there exists a (unique) probability measure $\mathbf{P}$ on $(\mathbb{R}^{[0,\infty)}, \mathscr{C}(\mathbb{R}^{[0,\infty)}))$ such that*

$$Q_{[t]}(A) = \mathbf{P}(\{\omega \in \mathbb{R}^{[0,\infty)} : (\omega(t_1), \ldots, \omega(t_n)) \in A\}), \quad \forall\, A \in \mathscr{B}(\mathbb{R}^n) \tag{2.3}$$

holds for all $[t] \in \mathcal{T}$.

We will not present the long and technical proof; we merely remark that the main theoretical tool used in the proof is the Carathéodory extension theorem. See [KS91, Section 2.2A] for the details.

2.5 Supplement: essential probabilistic tools

In this section, we collect some basic results in probability that will be essential to the analysis of stochastic processes.

Proposition 2.10 (Markov-Chebyshev inequalities). *Let $X : \Omega \to \mathbb{R}$ be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For every $1 \leq p < \infty$ and $\lambda > 0$, it holds that*

$$\mathbf{P}(|X| \geq \lambda) \leq \frac{1}{\lambda^p}\mathbf{E}|X|^p. \tag{2.4}$$

Proof. We observe that

$$\mathbf{E}|X|^p = \int_\Omega |X|^p \, d\mathbf{P} \geq \int_{(|X| \geq \lambda)} |X|^p \, d\mathbf{P} = \lambda^p \mathbf{P}(|X| \geq \lambda). \qquad \square$$

In probability and stochastic analysis, we often want to determine when something happens 'infinitely often'. If $(A_n)_{n\in\mathbb{N}}$ is a sequence of events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, we define

$$(A_n \text{ i.o.}) := \limsup A_n := \bigcap_{n=1}^\infty \bigcup_{m=n}^\infty A_m, \tag{2.5}$$

where i.o. is an abbreviation for 'infinitely often'. We have the following fundamental lemma.

Lemma 2.11 (Borel-Cantelli). *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. If $\sum_{n=1}^\infty \mathbf{P}(A_n) < \infty$, then*

$$\mathbf{P}(A_n \text{ i.o.}) = 0.$$

On the other hand, if $(A_n)_{n\in\mathbb{N}}$ is a sequence of independent *events and $\sum_{n=1}^\infty \mathbf{P}(A_n) = +\infty$, then*

$$\mathbf{P}(A_n \text{ i.o.}) = 1.$$

Proof. Observe that for every $n \in \mathbb{N}$, we have

$$\mathbf{P}(A_n \text{ i.o.}) \leq \mathbf{P}\left(\bigcup_{m=n}^\infty A_m\right) \leq \sum_{m=n}^\infty \mathbf{P}(A_m).$$

If the series $\sum_{n=1}^\infty \mathbf{P}(A_n)$ converges, then the right hand side of the above converges to 0 as $n \to \infty$. This proves the first assertion.

To prove the second assertion, we set $p_i := \mathbf{P}(A_i)$, and prove the equivalent statement $\mathbf{P}((A_n \text{ i.o.})^c) = 0$. Since the A_i are independent, for all $N > m \geq 1$ we have

$$\mathbf{P}\left(\bigcap_{n=m}^N A_n^c\right) = \prod_{n=m}^N (1 - p_n) \leq \prod_{n=m}^N e^{-p_n} = \exp\left(-\sum_{n=m}^N p_n\right) \longrightarrow 0$$

as $N \to \infty$, since the series of probabilities diverge to $+\infty$. By continuity of measures, we obtain

$$\mathbf{P}\left(\bigcap_{n=m}^\infty A_n^c\right) = \lim_{N\to\infty} \mathbf{P}\left(\bigcap_{n=m}^N A_n^c\right) = 0$$

for every $m \geq 1$. Hence

$$\mathbf{P}((A_n \text{ i.o.})^c) = \mathbf{P}\left(\bigcup_{m=1}^\infty \bigcap_{n=m}^\infty A_n^c\right) \leq \sum_{m=1}^\infty \mathbf{P}\left(\bigcap_{n=m}^\infty A_n\right) = 0$$

and indeed $\mathbf{P}(A_n \text{ i.o.}) = 1$. $\square$

add more as necessary

3 Brownian motion

3.1 Basic properties

We are now ready to introduce the mathematical definition of one-dimensional Brownian motion.

Definition 3.1. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with filtration $(\mathcal{F}_t)_{t\geq 0}$. An $\mathcal{F}_t$-adapted stochastic process $(W_t)_{t\geq 0}$, where each W_t takes values in $\mathbb{R}$, is a **(standard) Brownian motion** or **Wiener process** if

(W1) $W_0 = 0$ almost surely;

(W2) W has *independent increments*: for $0 \leq s < t$, $W_t - W_s$ is independent of $\mathcal{F}_s$;

(W3) $W_t - W_s \sim N(0, t-s)$, i.e. increments are normally distributed with mean 0 and variance $t - s$;

(W4) Sample paths are almost surely continuous, that is,

$$\mathbf{P}\big(\{\omega \in \Omega : t \mapsto W(t,\omega) \text{ is continuous on } [0,\infty)\}\big) = 1.$$

We can also define Brownian motion *starting at* $x \in \mathbb{R}$ by replacing condition (W1) with $W_0 = x$ almost surely. More generally, given a probability measure μ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, we define Brownian motion *with initial distribution* μ by requiring $\mathbf{P}(W_0 \in B) = \mu(B)$ for all Borel subsets $B \subseteq \mathbb{R}$ in place of (W1).

Remark 3.1. Condition (W2) implies in particular that $W_t - W_s$ is independent of any $\mathcal{F}_s$-measurable random variable X. It is also equivalent to the seemingly weaker condition:

(W2′) for all $n \in \mathbb{N}$ and indices $0 \leq t_0 < t_1 < \ldots < t_n < \infty$, the random variables $W_{t_0}, W_{t_1} - W_{t_0}, \ldots, W_{t_n} - W_{t_{n-1}}$ are independent.

In fact, if $(X_t)_{t\geq 0}$ is any stochastic process satisfying (W2′), then for $0 \leq s < t$ the increment $X_t - X_s$ is independent of $\mathcal{F}_s^X$. See [KS91, Chapter 2, Problem 1.4] for the proof of this general statement, which uses the Dynkin theorem (1.4).

A fundamental problem is naturally to determine if a Brownian motion exists! We will address this in detail in Section 3.2. For now, let us acquaint ourselves with elementary consequences of the definition of Brownian motion.

Exercise 3.1 (Important!). Let $(W_t, \mathcal{F}_t)$ be a Brownian motion. Show the following facts:

(i) $(-W_t, (\mathcal{F}_t)_{t\geq 0})$ is a Brownian motion.

(ii) $\operatorname{Cov}(W_t, W_s) = s \wedge t$.

(iii) W_t and $M_t := W_t^2 - t$ are $\mathcal{F}_t$-martingales.

Brownian motion is a particular example of various important types of stochastic processes. In the previous exercise, we showed that $(W_t)_{t\geq 0}$ is a *continuous martingale* (more precisely, a martingale such that almost every sample path is continuous). Another class of processes that appears frequently in applications is the following.

Definition 3.2. An $\mathbb{R}^d$-valued stochastic process $(X_t)_{t\geq 0}$ is called **Gaussian** if, for any $n \in \mathbb{N}$ and moments of time $0 \leq t_1 < t_2 < \ldots < t_n < \infty$, the random vector $(X_{t_1}, \ldots, X_{t_n})$ has a joint normal distribution. Furthermore, if the distribution of $(X_{t+t_1}, \ldots, X_{t+t_n})$ is independent of $t \geq 0$, we say that the process is **stationary**.

Remark 3.2. (i): The defining condition of Gaussian processes is equivalent to the following: for any $n \in \mathbb{N}$, moments of time $0 \leq t_1 < t_2 < \ldots < t_n < \infty$, and vector $(a_1, \ldots, a_n) \in \mathbb{R}^n$, the random variable $\sum_{k=1}^n a_k X_{t_k}$ has a normal distribution. Indeed, this is a direct consequence of Theorem 1.25.

(ii): By considering characteristic functions, it can be shown that a Gaussian process is completely determined by its expectation vector $m(t) := \mathbf{E}(X_t)$ and its (matrix-valued) covariance function

$$\varrho(s,t) := \mathbf{E}[(X_s - m(s))(X_t - m(t))^\top], \qquad s,t \geq 0.$$

We say that a process is *zero-mean* if $m(t) \equiv 0$.

Proposition 3.3. *The standard one-dimensional Brownian motion is a zero-mean Gaussian process. Conversely, if $X = (X_t)_{t\geq 0}$ is a zero-mean Gaussian process with continuous sample paths and covariance function $\varrho(s,t) = s \wedge t$, then X is a standard Brownian motion.*

Proof. The zero-mean property is obvious. To prove the Gaussian property, take arbitrary points in time $0 \leq t_1 < t_2 < \ldots < t_n$ and numbers $a_1, \ldots, a_n \in \mathbb{R}$. By Remark 3.2(i), we only need to show that $Z := \sum_{i=1}^n a_i W_{t_i}$ is normally distributed. The argument is inductive. Consider first $n = 2$, so let $Z = a_1 W_{t_1} + a_2 W_{t_2}$. Then

$$Z = a_1 W_{t_1} + a_2(W_{t_2} - W_{t_1} + W_{t_1}) = (a_1 + a_2)(W_{t_1} - W_0) + a_2(W_{t_2} - W_{t_1})$$

which is a linear combination of independent Gaussian random variables. This has a normal distribution by Proposition 1.24.

Now assume the claim holds for some $n \geq 2$, and let $0 \leq t_1 < t_2 < \ldots < t_{n+1}$ and $a_1, \ldots, a_n, a_{n+1} \in \mathbb{R}$ be arbitrary. Define $Z = \sum_{i=1}^{n+1} a_i W_{t_i}$, and observe that

$$\begin{aligned} Z &= a_1 W_{t_1} + \ldots + a_n W_{t_n} + a_{n+1}(W_{t_{n+1}} - W_{t_n} + W_{t_n}) \\ &= a_1 W_{t_1} + \ldots + (a_{n+1} - a_n) W_{t_n} + a_{n+1}(W_{t_{n+1}} - W_{t_n}). \end{aligned}$$

By the induction hypothesis, the sum of the first n terms in the above expression defines a Gaussian random variable, and note that it is independent of $W_{t_{n+1}} - W_{t_n}$. Hence Z is a Gaussian random variable.

Finally, the converse statement follows from Remark 3.2(ii). □

We now examine some simple transformations that preserve the properties of Brownian motion.

Proposition 3.4. *Let $(W_t, \mathcal{F}_t)$ be a standard one-dimensional Brownian motion.*

(i) (Time translation) *Fix $s \geq 0$, and define $B_t := W_{t+s} - W_s$. Then B_t is a Brownian motion with respect to the filtration $\mathcal{G}_t = \mathcal{F}_{t+s}$. Moreover, $(B_t)_{t\geq 0}$ is independent of $\{W_u : 0 \leq u \leq s\}$.*

(ii) (Time reversal) *Fix $T > 0$ and define $B_t = W_T - W_{T-t}$ for all $t \in [0,T]$. Then B_t is a Brownian motion with respect to the filtration $\mathcal{G}_t = \sigma(B_s : 0 \leq s \leq t)$.*

(iii) (Scale invariance) *Fix $c > 0$, and define $B_t = cW_{t/c^2}$. Then B_t is a Brownian motion with respect to the filtration $\mathcal{G}_t := \mathcal{F}_{t/c^2}$.*

(iv) (Time inversion) *Define*

$$B_t := \begin{cases} tW_{1/t} & t > 0 \\ 0 & t = 0 \end{cases}$$

and $\mathcal{G}_t = \sigma(W_{1/s} : 0 < s \leq t)$. Then $(B_t, \mathcal{G}_t)$ is a Brownian motion.

Proof. Property (W1) is trivial for all four processes defined in the proposition. The continuity on $[0, \infty)$ is clear for (i) and (iii), and likewise for the continuity on $[0, T]$ for (ii). For the time-inverted process in (iv), the continuity for $t > 0$ is immediate, but continuity up to $t = 0$ is quite non-trivial, so we defer this discussion to Section 3.3.

In (i), clearly B_t is adapted to $\mathcal{F}_{t+s} = \mathcal{G}_t$. Adaptedness is automatic in (ii), (iii) and (iv) from the definition of $\mathcal{G}_t$ in each case. It remains to verify properties (W2) and (W3) for each process.

(i) It is obvious that $\mathbf{E}(B_t) = 0$ for all $t \geq 0$, and if $0 \leq t_1 < t_2$, then

$$B_{t_2} - B_{t_1} = (W_{t_2+s} - W_s) - (W_{t_1+s} - W_s) = W_{t_2+s} - W_{t_1+s} \sim N(0, t - s).$$

The above calculation also shows that $B_{t_2} - B_{t_1}$ is independent of $\mathcal{F}_{t_1+s} = \mathcal{G}_{t_1}$.

(ii) If $0 \leq t_1 < t_2 \leq T$, we have

$$B_{t_2} - B_{t_1} = (W_T - W_{T-t_2}) - (W_T - W_{T-t_1}) = W_{T-t_1} - W_{T-t_2} \sim N(0, t_2 - t_1).$$

Note that $\mathcal{G}_{t_1} = \sigma(W_T - W_{T-s} : 0 \leq s \leq t_1)$. By property (W3) for $(W_t)_{t\in[0,T]}$, we see that $W_{T-t_1} - W_{T-t_2}$ is independent of $\mathcal{G}_{t_1}$.

(iii) If $0 \leq s < t$, then

$$B_t - B_s = c(W_{t/c^2} - W_{s/c^2}) \sim N(0, t - s),$$

and clearly $B_t - B_s$ is independent of $\mathcal{F}_{s/c^2} = \mathcal{G}_s$.

(iv) By Proposition 3.3, it suffices to show that $(B_t)_{t\geq 0}$ defines a zero-mean Gaussian process, and then compute the covariance function.

It is clear that $\mathbf{E}(B_t) = 0$ for every $t \geq 0$, and the Gaussian property for B_t also follows from the calculations of Proposition 3.3. If $0 < s \leq t$, we compute

$$\operatorname{Cov}(B_s, B_t) = \mathbf{E}(B_s B_t) = st\mathbf{E}(W_{1/s} W_{1/t}) = st\left(\frac{1}{s} \wedge \frac{1}{t}\right) = st \cdot \frac{1}{t} = s.$$

Hence $\mathbf{E}(B_t B_s) = s \wedge t$ for all $s, t \geq 0$. Finally, observe that $tW_{1/t} - sW_{1/s}$ is independent of any increment $W_{1/u} - W_{1/s}$ for all $0 < u \leq s$, and thus is independent of $\sigma(W_{1/u} : 0 < u \leq s) = \mathcal{G}_s$. □

For completeness, we define Brownian motion with values in $\mathbb{R}^d$.

Definition 3.3. Let μ be a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, and let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with filtration $(\mathcal{F}_t)_{t\geq 0}$. An $\mathcal{F}_t$-adapted stochastic process $B = (B_t)_{t\geq 0}$ on Ω with values in $\mathbb{R}^d$ is called a d**-dimensional Brownian motion (Wiener process)** with initial distribution μ if

(i) $\mathbf{P}[B_0 \in \Gamma] = \mu(\Gamma)$ for all Borel sets $\Gamma \subseteq \mathbb{R}^d$.

(ii) For $0 \leq s < t$, the increment $B_t - B_s$ is independent of $\mathcal{F}_s$.

(iii) For $0 \leq s < t$, $B_t - B_s$ is normally distributed with mean $0 \in \mathbb{R}^d$ and covariance matrix $(t - s)I_{d\times d}$.

(iv) The sample path $t \mapsto B_t(\omega)$ is continuous for $\mathbf{P}$-almost every ω.

If μ is the Dirac measure at $x \in \mathbb{R}^d$, then we say that B is a Brownian motion *starting at* x.

It is straightforward to check that if $\mathbf{W}_t = (W_t^1, \dots, W_t^d)$ is a vector consisting of independent 1-dimensional Brownian motions W_t^i, then $\mathbf{W}_t$ is a d-dimensional Brownian motion. (In particular, use Proposition 1.26 to verify the independence of increments). Conversely, if $\mathbf{W}_t$ is a d-dimensional Brownian motion, then the component processes $\{W_t^i : t \geq 0, i = 1, \dots, d\}$ form a family of independent 1-dimensional Brownian motions. In this context, independence means that W_t^i, W_s^j are independent for any $i, j \in \{1, \dots, d\}$ with $i \neq j$ and any $t, s \geq 0$.

3.2 Existence of Brownian motion

The objective in this section is to give a rigorous construction of one-dimensional standard Brownian motion. There are in fact several construction methods — we will first outline a classic, measure-theoretic approach via the Daniell-Kolmogorov extension theorem (2.9), then present in detail the Lévy-Ciesielski construction, which has a harmonic analysis flavour.

Measure-theoretic construction

complete later

Lévy-Ciesielski construction

We will give an explicit construction, due to Lévy and Ciesielski, of a standard one-dimensional Brownian motion on the time interval $[0, 1]$. Let us first explain why this is sufficient for the construction of Brownian motion on $[0, \infty)$. One starts with a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ whose σ-algebra is rich enough to support a countable family of independent $N(0, 1)$ (i.e. standard Gaussian) random variables. We re-index this family so that for each $n \in \mathbb{N}$, there is a countable family of independent $N(0, 1)$ random variables. Given that we can obtain for each $n \in \mathbb{N}$ a standard Brownian motion $(B_t^{(n)})_{t\in[0,1]}$, we define recursively

$$\begin{aligned} W_t &:= B_t^{(1)}, \quad t \in [0, 1], \\ W_t &:= W_n + B_{t-n}^{(n+1)}, \quad t \in [n, n + 1]. \end{aligned} \tag{3.1}$$

Thus we can glue together countably many Brownian motions on $[0, 1]$ to obtain a Brownian motion on $[0, \infty)$.

Exercise 3.2. Verify in detail that (3.1) defines a standard one-dimensional Brownian motion for $t \in [0, \infty)$.

The core idea of the Lévy-Ciesielski construction is to express Brownian motion as a series whose coefficients are independent $N(0, 1)$ random variables. This idea goes back to Paley and Wiener, who investigated Fourier series with random coefficients. Our presentation mainly follows [Eva13, Section 3.3].

Definition 3.4 (Haar wavelets)**.** On the space $L^2(0, 1)$, we define

$$h_0(t) \equiv 1,$$
$$h_1(t) := \begin{cases} 1 & 0 \leq t \leq \frac{1}{2} \\ -1 & \frac{1}{2} < t \leq 1. \end{cases}$$

Then for all $n \in \mathbb{N}$ and $2^n \leq k < 2^{n+1}$, define

$$h_k(t) := \begin{cases} 2^{n/2} & \frac{k-2^n}{2^n} \leq t \leq \frac{k-2^n+1/2}{2^n} \\ -2^{n/2} & \frac{k-2^n+1/2}{2^n} < t \leq \frac{k-2^n+1}{2^n} \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 3.5. *The collection* $\{h_k\}_{k=0}^{\infty}$ *of Haar wavelets form an orthonormal basis of* $L^2(0,1)$.

Proof. Clearly $\|h_0\|_2 = 1$ and

$$\int_0^1 |h_k(t)|^2\,dt = 2^n(2^{-(n+1)} + 2^{-(n+1)}) = 1$$

for all $k \geq 1$. If $\ell > k$, then either $h_\ell(t)h_k(t) = 0$ for all $t \in [0,1]$, or else the support of h_ℓ is contained in an interval on which h_k is constant. In the second case, we have

$$\int_0^t h_\ell(t)h_k(t)\,dt = \pm 2^{n/2}\int_0^1 h_\ell(t)\,dt = 0.$$

Hence $\{h_k\}_{k=0}^{\infty}$ is an orthonormal family in $L^2(0,1)$.

To prove that the family forms a basis, it suffices to prove that if $\langle f, h_k\rangle = 0$ for all $k \geq 0$, then $f = 0$ a.e. on $[0,1]$. For $k = 0$, we have $\int_0^1 f\,dt = 0$. For $k = 1$, the condition $\int_0^1 h_1 f\,dt = 0$ implies that

$$\int_0^{1/2} f\,dt = \int_{1/2}^1 f\,dt.$$

But then, both integrals above are equal to 0, since

$$0 = \int_0^1 f\,dt = \int_0^{1/2} f\,dt + \int_{1/2}^1 f\,dt.$$

By induction, one shows that

$$\int_{k/2^{n+1}}^{(k+1)/2^{n+1}} f\,dt = 0$$

for all $n \in \mathbb{N}$ and $0 \leq k < 2^{n+1}$. Consequently, $\int_0^1 \mathbf{1}_{[r,s]} f\,dt = 0$ for all intervals $[r,s]$ with dyadic rational endpoints. Since the collection of all dyadic intervals $[r,s] \subseteq [0,1]$ generates the Borel σ-algebra on $[0,1]$, thus the span of indicator functions $\mathbf{1}_{[r,s]}$ is dense in $L^2(0,1)$. It follows that $f = 0$ a.e., as required. □

Exercise 3.3. Fill in the details of the induction argument in Proposition 3.5.

Obviously the Haar functions are not continuous. To obtain a stochastic process with continuous sample paths, we require the following modification.

Definition 3.5. For each $k = 0, 1, 2\ldots$, define the k-th **Schauder function** by

$$S_k(t) := \int_0^t h_k(s)\,ds \qquad (0 \leq t \leq 1). \tag{3.2}$$

Observe that each S_k is continuous on $[0,1]$, and the graph of S_k, for $2^n \leq k < 2^{n+1}$, is a 'tent' of height $2^{-\frac{n}{2}-1}$ supported on the interval $[\frac{k-2^n}{2^n}, \frac{k-2^n+1}{2^n}]$. We obtain

$$\max_{t\in[0,1]} |S_k(t)| = 2^{-\frac{n}{2}-1} \quad \text{for } 2^n \leq k < 2^{n+1}. \tag{3.3}$$

The following technical result will be required later.

Lemma 3.6. *The Schauder functions satisfy*

$$\sum_{k=0}^{\infty} S_k(t)S_k(s) = t \wedge s$$

for all $0 \leq t, s \leq 1$.

Proof. For fixed $s \in [0,1]$, define

$$\phi_s(\tau) := \begin{cases} 1 & 0 \le \tau \le s \\ 0 & s < \tau \le 1. \end{cases}$$

If $s \le t$, observe that $\int_0^1 \phi_t \phi_s \, d\tau = s$. However, using Proposition 3.5, we can expand ϕ_t, ϕ_s in the Haar basis to obtain

$$\int_0^1 \phi_t h_k \, d\tau = \int_0^t h_k \, d\tau = S_k(t),$$

and likewise $\int_0^1 \phi_s h_k \, d\tau = S_k(s)$, for all $k \ge 0$. Therefore

$$\int_0^1 \phi_t \phi_s \, d\tau = \sum_{k=0}^{\infty} S_k(t) S_k(s)$$

which completes the proof. □

Our intention is to define Brownian motion on $[0,1]$ as $W(t) := \sum_{k=0}^{\infty} A_k S_k(t)$, where the A_k's are independent $N(0,1)$ random variables. The following two lemmas contain the technical work to ensure that such a series converges uniformly for $\mathbf{P}$-almost every $\omega \in \Omega$.

Lemma 3.7. *Let $(a_k)_{k=0}^{\infty}$ be a sequence of real numbers such that*

$$|a_k| \le C k^{\delta}, \quad k \ge 1,$$

for some constants $C > 0$ and $0 \le \delta < \frac{1}{2}$. Then the series

$$\sum_{k=0}^{\infty} a_k S_k(t)$$

converges uniformly for $t \in [0,1]$.

Proof. Let $\varepsilon > 0$ be arbitrary. Notice that by construction, the functions S_k for $2^n \le k < 2^{n+1}$ have disjoint supports. From the assumptions of the lemma, we have

$$b_n := \max_{2^n \le k < 2^{n+1}} |a_k| \le C(2^{n+1})^{\delta}.$$

Then for all $t \in [0,1]$, we use the above estimate together with (3.3) to find

$$\sum_{k=2^m}^{\infty} |a_k||S_k(t)| \le \sum_{n=m}^{\infty} b_n \max_{2^n \le k < 2^{n+1}} \max_{t \in [0,1]} |S_k(t)| \le C \sum_{n=m}^{\infty} (2^{n+1})^{\delta} 2^{-\frac{n}{2}-1} < \varepsilon$$

for sufficiently large $m \in \mathbb{N}$, since $0 \le \delta < \frac{1}{2}$. □

Lemma 3.8. *Let $(A_k)_{k=0}^{\infty}$ be a sequence of independent $N(0,1)$ random variables. Then for* $\mathbf{P}$*-almost every ω, it holds that*

$$|A_k(\omega)| = O(\sqrt{\log k}) \quad \text{as } k \to \infty.$$

Moreover, there exists an $\mathbb{N}$-valued random variable K such that $|A_k(\omega)| \le 4k^{1/4}$ for all $k \ge K(\omega)$ and almost every $\omega \in \Omega$.

Proof. For all $x > 0$ and $k \geq 2$, we have

$$\mathbf{P}(|A_k| > x) = \frac{2}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{y^2}{2}}\, dy \leq \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{4}} \int_x^\infty e^{-\frac{y^2}{4}}\, dy \leq \sqrt{2} e^{-\frac{x^2}{4}}$$

since $\frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-y^2/4}\, dy = \sqrt{2}$. Setting $x = 4\sqrt{\log k}$, we find

$$\mathbf{P}(|A_k| > 4\sqrt{\log k}) \leq \sqrt{2} e^{-4 \log k} = \frac{\sqrt{2}}{k^4}.$$

Since the series $\sum \frac{1}{k^4}$ converges, the Borel-Cantelli lemma (2.11) implies that

$$\mathbf{P}(|A_k| > 4\sqrt{\log k} \text{ i.o.}) = 0.$$

Therefore, for almost every ω, there exists $K(\omega) \in \mathbb{N}$ such that

$$|A_k(\omega)| \leq 4\sqrt{\log k} \leq 4k^{1/4}$$

for all $k \geq K(\omega)$. The last inequality follows from the elementary estimate $\log t \leq t^{1/2}$ for all $t \geq 1$. □

Theorem 3.9. *Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space on which there exists a sequence $(A_k)_{k=0}^\infty$ of independent $N(0,1)$ random variables. Then the series*

$$W(t, \omega) := \sum_{k=0}^\infty A_k(\omega) S_k(t), \quad t \in [0,1],$$

converges uniformly in t for $\mathbf{P}$-almost every $\omega \in \Omega$. Moreover, W is a Brownian motion on $[0,1]$ (with the natural filtration), and in particular, for a.e. ω the sample path $t \mapsto W(t,\omega)$ is continuous.

Proof. The uniform convergence of the series in t for almost every ω follows from Lemmas 3.7 and 3.8. Since the Schauder functions S_k are continuous, the uniform convergence also yields the continuity of the sample path $t \mapsto W(t,\omega)$ for almost every ω.

Now we prove that W is a Brownian motion on $[0,1]$. Clearly $W(0) = 0$. To show that each increment $W_t - W_s \sim N(0, t-s)$, we compute the characteristic function.

$$\begin{aligned}
\mathbf{E}[e^{i\lambda(W_t - W_s)}] &= \mathbf{E}\left[\exp\left(i\lambda \sum_{k=0}^\infty A_k(S_k(t) - S_k(s))\right)\right] \\
\text{(independence of } A_k\text{'s)} \quad &= \prod_{k=0}^\infty \mathbf{E}\left[e^{i\lambda A_k(S_k(t) - S_k(s))}\right] \\
\text{(since } A_k \sim N(0,1)\text{)} \quad &= \prod_{k=0}^\infty e^{-\frac{\lambda^2}{2}(S_k(t) - S_k(s))^2} \\
&= \exp\left(-\frac{\lambda^2}{2} \sum_{k=0}^\infty \left[S_k^2(t) - 2S_k(t)S_k(s) + S_k(s)^2\right]\right) \\
\text{(Lemma 3.6)} \quad &= e^{-\frac{\lambda^2}{2}(t - 2s + s)} = e^{-\frac{\lambda^2}{2}(t-s)}.
\end{aligned}$$

By uniqueness of characteristic functions, we obtain $W_t - W_s \sim N(0, t-s)$.

It remains to show that $W_t - W_s$ is independent of $\sigma(W_u : 0 \leq u \leq s)$. It suffices to prove the following: for all $m \in \mathbb{N}$ and $0 = t_0 < t_1 < \ldots < t_m \leq 1$, it holds that

$$\mathbf{E}\left[\exp\left(i \sum_{k=1}^m \lambda_k (W_{t_k} - W_{t_{k-1}})\right)\right] = \prod_{k=1}^m e^{-\frac{\lambda_k^2}{2}(t_k - t_{k-1})} \tag{3.4}$$

for all $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$. Indeed, the above implies that the random variables $W_{t_1}, W_{t_2} - W_{t_1}, \ldots, W_{t_m} - W_{t_{m-1}}$ are independent. The proof is then complete by Remark 3.1.

For the case $m = 2$, let $\lambda_1, \lambda_2 \in \mathbb{R}$ and $0 < t_1 < t_2 \le 1$ be arbitrary. We compute

$$\begin{aligned}
\mathbf{E}\big[e^{i(\lambda_1 W_{t_1} + \lambda_2(W_{t_2} - W_{t_1}))}\big] &= \mathbf{E}\big[e^{i(\lambda_1 - \lambda_2)W_{t_1} + \lambda_2 W_{t_2})}\big] \\
&= \mathbf{E}\left[\exp\left(i(\lambda_1 - \lambda_2)\sum_{k=0}^{\infty} A_k S_k(t_1) + i\lambda_2 \sum_{k=0}^{\infty} A_k S_k(t_2)\right)\right] \\
&= \prod_{k=0}^{\infty} \mathbf{E}\big(e^{iA_k[(\lambda_1 - \lambda_2)S_k(t_1) + \lambda_2 S_k(t_2)]}\big) \\
&= \prod_{k=0}^{\infty} e^{-\frac{1}{2}((\lambda_1 - \lambda_2)S_k(t_1) + \lambda_2 S_k(t_2))^2} \\
&= \exp\left(-\frac{1}{2}\sum_{k=0}^{\infty}[(\lambda_1 - \lambda_2)S_k(t_1) + \lambda_2 S_k(t_2)]^2\right).
\end{aligned}$$

Using Lemma 3.6 again, we obtain

$$\begin{aligned}
\sum_{k=0}^{\infty}[(\lambda_1 - \lambda_2)S_k(t_1) + \lambda_2 S_k(t_2)]^2 &= \sum_{k=0}^{\infty}[(\lambda_1 - \lambda_2)^2 S_k(t_1)^2 + 2(\lambda_1 - \lambda_2)\lambda_2 S_k(t_1)S_k(t_2) + \lambda_2^2 S_k(t_2)^2] \\
&= (\lambda_1 - \lambda_2)^2 t_1 + 2(\lambda_1 - \lambda_2)\lambda_2 t_1 + \lambda_2^2 t_2 \\
&= \lambda_1^2(t_1 - t_0) + \lambda_2^2(t_2 - t_1).
\end{aligned}$$

Therefore

$$\mathbf{E}\big[e^{i(\lambda_1 W_{t_1} + \lambda_2(W_{t_2} - W_{t_1}))}\big] = e^{-\frac{\lambda_1^2}{2}(t_1 - t_0)} e^{-\frac{\lambda_2^2}{2}(t_2 - t_1)},$$

which is (3.4) for $m = 2$. The general case follows by induction. □

3.3 Regularity of sample paths

The following important theorem, due to Kolmogorov (like so many important results in probability), gives a sufficient criterion for sample paths to be Hölder continuous. To state the theorem more precisely, we require a definition: given two stochastic processes $(X_t)_{t\in I}$ and $(Y_t)_{t\in I}$ defined on the same probability space, we say that Y is a **modification** of X (or a **version** of X) if

$$\mathbf{P}\big(\omega \in \Omega : X(t,\omega) = Y(t,\omega)\big) = 1 \quad \text{for every } t \in I. \tag{3.5}$$

Theorem 3.10 (Kolmogorov continuity criterion). *Let $X(\cdot)$ be an $\mathbb{R}^d$-valued stochastic process on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Suppose that*

$$\mathbf{E}|X_t - X_s|^\alpha \le C|t - s|^{1+\beta}, \qquad 0 \le s, t \le T \tag{3.6}$$

for some positive constants α, β, C. Then there exists a continuous version $\widetilde{X}$ of X which is uniformly γ-Hölder continuous for every $\gamma \in (0, \beta/\alpha)$ on $[0, T]$. More precisely,

$$\mathbf{P}\left(\omega \in \Omega : \exists\, M = M(\omega, \gamma, T) > 0 \textit{ such that } \sup_{s \ne t \in [0,T]} \frac{|\widetilde{X}_t(\omega) - \widetilde{X}_s(\omega)|}{|t - s|^\gamma} \le M\right) = 1.$$

Theorem 3.11 (Hölder continuity of Brownian motion). *Let $\mathbf{W}_t = (W_t^1, \ldots, W_t^d)$ be an $\mathbb{R}^d$-valued Brownian motion on $[0, T]$. For almost all ω, the sample path $t \mapsto \mathbf{W}(t,\omega)$ is uniformly Hölder continuous on $[0, T]$ for every exponent $0 < \gamma < \frac{1}{2}$.*

Proof. For all $s, t \in [0, T]$ such that $t - s > 0$ and for all $m \in \mathbb{N}$, we have

$$\begin{aligned}\mathbf{E}(|\mathbf{W}_t - \mathbf{W}_s|^{2m}) &= \frac{1}{(2\pi(t-s))^{d/2}} \int_{\mathbb{R}^d} |x|^{2m} e^{-\frac{|x|^2}{2(t-s)}}\, dx \\ &= \frac{(t-s)^m}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |y|^{2m} e^{-\frac{|y|^2}{2}}\, dy \\ &= C|t-s|^m.\end{aligned}$$

Thus the conditions of Theorem 3.10 are satisfied with $\alpha = 2m$ and $\beta = m - 1$. Consequently, for almost every ω, the process is uniformly Hölder continuous on $[0, T]$ for every exponent $0 < \gamma < \frac{m-1}{2m} = \frac{1}{2} - \frac{1}{2m}$. Since $m \in \mathbb{N}$ was arbitrary, the conclusion follows. □

to include: nowhere differentiability, completion of Prop.3.4

4 Stochastic integrals

We are working towards a rigorous understanding of stochastic differential equations of the form

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t \tag{4.1}$$

where $b, \sigma : \mathbb{R} \to \mathbb{R}$ are deterministic functions, and the 'noise term' dW_t intuitively represents an infinitesimal increment of Brownian motion. However, from the previous section, we know that almost all sample paths of Brownian motion are nowhere differentiable, so that dW_t cannot be interpreted in a classical way as $\frac{dW}{dt}$. Nevertheless, if we *formally* integrate equation (4.1), we arrive at the expression

$$X_t = X_0 + \int_0^t F(X_s)\, ds + \int_0^t \sigma(X_t)\, dW_s. \tag{4.2}$$

The object $\int_0^t \sigma(X_s)\, dW_s$ is a **stochastic integral** with respect to Brownian motion. We will see that such an object is well-defined for a large class of stochastic processes.

4.1 Total and quadratic variation

In order to construct a stochastic integral, we will need to examine some oscillation properties of Brownian motion. Let $f : [0,T] \to \mathbb{R}$ be a Borel measurable function, and fix a *partition* π of $[0,T]$, i.e. a finite sequence of points $0 = t_0 < t_1 < \ldots < t_n = T$. Consider the quantity

$$V_\pi(f) := \sum_{k=1}^n |f(t_k) - f(t_{k-1})|.$$

We use the standard notation $\|\pi\| := \max_{1 \le k \le n}(t_k - t_{k-1})$, which is called the *mesh size* of the partition.

Definition 4.1. Let Π denote the family of all partitions of $[0,T]$. The **total variation** of f is defined by

$$V_T(f) := \sup_{\pi \in \Pi} V_\pi(f).$$

If $V_T(f) < \infty$, we say that f has **bounded variation** and write $f \in BV([0,T])$.

For a given partition π, we also define

$$Q_\pi(f) := \sum_{k=1}^n |f(t_k) - f(t_{k-1})|^2,$$

and the **quadratic variation** is defined as

$$Q_T(f) := \lim_{\|\pi\| \to 0} Q_\pi(f). \tag{4.3}$$

Example 4.1. (i) The function $f(t) = \sin(1/t)$ on $(0,1]$ has $V(f) = \infty$.

(ii) Every bounded monotone function on $[0,T]$ is of bounded variation, with $V(f) = |f(T) - f(0)|$. For a given partition $\{0 = t_0 < t_1 < \ldots < t_n = T\}$, we clearly have

$$\sum_{k=1}^n |f(t_k) - f(t_{k-1})| = \left| \sum_{k=1}^n f(t_k) - f(t_{k-1}) \right| = |f(T) - f(0)|.$$

It is a classic result in analysis that every function $f \in BV([0,T])$ can be expressed as the difference of two monotone increasing functions. (Indeed, one can check that the decomposition $f = \frac{1}{2}(V_t(f) + f) - \frac{1}{2}(V_t(f) - f)$ works).

(iii) If f can be written as $f(t) = f(0) + \int_0^t f'(s)\,ds$ such that $\int_0^T |f'(s)|\,ds < \infty$, then $f \in BV([0,T])$. In fact, it can be shown that $V(f) = \int_0^T |f'(s)|\,ds$ in this case.

Exercise 4.1 (Important!)**.** Show that if $f \in BV([0,T]) \cap C([0,T])$, then

$$Q_T(f) := \lim_{\|\pi\| \to 0} Q_\pi(f) = 0.$$

The important observation to make from Exercise 4.1 is that a continuous function with non-zero quadratic variation cannot be of bounded variation. Understanding quadratic variation is the key to developing stochastic calculus.

Remark 4.2. Note that when (4.3) is applied to a stochastic process, the result is a *random variable*. If we compute the quadratic variation of a process on $[0,t]$ for all $t > 0$, we obtain the *quadratic variation process*.

Theorem 4.3. *Let $(W_t, \mathcal{F}_t)$ be a standard one-dimensional Brownian motion defined on $[0,T]$. Then*

$$Q_T(W) = T$$

almost surely. In particular, almost every sample path of Brownian motion is not *of bounded variation on any interval $[0,T]$.*

Proof. We will prove that

$$\mathbf{E}|Q_\pi(W) - T|^2 \longrightarrow 0 \quad \text{as } \|\pi\| \to 0.$$

Fix a partition $\pi = \{0 = t_0 < t_1 < \ldots < t_n = T\}$ of $[0,T]$. Then

$$\mathbf{E}(Q_\pi(W)) = \sum_{k=1}^n \mathbf{E}|W_{t_k} - W_{t_{k-1}}|^2 = \sum_{k=1}^n (t_k - t_{k-1}) = T$$

and consequently

$$\mathbf{E}|Q_\pi(W) - T|^2 = \mathbf{E}(Q_\pi(W)^2) - (\mathbf{E}Q_\pi(W))^2 = \mathbf{E}(Q_\pi(W)^2) - T^2.$$

It remains to calculate $\mathbf{E}(Q_\pi(W)^2)$. For convenience of notation, let us denote Brownian increments by $\Delta W_k := W_{t_k} - W_{t_{k-1}}$. Then ΔW_j and ΔW_k are independent if $j < k$. We also recall from Proposition 1.23 that if $X \sim N(0,t)$, then $\mathbf{E}(X^4) = 3t^2$. Now we compute

$$\begin{aligned}
\mathbf{E}(Q_\pi(W)^2) = \mathbf{E}\left[\left(\sum_{k=1}^n \Delta W_k^2\right)^2\right] &= \sum_{k=1}^n \mathbf{E}(\Delta W_k^4) + 2\sum_{j<k} \mathbf{E}(\Delta W_j^2 \Delta W_k^2) \\
&= 3\sum_{k=1}^n (t_k - t_{k-1})^2 + 2\sum_{j<k} (t_j - t_{j-1})(t_k - t_{k-1}) \\
&= 2\sum_{k=1}^n (t_k - t_{k-1})^2 + \left[\sum_{k=1}^n (t_k - t_{k-1})\right]^2 \\
&= 2\sum_{k=1}^n (t_k - t_{k-1})^2 + T^2.
\end{aligned}$$

Let $\{\pi_m\}_{m \geq 1}$ be any increasingly fine sequence of partitions of $[0,T]$. We may conclude

$$\mathbf{E}|Q_{\pi_m}(W) - T|^2 = \mathbf{E}(Q_{\pi_m}(W)^2) - T^2 = 2\sum_{k=1}^n (t_k - t_{k-1})^2 \leq 2\|\pi_m\| T \longrightarrow 0$$

as $m \to \infty$. The convergence in $L^2(\Omega, \mathcal{F}, \mathbf{P})$ implies almost-sure convergence up to a subsequence, and hence $Q_T(W) = T$ in L^2. The last assertion follows from Exercise 4.1. □

Definition 4.2. For a real-valued, adapted stochastic process $(X_t, \mathcal{F}_t)$, we denote the quadratic variation process of X by $\langle X \rangle_t$.

Now let X_t, Y_t be $\mathbb{R}$-valued stochastic processes. Given a partition $\pi = \{0 = t_0 < t_1 < \ldots < t_n = T\}$ of $[0, T]$, we define

$$Q_\pi(X, Y) := \sum_{k=1}^{n} (X_{t_k} - X_{t_{k-1}})(Y_{t_k} - Y_{t_{k-1}}).$$

The **cross-variation** of X and Y on the interval $[0, T]$ is then defined as

$$\langle X, Y \rangle_T := \lim_{\|\pi\| \to 0} Q_\pi(X, Y), \tag{4.4}$$

where the limit is taken in $L^2(\Omega, \mathcal{F}, \mathbf{P})$.

Finally, if $X_t = (X_t^1, \ldots, X_t^d)$ is an $\mathbb{R}^d$-valued stochastic process, the cross-variation process $\langle X \rangle_t$ of X is a matrix-valued process with entries defined as

$$\langle X \rangle_T^{ij} := \langle X^i, X^j \rangle_T \,.$$

Observe that $\langle X, X \rangle_T$ is just the quadratic variation of X on $[0, T]$. The cross-variation bracket has many properties of an inner product.

Proposition 4.4. *The cross-variation bracket $\langle \cdot, \cdot \rangle$ satisfies the following properties, where all (in)equalities are understood to hold $\mathbf{P}$-almost surely and for all $t \in [0, T]$:*

(i) $\langle X \rangle_t = \langle X, X \rangle_t \geq 0$.

(ii) $\langle X, Y \rangle_t = \langle Y, X \rangle_t$.

(iii) $\langle aX + bY, Z \rangle_t = a \langle X, Z \rangle_t + b \langle Y, Z \rangle_t$ *for all* $a, b \in \mathbb{R}$.

(iv) (Polarisation) $\langle X, Y \rangle_t = \frac{1}{2}(\langle X + Y \rangle_t - \langle X \rangle_t - \langle Y \rangle_t) = \frac{1}{4}(\langle X + Y \rangle_t - \langle X - Y \rangle_t)$.

(v) (Cauchy-Schwarz) $| \langle X, Y \rangle_t |^2 \leq \langle X \rangle_t \langle Y \rangle_t$.

If one assumes that the convergence (4.4) holds, then the proofs are easy. Nevertheless it is quite a challenge to prove that the cross-variation exists for a sufficiently large class of processes. Moreover, the definition using partitions is certainly unwieldy, and there is a much more 'natural' characterisation of quadratic variation from the point of view of martingale theory. We take these considerations for granted in the meantime, and direct the interested reader to well-known references such as [RY99, Chapter IV] and [KS91, Chapters I, III].

Observe that $\langle \cdot, \cdot \rangle$ is a bilinear form (on an appropriate space of processes), but fails to be a genuine inner product, since $\langle X \rangle_t = 0$ need not imply that $X_t = 0$ almost surely for all $t \in [0, T]$. Indeed, a counterexample is given by the trivial process $X_t = 1$ for all $\omega \in \Omega$ and $t \in [0, T]$. More generally, by Exercise 4.1, any stochastic process X for which almost every sample path belongs to $BV([0, T]) \cap C([0, T])$ satisfies $\langle X \rangle_t = 0$.

In Theorem 4.3, we computed $\langle W \rangle_t = t$ for all $t \geq 0$. This is a rather special result. It turns out that Brownian motion is essentially the unique process in the class of continuous martingales with this property. We state the following beautiful characterisation due to Lévy. To include: Levy's martingale characterisation of BM, but maybe defer to chapter 6?

4.2 Construction of the Itô integral

The result of Theorem 4.3 shows the impossibility of defining an integral with respect to Brownian motion in a 'pathwise' manner. To be precise, if we fix a sample point ω, then the function $t \mapsto W_t(\omega)$ is continuous on $[0, \infty)$. However, since it is not of bounded variation on any interval $[0, T]$, we cannot interpret

$$\int_0^T X_t(\omega) dW_t(\omega)$$

literally in the Riemann-Stieltjes sense. Nevertheless, not all is lost, and some ideas from the Riemann-Stieltjes construction are still useful. Note that it is possible to define stochastic integrals of the type

$$\int_0^T F_t \, dM_t$$

for more general processes $(M_t)_{t\geq 0}$ (namely, continuous semimartingales), but for simplicity we will mainly consider the case of Brownian motion in these notes.

We begin by defining a class of elementary processes for which there is a natural definition of stochastic integral. Throughout this section, we consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ where the filtration satisfies the *usual conditions* (recall Definition 2.4).

Definition 4.3. A stochastic process $(X_t)_{t\geq 0}$ is called **progressively measurable** if for any $t \geq 0$, the mapping

$$[0, t] \times \Omega \ni (s, \omega) \mapsto X_s(\omega) \in \mathbb{R}$$

is measurable with respect to the product σ-algebra $\mathscr{B}([0, t]) \otimes \mathcal{F}_t$.

Remark 4.5. A stochastic process is called **measurable** if the mapping $(t, \omega) \mapsto X_t(\omega)$ is measurable with respect to $\mathscr{B}([0, \infty)) \otimes \mathcal{F}$. A progressively measurable process is evidently measurable, and one can check that it is also $\mathcal{F}_t$-adapted. The converse statement 'almost' holds, due to the following theorem of Chung and Doob: if X is a measurable, $\mathcal{F}_t$-adapted process, then there exists a progressively measurable version $\widetilde{X}$ of X. This is a non-trivial result, and we will not include the technical proof, see [Mey66, Theorem T46, p. 68].

Definition 4.4. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space with a filtration $(\mathcal{F}_t)_{t\geq 0}$ satisfying the usual conditions. We say that $(F_t)_{t\geq 0}$ is an **elementary process** if it is of the form

$$F_t = \sum_{i=1}^n A_{i-1} \mathbf{1}_{(t_{i-1}, t_i]}(t)$$

for some partition $\{t_0 = 0 < t_1 < t_2 < \ldots < t_n = T\}$ of $[0, T]$, where for each $1 \leq i \leq n$, the random variable $A_{i-1} : \Omega \to \mathbb{R}$ is $\mathcal{F}_{t_{i-1}}$-measurable. We denote by $\mathcal{E}^2(0, T)$ the set of all elementary processes on $[0, T]$ satisfying $\mathbf{E}|F_t|^2 < \infty$ for all $t \in [0, T]$.

Exercise 4.2. Verify that $\mathcal{E}^2(0, T)$ is a real vector space, and that every elementary process $F \in \mathcal{E}^2(0, T)$ is progressively measurable.

For each $F \in \mathcal{E}^2(0, T)$, we observe that

$$|F_t|^2 = \sum_{i=1}^n A_{i-1}^2 \mathbf{1}_{(t_{i-1}, t_i]}(t),$$

since the indicator functions $\mathbf{1}_{(t_{i-1},t_i]}$ are disjoint. If we define the functional

$$\|F\|_{\mathbb{L}^2} := \left(\mathbf{E}\int_0^T |F_t|^2\,dt\right)^{1/2}, \tag{4.5}$$

then clearly $\|F\|_{\mathbb{L}^2} < \infty$ for all $F \in \mathcal{E}^2(0,T)$, and (4.5) defines a norm on $\mathcal{E}^2(0,T)$.

Definition 4.5 (Itô integral for elementary processes)**.** Let $(W_t)_{t\geq 0}$ be a standard one-dimensional Brownian motion defined on the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$. Let $F \in \mathcal{E}^2(0,T)$ be given, so that $F_t = \sum_{i=1}^n A_{i-1}\mathbf{1}_{(t_{i-1},t_i]}(t)$ for some partition $0 = t_0 < t_1 < \ldots < t_n = T$ and $\mathcal{F}_{t_{i-1}}$-measurable random variables A_{i-1}. We define

$$\int_0^T F_t\,dW_t := I_T(F) := \sum_{i=1}^n A_{i-1}(W_{t_i} - W_{t_{i-1}}). \tag{4.6}$$

It is essential to recognise that $I_T(F)$ is a *random variable*. The basic properties of this integral are collected below.

Lemma 4.6. *Let $F, G \in \mathcal{E}^2(0,T)$. The following assertions hold:*

(i) *$I_T(F)$ is $\mathcal{F}_T$-measurable, and $\mathbf{E}[I_T(F)] = 0$.*

(ii) (Isometry) *The map*

$$\mathcal{E}^2(0,T) \ni F \mapsto I_T(F) \in L^2(\Omega, \mathcal{F}, \mathbf{P})$$

is a linear isometry, and $\mathbf{E}|I_T(F)|^2 = \|F\|^2_{\mathbb{L}^2}$.

(iii) (Inner product) *$\langle F, G\rangle_{\mathbb{L}^2} := \mathbf{E}[I_T(F)I_T(G)]$ is well-defined, and*

$$\langle F, G\rangle_{\mathbb{L}^2} = \mathbf{E}\int_0^T F_t \cdot G_t\,dt.$$

Proof. For convenience, we write $\Delta W_i := W_{t_i} - W_{t_{i-1}}$ for Brownian increments.

(i): It is clear that $I_T(F)$ is $\mathcal{F}_T$-measurable. Since each A_{i-1} is $\mathcal{F}_{t_{i-1}}$-measurable, in particular it is independent of the Brownian increment ΔW_i. Hence

$$\mathbf{E}[I_T(F)] = \sum_{i=1}^n \mathbf{E}(A_{i-1}\Delta W_i) = \sum_{i=1}^n \mathbf{E}(A_{i-1})\mathbf{E}(\Delta W_i) = 0.$$

(ii): It is immediate from the definition (4.6) that $I_T(\lambda F) = \lambda I_T(F)$ for any $\lambda \in \mathbb{R}$. Additivity of the integral is left as an exercise (the details are similar to Exercise 4.2).

To prove the isometry property, we observe firstly that for positive integers $j < k$, the random variable $A_{j-1}A_{k-1}\Delta W_j$ is $\mathcal{F}_j$-measurable, and hence independent of ΔW_k. Then we compute

$$\begin{aligned}
\mathbf{E}|I_T(F)|^2 &= \mathbf{E}\left(\sum_{k=1}^n A_{k-1}\Delta W_k\right)^2 \\
&= \sum_{j,k=1}^n \mathbf{E}(A_{j-1}A_{k-1}\Delta W_j \Delta W_k) \\
&= \sum_{k=1}^n \mathbf{E}[A_{k-1}^2(\Delta W_k)^2] + 2\sum_{j<k}\mathbf{E}(A_{j-1}A_{k-1}\Delta W_j\Delta W_k) \\
&= \sum_{k=1}^n \mathbf{E}(A_{k-1}^2)\mathbf{E}((\Delta W_k)^2) + 2\sum_{j<k}\mathbf{E}(A_{j-1}A_{k-1}\Delta W_j)\mathbf{E}(\Delta W_k) \\
&= \sum_{k=1}^n \mathbf{E}(A_{k-1}^2)(t_k - t_{k-1}) + 0 = \mathbf{E}\int_0^T |F_t|^2\,dt = \|F\|^2_{\mathbb{L}^2}.
\end{aligned}$$

(iii): This follows from (ii) via the polarisation formula $ab = \frac{1}{4}[(a+b)^2 - (a-b)^2]$, and the reader may easily verify the details. □

The isometry property of the integral (4.6) is extremely important. It allows us to extend the definition to a larger class of stochastic processes.

Definition 4.6. Two adapted stochastic processes X and Y defined on $[0,\infty)$ are **equivalent** if

$$X_t(\omega) = Y_t(\omega) \quad \text{for } \lambda^1 \otimes \mathbf{P}\text{-almost every } (t,\omega),$$

where λ^1 is the Lebesgue measure on $[0,\infty)$. We define $\mathbb{L}^2_W(0,T)$ to be the space of equivalence classes of processes $X : [0,\infty) \times \Omega \to \mathbb{R}$ such that

(i) X is progressively measurable (see Definition 4.3); and

(ii) $\|X\|^2_{\mathbb{L}^2(0,T)} = \mathbf{E}\int_0^T |X_t|^2\, dt < \infty$.

Remark 4.7. The definition of equivalence introduced above actually depends on the quadratic variation of the process for which we will construct the stochastic integral. This is why we write $\mathbb{L}^2_W(0,T)$ to emphasise the dependence on Brownian motion. Recall that $\langle W\rangle_t = t$, and so the integral with respect to dt is simply the usual Lebesgue integral. For integration with respect to a more general continuous martingale M, we need to consider the measure $d\langle M\rangle_t$. Details can be found on [KS91, p. 130].

Since we mainly consider stochastic integrals with respect to Brownian motion in these notes, we will drop the subscript W and simply write $\mathbb{L}^2(0,T)$. Clearly this space embeds into the Hilbert space

$$\mathcal{H}_T := L^2([0,T] \times \Omega, \mathscr{B}([0,T]) \otimes \mathcal{F}, \lambda^1 \otimes \mathbf{P}).$$

Lemma 4.8. *For any $0 < T \le \infty$, the space $\mathbb{L}^2(0,T)$ is a closed subspace of $\mathcal{H}_T$, and hence is a Hilbert space with respect to the norm* (4.5).

Proof. Let $(X^{(n)})_{n\ge 1}$ be a convergent sequence in $\mathbb{L}^2(0,T)$ with limit $X \in \mathcal{H}_T$. By passing to a subsequence if necessary, we may assume that $X^{(n)}$ converges to X for $\lambda^1 \otimes \mathbf{P}$-almost every (t,ω). Then X is $\mathscr{B}([0,T]) \otimes \mathcal{F}$-measurable, but *a priori* we do not know if it is progressively measurable. To rectify this, define

$$A := \left\{(t,\omega) \in [0,T] \times \Omega : \lim_{n\to\infty} X^{(n)}_t(\omega) \text{ exists in } \mathbb{R}\right\}$$

and

$$Y_t(\omega) := \mathbf{1}_A(t,\omega) \lim_{n\to\infty} X^{(n)}_t(\omega).$$

By construction, Y is equivalent to X, and inherits progressive measurability from the sequence $(X^{(n)})_{n\ge 1}$. Thus we conclude $X \in \mathbb{L}^2(0,T)$. □

Theorem 4.9. *The space of elementary processes $\mathcal{E}^2(0,T)$ is dense in $\mathbb{L}^2(0,T)$.*

Proof. Let $X \in \mathbb{L}^2(0,T)$ be given. The proof requires a 3-step approximation.

Step 1: We first approximate X by bounded processes. This is accomplished easily by defining

$$X^{(n)}_t(\omega) := X_t(\omega)\mathbf{1}_{[-n,n]}(X_t(\omega)).$$

Obviously $X^{(n)}_t(\omega) \to X_t(\omega)$ pointwise in (t,ω) as $n \to \infty$. By the dominated convergence theorem, we have that $\|X^{(n)} - X\|_{\mathbb{L}^2(0,T)} \to 0$ as $n \to \infty$.

Step 2: Next, we show that a bounded process $Y \in \mathbb{L}^2(0,T)$ can be approximated by bounded processes with continuous paths. This is the tricky step. We can construct a sequence $(\psi_n)_{n\ge 1}$ of non-negative continuous functions satisfying

(i) $\mathrm{spt}(\psi_n) \subseteq [-\frac{1}{n}, 0]$; and

(ii) $\int_{\mathbb{R}} \psi_n(x)\,dx = 1$ for all $n \geq 1$.

Then define the processes $Y^{(n)}$ by convolutions:

$$Y^{(n)}(t,\omega) := \int_{\mathbb{R}} \psi_n(s-t) Y(s,\omega)\,ds = \int_0^t \psi_n(s-t) Y(s,\omega)\,ds, \tag{4.7}$$

where we trivially extend $Y(s,\omega) = 0$ for $s \leq 0$, for all ω. Exercise 4.3 below shows that each $Y^{(n)}$ is a bounded process with continuous paths. Importantly, each $Y^{(n)}_t$ is progressively measurable, but this not immediate—see Remark 4.10 below. For each ω, we have that

$$\int_0^T |Y^{(n)}(s,\omega) - Y(s,\omega)|^2\,ds \to 0$$

as $n \to \infty$, since $(\psi)_{n\geq 1}$ forms an approximate identity. Hence $\|Y^{(n)} - Y\|_{\mathbb{L}^2(0,T)} \to 0$ as $n \to \infty$, by dominated convergence.

Step 3: Finally, every continuous, bounded, progressively measurable process Z can be approximated by elementary processes. Given a sequence of partitions $\pi_m = \{0 = t_0 < t_1 < \ldots < t_{m_n} = T\}$ of $[0,T]$, we define

$$Z^{(n)}(t,\omega) := \sum_{k=1}^{m_n} Z(t_{k-1},\omega) \mathbf{1}_{(t_{k-1},t_k]}(t).$$

Then each $Z^{(n)}$ is elementary, since $Z_{t_{k-1}}$ is $\mathcal{F}_{t_{k-1}}$-measurable. Moreover, the continuity of $t \mapsto Z(t,\omega)$ for each ω yields that

$$\int_0^T |Z^{(n)}(t,\omega) - Z(t,\omega)|^2\,dt$$

as $n \to \infty$, for each ω. Thus we conclude $\|Z^{(n)} - Z\|_{\mathbb{L}^2(0,T)} \to 0$ as $n \to \infty$ by dominated convergence. This completes the proof. □

Exercise 4.3. (i): Consider the continuous function $\psi : \mathbb{R} \to [0,\infty)$ defined by

$$\psi(s) := 2(1 - |2s+1|) \vee 0.$$

Verify that the sequence $\psi_n(x) := n\psi(nx)$ satisfies the conditions required for Step 2 in the proof of Theorem 4.9.

(ii): Assume that $|Y(t,\omega)| \leq M$ for all (t,ω). Show that, for every $n \geq 1$, the map $t \mapsto Y^{(n)}(t,\omega)$ is continuous for every ω, and $|Y^{(n)}(t,\omega)| \leq M$.

Remark 4.10. (i): We quote the following result from [KS91, Problem 1.2.19]: let $(X_t)_{t\geq 0}$ be an $\mathbb{R}^d$-valued progressively measurable process, and $f : [0,\infty) \times \mathbb{R}^d \to \mathbb{R}$ be a bounded, $\mathscr{B}([0,\infty)) \otimes \mathscr{B}(\mathbb{R}^d)$ measurable function. Then the process

$$Y_t := \int_0^t f(s, X_s)\,ds, \quad t \geq 0$$

is progressively measurable.

(ii): In many advanced texts on stochastic calculus, stochastic integration is usually developed in the context of a general theory of continuous martingales. In Theorem 4.9 we adopt the more 'hands-on' analytic approach, as seen for example in [Øks03, Section 3.1].

Using Theorem 4.9, we can define the Itô integral for all $X \in \mathbb{L}^2(0,T)$.

Definition 4.7. Let $X \in \mathbb{L}^2(0,T)$. The **Itô integral** of X on $[0,T]$ is defined by

$$\int_0^T X_t dW_t := \lim_{n\to\infty} \int_0^T F_t^{(n)}\, dW_t, \quad \text{limit in } L^2(\Omega, \mathcal{F}, \mathbf{P}), \tag{4.8}$$

for any sequence $(F^{(n)})_{n\geq 1} \subset \mathcal{E}^2(0,T)$ of elementary processes such that $\|F^{(n)} - X\|_{\mathbb{L}^2(0,T)} \to 0$ as $n \to \infty$.

Of course, the definition does not depend on the choice of approximating sequence. Moreover, the integral satisfies all the properties listed in Lemma 4.6. We restate the properties below for the reader's convenience.

Proposition 4.11. *Let $F, G \in \mathbb{L}^2(0,T)$. The following assertions hold:*

(i) *$I_T(F)$ is $\mathcal{F}_T$-measurable, and $\mathbf{E}[I_T(F)] = 0$.*

(ii) (Isometry) *The map*

$$\mathbb{L}^2(0,T) \ni F \mapsto I_T(F) \in L^2(\Omega, \mathcal{F}, \mathbf{P})$$

is a linear isometry, and $\mathbf{E}|I_T(F)|^2 = \|F\|^2_{\mathbb{L}^2}$.

(iii) (Inner product) *$(F,G)_{\mathbb{L}^2} := \mathbf{E}[I_T(F) I_T(G)]$ is well-defined, and*

$$(F,G)_{\mathbb{L}^2} = \mathbf{E} \int_0^T F_t \cdot G_t \, dt.$$

The proofs follow from simple reasoning with limits. We observe in particular that because the isometry property holds on the dense subspace $\mathcal{E}^2(0,T)$, the linear map $F \mapsto I_T(F)$ extends in a unique way to $\mathbb{L}^2(0,T)$.

Finally, we define indefinite stochastic integrals and 'stopped' integrals.

Definition 4.8. Let $0 < T \leq \infty$ and $X \in \mathbb{L}^2(0,T)$. Given a stopping time τ such that $\tau \leq T$ almost surely, define

$$\int_0^\tau X_t \, dW_t := \int_0^T \mathbf{1}_{(t\leq\tau)}(t) X_t \, dW_t. \tag{4.9}$$

As a special case, if $\tau = t$ (i.e. a deterministic time), we obtain the **indefinite Iô integral**

$$I_t(X) = \int_0^t X_s \, dW_s, \qquad t \geq 0.$$

We note that (4.9) is well-defined, since $\mathbf{E} \int_0^T |\mathbf{1}_{(t\leq\tau)} X_t|^2 \, dt \leq \mathbf{E} \int_0^T |X_t|^2 \, dt < \infty$, and it can be shown that $\mathbf{1}_{(t\leq\tau)}(t) X_t$ is progressively measurable. Hence $\mathbf{1}_{(t\leq\tau)}(\cdot) X \in \mathbb{L}^2(0,T)$.

Integrals over other time intervals are defined in a way similar to (4.9). If $0 \leq s < t \leq T$, then we set

$$\int_s^t X_r \, dW_r := \int_0^T \mathbf{1}_{[s,t]}(r) X_r \, dW_r$$

for all $X \in \mathbb{L}^2(0,T)$. It is then clear that

$$\int_s^t X_r \, dW_r = \int_s^u X_r \, dW_r + \int_u^t X_r \, dW_r$$

for all $0 \leq s < u < t \leq T$.

One of the most important properties of the indefinite Itô integral is that it is a continuous martingale.

Theorem 4.12. *Let $F \in \mathbb{L}^2(0,T)$. Then the process $I_t(F) = \int_0^t F_s\,dW_s$ is an $\mathcal{F}_t$-martingale with a continuous version.*

Proof. We prove the martingale property for $F \in \mathcal{E}^2(0,T)$ first. Suppose that F has the decomposition

$$F_t = \sum_{k=1}^{m} A_{k-1}\mathbf{1}_{(t_{k-1},t_k]}(t).$$

The proof of Theorem 4.9 shows that we may assume F_t is almost surely bounded for each t. In particular, we may assume that each A_j is almost surely bounded.

For $s < t$, we have

$$\begin{aligned}\mathbf{E}\left(\int_0^t F_u\,dW_u\middle|\mathcal{F}_s\right) &= \mathbf{E}\left(\int_0^s F_u\,dW_u + \int_s^t F_u\,dW_u\middle|\mathcal{F}_s\right)\\ &= I_s(F) + \mathbf{E}\left(\int_s^t F_u\,dW_u\middle|\mathcal{F}_s\right),\end{aligned}$$

using the fact that $I_s(F) = \int_0^s F_u\,dW_u$ is $\mathcal{F}_s$-measurable. Next, we write $\Delta W_k = W_{t_k} - W_{t_{k-1}}$, and thus

$$\int_s^t F_u\,dW_u = \sum_{s\le t_{k-1}<t_k\le t} A_{k-1}\Delta W_k,$$

where the sum only involves points of the partition $\{0 = t_0 < t_1 < \ldots < t_m = T\}$ contained in the interval $[s,t]$. Using that $A_{k-1}\Delta W_k$ is $\mathcal{F}_s$-measurable and Proposition 1.16 ('taking out what is known'), we compute

$$\begin{aligned}\mathbf{E}\left(\int_s^t F_u\,dW_u\middle|\mathcal{F}_s\right) &= \sum_{s\le t_{k-1}<t_k\le t} \mathbf{E}[A_{k-1}\Delta W_k|\mathcal{F}_s]\\ &= \sum_{s\le t_{k-1}<t_k\le t} \mathbf{E}[\mathbf{E}(A_{k-1}\Delta W_k|\mathcal{F}_{t_{k-1}})|\mathcal{F}_s]\\ &= \sum_{s\le t_{k-1}<t_k\le t} \mathbf{E}[A_{k-1}\mathbf{E}(\Delta W_k|\mathcal{F}_{t_{k-1}})|\mathcal{F}_s] = 0.\end{aligned}$$

In the last step, we recalled that ΔW_k is independent of $\mathcal{F}_{t_{k-1}}$, and hence $\mathbf{E}(\Delta W_k|\mathcal{F}_{t_{k-1}}) = \mathbf{E}(\Delta W_k) = 0$. We have thus shown the martingale property

$$\mathbf{E}(I_t(F)|\mathcal{F}_s) = I_s(F)$$

for all $F \in \mathcal{E}^2(0,T)$. The integrability of $I_t(F)$ follows from the Hölder inequality and the Itô isometry:

$$\mathbf{E}|I_t(F)| \le \mathbf{E}|I_t(F)|^2 = \mathbf{E}\int_0^t |F_s|^2\,ds \le \|F\|^2_{\mathbb{L}^2(0,T)} < \infty.$$

Let $(F_t^{(n)})_{t\ge 0}$ be a sequence of elementary processes such that $\|F^{(n)} - F\|_{\mathbb{L}^2(0,T)} \to 0$ as $n \to \infty$. proof of continuity: requires Borel-Cantelli and Doob maximal inequality, complete later □

5 Stochastic calculus and the Itô formula

Experience with Riemann integration shows that it is one thing to develop a theory of integration and another problem entirely to compute explicit integrals. No doubt a large part of the success of the theory stochastic integration is due to the robust computational tools available, namely the *stochastic calculus*. The fundamental tool of this calculus is the *Itô formula*, or the *stochastic chain rule*.

Throughout this chapter, we employ the same notations and conventions as Chapter 4.

5.1 Computations from first principles

Before we develop the basic rules of stochastic calculus, it is important to compute some special cases of stochastic integrals from first principles.

Lemma 5.1. *Let $(W_t)_{t\geq 0}$ be a standard one-dimensional Brownian motion, and let $\pi^n = \{0 = t_0^n < t_1^n < \ldots < t_{m_n}^n = T\}$ be a sequence of partitions of $[0,T]$ such that $\|\pi^n\| \to 0$ as $n \to \infty$. Define*

$$R_\lambda^n := \sum_{k=1}^{m_n} W(\tau_k^n)[W(t_k^n) - W(t_{k-1}^n)].$$

where $\tau_k^n := (1-\lambda)t_{k-1}^n - \lambda t_k^n$ is an intermediate point. Then

$$\lim_{n\to\infty} R_n^\lambda = \frac{W(T)^2}{2} + \left(\lambda - \frac{1}{2}\right)T, \quad \textit{limit in } L^2(\Omega, \mathcal{F}, \mathbf{P}). \tag{5.1}$$

Proof. We employ a very sneaky algebraic trick:

$$b(c-a) = (b-a)(c-a) + \frac{1}{2}(c^2 - a^2) - \frac{1}{2}(c-a)^2.$$

Then

$$\begin{aligned} R_\lambda^n &= \sum_{k=1}^{m_n}[W(\tau_k^n) - W(t_{k-1}^n)][W(t_k^n) - W(t_{k-1}^n)] + \frac{1}{2}\sum_{k=1}^{m_n}[W(t_k^n)^2 - W(t_{k-1}^n)^2] \\ &\quad - \frac{1}{2}\sum_{k=1}^{m_n}[W(t_k^n) - W(t_{k-1}^n)]^2 \\ &=: A_n + B_n - C_n. \end{aligned}$$

The sum appearing in B_n is telescoping, and hence we obtain

$$B_n = \frac{1}{2}[W(t_{m_n}^n)^2 - W(t_0^n)^2] = \frac{W(T)^2}{2}$$

for all $n \geq 1$. The sum in C_n converges in $L^2(\Omega)$ to the quadratic variation of W on $[0,T]$ as $n \to \infty$, and thus $C_n \to \frac{T}{2}$ as $n \to \infty$ (see Theorem 4.3).

Next, we rewrite

$$\begin{aligned} A_n &= \sum_{k=1}^{m_n}[W(\tau_k^n) - W(t_{k-1}^n)]^2 + \sum_{k=1}^{m_n}[W(\tau_k^n) - W(t_{k-1}^n)][W(t_k^n) - W(\tau_k^n)] \\ &=: D_n + E_n. \end{aligned}$$

By adapting the proof of Theorem 4.3, one shows easily that $D_n \to \lambda T$ in $L^2(\Omega)$ as $n \to \infty$. We also leave it as a simple (but slightly tedious) exercise to show that $E_n \to 0$ in $L^2(\Omega)$ as $n \to \infty$, using the independence of Brownian increments. □

Lemma 5.1 shows that the limit of the Riemann sums R^n_λ *depends on the choice* of intermediate points! Observe that the Itô integral corresponds to $\lambda = 0$ (left end-points). Thus Lemma 5.1 establishes the result

$$\int_0^T W_t \, dW_t = \frac{W(T)^2}{2} - \frac{T}{2}$$

from first principles. There is an extra term $\frac{T}{2}$ — the *Itô correction* — which does not appear in deterministic integration. A more useful way to rewrite the formula above is to recall that the T is the quadratic variation of W on $[0, T]$. Thus

$$\int_0^t W_s \, dW_s = \frac{1}{2}(W_t^2 - \langle W \rangle_t), \qquad t \in [0, T]. \tag{5.2}$$

Formula (5.2) suggests the formal derivative

$$\text{“}d(W^2) = 2W dW + d\, \langle W \rangle_t \text{”}.$$

We will see later that Itô's chain rule essentially says that the stochastic differential (to be defined shortly) of a composition always looks like the deterministic chain rule plus a correction involving the quadratic variation of the integrand.

We now give a formal meaning to the *stochastic differential*. To do so, we introduce the space $\mathbb{L}^1(0, T)$, which is defined similarly to $\mathbb{L}^2(0, T)$. Namely, it is the (Banach) space of progressively measurable processes X such that

$$\|X\|_{\mathbb{L}^1(0,T)} := \mathbf{E} \int_0^T |X_t| \, dt < \infty.$$

Definition 5.1 (Stochastic differentials)**.** Let X be a real-valued stochastic process satisfying

$$X_t = X_0 + \int_0^t F_s \, ds + \int_0^t G_s \, dW_s \quad \mathbf{P}\text{-almost surely}$$

for some $F \in \mathbb{L}^1(0, T)$ and $G \in \mathbb{L}^2(0, T)$, and for all $0 \le t \le T$. We say that X has stochastic differential

$$dX = F dt + G dW \qquad \text{on } [0, T]. \tag{5.3}$$

It is important to note that the symbols dt, dW have no meaning alone — they are *defined* as abbreviations of the integral expression.

We compute another special case of a stochastic differential.

Lemma 5.2. *Let $f \in BV([0, T])$ (recall Definition 4.1). Then*

$$d(fW) = W df + f dW,$$

where the term $W df$ is interpreted as the usual Riemann-Stieltjes integral.

Proof. Let $0 \le r < s \le T$. The stochastic differential, by definition, means that

$$W(s)f(s) - W(r)f(r) = \int_r^s W \, df + \int_r^s f \, dW.$$

Fix a partition $\pi = \{r = t_0 < t_1 < \ldots < t_n = s\}$ of $[r, s]$. We consider the telescoping sum

$$\begin{aligned} W(s)f(s) - W(r)f(r) &= \sum_{k=1}^n W(t_k)f(t_k) - W(t_{k-1})f(t_{k-1}) \\ &= \sum_{k=1}^n f(t_{k-1})[W(t_k) - W(t_{k-1})] + \sum_{k=1}^n W(t_k)[f(t_k) - f(t_{k-1})] \\ &=: A + B. \end{aligned}$$

Since f is a bounded deterministic function, it can be trivially considered as an element of $\mathbb{L}^2(0,T)$. Hence, by the definition of the Itô integral, we have that A converges in $L^2(\Omega)$ to $\int_r^s f\,dW$ as $\|\pi\| \to 0$.

Observe that in the term B, we evaluate W at the *right* end-point of each partition interval. However, since W is almost-surely continuous and f is of bounded variation, this is just a 'usual' Riemann sum. Hence B converges almost-surely, and also in $L^2(\Omega)$ by dominated convergence, to the Riemann-Stieltjes integral $\int_r^s W\,df$ as $\|\pi\| \to 0$. □

5.2 The product rule

In this section, we prove the stochastic product rule (or integration by parts).

Theorem 5.3. *Suppose X, Y are real-valued stochastic processes on $[0,T]$ with differentials given by*

$$dX = F_1 dt + G_1 dW, \qquad dY = F_2 dt + G_2 dW,$$

where $F_i \in \mathbb{L}^1(0,T)$ and $G_i \in \mathbb{L}^2(0,T)$. Then

$$d(XY) = XdY + YdX + G_1 G_2 dt. \tag{5.4}$$

Theorem 5.3 is often derived as a simple consequence of the (multidimensional version of the) Itô chain rule. However, it can also be established directly using martingale theory, and it is perhaps more useful to view it this way. While we will not dive into the technical details in this section, let us discuss the key ideas. Firstly, we compute the cross-variation process $\langle X, Y\rangle$ (recall Definition 4.2). Write

$$X_t = X_0 + \int_0^t F_1\,dt + \int_0^t G_1\,dW =: X_0 + A_t + M_t$$

and similarly $Y_t = Y_0 + B_t + N_t$ with analogous definitions. Importantly, the processes A, B are continuous and have bounded variation on $[0,T]$. Indeed, since $F_i \in \mathbb{L}^1(0,T)$, the functions $t \mapsto F^i(t,\omega)$ belong to (the usual!) $L^1(0,T)$ for almost every ω, and then one can verify that primitives of $L^1(0,T)$ functions have the stated properties. Note also that M, N are continuous martingales (recall Theorem 4.12).

Since the cross-variation is a bilinear form, and clearly the cross-variation of a time-independent random variable with any stochastic process is 0 almost surely, we obtain

$$\begin{aligned}\langle X, Y\rangle_t &= \langle X_0 + A + M, Y_0 + B + N\rangle_t \\ &= \langle A, B\rangle_t + \langle A, N\rangle_t + \langle M, B\rangle_t + \langle M, N\rangle_t .\end{aligned}$$

We leave it as an exercise to show that all terms except $\langle M, N\rangle_t$ are equal to 0 almost surely.

Exercise 5.1. (i) Let $f \in L^1(0,T)$. Verify in detail that the function

$$F(t) := \int_0^t f(s)\,ds$$

belongs to $BV(0,T) \cap C([0,T])$.

(ii) (Important!) Let X be a process for which almost every sample path belongs to $BV(0,T) \cap C([0,T])$, and let $Y \in \mathbb{L}^2(0,T)$. Then

$$\langle X, Y\rangle_t = 0 \quad \text{a.s. for all } t \in [0,T].$$

[*Hint*: Cauchy-Schwarz.]

Exercise 5.1 shows that $\langle X, Y\rangle_t = \langle M, N\rangle_t$, and thus our problem is reduced to the computation of the cross-variation of two Itô integrals.

Proposition 5.4. *For any process $F \in \mathbb{L}^2(0,T)$, the process $X_t := \int_0^t F\,dW$ has quadratic variation*

$$\langle X\rangle_t = \int_0^t F_s^2\,d\langle W\rangle_s = \int_0^t F_s^2\,ds \quad \text{on } [0,t] \tag{5.5}$$

for every $t \in [0,T]$.

A complete proof of Proposition 5.4 is rather involved, since we need to establish that the quadratic variation actually exists for a general continuous martingale. Although we have the density result of Theorem 4.9, recall that $\int_0^t F\,dW$ is an object in its own right, so that the problem is not simply a matter of passing a limit under an integral.

We will merely give some indication of why the result is true by verifying the conclusion for elementary processes. If $F \in \mathcal{E}^2(0,T)$ with decomposition

$$F_t = \sum_{k=1}^n A_{k-1}\mathbf{1}_{(t_{k-1},t_k]}(t)$$

on $[0,T]$, we compute

$$\begin{aligned} I_t(F) = \int_0^t F_s\,dW_s &= \sum_{0\le t_{k-1}<t_k\le t} A_{k-1}(W_{t_k} - W_{t_{k-1}}) \\ &= \sum_{k=1}^n A_{k-1}(W_{t_k\wedge t} - W_{t_{k-1}\wedge t}). \end{aligned}$$

Denote $M_t^{(k)} := W_{t_k\wedge t} - W_{t_{k-1}\wedge t}$. Observe that for $t \le t_{k-1}$, we have $M_t^{(k)} = 0$, and for $t \ge t_k$, we have $M_t^{(k)} = W_{t_k} - W_{t_{k-1}}$. Hence $M^{(k)}$ is almost surely constant except on the interval $[t_{k-1}, t_k]$. Moreover, if $j \ne k$, then $M^{(k)}$ and $M^{(j)}$ are *orthogonal* in the sense that the cross-variation bracket vanishes: $\langle M^{(k)}, M^{(j)}\rangle_t = 0$ for all $t \in [0,T]$. This follows from the disjointness of the intervals $[t_{k-1}, t_k]$ and the observation that $M^{(k)}$ is constant outside of this interval. Hence we obtain

$$\begin{aligned} \langle I(F)\rangle_t &= \left\langle \sum_{k=1}^n A_{k-1}M^{(k)} \right\rangle_t \\ &= \sum_{k=1}^n \langle A_{k-1}M^{(k)}\rangle_t + 2\sum_{j<k}\langle A_{j-1}M^{(j)}, A_{k-1}M^{(k)}\rangle_t \\ &= \sum_{k=1}^n A_{k-1}^2\langle M^{(k)}\rangle_t + 2\sum_{j<k} A_{j-1}A_{k-1}\underbrace{\langle M^{(j)}, M^{(k)}\rangle_t}_{=0} = \sum_{k=1}^n A_{k-1}^2\langle M^{(k)}\rangle_t\,. \end{aligned}$$

We leave it as a simple exercise to check that

$$\langle M^{(k)}\rangle_t = \langle W_{t_k\wedge(\cdot)}\rangle_t - \langle W_{t_{k-1}\wedge(\cdot)}\rangle_t = \langle W\rangle_{t_k\wedge t} - \langle W\rangle_{t_{k-1}\wedge t}\,.$$

Therefore

$$\langle I(F)\rangle_t = \sum_{k=1}^n A_{k-1}^2\left(\langle W\rangle_{t_k\wedge t} - \langle W\rangle_{t_{k-1}\wedge t}\right) = \int_0^t F^2\,d\langle W\rangle\,,$$

which is (5.5) for elementary processes.

Taking Proposition 5.4 and some martingale theory as a black box, the proof of the product rule is now straightforward.

Proof of Theorem 5.3. We treat the special case where $X = Y$; that is, for $dX = Fdt + GdW$ we establish

$$d(X^2) = 2X dX + G^2\, dt.$$

Recall that this means

$$X_t^2 = X_0^2 + 2\int_0^t X\, dX + \int_0^t G^2\, dt \quad \text{for all } t \in [0,T]. \tag{5.6}$$

If $\pi = \{0 = t_0 < t_1 < \ldots < t_n = t\}$ is a partition of $[0,t]$, then observe that

$$\begin{aligned}\sum_{k=1}^{n}(X_{t_k} - X_{t_{k-1}})^2 &= \sum_{k=1}^{n}[X_{t_k}^2 - X_{t_{k-1}}^2 - 2X_{t_{k-1}}(X_{t_k} - X_{t_{k-1}})] \\ &= X_t^2 - X_0^2 - 2\sum_{k=1}^{n} X_{t_{k-1}}(X_{t_k} - X_{t_{k-1}}).\end{aligned}$$

As $\|\pi\| \to 0$, the left hand side converges to $\langle X\rangle_t = \int_0^t G^2\, dt$ by Proposition 5.4. It can also be shown that the sum on the right hand side converges to $\int_0^t X\, dX$. Thus

$$\int_0^t G^2\, dt = X_t^2 - X_0^2 - 2\int_0^t X\, dX,$$

and (5.6) is proved.

The general case follows by the polarisation identity:

$$\begin{aligned}d(XY) &= \frac{1}{2}d((X+Y)^2) - \frac{1}{2}d(X^2) - \frac{1}{2}d(Y^2) \\ &= (X+Y)d(X+Y) + \frac{1}{2}(G_1 + G_2)^2\, dt \\ &\qquad - \left(XdX + \frac{1}{2}G_1^2\, dt\right) - \left(YdY + \frac{1}{2}G_2^2\, dt\right) \\ &= XdY + YdX + G_1G_2 dt.\end{aligned}$$

□

Remark 5.5. To highlight the important role of quadratic variation in stochastic calculus, we often write

$$d(XY) = XdY + YdX + d\langle X, Y\rangle.$$

The cross-variation term can then be 'expanded', as described above, to produce $d\langle X, Y\rangle = G_1G_2 dt$. This is a good way to remember the stochastic product rule: it is the 'usual' product rule plus a 'stochastic correction' given by the cross-variation.

5.3 The chain rule: Itô's formula

We are now in a position to state and prove the fundamental result of stochastic calculus, the Itô chain rule. It is essential for computations with stochastic processes, but moreover it provides a concrete link between stochastic differential equations (SDE) and partial differential equations (PDE).

A function $u : [0,T] \times \mathbb{R} \to \mathbb{R}$ is said to be of class $C^{1,2}([0,T] \times \mathbb{R})$ if u is continuous, and the partial derivatives u_t, u_x, u_{xx} exist and are continuous.

Theorem 5.6. *Suppose X has stochastic differential*

$$dX = Fdt + GdW \quad \textit{on } [0,T]$$

for some $F \in \mathbb{L}^1(0,T)$ and $G \in \mathbb{L}^2(0,T)$. Assume that $u \in C^{1,2}([0,T] \times \mathbb{R})$. Then the composition $Y_t := u(t, X_t)$ has the stochastic differential

$$\begin{aligned} dY &= \frac{\partial u}{\partial t}(t, X_t)dt + \frac{\partial u}{\partial x}(t, X_t)dX + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t, X_t)G^2 dt \\ &= \left(\frac{\partial u}{\partial t}(t, X_t) + \frac{\partial u}{\partial x}(t, X_t)F + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t, X_t)G^2\right) dt + \frac{\partial u}{\partial x}(t, X_t)GdW. \end{aligned} \tag{5.7}$$

Proof. The proof proceeds in three steps.

Step 1: We begin with the case $u(x) = x^m$ for $m \in \mathbb{N}$. Thus we need to show

$$d(X^m) = mX^{m-1}dX + \frac{1}{2}m(m-1)X^{m-2}G^2\, dt. \tag{5.8}$$

This is trivial for $m = 1$, and the case $m = 2$ was proved already in Theorem 5.3. The general case follows by induction. Assume that (5.8) holds for some $m \geq 1$. We prove it for $m+1$, using the product rule:

$$\begin{aligned} d(X^{m+1}) &= d(X \cdot X^m) \\ &= Xd(X^m) + X^m dX + d\langle X, X^m\rangle \\ &= X\left(mX^{m-1}dX + \frac{1}{2}m(m-1)X^{m-2}G^2 dt\right) + X^m dX + d\langle X, X^m\rangle \\ &= (m+1)X^m dX + \frac{1}{2}m(m-1)X^{m-1}G^2 dt + d\langle X, X^m\rangle. \end{aligned}$$

Since $dX = Fdt + GdW$ and

$$\begin{aligned} d(X^m) &= mX^{m-1}(Fdt + GdW) + \frac{1}{2}m(m-1)X^{m-2}G^2 dt \\ &= (\text{bounded variation part})\, dt + mX^{m-1}GdW, \end{aligned}$$

it follows that

$$d\langle X, X^m\rangle = mX^{m-1}G^2 dt.$$

Therefore

$$\begin{aligned} d(X^{m+1}) &= (m+1)X^m dX + \left(\frac{1}{2}m(m-1) + m\right) X^{m-1}G^2 dt \\ &= (m+1)X^m dX + \frac{1}{2}(m+1)mX^{m-1}G^2 dt, \end{aligned}$$

which proves (5.8) for $m+1$.

Hence Itô's chain rule is valid for all monomials x^m, $m = 0, 1, 2, \ldots$, and consequently it holds for all polynomials by linearity of the d operator.

Step 2: Next, consider functions of the form $u(t, x) = g(t)f(x)$, where f, g are polynomials. We compute

$$\begin{aligned} du(t, X_t) &= d(g \cdot f(X)) \\ &= f(X)dg + gd(f(X)) \\ &= f(X)g'dt + g\left(f'(X)dX + \frac{1}{2}f''(X)G^2 dt\right) \\ &= \frac{\partial u}{\partial t}(t, X_t)dt + \frac{\partial u}{\partial x}(t, X_t)dX + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}(t, X_t)G^2 dt, \end{aligned}$$

and thus the Itô chain rule holds for all functions of the form

$$u(t,x) = \sum_{i=1}^{m} g^i(t) f^i(x),$$

namely, all polynomials in the two variables t, x.

Step 3: Given $u \in C^{1,2}([0,T] \times \mathbb{R})$, there exists a sequence of polynomials $(u^{(n)})_{n\in\mathbb{N}}$ in the two variables (t,x) such that

$$u^{(n)} \to u, u_t^{(n)} \to u_t, u_x^{(n)} \to u_x, u_{xx}^{(n)} \to u_{xx}$$

uniformly on compact subsets of $[0,T] \times \mathbb{R}$. From Step 2, we have

$$\begin{aligned} u^{(n)}(s,X_s) - u^{(n)}(0,X_0) = \int_0^s \left(\frac{\partial u^{(n)}}{\partial t} + \frac{\partial u^{(n)}}{\partial x} F + \frac{1}{2} \frac{\partial^2 u^{(n)}}{\partial x^2} G^2 \right) dt \\ + \int_0^s \frac{\partial u^{(n)}}{\partial x} G \, dW \quad \mathbf{P}\text{-almost surely,} \end{aligned}$$

for all $0 \le s \le T$, where we have omitted the argument (t, X_t) for brevity. To obtain Itô's formula in the general case, we need to take $n \to \infty$ in the above identity. (complete later) □

We conclude this section by stating the multidimensional generalisation of the Itô formula. We say that a function $u : [0,T] \times \mathbb{R}^n \to \mathbb{R}$ is of class $C^{1,2}([0,T] \times \mathbb{R}^n)$ if u is continuous, u_t exists and is continuous, and all spatial derivatives up to order 2 ($u_{x_i}, u_{x_i x_j}$) exist and are continuous.

Theorem 5.7. *Let $W = (W^1, \dots, W^d)$ be a d-dimensional Brownian motion, and assume the $\mathbb{R}^n$-valued process $X = (X^1, \dots, X^n)$ is given by*

$$dX = F dt + G dW,$$

where $F \in \mathbb{L}^1(0,T;\mathbb{R}^n)$ and $G \in \mathbb{L}^2(0,T;\mathbb{R}^{n\times d})$. Then, for every real-valued $u \in C^{1,2}([0,T] \times \mathbb{R}^n)$, it holds that

$$du(t,X_t) = \frac{\partial u}{\partial t}(t,X_t) + [\nabla u](t,X_t) \cdot dX + \frac{1}{2} \operatorname{tr}([D^2 u](t,X_t) G G^\top) \, dt \tag{5.9}$$

where ∇u and $D^2 u$ denote the gradient vector and Hessian matrix of u respectively.

Remark 5.8. The generalised Itô formula provides a direct link between stochastic processes and PDE. Expanding out (5.9) and suppressing the argument (t, X_t) yields the expression

$$\begin{aligned} du(t,X_t) &= \left(\frac{\partial u}{\partial t} + F \cdot \nabla u + \frac{1}{2} \operatorname{tr}(D^2 u G G^\top) \right) dt + \nabla u \cdot G dW \\ &= \left(\frac{\partial u}{\partial t} + L u \right) dt + \nabla u \cdot G dW, \end{aligned}$$

where L is the second-order differential operator given by

$$L v := F \cdot \nabla v + \frac{1}{2} \operatorname{tr}(D^2 v G G^\top).$$

In typical applications, G is an invertible matrix, and then L is *uniformly elliptic*.

In the special case that $X_t = W_t$ (so $F \equiv 0$ and $G = I$, the identity matrix), and $u \in C^2(\mathbb{R}^n)$ is time independent, we have

$$u(W_t) = u(W_0) + \int_0^t \nabla u(W_s) \cdot dW_s + \frac{1}{2}\int_0^t \Delta u(W_s)\, ds,$$

where $\Delta = \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2}$ is the Laplacian.

Example 5.9. (i) Consider the real-valued process given by

$$S_t = e^{\sigma W_t - \frac{\sigma^2}{2}t},$$

where W is a standard one-dimensional Brownian motion and $\sigma > 0$ is a constant. This is the product of two processes: $Y_t = e^{\sigma W_t}$ and $A_t = e^{-\frac{\sigma^2}{2}t}$. Note that $dA = -\frac{\sigma^2}{2}Adt$ (in the sense of ordinary calculus). The product rule yields

$$dS = d(YA) = YdA + AdY + d\langle Y, A\rangle = YdA + AdY,$$

since $\langle Y, A\rangle = 0$ (why?). Then we use the Itô chain rule to calculate

$$dY = d(e^{\sigma W}) = \sigma e^{\sigma W}dW + \frac{1}{2}\sigma^2 e^{\sigma W}dt = Y\left(\sigma dW + \frac{\sigma^2}{2}dt\right)$$

Therefore

$$\begin{aligned} dS &= Y\left(-\frac{\sigma^2}{2}Adt\right)dt + AY\left(\sigma dW + \frac{\sigma^2}{2}dt\right) \\ &= -\frac{\sigma^2}{2}Sdt + \sigma SdW + \frac{\sigma^2}{2}Sdt = \sigma SdW. \end{aligned}$$

We discover that S solves the stochastic differential equation $dS = \sigma SdW$. Observe in particular that the solution is *not* simply given by $e^{\sigma W_t}$.

(ii) Let X satisfy the stochastic differential equation $dX = bXdt + \sigma XdW$ on $[0, T]$, where $b \in \mathbb{R}$ and $\sigma > 0$ are constants. Therefore, for all $t \in [0, T]$, we have

$$\int_0^t \frac{dX_s}{X_s} = \int_0^t b\, ds + \int_0^t \sigma dW_s = bt + \sigma W_t.$$

On the other hand, motivated by the term $\frac{dX}{X}$, we compute the following using the Itô formula:

$$d(\ln X) = \frac{1}{X}dX - \frac{1}{2}\frac{1}{X^2}(\sigma^2 X^2)dt = \frac{1}{X}dX - \frac{1}{2}\sigma^2 dt,$$

which by definition means that

$$\ln\left(\frac{X_t}{X_0}\right) = \int_0^t \frac{dX_s}{X_s} - \frac{1}{2}\int_0^t \sigma^2\, dt = \int_0^t \frac{dX_s}{X_s} - \frac{1}{2}\sigma^2 t.$$

This shows that

$$\int_0^t \frac{dX_s}{X_s} = \ln\left(\frac{X_t}{X_0}\right) + \frac{1}{2}\sigma^2 t = bt + \sigma W_t,$$

and hence, we have found a solution to the original SDE:

$$X_t = X_0 \exp\left[\left(b - \frac{1}{2}\sigma^2\right)t + \sigma W_t\right], \quad t \in [0, T].$$

This represents a very simple model for a stock price with *drift* b, *volatility* σ, and initial price X_0.

Exercise 5.2. (i) Derive the product rule using the multidimensional Itô formula with $u(x,y) = xy$.

(ii) For a 1-dimensional Brownian motion, prove the identity

$$\int_0^t W_s^2 \, dW_s = \frac{1}{3} W_t^3 - \int_0^t W_s \, ds.$$

(iii) Let $W = (W^1, W^2)$ be a 2-dimensional Brownian motion. Compute the stochastic differential of the process $Y_t := |W_t|^2 = (W_t^1)^2 + (W_t^2)^2$.

5.4 Supplement: Stratonovich integral

(include later)

6 Semimartingales and Itô processes

This chapter could have been placed much earlier in these notes (after Chapter 2, for instance). It appears in its current position simply because we wished to give a rather direct approach to stochastic calculus, with minimal development of the general theory of martingales. However, it would be foolish to omit this material, since martingales are such a fundamental part of stochastic analysis. Despite this, the impatient reader could skip ahead to the next chapter on first reading without spoiling the overall story.

We continue with the conventions and notations of the previous two chapters. In particular, we are always working on a filtered probability space, and for convenience, we will assume that the filtration $(\mathcal{F})_{t\geq 0}$ satisfies the usual conditions. In addition, we consider exclusively processes with continuous sample paths.

6.1 Local martingales

7 Introduction to stochastic differential equations

References

[Bas98] Richard F. Bass. *Diffusions and elliptic operators*. Probability and its Applications (New York). Springer-Verlag, New York, 1998.

[Eva13] Lawrence C. Evans. *An introduction to stochastic differential equations*. American Mathematical Society, Providence, RI, 2013. doi:10.1090/mbk/082.

[KS91] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991. doi:10.1007/978-1-4612-0949-2.

[LG16] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, [Cham], french edition, 2016. doi:10.1007/978-3-319-31089-3.

[Mey66] Paul-A. Meyer. *Probability and potentials*. Blaisdell Publishing Co. [Ginn and Co.], Waltham, Mass.-Toronto, Ont.-London, 1966.

[Øks03] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications. doi:10.1007/978-3-642-14394-6.

[RY99] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999. doi:10.1007/978-3-662-06400-9.

[Wil91] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. doi:10.1017/CBO9780511813658.