Stochastic Calculus

Quasar Chunawala

July 4, 2024

Abstract

The most important results and ideas in basic mathematical finance.

1 Measure.

1.1 Null Sets.

Definition 1.1. *(Null Set)*. A *null set* is a set that can be covered by a sequence of intervals of arbitrarily small total length. Given any $\epsilon > 0$, there exists a sequence of intervals $(I_n)_{n\geq 1}$ such that:

$$A \subseteq \bigcup_{n=1}^{\infty} I_n$$

and

$$\sum_{n=1}^{\infty} l(I_n) < \epsilon$$

Problem 1.1. Show that we get an equivalent notion if in the above definition we replace the word intervals by any of these: open-intervals, closed-intervals, intervals of the form $(a, b]$ or intervals of the form $[a, b)$.

Proof. Let A be a null set. Then, we can cover it by a sequence of intervals, such that total length of the cover can be made as small as we please. Mathematically,

$$\forall \epsilon > 0, \exists (I_n)_{n\geq 1}, A \subseteq \bigcup_{n=1}^{\infty} I_n, \text{ such that } \sum_{n=1}^{\infty} l(I_n) < \frac{\epsilon}{2}$$

Let $I_n = [a_n, b_n]$ and define:

$$J_n := \left(a_n - \frac{\epsilon}{2^{n+2}}, b_n + \frac{\epsilon}{2^{n+2}}\right)$$

Since, $I_n \subseteq J_n$, it follows that:

$$A \subseteq \bigcup_{n=1}^{\infty} I_n \subseteq \bigcup_{n=1}^{\infty} J_n$$

Moreover,

$$\begin{aligned}
l(J_n) &= l(I_n) + \frac{\epsilon}{2^{n+1}} \\
\sum_{n=1}^{\infty} l(J_n) &= \sum_{n=1}^{\infty} l(I_n) + \frac{\epsilon}{2^2}\left(1 + \frac{1}{2} + \frac{1}{2^2} + \dots\right) \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2^2} \cdot \frac{1}{1 - 1/2} \\
&= \epsilon
\end{aligned}$$

Consequently, A can be covered by a sequence of open intervals, whose total length can be made arbitrarily small. This closes the proof. $\square$

Theorem 1.1. *If $(N_n)_{n\geq 1}$ is a sequence of null sets, then their countable union*

$$N = \bigcup_{n=1}^{\infty} N_n$$

is also null.

Proof. Since N_1 is null, there exists a sequence of intervals (I_k^1) such that $N_1 \subseteq \bigcup_{k=1}^{\infty} I_k^1$ and $\sum_{k=1}^{\infty} l(I_k^1) < \frac{\epsilon}{2^2}$.

Since N_2 is null, there exists a sequence of intervals (I_k^2) such that $N_2 \subseteq \bigcup_{k=1}^{\infty} I_k^2$ and $\sum_{k=1}^{\infty} l(I_k^2) < \frac{\epsilon}{2^3}$.

Since N_j is null, there exists a sequence of intervals (I_k^j) such that $N_j \subseteq \bigcup_{k=1}^{\infty} I_k^j$ and $\sum_{k=1}^{\infty} l(I_k^j) < \frac{\epsilon}{2^{2+j}}$.

Clearly, we have:

$$\bigcup_{j=1}^{\infty} N_j \subseteq \bigcup_{j=1}^{\infty} \bigcup_{k}^{\infty} I_k^j$$

Moreover,

$$\begin{aligned}
\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} l(I_k^j) &< \frac{\epsilon}{2^2} + \frac{\epsilon}{2^3} + \dots \\
&= \frac{\epsilon}{2^2}\left[1 + \frac{1}{2} + \frac{1}{2^2} + \dots\right] \\
&= \frac{\epsilon}{2} < \epsilon
\end{aligned}$$

Consequently, N is a null set. $\square$

A singleton set $\{x\}$ is a null set - let $I_1 = [x - \frac{\epsilon}{4}, x + \frac{\epsilon}{4}], I_n = [x, x]$ for $n \geq 2$. Thus, any countable set is a null set, and null sets appear to be closely related to countable sets - this is no surprise, as any proper interval is uncountable, so any countable subset is quite sparse when compared with an interval, hence makes no real contribution to its *length*.

However, uncountable sets can be null, provided their points are sufficiently *sparsely distributed*, as the following example due to Cantor shows:

1. Start with the interval $C_0 = [0, 1]$, remove the open middle one-third, that is the interval $\left(\frac{1}{3}, \frac{2}{3}\right)$, ontaining C_1 which consists of two intervals $[0, \frac{1}{3}]$ and $[\frac{2}{3}, 1]$.

2. Next, remove the middle third of each of these two intervals leaving C_2, consisting of four intervals $[0, \frac{1}{9}]$, $[\frac{2}{9}, \frac{3}{9}]$, $[\frac{6}{9}, \frac{7}{9}]$ and $[\frac{8}{9}, 1]$.

3. At the nth stage, we have a set C_n consisting of 2^n disjoint closed intervals, each of length $\frac{1}{3^n}$. Thus, the total length of C_n is $\left(\frac{2}{3}\right)^n$.

We call

$$C = \bigcap_{n=1}^{\infty} C_n$$

the *Cantor set*. Now, we show that C is null as promised.

Given any $\epsilon > 0$, choose n such that $\left(\frac{2}{3}\right)^n < \epsilon$. Since, $C \subseteq C_n$ and C_n is a union of disjoint intervals of total length less than ϵ, we see that C is a null set. All that remains to be checked is that C is an uncountable set.

Problem 1.2. Prove that C is uncountable.

Proof. Let $x \in C$ be an arbitrary point.

Starting with $I_0 = [0, 1]$, for all $n \in \mathbf{N}$, define the sequence of intervals (I_n), where $I_n = [a_n, b_n]$, (L_n) and (R_n) as:

$$L_{n+1} = \left[a_n, a_n + \frac{1}{3^{n+1}}\right]$$
$$R_{n+1} = \left[b_n - \frac{1}{3^{n+1}}, b_n\right]$$
$$I_{n+1} = \begin{cases} L_{n+1} & \text{if } x \in L_{n+1} \\ R_{n+1} & \text{otherwise} \end{cases}$$

Clearly, the left-end point of I_{n+1}, is a_n, if $x \in L_{n+1}$, otherwise it is $b_n - \frac{1}{3^{n+1}} = a_n + \frac{1}{3^n} - \frac{1}{3^{n+1}} = a_n + \frac{2}{3^{n+1}}$. To summarize:

$$a_{n+1} = \begin{cases} a_n + \frac{0}{3^{n+1}} & \text{if } x \in L_{n+1} \\ a_n + \frac{2}{3^{n+1}} & \text{if } x \in R_{n+1} \end{cases}$$

We have that, since $C \subseteq [0, 1]$, it implies $x \in I_0$. By construction, if $x \in I_n$, then $x \in I_{n+1}$.

Hence,

$$a_{n+1} = \sum_{i=1}^{n+1} \frac{x_k}{2^k}$$

where $x_k \in \{0, 2\}$ and (by induction)

$$x \in I_n, \quad \forall n \in \mathbf{N}$$

Pick an arbitrary $\epsilon > 0$. We can choose N such that $l(I_N) = \frac{1}{3^N} < \epsilon$. Consequently, for all $n \geq N$, $|a_n - x| < l(I_n) < \epsilon$. Hence, $(a_n) \to x$.

Thus, x can be written in the ternary system as an infinite-length (non-terminating) string of $0s$ and 2s. That is $x = (0.x_1x_2x_3\ldots)_3$.

By Cantor's diagonal argument, the collection of all infinite-length (non-terminating) binary strings consisting of $0s$ and $2s$ is uncountable. So, C is uncountable. $\square$

1.2 Outer Measure.

Definition 1.2. (*Outer measure*) The *outer measure* of any set $A \subseteq \mathbf{R}$ is given by:

$$\mu^*(A) = \inf Z_A$$

where

$$Z_A = \left\{ \sum_{n=1}^{\infty} l(I_n) : I_n \text{ are intervals }, A \subseteq \bigcup_{n=1}^{\infty} I_n \right\}$$

We say that the $(I_n)_{n\geq 1}$ covers the set A. So, the outer measure is the infimum of lengths of all possible covers of A (Note again, that some of the I_n may be empty; this avoids having to worry whether the sequence (I_n) has finitely or infinitely many different members.)

Clearly, $\mu^*(A) \geq 0$ for any $A \subseteq \mathbf{R}$. For some sets A, the series $\sum_{n=1}^{\infty} l(I_n)$ may diverge for any covering of A, so $\mu^*(A)$ may be equal to ∞. Since we wish to be able to add the outer measures of various sets we have to adopt a convention to deal with infinity. An obvious choice is $a + \infty = \infty$, $\infty + \infty = \infty$ and a less obvious but quite practical assumption is $0 \times \infty = 0$, as we have already seen.

The set Z_A is bounded from below by 0 so that the infimum always exists. If $r \in Z_A$, then $[r, +\infty] \subseteq Z_A$ (clearly we may expand the first interval of any cover to increase the total length by any number). This shows that Z_A is either $+\infty$ or the interval (x, ∞) or $[x, \infty]$ for some real number x. So, the infimum of Z_A is just x.

First, we show that the concept of a null set is consistent with that of Outer measure.

Theorem 1.2. *A set $A \subseteq \mathbf{R}$ is a null set if and only if $\mu^\star(A) = 0$.*

Proof. ($\Longrightarrow$ direction).

Suppose that A is a null set. We wish to show that $\inf Z_A = 0$. To this end, our claim is, that given any $\epsilon > 0$, there exists $z \in Z_A$ such that $0 < z < \epsilon$.

By definition of a null set, we can find a sequence of intervals $(I_n)_{n\geq 1}$ covering A such that $\sum_{n=1}^{\infty} l(I_n) < \epsilon/2$ and so $\sum_{n=1}^{\infty} l(I_n)$ is an element of Z_A.

($\Longleftarrow$direction).

Suppose that A is a set such that $\mu^*(A) = 0$. That is, $\inf Z_A = 0$. Pick an arbitrary $\epsilon > 0$. By the definition of inf, there exists $z \in Z_A$, such that $z < \epsilon$. But, a member of Z_A is the total length of some covering of A. That is, there exists a covering (I_n) of A, with total length smaller than ϵ. Since, $\epsilon > 0$ was arbitrary to begin with, this is true for all $\epsilon > 0$. Hence, A is a null set. $\square$

This combines our general *outer measure* with the special case of zero measure. Note that, $\mu^*(\emptyset) = 0$ and $\mu^*(\{x\}) = 0$ and $\mu^*(\mathbf{Q}) = 0$.

Next, we observe that $\mu^\star$ is monotone: the bigger the set, the greater is its outer measure.

Proposition 1.1. *If $A \subset B$, then $\mu^*(A) \leq \mu^*(B)$.*

Proof. Let (I_n) be an arbitrary covering for B. Then, $B \subseteq \bigcup_{n=1}^{\infty} I_n$. Since $A \subset B$, it follows that $A \subset \bigcup_{n=1}^{\infty} I_n$. Thus, $(I_n)_{n\geq 1}$ covers A. So, every cover for B covers A. Consequently, $Z_B \subseteq Z_A$.

Now, $B \setminus A$ is non-empty, so let $x \in B \setminus A$.

Now, let (J_n) be a covering for A, where $J_n = (a_n, b_n)$. Define:

$$J_n' = \begin{cases} (a_n, x) \cup (x, b_n) & \text{if } x \in J_n \\ J_n & \text{otherwise} \end{cases}$$

$(J_n')_{n\geq 1}$ covers A, but not B. Let $z = \sum_{n=1}^{\infty} l(J_n')$. Thus, there exists $z \in Z_A$, such that $z \notin Z_B$. So, $Z_B \subset Z_A$. By the properties of inf, it follows that $\inf Z_A \leq \inf Z_B$. Thus, $\mu^*(A) \leq \mu^*(B)$. $\square$

Theorem 1.3. *The outer measure of an interval equals its length.*

If I is an interval, we have:

$$\mu^\star(I) = l(I)$$

Proof. If I is unbounded then, it is clear that it cannot be covered by a system of intervals of with finite total length. This shows that $\mu^*(I) = \infty$ and so $\mu^*(I) = l(I) = \infty$.

So we restrict ourselves to bounded intervals.

Step 1. $\mu^\star(I) \leq l(I)$.

Take the following sequence of intervals. $I_1 = I$, $I_n = [0, 0]$ for all $n \geq 2$. Then, $\sum_{n=1}^{\infty} l(I_n) = l(I)$. So, $l(I) \in Z_I$. But, $\mu^\star(I) = \inf Z_I \leq l(I)$.

Step II. $l(I) \leq \mu^*(I)$.

(i) $I = [a, b]$. We shall show that for any $\epsilon > 0$:

$$l([a, b] \leq \mu^*([a, b]) + \epsilon \tag{1.1}$$

Pick an arbitrary $\epsilon > 0$. By the definition of outer measure, there exists a sequence of intervals (I_n) such that :

$$\inf Z_I = \mu^*(I) \leq \sum_{n=1}^{\infty} l(I_n) < \mu^*(I) + \frac{\epsilon}{2} \tag{1.2}$$

We shall slightly increase each of the intervals to an open one. Let the endpoints of I_n be a_n, b_n and we take:

$$J_n = \left(a_n - \frac{\epsilon}{2^{n+2}}, b_n + \frac{\epsilon}{2^{n+2}}\right)$$

It is clear that

$$l(I_n) = l(J_n) - \frac{\epsilon}{2^{n+1}}$$

so that:

$$\sum_{n=1}^{\infty} l(I_n) = \sum_{n=1}^{\infty} l(J_n) - \frac{\epsilon}{2}$$

We insert this in 1.2, and we have:

$$\sum_{n=1}^{\infty} l(J_n) \leq \mu^*([a,b]) + \epsilon \tag{1.3}$$

The new sequence of intervals cover $[a,b]$, so by the Heine Borel theorem, we can choose a finite number of J_n to cover $[a,b]$ (the set $[a,b]$ is compact in $\mathbf{R}$). We can add some intervals to this finite family to form an initial segment of the sequence - just for the simplicity of notation. So, for some finite index m we have:

$$[a,b] \subseteq \bigcup_{n=1}^{m} J_n$$

Let $J_n = [c_n, d_n]$. Put $c = \min\{c_1, \ldots, c_m\}$ and $d = \max\{d_1, \ldots, d_m\}$. Then, the above covering means that $c < a$ and $b < d$ and hence $l([a,b]) < d - c$.

Next, the number $d - c$ is certainly smaller than the total length of J_n, $n = 1, 2, 3, \ldots, m$ (some overlapping takes place) and

$$l(a,b) < d - c < \sum_{j=1}^{m} l(J_n) \tag{1.4}$$

Now, it is sufficient to put (1.3) and (1.4) together to deduce (1.1). (The finite sum is less than equal to the sum of the series, since all terms are non-negative) Letting $\epsilon \to 0$, we have the desired result. $l([a,b]) \leq \mu([a,b])$.

(ii) What if $I = (a,b)$?

Fix an arbitrary $\epsilon > 0$ as before. As before it is sufficient to show (1.1). We have:

$$\begin{aligned} l((a,b) &= l\left(\left[a + \frac{\epsilon}{2}, b - \frac{\epsilon}{2}\right]\right) + \epsilon \\ &= \mu^*\left(\left[a + \frac{\epsilon}{2}, b - \frac{\epsilon}{2}\right]\right) + \epsilon \\ &\quad \{\text{ From part I }\} \\ &\leq \mu^*((a,b)) + \epsilon \\ &\quad \{\text{ By monotonicity of outer measure (1.1)}\} \end{aligned}$$

(iii) $I = (a,b]$ or $I = [a,b)$.

$$\begin{aligned} l(I) = l((a,b)) &\leq \mu^*((a,b)) && \{\text{ From part II }\} \\ &\leq \mu^*(I) && \{\text{ Monotonicity of Lebesgue Measure }\} \end{aligned}$$

This closes the proof. $\square$

Theorem 1.4. *(Countable Subadditivity). The outer measure is countably subadditive.*

For all sequences of sets (E_n), we have:

$$\mu^\star(\bigcup_{n=1}^{\infty} E_n) \leq \sum_{n=1}^{\infty} \mu^\star(E_n)$$

(Note that both sides might be infinite here.)

Proof. (A warm-up)

Let's first prove a simpler statement:

$$\mu^*(E_1 \cup E_2) \le \mu^*(E_1) + \mu^*(E_2)$$

Take an $\epsilon > 0$ and we show an even easier inequality:

$$\mu^*(E_1 \cup E_2) \le \mu^*(E_1) + \mu^*(E_2) + \epsilon$$

By the definition of outer measure,

There exists a sequence of intervals (I_n^1) covering E_1 such that:

$$\mu^*(E_1) < \sum_{n=1}^{\infty} l(I_n^1) < \mu^*(E_1) + \frac{\epsilon}{2}$$

There exists a sequence of intervals (I_n^2) covering E_2 such that:

$$\mu^*(E_2) < \sum_{n=1}^{\infty} l(I_n^2) < \mu^*(E_2) + \frac{\epsilon}{2}$$

Now, the sequence of intervals $I_1^1, I_1^2, I_2^1, I_2^2, \ldots$ covers $E_1 \cup E_2$. Hence,

$$\begin{aligned} \mu^*(E_1 \cup E_2) &\le \sum_{n=1}^{\infty} \left(l(I_n^1) + l(I_n^2) \right) \\ &\le \mu^*(E_1) + \frac{\epsilon}{2} + \mu^*(E_2) + \frac{\epsilon}{2} \\ &= \mu^*(E_1) + \mu^*(E_2) + \epsilon \end{aligned}$$

Since ϵ was arbitrary, this is true for all $\epsilon > 0$.

Choosing $\epsilon = \frac{1}{n}$, passing to the limit as $n \to \infty$, we have:

$$\mu^*(E_1 \cup E_2) \le \mu^*(E_1) + \mu^*(E_2)$$

(*Proof of the theorem.*)

If the right-hand side is infinite, then inequality is of course true. So, suppose that $\sum_{k=1}^{\infty} \mu^*(E_n) < \infty$. For each given $\epsilon > 0$ and $k \ge 1$, find a covering sequence (I_n^k) of E_k with :

$$\sum_{n=1}^{\infty} l(I_n^k) < \mu^*(E_k) + \frac{\epsilon}{2^k}$$

The iterated series

$$\sum_{k=1}^{\infty}\left(\sum_{n=1}^{\infty} l(I_n^k)\right) < \sum_{k=1}^{\infty} \mu^*(E_k) + \epsilon < \infty$$

Now, $I_1^1, I_2^1, I_1^2, I_3^1, I_2^2, I_1^3, \ldots$ is a countable sequence (since $\mathbf{N} \times \mathbf{N}$ is countable) that covers $\bigcup_{k=1}^{\infty} E_k$. So,

$$\mu^*\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty}\left(\sum_{n=1}^{\infty} l(I_n^k)\right) < \sum_{k=1}^{\infty} \mu^*(E_k) + \epsilon$$

To complete the proof, we simply let $\epsilon \to 0$. $\square$

Problem 1.3. Prove that if $\mu^\star(A) = 0$ then for each B, $\mu^\star(A \cup B) = \mu^\star(B)$.

Proof. Let B be an arbitrary set. By countable additivity of outer-measure, we have:

$$\begin{aligned} \mu^*(A \cup B) &\leq \mu^*(A) + \mu^*(B) \\ &= \mu^*(B) \end{aligned}$$

Since $B \subseteq A \cup B$, by the monotonicity of outer-measure,

$$\mu^*(B) \leq \mu^*(A \cup B)$$

From the above discussion, it follows that, $\mu^*(A \cup B) = \mu^*(B)$. Since, B was arbitrary, this must be true for all sets B. This closes the proof. $\square$

Problem 1.4. Prove that if $\mu^\star(A \triangle B) = 0$, then $\mu^\star(A) = \mu^\star(B)$.

Proof. We know that, $A \subseteq B \cup (A \triangle B)$. Hence:

$$\begin{aligned} \mu^*(A) &\leq \mu^*(B \cup (A \triangle B)) \\ &\quad \{ \text{ Monotonicity of Outer Measure } \} \\ &\leq \mu^*(B) + \mu^*(A \triangle B) \\ &\quad \{ \text{ Countable Subadditivity } \} \\ &= \mu^*(B) \end{aligned}$$

On the other hand, $B \subseteq A \cup (A \triangle B)$. Hence, $\mu^*(B) \leq \mu^*(A)$. Consequently, it follows that

$$\mu^*(A) = \mu^*(B)$$

$\square$

Proposition 1.2. *The outer measure is translation invariant.*

$$\mu^*(A) = \mu^*(A + t)$$

for each A and t.

Proof. Let $A \subset \mathbf{R}$ and t be a fixed real.

Let (I_n) be any sequence of intervals covering A. Then, $I_n' = [a_n + t, b_n + t]$ is a covering for $A + t$. Now, $l(I_n') = l(I_n)$ for all $n \in \mathbf{N}$. So, $\sum_{n=1}^{\infty} l(I_n') = \sum_{n=1}^{\infty} l(I_n)$. Hence, if $z \in Z_A$, it follows that $z \in Z_{A+t}$ and vice-versa. Consequently, $Z_A = Z_{A+t}$. So, $\inf Z_A = \inf Z_{A+t}$. Therefore, $\mu^*(A) = \mu^*(A + t)$. $\square$

1.3 Lebesgue measurable sets and Lebesgue measure.

With the outer measure, subadditivity as in Theorem (1.4) is as far as we can get. We wish to however, ensure, that, if the sets (E_n) are pairwise disjoint (that is $E_i \cap E_j = \emptyset$, $i \neq j$) then the inequality in Theorem (1.4) becomes an equality. It turns out that this will not in general be true for the outer-measure. But our wish is entirely a reasonable one: any length function should atleast be finitely additive, since decomposing a set into finitely many disjoint pieces, should not alter it's length. Moreover, since we constructed our length function via the approximation of complicated sets by simpler sets (that is intervals), it seems fair to demand a *continuity property* : if pairwise disjoint E_n have union E, then the lengths of sets $B_n = E \setminus \bigcup_{k=1}^{n} E_k$ may be expected to decrease to 0 as $n \to \infty$. Combining this with finite additivity leads quite naturally to demand that length be countably additive, that is:

$$\mu^{\star}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu^{\star}(E_n) \quad \text{when } E_i \cap E_j = \emptyset \text{ for } i \neq j \tag{1.5}$$

We therefore turn to the task of finding the class of sets in **R** which have this property.

Definition 1.3. A set E is Lebesgue *measurable* if for every set $A \subseteq \mathbf{R}$ we have:

$$\mu^{\star}(A) = \mu^{\star}(A \cap E) + \mu^{\star}(A \cap E^C) \tag{1.6}$$

We write $E \subset \mathcal{F}$.

We obviously have $A = (A \cap E) \cup (A \cap E^C)$, hence by countable subadditivity (1.4), we have:

$$\mu^{*}(A) \leq \mu^{*}(A \cap E) + \mu^{*}(A \cap E^C)$$

for any A and E. So, our future task of verifying countable additivity property (1.5) has simplified: $E \in \mathcal{F}$ if and only the following inequality holds:

$$\mu^{*}(A) \geq \mu^{*}(A \cap E) + \mu^{*}(A \cap E^C)$$

for all $A \subseteq \mathbf{R}$.

Now, we give examples of measurable sets.

Theorem 1.5. *(i) Any null set is measurable.*

(ii) Any interval is measurable.

Proof. (i) If N is a null set, then by Theorem (1.2), the null set has outer measure zero, so $\mu^{*}(N) = 0$. For all $A \subseteq \mathbf{R}$, since $A \cap N \subseteq N$ and $A \cap N^C \subseteq A$. Thus,

$$\mu^{*}(A \cap N) + \mu^{*}(A \cap N^C) \leq \mu^{*}(N) + \mu^{*}(A)$$
$$\mu^{*}(A \cap N) + \mu^{*}(A \cap N^C) \leq \mu^{*}(A)$$

(ii) Let $E = I$ be an interval. Suppose, for example, $I = [a, b]$. Take any $A \subseteq \mathbf{R}$ and $\epsilon > 0$. Find a covering of A with:

$$\mu^{\star}(A) \leq \sum_{n=1}^{\infty} l(I_n) \leq \mu^{\star}(A) + \epsilon$$

Clearly, the intervals $I'_n = I_n \cap [a,b]$ cover $A \cap [a,b]$ and hence $\sum l(I'_n) \in Z_{A\cap[a,b]}$, that is,

$$\mu^\star(A \cap [a,b]) \leq \sum_{n=1}^{\infty} l(I'_n)$$

The intervals $I''_n = I_n \cap (-\infty, a)$ and $I'''_n = I_n \cap (b, +\infty)$ cover $A \cap [a,b]^c$, so:

$$\mu^*(A \cap [a,b]^c) \leq \sum_{n=1}^{\infty} l(I''_n) + l(I'''_n)$$

Since, the intervals $I'_n \cup I''_n \cup I'''_n$ cover A, it follows that:

$$\begin{aligned}
\mu^*(A \cap [a,b]) + \mu^*(A \cap [a,b]^c) &\leq \sum_{n=1}^{\infty} l(I'_n) + l(I''_n) + l(I'''_n) \\
&= \sum_{n=1}^{\infty} l(I_n) \\
&\leq \mu^*(A) + \epsilon
\end{aligned}$$

Letting $\epsilon \to 0$, we have the desired result. □

The fundamental properties of the class $\mathcal{F}$ of all Lebesgue measurable subsets of $\mathbf{R}$ can now be proved. They fall into two categories: first we show that certain set operations on $\mathcal{F}$ produce sets in $\mathcal{F}$(these are what we call closure properties) and second we prove that for sets in $\mathcal{F}$ the outer measure μ^* has the property of countable additivity announced above.

Theorem 1.6. *(Closure properties of $\mathcal{F}$)*

(i) $\mathbf{R} \in \mathcal{F}$.

(ii) If $E \in \mathcal{F}$, then $E^C \in \mathcal{F}$.

(iii) If $E_n \in \mathcal{F}$, for all $n = 1, 2, 3, \ldots$ then $\bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$.

Moreover, if $E_n \in \mathcal{F}$, for all $n = 1, 2, 3, \ldots$ and $E_i \cap E_j = \emptyset$ for $i \neq j$, then:

$$\mu^\star(\bigcup_{n=1}^{\infty} E_n) = \sum_{n=1}^{\infty} \mu^\star(E_n) \tag{1.7}$$

Remark. This result is the most important theorem in this chapter and provides the basis for all that follows. It also allows us to give names to the quantities under discussion.

Conditions (i)-(iii) mean that $\mathcal{F}$ is a sigma-algebra. In other words, we say that a family of sets is a sigma-algebra, if it contains the base set and is closed under countable unions, and complements. A $[0, \infty)$-valued function defined on a sigma-algebra is called a measure if it satisfies countable additivity (1.7) for pairwise disjoint sets.

An alternative, rather more abstract and general approach to measure theory is to begin with the above properties as axioms, i.e. to call the the triple $(\Omega, \mathcal{F}, \mu)$ a measure space, if Ω is an abstractly given set, $\mathcal{F}$ is a sigma-algebra of the subsets of Ω and $\mu : \mathcal{F} \to [0, \infty]$ is a function satisfying countable additivity. The task of defining the Lebesgue measure on $\mathbf{R}$ then becomes that of verifying, with $\mathcal{F}$ and $\mu = \mu^\star$ on $\mathcal{F}$ defined above, that the triple $(\Omega, \mathcal{F}, \mu)$ satisfies these axioms.

Although the requirements of probability theory will mean that we have to consider such general measure spaces in due course, we have chosen our more concrete approach to the fundamental example of Lebesgue measure in order to demonstrate how this important measure space arises quite naturally from the considerations of the *lengths* of sets in $\mathbf{R}$ and leads to a theory of integration which greatly extends that of Riemann. It is also sufficient to allow us to develop most of the important examples of probability distributions.

Proof. (1) Let $A \subseteq \mathbf{R}$. Note that $A \cap \mathbf{R} = A$, $\mathbf{R}^C = \emptyset$, so that $A \cap \mathbf{R}^C = \emptyset$. Thus, the equation (1.6) now reads, $\mu^\star(A) = \mu^\star(A) + \mu^\star(\emptyset)$ which is obviously true, since $\emptyset$ is a null set and $\mu^\star(\emptyset) = 0$.

(2) Suppose $E \in \mathcal{F}$ and take any arbitrary $A \subseteq \mathbf{R}$. We have to show (1.6) for E^C, that is:

$$\mu^\star(A) = \mu^\star(A \cap E^C) + \mu^\star(A \cap (E^C)^C) \tag{1.8}$$

but since $(E^C)^C = E$, this reduces to the condition for E which holds by hypothesis.

(3) We split the proof (iii) into several steps. But first:

A warm up. Suppose that $E_1 \cap E_2 = \emptyset$, $E_1, E_2 \in \mathcal{F}$. We shall show that $E_1 \cup E_2 \in \mathcal{F}$ and $\mu^\star(E_1 \cup E_2) = \mu^\star(E_1) + \mu^\star(E_2)$.

Let $A \subseteq \mathbf{R}$. We have the condition for E_1:

$$\mu^\star(A) = \mu^\star(A \cap E_1) + \mu^\star(A \cap E_1^C) \tag{1.9}$$

Now, we apply (1.6) for E_2 with $A \cap E_1^C$ in place of A.

$$\begin{aligned}\mu^\star(A \cap E_1^C) &= \mu^\star(A \cap E_1^C \cap E_2) + \mu^\star(A \cap E_1^C \cap E_2^C) \\ &= \mu^\star(A \cap (E_1^C \cap E_2)) + \mu^\star(A \cap (E_1^C \cap E_2^C))\end{aligned}$$

The situation is depicted in the figure below.

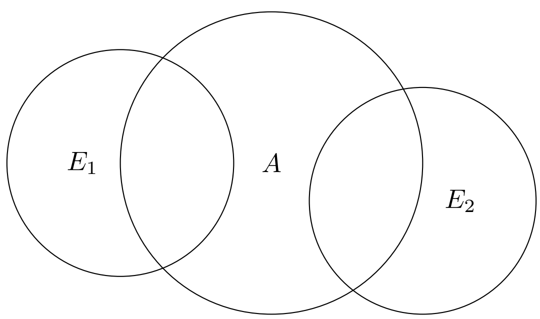


Figure. The sets A, E_1 and E_2.

Since E_1 and E_2 are disjoint, $E_1^C \cap E_2 = E_2$. By De-Morgan's laws, $E_1^C \cap E_2^C = (E_1 \cup E_2)^C$. We substitute and we have:

$$\mu^\star(A \cap E_1^C) = \mu^\star(A \cap E_2) + \mu^\star(A \cap (E_1 \cup E_2)^C)$$

Substituting this into (1.9), we get:

$$\mu^\star(A) = \mu^\star(A \cap E_1) + \mu^\star(A \cap E_2) + \mu^\star(A \cap (E_1 \cup E_2)^C) \tag{1.10}$$

Now, by the subadditivity property of $\mu^\star$, we have:

$$\begin{aligned}\mu^\star(A \cap E_1) + \mu^\star(A \cap E_2) &\geq \mu^\star((A \cap E_1) \cup (A \cap E_2)) \\ &= \mu^\star(A \cap (E_1 \cup E_2))\end{aligned}$$

So, (1.10) gives:

$$\mu^\star(A) \geq \mu^\star(A \cap (E_1 \cup E_2)) + \mu^\star(A \cap (E_1 \cup E_2)^C)$$

which is sufficient for $E_1 \cup E_2$ to belong to $\mathcal{F}$.

Finally, let $A = E_1 \cup E_2$. Then, the equation (1.10) yields:

$$\mu^\star(E_1 \cup E_2) = \mu^\star(E_1) + \mu^\star(E_2)$$

We return to the main proof of (iii).

Step 1. Our claim is: if pariwise disjoint E_k, $k = 1, 2, \ldots$ are in $\mathcal{F}$ then their countable union is in $\mathcal{F}$ and countable additivity (1.5) holds.

We begin as in the proof of the **Warm Up** and we have:

$$\begin{aligned} \mu^\star(A) &= \mu^\star(A \cap E_1) + \mu^\star(A \cap E_1^C) \\ \mu^\star(A) &= \mu^\star(A \cap E_1) + \mu^\star(A \cap E_2) + \mu^\star(A \cap (E_1 \cup E_2)^C) \end{aligned} \tag{1.11}$$

(See (1.10)).

E_3 is also measurable. Let $A = A \cap E_1^C \cap E_2^C$. Then:

$$\mu^\star(A \cap E_1^C \cap E_2^C) = \mu^\star(A \cap E_1^C \cap E_2^C \cap E_3) + \mu^\star(A \cap E_1^C \cap E_2^C \cap E_3^C)$$

But, $E_1^C \cap E_2^C \cap E_3 = E_3$ since they are pairwise disjoint. So,

$$\mu^\star(A \cap (E_1 \cup E_2)^C) = \mu^\star(A \cap E_3) + \mu^\star(A \cap (E_1 \cup E_2 \cup E_3)^C) \tag{1.12}$$

Substituting the value of (1.12) in equation (1.11), we get after $n = 3$ steps:

$$\mu^\star(A) = \sum_{k=1}^{3} \mu^\star(A \cap E_k) + \mu^\star\left(A \bigcap \left(\bigcup_{k=1}^{3} E_k\right)^C\right) \tag{1.13}$$

We proceed by mathematical induction. We induct on k. Our hypothesis is, that after n steps, we expect:

$$\mu^\star(A) = \sum_{k=1}^{n} \mu^\star(A \cap E_k) + \mu^\star\left(A \bigcap \left(\bigcup_{k=1}^{n} E_k\right)^C\right) \tag{1.14}$$

Let's assume that

$$\mu^\star(A) = \sum_{k=1}^{n-1} \mu^\star(A \cap E_k) + \mu^\star\left(A \bigcap \left(\bigcup_{k=1}^{n-1} E_k\right)^C\right) \tag{1.15}$$

is true.

Since $E_n \in \mathcal{F}$, we may apply the definition (1.6) with $A = A \bigcap \left(\bigcup_{k=1}^{n-1} E_k \right)^C$:

$$\mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n-1} E_k \right)^C \right) = \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n-1} E_k \right)^C \bigcap E_n \right) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n-1} E_k \right)^C \bigcap E_n^C \right) \tag{1.16}$$

Now we make the same observations as in the **Warm Up**:

$$\begin{aligned} \left(\bigcup_{k=1}^{n-1} E_k \right)^C \bigcap E_n &= E_n \quad \{E_i \text{ are pairwise disjoint}\} \\ \left(\bigcup_{k=1}^{n-1} E_k \right)^C \bigcap E_n^C &= \left(\bigcup_{k=1}^{n} E_n \right)^C \quad \{\text{De-Morgan's laws}\} \end{aligned}$$

Inserting these into equation (1.16), we get:

$$\mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n-1} E_k \right)^C \right) = \mu^\star (A \cap E_n) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n} E_n \right)^C \right)$$

and inserting this into (1.15), we get :

$$\mu^\star(A) = \sum_{k=1}^{n-1} \mu^\star(A \cap E_k) + \mu^\star(A \cap E_n) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n} E_n \right)^C \right)$$

This proves the induction hypothesis.

As will be seen at the next step, the fact that E_k are pairwise disjoint is not necessary in order to ensure that their union belongs to $\mathcal{F}$. However, with this assumption we have equality in (1.14)which does not hold otherwise. This equality will allow us to prove countable additivity (1.7).

Since:

$$\left(\bigcup_{k=1}^{n} E_k \right)^C \supseteq \left(\bigcup_{k=1}^{\infty} E_k \right)^C$$

from (1.14) by monotonicity of measure, we get:

$$\begin{aligned} \mu^\star(A) &= \sum_{k=1}^{n} \mu^\star(A \cap E_k) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{n} E_k \right)^C \right) \\ &\geq \sum_{k=1}^{n} \mu^\star(A \cap E_k) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \end{aligned}$$

By the Order Limit Theorem, the inequality remains true, if we pass to the limit, as $n \to \infty$:

$$\mu^\star(A) \geq \sum_{k=1}^{\infty} \mu^\star(A \cap E_k) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \tag{1.17}$$

By countable sub-additivity of $\mu^\star$ (Theorem (1.4)) :

$$\sum_{k=1}^{\infty} \mu^\star(A \cap E_k) \geq \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right) \right)$$

and so:

$$\mu^\star(A) \geq \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right) \right) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \tag{1.18}$$

So, we have shown that $\bigcup_{k=1}^{\infty} E_k \in \mathcal{F}$ and hence the two sides of (1.18) are equal.

The right hand side of (1.17) is squeezed between the left and right of (1.18). That is:

$$\begin{aligned}
\mu^\star(A) &= \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right) \right) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \\
&\leq \sum_{k=1}^{\infty} \mu^\star(A \cap E_k) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \\
&\leq \mu^\star(A)
\end{aligned}$$

Consequently,

$$\mu^\star(A) = \sum_{k=1}^{\infty} \mu^\star(A \cap E_k) + \mu^\star \left(A \bigcap \left(\bigcup_{k=1}^{\infty} E_k \right)^C \right) \tag{1.19}$$

The equality here is a consequence of the assumption that E_k are pairwise disjoint. It holds for any set A so we may insert $A = \bigcup_{j=1}^{\infty} E_j$. The last term on the right is zero, because the length of the empty set, $\mu^\star(\emptyset) = 0$. And, since the E_i are disjoint, $\left(\bigcup_{j=1}^{\infty} E_j\right) \bigcap E_k = E_k$. As a result, we have:

$$\mu^\star \left(\bigcup_{j=1}^{\infty} E_j \right) = \sum_{j=1}^{\infty} \mu^\star(E_j)$$

Step 2. Our claim is, if $E_1, E_2 \in \mathcal{F}$, then $E_1 \cup E_2 \in \mathcal{F}$ (not necessarily disjoint).

Again we begin as in the **Warm Up**:

$$\mu^\star(A) = \mu^\star(A \cap E_1) + \mu^\star(A \cap E_1^C) \tag{1.20}$$

Next applying the definition (1.6) to E_2 and with $A \cap E_1^C$ in place of A we get:

$$\mu^\star(A \cap E_1^C) = \mu^\star(A \cap E_1^C \cap E_2) + \mu^\star(A \cap E_1^C \cap E_2^C)$$

We insert this into (1.20) to get:

$$\mu^\star(A) = \mu^\star(A \cap E_1) + \mu^\star(A \cap E_1^C \cap E_2) + \mu^\star(A \cap E_1^C \cap E_2^C) \tag{1.21}$$

By DeMorgan's law, $E_1^C \cap E_2^C = (E_1 \cup E_2)^C$ so as before:

$$\mu^\star(A \cap E_1^C \cap E_2^C) = \mu^\star(A \cap (E_1 \cup E_2)^C)$$

Now, consider the set $(A \cap E_1) \cup (A \cap E_1^C \cap E_2)$. This, can be written as: $A \cap (E_1 \cup (E_1^C \cap E_2)) = A \cap (E_1 \cup E_1^C) \cap (E_1 \cup E_2) = A \cap (E_1 \cup E_2)$.

By Countable Subadditivity of $\mu^\star$, we have:

$$\mu^\star(A \cap E_1) + \mu^\star(A \cap E_1^C \cap E_2) \geq \mu^\star(A \cap (E_1 \cup E_2))$$

Inserting these facts into (1.21), we get:

$$\mu^\star(A) \geq \mu^\star(A \cap (E_1 \cup E_2)) + \mu^\star(A \cap (E_1 \cup E_2)^C)$$

as required.

Step 3. Our claim is, if $E_k \in \mathcal{F}$, $k = 1, 2, \dots, n$, then the finite union $E_1 \cup E_2 \cup \ldots \cup E_n \in \mathcal{F}$. (not necessarily disjoint)

We argue by induction. Suppose that the claim is true for $n - 1$. Then,

$$E_1 \cup E_2 \cup \ldots \cup E_n = (E_1 \cup \ldots \cup E_{n-1}) \cup E_n$$

so that the result follows from **Step 2**.

Step 4. If $E_1, E_2 \in \mathcal{F}$, then $E_1 \cap E_2 \in \mathcal{F}$.

We have $E_1^C, E_2^C \in \mathcal{F}$ by (ii), $E_1^C \cup E_2^C \in \mathcal{F}$ by step 2, and $(E_1^C \cup E_2^C)^C \in \mathcal{F}$ by (ii) again. But, by De-Morgan's laws, this is $(E_1^C \cup E_2^C)^C = E_1 \cap E_2$.

Step 5. The general case: if $E_1, E_2, \ldots$ are in $\mathcal{F}$, then so is the countably infinite union $\bigcup_{k=1}^\infty E_k$.

Let $E_k \in \mathcal{F}$, $k = 1, 2, \ldots$. We define the auxiliary sequence of pairwise disjoint sets F_k with the same union as E_k:

$$\begin{aligned}
&F_1 = E_1 \\
&F_2 = E_2 \setminus E_1 = E_2 \cap E_1^C \\
&F_3 = E_3 \setminus (E_1 \cup E_2) = E_3 \cap (E_1 \cup E_2)^C \\
&\quad\vdots \\
&F_k = E_k \setminus (E_1 \cup E_2 \cup \ldots \cup E_{k-1}) = E_k \cap (E_1 \cup \ldots \cup E_{k-1})^C
\end{aligned}$$

By steps 3 and 4, we know that all F_k are in $\mathcal{F}$. By the very construction, they are pairwise disjoint, so by step 1, their union is in $\mathcal{F}$. We shall show that:

$$\bigcup_{k=1}^{\infty} F_k = \bigcup_{k=1}^{\infty} E_k$$

The inclusion:

$$\bigcup_{k=1}^{\infty} F_k \subseteq \bigcup_{k=1}^{\infty} E_k$$

is obvious since for each k, $F_k \subseteq E_k$ by definition. For the inverse, let $a \in \bigcup_{k=1}^{\infty} E_k$. Put $S = \{n \in \mathbf{N} : a \in E_n\}$ which is non-empty since a belongs to the union. Let $n_0 = \min S \in S$. If $n_0 = 1$, then $a \in E_1 = F_1$. Suppose $n_0 > 1$. So, $a \in E_{n_0}$ and by definition of n_0, $a \notin E_1, \ldots, a \notin E_{n_0-1}$. By the definition of F_{n_0}, this means that $a \in F_{n_0}$ so a is in $\bigcup_{k=1}^{\infty} F_k$. This closes the proof. □

Using De-Morgan's laws, we can easily verify an additional property of $\mathcal{F}$.

Proposition 1.3. *If* $E_k \in \mathcal{F}$, $k = 1, 2, \ldots$, *then*

$$E = \bigcap_{k=1}^{\infty} E_k \in \mathcal{F}$$

Proof. $\mathcal{F}$ is closed under complementation. Thus, $E_k \in \mathcal{F} \implies E_k^C \in \mathcal{F}$. Since, $\mathcal{F}$ is closed under countable unions, $\bigcup_{k=1}^{\infty} E_k^C \in \mathcal{F}$. And it follows that, $\left(\bigcup_{k=1}^{\infty} E_k^C\right)^C \in \mathcal{F}$. By De-Morgan's laws, $\left(\bigcup_{k=1}^{\infty} E_k^C\right)^C = \bigcap_{k=1}^{\infty} E_k$. This closes the proof. □

We can therefore summarize the properties of the family $\mathcal{F}$ of Lebesgue measurable sets as follows:

$\mathcal{F}$ is closed under countable unions, countable intersections and complements. It contains intervals and null sets.

Definition 1.4. (*Lebesgue Measure*). We shall write $\mu(E)$ instead of $\mu^\star(E)$ for any E in $\mathcal{F}$ and call $\mu(E)$ the Lebesgue measure of the set E.

The Lebesgue measure $\mu : \mathcal{F} \to [0, \infty]$ is a countably additive set function defined on the sigma-algebra $\mathcal{F}$ of measurable sets. The Lebesgue measure of an interval is equal to its length. The Lebesgue measure of a null-set is zero.

1.4 Basic Properties of Lebesgue Measure.

Since Lebesgue measure is nothing else than the outer measure restricted to a special class of sets $\mathcal{F}$, some properties of the outer measure are automatically inherited by the Lebesgue measure.

Proposition 1.4. *Suppose that* $A, B \in \mathcal{F}$.

(1) If $A \subset B$, *then* $\mu(A) \le \mu(B)$.

(2) If $A \subset B$ *and* $\mu(A)$ *is finite, then* $\mu(B \setminus A) = \mu(B) - \mu(A)$.

(3) μ *is translation invariant.*

Since the empty set $\emptyset \in \mathcal{F}$, *we can take* $E_i = \emptyset$ *for all* $i > n$ *in (1.7) to conclude that Lebesgue measure is finitely additive: if* $E_i \in \mathcal{F}$ *are pairwise disjoint, then:*

$$\mu\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mu(E_i)$$

Remark. Property (2) is derived as follows. Since $B = (B \setminus A) \cup A$ and $B \setminus A$ and A are disjoint, $\mu(B) = \mu(B \setminus A) + \mu(A)$. Consequently, $\mu(B \setminus A) = \mu(B) - \mu(A)$.

Problem 1.5. Find a formula describing $\mu(A \cup B)$ and $\mu(A \cup B \cup C)$ in terms of measures of the individual sets and their intersections (we do not assume that the sets are pairwise disjoint).

Proof. We have:

$$A \cup B = \left(A \cap (A \cap B)^C\right) \cup \left(B \cap (A \cap B)^C\right) \cup (A \cap B)$$

The three sets $A \setminus (A \cap B)$, $B \setminus (A \cap B)$ and $A \cap B$ are pairwise disjoint. Consequently, by finite additivity of the Lebesgue measure:

$$\begin{aligned}
\mu(A \cup B) &= \mu(A \setminus (A \cap B)) + \mu(B \setminus (A \cap B)) + \mu(A \cap B) \\
&= \mu(A) - \mu(A \cap B) + \mu(B) - \mu(A \cap B) + \mu(A \cap B) \\
&\quad \{\because (A \cap B) \subseteq A, \mu(A \setminus (A \cap B)) = \mu(A) - \mu(A \cap B)\} \\
&= \mu(A) + \mu(B) - \mu(A \cap B)
\end{aligned}$$

Let $B = B \cup C$

$$\begin{aligned}
\mu(A \cup (B \cup C)) &= \mu(A) + \mu(B \cup C) - \mu(A \cap (B \cup C)) \\
&= \mu(A) + \mu(B) + \mu(C) - \mu(B \cap C) - \mu((A \cap B) \cup (A \cap C)) \\
&= \mu(A) + \mu(B) + \mu(C) - \mu(B \cap C) - \mu(A \cap B) - \mu(A \cap C) \\
&+ \mu(A \cap B \cap C)
\end{aligned}$$

□

Recalling that the *symmetric difference* $A \Delta B$ of two sets is defined by $A \Delta B = (A \setminus B) \cup (B \setminus A)$ the following result is also easy to check:

Proposition 1.5. *If $A \in \mathcal{F}$, and $\mu(A \Delta B) = 0$, then $B \in \mathcal{F}$ and $\mu(A) = \mu(B)$.*

Proof. Null sets belong to $\mathcal{F}$. Since $A \Delta B$ is a null set, it belongs to $\mathcal{F}$. Now, $A \cap B^C$ and$A^C \cap B$ are subsets of $A \Delta B$, they are also null sets and belong to $\mathcal{F}$. We have:

$$\begin{aligned}
\mu(A) &= \mu(A \cap (B \cup B^C)) \\
&= \mu(A \cap B) + \mu(A \cap B^C) \\
&= \mu(A \cap B)
\end{aligned}$$

And likewise, $\mu(B) = \mu(A \cap B)$. Hence, $\mu(A) = \mu(B)$. □

Lemma 1.1. *Let $(A_n)_{n=1}^{\infty}$, $(B_n)_{n=1}^{\infty}$ be a sequence of sets. The difference of the union of sets is contained in the union of the difference of sets. We have:*

$$\left(\bigcup_{n \geq 1} A_n\right) - \left(\bigcup_{n \geq 1} B_n\right) \subset \bigcup_{n \geq 1} (A_n - B_n)$$

Proof. We have:

$$\begin{aligned}
A - \left(\bigcup_{n\geq 1} B_n\right) &= A \bigcap \left(\bigcup_{n\geq 1} B_n\right)^C \\
&= A \bigcap \left(\bigcap_{n\geq 1} B_n^C\right) \\
&= \bigcap_{n\geq 1} \left(A \cap B_n^C\right) \\
&= \bigcap_{n\geq 1} (A - B_n)
\end{aligned}$$

Thus,

$$\begin{aligned}
\left(\bigcup_{n\geq 1} A_n\right) - \left(\bigcup_{n\geq 1} B_n\right) &= \left(\bigcup_{n\geq 1} A_n\right) \bigcap \left(\bigcup_{n\geq 1} B_n\right)^C \\
&= \left(\bigcup_{m\geq 1} A_m\right) \bigcap \left(\bigcap_{n\geq 1} B_n^C\right) \\
&= \bigcup_{m\geq 1} \left(A_m \cap \left(\bigcap_{n\geq 1} B_n^C\right)\right) \\
&= \bigcup_{m\geq 1} \bigcap_{n\geq 1} (A_m \cap B_n^C) \\
&= \bigcup_{m\geq 1} \left\{\bigcap_{n\geq 1} (A_m - B_n)\right\} \\
&\subset \bigcup_{m\geq 1} \{A_m - B_m\}
\end{aligned}$$

□

Every open set in **R** can be expressed as the union of a countable number of open intervals. This ensures that open sets in **R** are Lebesgue measurable, since $\mathcal{F}$ contains intervals and is closed under countable unions. We can approximate the measure of any $A \in \mathcal{F}$ from the above by the measures of a sequence of open sets containing A. This is clear from the below result:

Theorem 1.7. *(i)For any $\epsilon > 0$, $A \in$ **R**, we can find an open set O such that :*

$$A \subset O, \quad \mu(O) \leq \mu^\star(A) + \epsilon$$

Consequently, for any $E \in \mathcal{F}$ we can find an open set O containing E such that $\mu(O \setminus E) < \epsilon$.

*(ii) For any $A \subset$ **R**, we can find a sequence of open sets O_n,such that:*

$$A \subset \bigcap_{n=1}^{\infty} O_n, \quad \mu\left(\bigcap_{n=1}^{\infty} O_n\right) = \mu^\star(A)$$

Proof. (i) By definition of $\mu^\star(A)$ we can find a sequence (I_n) of intervals with $A \subset \bigcup_{n=1}^\infty I_n$ and $\sum_{n=1}^\infty l(I_n) \leq \mu^\star(A) + \epsilon/2$. That is,

$$\exists (I_n)_{n=1}^\infty, \quad A \subset \bigcup_n I_n, \quad \sum_{n=1}^\infty l(I_n) - \frac{\epsilon}{2} \leq \mu^\star(A)$$

Each I_n is contained in an open interval whose length is very close to that of I_n; if the left and right end-points of I_n are a_n and b_n respectively, let $J_n = \left(a_n - \frac{\epsilon}{2^{n+2}}, b_n + \frac{\epsilon}{2^{n+2}}\right)$. Set $O = \bigcup_{n=1}^\infty J_n$, which is open. Remember, that J_n's are overlapping. Then, $A \subset O$ and

$$\mu(O) \leq \sum_{n=1}^\infty l(J_n) = \sum_{n=1}^\infty l(I_n) + \frac{\epsilon}{2} \leq \mu^\star(A) + \epsilon$$

When $\mu(E) < \infty$ the final statement follows at once from (ii) in proposition (1.4), since $\mu(O \setminus E) = \mu(O) - \mu(E) \leq \epsilon$.

When $\mu(E) = \infty$ we first write **R** as the countable union of the finite intervals: $\mathbf{R} = \bigcup_n (-n, n)$. Now, $E_n = E \cap (-n, n)$ has finite measure, so we can find an open set $O_n \supset E_n$ with $\mu(O_n \setminus E_n) \leq \frac{\epsilon}{2^n}$. The set $O = \bigcup_n O_n$ is open and contains E. Now,

$$\begin{aligned} O \setminus E &= \left(\bigcup_n O_n\right) \setminus \left(\bigcup_n E_n\right) \\ &\subset \bigcup_n (O_n \setminus E_n) \end{aligned}$$

so that $\mu(O \setminus E) \leq \sum_n \mu(O_n \setminus E_n) \leq \epsilon$.

(ii) In (i) use $\epsilon = \frac{1}{n}$ and let O_n be the open set so obtained. With $E = \bigcap_n O_n$ we obtain a measurable set containing A such that $\mu(E) < \mu(O_n) \leq \mu^\star(A) + \frac{1}{n}$ for each n, hence the result follows. $\square$

Remark. Theorem (1.7) shows how the freedom of movement allowed by the closure properties of the sigma-field $\mathcal{F}$ can be exploited by producing, for any set $A \subset \mathbf{R}$, a measurable set $O \supset A$ which is obtained from open intervals using two operations and whose measure(length) equals the outer measure of A.

Theorem 1.8. *(Continuity Property of the Lebesgue measure) The Lebesgue measure μ preserves limits.*

(1) Suppose that $(A_n)_{n=1}^\infty$ is a sequence of measurable sets in $\mathcal{F}$. Then, we have:

$$\lim_{m\to\infty} \mu\left(\bigcup_{i=1}^m A_i\right) = \mu\left(\lim_{m\to\infty} \bigcup_{i=1}^m A_i\right) = \mu\left(\bigcup_{i=1}^\infty A_i\right) \tag{1.22}$$

(2) If $A_n \subset A_{n+1}$ is a monotonically increasing sequence of sets in $\mathcal{F}$, then we have:

$$\lim_{m\to\infty} \mu(A_m) = \mu\left(\bigcup_{m=1}^\infty A_m\right) \tag{1.23}$$

(3) If $A_n \supset A_{n+1}$ is a monotonically decreasing sequence of sets in $\mathcal{F}$, then we have:

$$\lim_{m\to\infty} \mu(A_m) = \mu\left(\bigcap_{m=1}^\infty A_m\right) \tag{1.24}$$

Proof. (1) Define a new family of sets $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, . . ., $B_n = A_n \setminus \bigcup_{n=1}^{\infty} A_i$ and so forth. Then, we make the following claims:

Claim I.$B_i \cap B_j = \emptyset$, for all $i \neq j$.

We proceed by contradiction. Let $m < n$. Assume that there exists an element $x \in B_m \cap B_n$. It follows that:

$$\begin{aligned} x \in (B_m \cap B_n) &\iff (x \in B_m) \wedge (x \in B_n) \\ &\iff \left(x \in \left(A_m \setminus \bigcup_{i=1}^{m-1} A_i\right)\right) \wedge \left(x \in \left(A_n \setminus \bigcup_{j=1}^{n-1} A_j\right)\right) \end{aligned}$$

In words, x belongs to both A_m and the set $\left(\bigcup_{j=1}^{n-1} A_j\right)^C$. Since, $m, n \in \mathbf{Z}_+$, and $m < n$, we must have $m \leq n - 1$. If $x \in A_m$, then it must belong to $\bigcup_{j=1}^{n-1} A_j$. This is a contradiction. Hence, our initial assumption is false. $B_m \cap B_n$ is disjoint.

Claim II. $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

We proceed by mathematical induction. The claim is vacuously true for $n = 1$, since $B_1 = A_1$ by construction. For $n = 2$, we have:

$$\begin{aligned} A_1 \cup A_2 &= (A_2 \cup A_1) \cap (A_1 \cup A_1^C) \\ &= ((A_2 \cup A_1) \cap A_1) \cup ((A_2 \cup A_1) \cap A_1^C) \\ &= A_1 \cup ((A_2 \cap A_1^C) \cup \emptyset) \\ &= A_1 \cup (A_2 \setminus A_1) \\ &= B_1 \cup B_2 \end{aligned}$$

Assume that the claim is true for $n - 1$. Define $S = \left(\bigcup_{i=1}^{n-1} A_i\right)$We have:

$$\begin{aligned} \bigcup_{i=1}^{n} A_i &= (A_n \cup S) \cap \left(S \cup S^C\right) \\ &= S \cup (A_n \setminus S) \\ &= \left(\bigcup_{i=1}^{n-1} A_i\right) \bigcup \left(A_n \setminus \left(\bigcup_{i=1}^{n-1} A_i\right)\right) \\ &= \left(\bigcup_{i=1}^{n-1} B_i\right) \bigcup B_n \\ &\quad \{\text{since the claim holds for } n-1\} \\ &= \bigcup_{i=1}^{n} B_i \end{aligned}$$

Hence, the proposition holds true for all n. Passing to the limit as $n \to \infty$, we have the desired result. This closes the proof.

Since $\{B_i, i \geq 1\}$ is a disjoint sequence of events, and using the above claims, we get:

$$
\begin{aligned}
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} B_i\right) \\
&= \sum_{i=1}^{\infty} \mu(B_i) \\
&\quad \{\text{Countable additivity}\}
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) &= \sum_{i=1}^{\infty} \mu(B_i) \\
&= \lim_{m\to\infty} \sum_{i=1}^{m} \mu(B_i) \\
&\quad \{\text{An infinite series converges to the limit of the sequence of partial sums.}\} \\
&= \lim_{m\to\infty} \mu\left(\bigcup_{i=1}^{m} B_i\right) \\
&\quad \{\text{Finite additivity}\} \\
&= \lim_{m\to\infty} \mu\left(\bigcup_{i=1}^{m} A_i\right) \\
&\quad \{\text{By construction}\}
\end{aligned}
$$

(2) If $A_n \subset A_{n+1}$, then $\bigcup_{i=1}^{m} A_i = A_m$. Consequently,

$$
\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{m\to\infty} \mu\left(\bigcup_{i=1}^{m} A_i\right) = \lim_{m\to\infty} \mu(A_m)
$$

(3) If $A_n \supset A_{n+1}$, then $A_1 \setminus A_n \subset A_1 \setminus A_{n+1}$. Thus, $\{A_1 \setminus A_n, n \geq 1\}$ is an increasing sequence of sets. From (2), it follows that:

$$\begin{aligned}
\lim_{m\to\infty} \mu(A_1 \setminus A_m) &= \mu\left(\lim_{m\to\infty} \bigcup_{i=1}^{m} A_1 \setminus A_i\right) \\
&= \mu\left(\lim_{m\to\infty} \bigcup_{i=1}^{m} (A_1 \cap A_i^C)\right) \\
&= \mu\left(\lim_{m\to\infty} A_1 \bigcap \left(\bigcup_{i=1}^{m} A_i^C\right)\right) \\
&= \mu\left(\lim_{m\to\infty} A_1 \bigcap \left(\bigcap_{i=1}^{m} A_i\right)^C\right) \\
\lim_{m\to\infty} \mu(A_1) - \lim_{m\to\infty} \mu(A_m) &= \mu(A_1) - \mu\left(\lim_{m\to\infty} \bigcap_{i=1}^{m} A_i\right) \\
\implies \lim_{m\to\infty} \mu(A_m) &= \mu\left(\lim_{m\to\infty} \bigcap_{i=1}^{m} A_i\right)
\end{aligned}$$

□

Remark. The proof of theorem (1.8) simply relies on countable additivity of μ and on the definition of the sum of an infinite series in $[0, \infty]$, i.e. that:

$$\sum_{i=1}^{\infty} \mu(A_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mu(A_i)$$

Consequently, this result is true not only for the set function μ, but any countably additive set function defined on a sigma-field. It also leads us to the following claim, which, though, we consider it here only for μ, actually characterizes countably additive set functions.

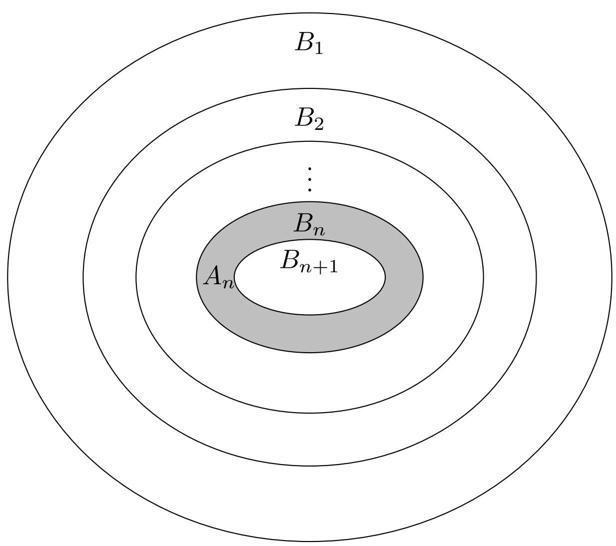


Figure. The sets B_n and A_n(light-gray).

Theorem 1.9. *The set function μ satisfies:*

(1) μ is finitely additive, that is, for pairwise disjoint sets (A_i) we have:

$$\mu\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mu(A_i)$$

for each n;

(ii) μ is continuous at $\emptyset$, that is, if (B_n) decrease to $\emptyset$, $\mu(B_n)$ decreases to 0.

Proof. To prove this claim, recall that $\mu : \mathcal{F} \to [0, \infty]$ is countably additive. This implies (i), as we have already seen. To prove (ii), consider a sequence (B_n) in $\mathcal{F}$ which decreases to $\emptyset$. Then, $A_n = B_n \setminus B_{n+1}$ defines a disjoint sequence in $\mathcal{F}$ and $\bigcup_{n=1}^{\infty} A_n = B_1$. We may assume that B_1 is bounded, so that $\mu(B_n)$ is finite for all n, so that, $\mu(A_n) = \mu(B_n \setminus B_{n+1}) = \mu(B_n) - \mu(B_{n+1}) \geq 0$ and hence we have:

$$\begin{aligned}
\mu(B_1) &= \sum_{n=1}^{\infty} \mu(A_n) \\
&= \sum_{n=1}^{\infty} \mu(B_n) - \mu(B_{n+1}) \\
&= \lim_{n\to\infty} (\mu(B_1) - \mu(B_n))
\end{aligned}$$

which shows that $\lim_{n\to\infty} \mu(B_n) \to 0$. □

1.5 Borel Sets.

The definition of $\mathcal{F}$ does not lend itself easily to the verification that a particular set belongs to $\mathcal{F}$; in our proofs we have had to work quite hard to show that $\mathcal{F}$ is closed under various operations. It is therefore useful to add another construction to our armoury; one which shows more directly, how open sets(and indeed open intervals) and the structure of sigma-fields lie at the heart of many of the concepts we have developed. We begin with an auxiliary construction enabling us to produce new sigma-fields.

Theorem 1.10. *The intersection of a family of σ-fields is a σ-field.*

Proof. Let $\mathcal{F}_\alpha$ be σ-fields for $\alpha \in \Lambda$ (the index set Λ can be arbitrary). Put

$$\mathcal{F} = \bigcap_{\alpha\in\Lambda} \mathcal{F}_\alpha$$

We verify the conditions of the definition.

1. $\mathbf{R} \in \mathcal{F}_\alpha$ for all $\alpha \in \Lambda$ so $\mathbf{R} \in \mathcal{F}$.

2. If $E \in \mathcal{F}$, then $E \in \mathcal{F}_\alpha$ for all $\alpha \in \Lambda$. Since $\mathcal{F}_\alpha$ is a σ-field, it is closed under complementation, so E^C belongs to $\mathcal{F}_\alpha$ for all $\alpha \in \Lambda$. Hence, $E^C \in \mathcal{F}$.

3. If E_k belongs to $\mathcal{F}$ for $k = 1, 2, 3, \ldots$, then $E_k \in \mathcal{F}_\alpha$ for all α, k hence, $\bigcup_{k=1}^{\infty} E_k \in \mathcal{F}_\alpha$ for all α and so $\bigcup_{k=1}^{\infty} E_k \in \mathcal{F}$. □

Definition 1.5. Put

$$\mathcal{B} = \bigcap \{\mathcal{F} : \mathcal{F} \text{ is a sigma-field containing all intervals}\} \tag{1.25}$$

We say that $\mathcal{B}$ is the σ-field generated by all the intervals and we call the elements of $\mathcal{B}$ - *Borel sets* (after Emile Borel 1871-1956). It is obviously the smallest σ-field containing all the intervals. In general, we say that $\mathcal{G}$ is the σ-field generated by a family of sets $\mathcal{A}$ if $\mathcal{G} = \bigcap\{\mathcal{F} : \mathcal{F}$ is a sigma-field such that $\mathcal{F} \supset A\}$.

Example 1.1. (Borel Sets) The following examples illustrate how the closure properties of the σ-field $\mathcal{B}$ may be used to verify that most familiar sets in $\mathbf{R}$ belong to $\mathcal{B}$.

(1) By construction, all intervals belong to $\mathcal{B}$ and since $\mathcal{B}$ is a σ-field, all open sets must belong to $\mathcal{B}$, as any open set is the countable union of open intervals.

(2) Countable sets are Borel sets, since each set is a countable union of closed intervals of the form $[a, a]$; in particular $\mathbf{N}$ and $\mathbf{Q}$ are Borel sets. Since, $\mathcal{B}$ is a σ-field, it is closed under complementation. So, $\mathbf{R} \setminus \mathbf{Q}$ - the set of irrational numbers belongs to $\mathcal{B}$ and it is a borel set. Similarly, finite sets are also Borel sets.

The definition of $\mathcal{B}$ is also very flexible - as long as we start will all intervals of a particular type, these collections generate the same Borel σ-field:

Theorem 1.11. *If instead of all intervals, we take all open intervals, all closed intervals, all intervals of the form* (a, ∞) *(or of the form* $[a, \infty)$*,* $(-\infty, b)$ *or* $(-\infty, b]$*), all open sets, or all closed sets, then the* σ*-field generated by them is the same as* $\mathcal{B}$*.*

Proof. Let I be the set of all intervals and O be the set of all open intervals. Consider for example the σ-field generated by the family of open intervals O and denote it by $\mathcal{C}$:

$$\mathcal{C} = \bigcap \{\mathcal{F} \supset O, \mathcal{F}\text{is a sigma-field}\}$$

We have to show that $\mathcal{B} = \mathcal{C}$. Since open intervals are intervals, $O \subset I$ (the family of all intervals), then :

$$\{\mathcal{F} \supset I\} \subset \{\mathcal{F} \supset O\}$$

that is the collection of all σ-fields $\mathcal{F}$ which contain I is smaller than the collection of all σ-fields which contain the smaller family O, since it is a more demanding requirement to contain a bigger family, so there are fewer such objects. The inclusion is reversed after we take the intersection on both sides, thus $\mathcal{C} \subset \mathcal{B}$ (the intersection of a smaller family is bigger, as the requirement of belong to each of its members is a less stringent one).

We shall show that $\mathcal{C}$ contains all the intervals. This will be sufficient, since $\mathcal{B}$ is the intersection of such σ-fields, so it is contained in each, and therefore $\mathcal{B} \subset \mathcal{C}$.

To this end, consider the intervals $[a, b)$, $[a, b]$, (a, b) (the intervals of the form (a, b) are in $\mathcal{C}$ by definition):

$$[a, b) = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b\right)$$

$$[a, b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b + \frac{1}{n}\right)$$

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)$$

$\mathcal{C}$ as a σ-field is closed with respect to countable intersection, so it contains the sets on the right. The argument for unbounded intervals is similar:

$$(a,\infty) = \bigcup_{n=1}^{\infty} (a,n)$$

and

$$(-\infty,b) = \bigcup_{n=1}^{\infty} (-n,b)$$

The proof is complete. □

Remark. Since $\mathcal{F}$ is a σ-field containing all the intervals, and $\mathcal{B}$ is the smallest such σ-field, we have the inclusion $\mathcal{B} \subset \mathcal{F}$, that is every Borel set in **R** is Lebesgue measurable. The question therefore arises whether these σ-fields might be the same. In fact, the inclusion is proper. It is not altogether straightforward to construct a set in $\mathcal{F} \setminus \mathcal{B}$. However, by theorem 1.7 (ii), given any $E \in \mathcal{F}$, we can find a Borel set $B \supset E$ of the form $B = \bigcap_n O_n$, where the O_n are open sets, such that $\mu(E) = \mu(B)$. In particular,

$$\mu(B \Delta E) = \mu(B \setminus E) = 0$$

Hence, μ cannot distinguish between the measurable set E and the Borel set B we have constructed.

Thus, given a Lebesgue measurable set E we can find a Borel set B such that their symmetric difference $E \Delta B$ is a null set. Now, we know that $E \Delta B \in \mathcal{F}$, and it is obvious that subsets of null sets are also null, and hence in $\mathcal{F}$. However, we cannot conclude that every null set will be a Borel set (if $\mathcal{B}$ did contain all the null sets then by theorem 1.7 (ii), we should have

2 Expectation.

The goal of this section is to define the expectation of random variables and establish it's basic properties.

2.1 Lebesgue-measurable functions.

Integration is concerned with the process of approximation. In the Riemann integral, we split the interval $I = [a,b]$ over which we integrate into a partition $\{x_0 = a < x_1 < x_2 < \ldots < x_n = b\}$. Define $I_n := [x_{n-1}, x_n]$. Then, we construct approximating sums by multiplying the lengths of small subintervals by certain numbers a_n (related to the values of the function in question; for example $a_n = \sup_{I_n} f(x)$, $a_n = \inf_{I_n} f(x)$):

$$\sum_{n=1}^{\infty} a_n l(I_n) \tag{2.1}$$

For large n, this sum is close to the Riemann integral $\int_a^b f(x)dx$.

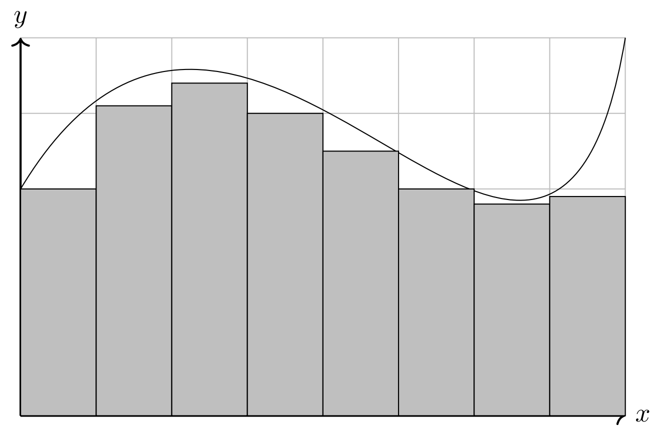


Figure. Riemann sums.

The approach to the Lebesgue integral is similar but there is a crucial difference. Instead of splitting the integration domain into various parts, we decompose the range of the function. Again, a simple way is to introduce short intervals J_n of equal length.

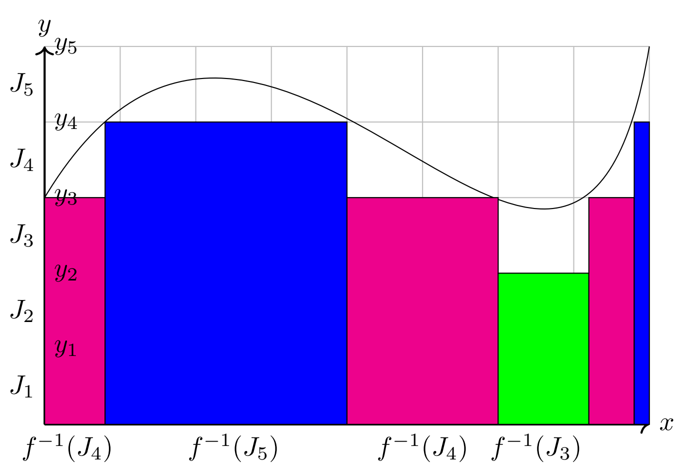


Figure. Lebesgue sums.

To build the approximating sums, we first take the inverse images of J_n by f, that is by $f^{-1}(J_n)$. These may be complicated sets, not necessarily intervals. Here the theory of measure developed previously comes into its own. We are able to measure sets provided they are measurable i.e. they are in $\mathcal{F}$. Given that, we compute:

$$\sum_{n=1}^{N} a_n \mu(f^{-1}(J_n)) \tag{2.2}$$

where $a_n \in J_n$ or $a_n = \inf J_n = y_{n-1}$ for example. The following definition guarantees that the above procedure makes sense.

Definition 2.1. Suppose that E is a measurable set. We say that a function $f : E \to \mathbf{R}$ is *(Lebesgue)*-measurable if for any interval $I \subseteq \mathbf{R}$

$$f^{-1}(I) = \{x \in \mathbf{R} : f(x) \in I\} \in \mathcal{F}$$

In what follows, the term *measurable* (without qualification) will refer to Lebesgue-measurable functions.

If all the sets $f^{-1}(I) \in \mathcal{B}$, that is, if they are Borel sets, we call f *Borel-measurable*, or simply a Borel function.

The underlying philosophy is one which is common for various mathematical notions : the inverse image of a *nice set* is *nice*. Remember continous functions, for example, where the inverse image of an open set is open. The actual meaning of the word nice depends on the particular branch of mathematics.

Remark. The terminology is unfortunate. *Measurable objects* should be measured (as with measurable sets). However, measurable functions will be integrated. The confusion here stems from the fact that the word *integrable* which would probably best fit here, carries a more restricted meaning as we shall see later.

2.2 Simple Random Variables.

In the special case of probability spaces we use the phrase *random variable* to mean a measurable function. That is, if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then $X : \Omega \to \mathbf{R}$ is a random variable if for all $x \in \mathbf{R}$, the set $X^{-1}((-\infty, x])$ is in $\mathcal{F}$:

$$\{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

A function $f : \Omega \to \mathbf{R}$ is called *simple* if its image $f(\Omega)$ is a finite-set. That is, f can be written as a finite linear-combination of indicator functions. We can write:

$$f(\omega) = \sum_{i=1}^{n} a_i I_{\omega \in A_i}(\omega)$$

for all $\omega \in \Omega$, for some distinct $a_1, \ldots, a_n \geq 0$ (values) and sets $A_1, \ldots, A_n$ which form a partition of Ω.

A random variable $X : \Omega \to \mathbf{R}$ is called simple, if its image $X(\Omega)$ takes a finite set of values. That is, X can be written as a finite linear-combination of indicator random variables. We can write:

$$X(\omega) = \sum_{i=1}^{n} a_i I_{A_i}(\omega)$$

for all $\omega \in \Omega$, for some distinct $a_1, \ldots, a_n \geq 0$ and events $A_1, \ldots, A_n$which form a partition of Ω. Note that: $X \geq 0$.

The abstract(Lebesgue) integral of a simple function f (with respect to the measure μ), denoted $\int f d\mu$ is defined as:

$$\int f d\mu = \sum_{k=1}^{n} a_k \mu(A_k)$$

The **expectation** of the simple random variable X, denoted by EX is defined as :

$$\int X d\mathbb{P} = \mathbb{E}X = \sum_{k=1}^{n} x_k \mathbb{P}(A_k)$$

This equates to discretising the y-axis.

The expectation of a non-negative random variable X is defined as :

$$\mathbb{E}X = \sup\{\mathbb{E}Z : Z \text{ is simple and } Z \leq X\}$$

Note that, we can always take $Z = 0$, so that, $\mathbb{E}Z = 0$ and therefore $\mathbb{E}X$ is bounded below by 0. That is, $\mathbb{E}X \geq 0$.

The abstract(Lebesgue) integral of a non-negative function f is defined as:

$$\int f d\mu = \sup\{\int q d\mu : q \text{ is simple and } q \leq f\}$$

Again, we can always take $q = 0$, so that $\int q d\mu = 0 \cdot I_\Omega = 0$. Therefore, $\int f d\mu$ is bounded below by zero and $\int f d\mu \geq 0$.

For an arbitrary random variable X, we can always write :

$$X = X^+ - X^-$$

where $X^+ = \max\{X, 0\} = X \cdot I_{\{X \geq 0\}}$ and $X^- = \max\{-X, 0\} = -X \cdot I_{\{X \leq 0\}}$.

These are non-negative random variables and the expectation of X is defined as:

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$$

Theorem 2.1. *Let X and Y be simple random variables. Then,* $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

Proof. Let $X = \sum_{k=1}^{m} x_k I_{A_k}$ and $Y = \sum_{l=1}^{n} y_l I_{B_l}$ for some non-negative numbers x_k, y_l and events A_k and B_l are such that the A_k and B_l partition Ω. Then, the events $A_k \cap B_l$ partition Ω and

$$\begin{aligned}
\mathbb{E}(X + Y) &= \sum_{k \leq m, l \leq n} (x_k + y_l)\mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m, l \leq n} x_k \mathbb{P}(A_k \cap B_l) + \sum_{k \leq m, l \leq n} y_l \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k \sum_{l \leq n} \mathbb{P}(A_k \cap B_l) + \sum_{l \leq n} y_l \sum_{k \leq m} \mathbb{P}(A_k \cap B_l) \\
&= \sum_{k \leq m} x_k (\mathbb{P}(A_k \cap B_1) + \mathbb{P}(A_k \cap B_2) + \ldots + \mathbb{P}(A_k \cap B_n)) \\
&+ \sum_{l \leq n} y_l (\mathbb{P}(A_1 \cap B_l) + \mathbb{P}(A_2 \cap B_l) + \ldots + \mathbb{P}(A_m \cap B_l)) \\
&= \sum_{k \leq m} x_k \mathbb{P}(A_k) + \sum_{l \leq n} y_l \mathbb{P}(B_l) \\
&= \mathbb{E}X + \mathbb{E}Y
\end{aligned}$$

□

2.3 Non-negative Random Variables.

Our main goal is to prove the linearity of expectation. We first establish a few basic properties of expectation for non-negative random variables.

Theorem 2.2. *Let X and Y be non-negative random variables. We have:*

(a) If $A \in \mathcal{F}$, then $\mathbb{E}I_A = \int I_A \cdot d\mathbb{P} = \mathbb{P}(A)$.

(b) (Monotonicity). If $X \leq Y$, then $EX \leq EY$.

(c) (Translation and Scaling) For $a \geq 0$, $\mathbb{E}(a + X) = a + \mathbb{E}X$ and $\mathbb{E}(aX) = a\mathbb{E}X$.

(d) If $\mathbb{E}X = 0$, then $X = 0$ almost surely (that is $\mathbb{P}\{X = 0\} = 1$).

(e) If A and B are events such that $A \subset B$, then $\mathbb{E}X1_A \leq \mathbb{E}X1_B$.

Proof. (a) $I_A$is a simple random variable. Then, by the definition of the Lebesgue integral, $\mathbb{E}I_A = \int I_A d\mathbb{P} = 1 \cdot \mathbb{P}(A)$.

(b) Let S_X, S_Y be the set of all simple random variables which are less than or equal to X, Y respectively. Since $X \leq Y$, every simple random variable which is less than or equal to X is also less than or equal Y. But, there exists simple random variables that are less than or equal to Y but greater than X. Consequently, $S_X \subseteq S_Y$. Thus, $\{\mathbb{E}Z : \text{Z is simple and } Z \leq X\} \subseteq \{\mathbb{E}Z : Z \text{ is simple and } Z \leq Y\}$. Therefore, it follows that $\sup\{\mathbb{E}Z : \text{Z is simple and } Z \leq X\} \leq \sup\{\mathbb{E}Z : \text{Z is simple and } Z \leq Y\}$. Consequently, $\mathbb{E}X \leq \mathbb{E}Y$.

(c) Let Z be an arbitrary simple random variable which is less than or equal to X. Then, $Z = \sum_{k=1}^{m} x_k I_{A_k}$ where $x_k \geq 0$. We have:

$$\begin{aligned}
\mathbb{E}(a + Z) &= \sum_{k=1}^{m} (a + x_k)\mathbb{P}(A_k) \\
&= a \sum_{k=1}^{m} \mathbb{P}(A_k) + \sum_{k=1}^{m} x_k \mathbb{P}(A_k) \\
&= a + \mathbb{E}Z
\end{aligned}$$

Note that, for all simple random variables $Z \leq X \Longleftrightarrow a + Z \leq a + X$.

$$\begin{aligned}
\mathbb{E}(a + X) &= \sup\{\mathbb{E}(a + Z) : a + Z \text{ is a simple random variable and } a + Z \leq a + X\} \\
&= \sup\{a + \mathbb{E}Z : Z \text{ is a simple random variable and } Z \leq X\} \\
&= a + \sup\{\mathbb{E}Z : Z \text{ is a simple random variable and } Z \leq X\} \\
&= a + \mathbb{E}X
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}(aZ) &= \sum_{k=1}^{m} a x_k \mathbb{P}(A_k) \\
&= a \sum_{k=1}^{m} x_k \mathbb{P}(A_k) \\
&= a\mathbb{E}Z
\end{aligned}$$

($\forall$ simple random variables Z)($Z \leq X$)$\Longleftrightarrow aZ \leq aX$.

$$
\begin{aligned}
\mathbb{E}(aX) &= \sup\{\mathbb{E}aZ : aZ \text{ is a simple random variable and } aZ \le aX\} \\
&= \sup\{a\mathbb{E}Z : Z \text{ is a simple random variable and } Z \le X\} \\
&= a\sup\{\mathbb{E}Z : Z \text{ is a simple random variable and } Z \le X\} \\
&= a\mathbb{E}X
\end{aligned}
$$

(d) For $n \ge 1$, we have $X \ge XI_{\{X \ge \frac{1}{n}\}} \ge \frac{1}{n}I_{\{X \ge \frac{1}{n}\}}$. So, by (a) and (b), we have:

$$0 = \mathbb{E}X \ge \frac{1}{n}\mathbb{E}I_{\{X \ge \frac{1}{n}\}} = \frac{1}{n}\mathbb{P}\{X \ge \frac{1}{n}\}$$

But since $\mathbb{P}\{X \ge \frac{1}{n}\} \ge 0$, we conclude that $\mathbb{P}\{X \ge \frac{1}{n}\} = 0$.

Now,

$$\mathbb{P}(X > 0) = \mathbb{P}(\bigcup_{n=1}^{\infty}\{X \ge \frac{1}{n}\}) = \mathbb{P}(\lim\{X \ge \frac{1}{n}\}) = \lim\left(\mathbb{P}\left\{X \ge \frac{1}{n}\right\}\right) = 0$$

(e) Clearly, if $A \subset B$, then $X \cdot I_A \le X \cdot I_B$. Thus, by the monotonicity property, $\mathbb{E}X1_A \le \mathbb{E}X1_B$. $\square$

The following lemma gives a way to approximate non-negative random variables with monotone sequences of simple ones.

Lemma 2.1. *If X is a random variable, then there is a sequence (Z_n) of non-negative simple random variables such that for every $\omega \in \Omega$, $Z_n(\omega) \le Z_{n+1}(\omega)$ and $Z_n(\omega) \to X(\omega)$ pointwise.*

Proof. For each positive integer n, define

$$Z_n = \sum_{k=1}^{n\cdot 2^n} \frac{k-1}{2^n}\mathbf{1}_{\{\frac{k-1}{2^n} < X < \frac{k}{2^n}\}} + n \cdot \mathbf{1}_{\{X \ge n\}}$$

Essentially, we are dividing the interval $(0, n)$ on the $y-$axis into $n \cdot 2^n$ strips, each of size $1/2^n$. Beyond the point $X \ge n$, Z_n takes the constant value n.

If $n = 2$, this is what $Z_2(\omega)$ looks like. It chops the interval $[0, 2]$ on the $Y-$axis into 8 sub-intervals.

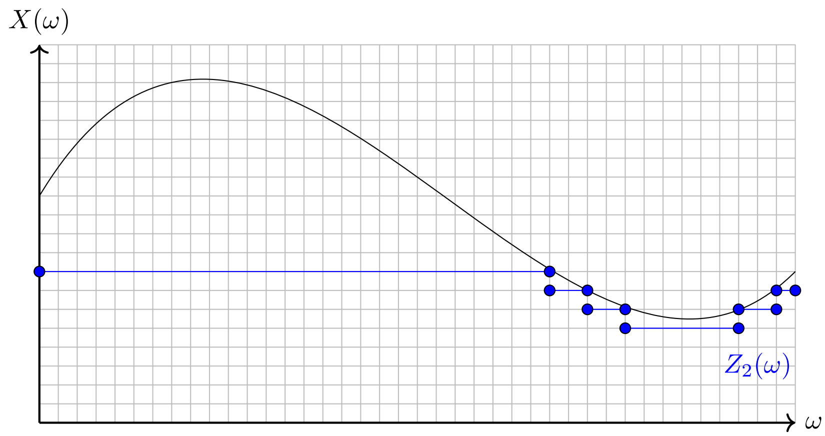


Figure. The step function $Z_2(\omega)$.

As n increases, $Z_n(\omega)$ better approximates of $X(\omega)$.

Pick any arbitrary $\omega \in \Omega$. Let $\epsilon > 0$.

By the Archimedean property, there exists a natural number $N_1 \in \mathbf{N}$, such that $N_1 > X(\omega)$.

We have that $X(\omega)$ lies in an interval I_n, that is $\frac{k-1}{2^n} < X(\omega) < \frac{k}{2^n}$ for some $1 \leq k \leq n \cdot 2^n$, $k \in \mathbf{Z}^+$, for all $n \geq N_1$.

There exists $N_2 \in \mathbf{N}$, such that $l(I_n) = \frac{1}{2^n} < \epsilon$ for all $n \geq N_2$.

Pick $N = \max\{N_1, N_2\}$. Then, for all $n \geq N$, $|Z_n(\omega) - X(\omega)| < \epsilon$.

Thus, $(Z_n(\omega))$ converges pointwise to $X(\omega)$ for all $\omega \in \Omega$.

Note that, the partition points at stage $(n+1)$ include the partition points at stage n and new partition points at the mid-points of the old ones. Because of this, (Z_n) is a monotonically increasing sequence. $\square$

Lemma 2.2. *If X is a positive random variable, and if $(X_n)_{n=1}^{\infty}$ is any sequence of positive simple random variables increasing to X, then $\mathbf{E}[X_n]$ increases to $\mathbf{E}[X]$.*

Proof. Suppose that $X \geq 0$ is a random variable and let (X_n) be a sequence of positive simple random variables, $X_n \geq 0$ such that $X_n \uparrow X$. We would like to show that $\mathbf{E}X_n \to \mathbf{E}X$. Assume that $\mathbf{E}X_n \to a$. We have $X_n \leq X$. By montonicity of expectations $\mathbf{E}X_n \leq \mathbf{E}X$. By the order limit theorem, $\lim \mathbf{E}X_n \leq \mathbf{E}X$, so $a \leq \mathbf{E}X$.

We are interested to show that $a = \mathbf{E}X$. To prove this, we must show that $\mathbf{E}X \leq a$. But, by definition of expectation:

$$\mathbf{E}X = \sup\{\mathbf{E}Y : Y \text{ is a simple random variable and } 0 \leq Y \leq X\}$$

Therefore, it is sufficient to prove that if Y is any simple random variable satisfying $\mathbf{E}Y \leq a$, then a is an upper bound for the set $\{\mathbf{E}Y : Y \text{ is a simple random variable and } 0 \leq Y \leq X\}$. By the definition of supremum,

$$\mathbf{E}X = \sup\{\mathbf{E}Y : Y \text{ is a simple random variable and } 0 \leq Y \leq X\} \leq a$$

To this end, let Y an arbitrary simple random variable satisfying $0 \leq Y \leq X$ and suppose it takes on a finite set of values $\{a_1, \ldots, a_m\}$:

$$Y = \sum_{k=1}^{m} a_k 1_{Y(\omega)=a_k}$$

Take $A_k = \{\omega : Y(\omega) = a_k\}$. Let $0 \leq \epsilon \leq 1$, and consider the random variable (a shifted step function):

$$Y_{n,\epsilon} = (1-\epsilon)Y \cdot 1_{\{(1-\epsilon)Y \leq X_n}$$

Now,

$$Y_{n,\epsilon} = (1-\epsilon)Y$$

on the set $A_k \cap \{\omega : (1-\epsilon)Y(\omega) \leq X_n(\omega)\} = A_{k,n,\epsilon}$ and that $Y_{n,\epsilon} = 0$ on the set $\{\omega : (1-\epsilon)Y(\omega) > X_n(\omega)\}$. Clearly, $Y_{n,\epsilon} \leq X_n$ and so:

$$\begin{aligned}
\mathbf{E}Y_{n,\epsilon} &= \sum_{k=1}^{m}(1-\epsilon)a_k \mathbb{P}(A_{k,n,\epsilon}) \\
&= (1-\epsilon)\sum_{k=1}^{m} a_k \mathbb{P}(A_{k,n,\epsilon}) \\
&\leq \mathbf{E}X_n
\end{aligned}$$

We will show that $A_{k,n,\epsilon}$ increases to A_k. Since $X_n \leq X_{n+1}$ and $X_n \uparrow X$, we conclude that:

$$\{(1-\epsilon)Y(\omega) \leq X_n(\omega)\} \subseteq \{(1-\epsilon)Y(\omega) \leq X_{n+1}(\omega)\} \subseteq \ldots \subseteq \{(1-\epsilon)Y(\omega) \leq X(\omega)\} = \Omega$$

So,

$$A_k \cap \{(1-\epsilon)Y(\omega) \leq X_n(\omega)\} \subseteq A_k \cap \{(1-\epsilon)Y(\omega) \subseteq X_{n+1}(\omega)\} \subseteq \ldots \subseteq A_k$$

That is:

$$A_{k,n,\epsilon} \subseteq A_{k,n+1,\epsilon} \subseteq \ldots \subseteq A_k$$

So,

$$\bigcup_{n=1}^{\infty} A_{k,n,\epsilon} = A_k$$

By continuity of probability:

$$\lim \mathbb{P}(A_{k,n,\epsilon}) = \mathbb{P}(\lim A_{k,n,\epsilon}) = \mathbb{P}(A_k)$$

Therefore, taking limits on both sides of the expression:

$$\begin{aligned}
\lim \mathbf{E}Y_{n,\epsilon} &= (1-\epsilon)\sum_{k=1}^{m} a_k \lim \mathbb{P}(A_{k,n,\epsilon}) \\
&= (1-\epsilon)\sum_{k=1}^{m} a_k \mathbb{P}(A_k) \\
&= (1-\epsilon)\mathbf{E}Y
\end{aligned}$$

Now, since $\mathbf{E}X_n \uparrow a$, so $\mathbf{E}X_n \leq a$. And from above $\mathbf{E}Y_{n,\epsilon} \leq \mathbf{E}X_n$. So:

$$\begin{aligned}
\lim \mathbf{E}Y_{n,\epsilon} &\leq \lim \mathbf{E}X_n \\
(1-\epsilon)\mathbf{E}Y &\leq a
\end{aligned}$$

Letting $\epsilon \to 0$, we have $\mathbf{E}Y \leq a$, which as noted earlier is sufficient to conclude that:

$$\mathbf{E}X \leq a$$

Then:

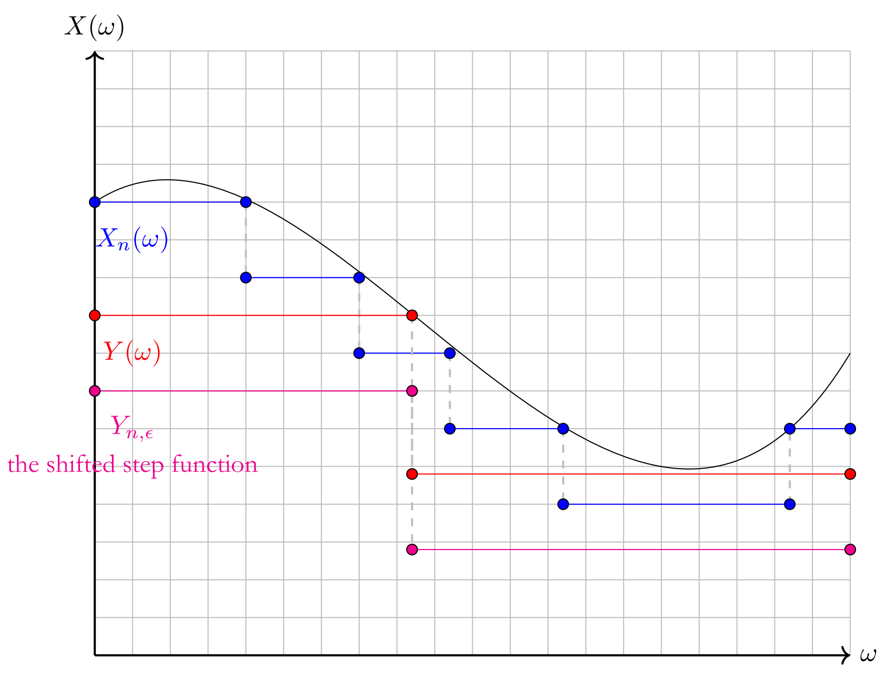


Figure. The shifted step function $Y_{n,\epsilon} = (1-\epsilon)Y\mathbf{1}_{(1-\epsilon)Y \leq X_n}$.

$\square$

Theorem 2.3. *(Monotone Convergence Theorem). Let $X_1, X_2, \ldots, X_n$ be a sequence of non-negative random variables converging increasingly to another real valued random variable X. Meaning, if*

$$X_n \geq 0, \quad X_n \leq X_{n+1}$$

and

$$X_n \uparrow X$$

almost surely, then it follows that:

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}(\lim_{n\to\infty} X_n) = \mathbb{E}X$$

That is, expectation preserves limits.

Proof. Using lemma (2.1), for each n, we can choose an increasing sequence $Y_{n,k}$, $k = 1, 2, 3, \ldots$ of positive simple random variables increasing to X_n and set:

$$Z_k = \max_{n\leq k} Y_{n,k}$$

Essentially, we have these sequences of positive increasing random variables $(Y_{1,k}) \to X_1$, $(Y_{2,k}) \to X_2$, ..., $(Y_{n,k}) \to X_n$. And now, we construct a sequence (Z_k) by taking Z_k as the maximum of the r.v.'s$\{Y_{k,1}, Y_{k,2}, \ldots, Y_{k,k},\}$. Thus, $(Z_k : k \geq 1)$ is a non-decreasing sequence of positive simple random variables and thus it has a limit $Z = \lim_{k\to\infty} Z_k$. Also,

$$Y_{n,k} \leq Z_k \leq X_k \leq X \quad \text{almost surely } \forall n \leq k$$

Hence,

$$\lim_{k\to\infty} Y_{n,k} \leq Z_k \leq X \quad \text{almost surely}$$

In other words, by the Squeeze Theorem, $\lim Z_k = Z$ exists and

$$X_n \leq Z \leq X \quad \text{almost surely}$$

Next, if we let $n \to \infty$, we have:

$$X = Z \quad \text{almost surely}$$

Since the expectation is a positive operator, we have:

$$\mathbf{E}[Y_{n,k}] \leq \mathbf{E}[Z_k] \leq \mathbf{E}[X_k] \quad \text{for } n \leq k$$

Fix n and let $k \to \infty$. Taking limits on both sides of the inequality and using lemma (2.2), we obtain:

$$\mathbf{E}[X_n] \leq \mathbf{E}[Z_k] \leq \lim_{k\to\infty} \mathbf{E}[X_k]$$

Now, let $n \to \infty$ on both sides to obtain:

$$\lim_{n\to\infty} \mathbf{E}[X_n] \leq \mathbf{E}[Z_k] \leq \lim_{k\to\infty} \mathbf{E}[X_k]$$

By the squeeze theorem, $\lim \mathbf{E}[Z_k]$ exists and $\lim \mathbf{E}[Z_k] = \lim \mathbf{E}[X_n]$. But, (Z_k) is a sequence of positive simple random variables and $Z_k \uparrow X$. So, $\mathbf{E}[Z_k] \uparrow \mathbf{E}[X]$. Since, $\mathbf{E}[X_n] \leq \mathbf{E}[X_{n+1}]$, it follows that $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$.

$\square$

Theorem 2.4. *(Linearity of Expectations) Let X and Y be non-negative random variables. Then,*

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$$

Proof. By lemma 2.1, there exists monotonic sequences of non-negative random variables $(X_n)_{n=1}^{\infty}$ and $(Y_n)_{n=1}^{\infty}$ such that $(X_n) \to X$ and $(Y_n) \to Y$. Then, the sequence $X_n + Y_n$ is also monotone, and by the Algebraic limit theorem for sequences, $X_n + Y_n \to X + Y$. By theorem 2.1,

$$\mathbb{E}(X_n + Y_n) = \mathbb{E}X_n + \mathbb{E}Y_n$$

Passing to the limits, we get:

$$\lim \mathbb{E}(X_n + Y_n) = \lim \mathbb{E}X_n + \lim \mathbb{E}Y_n$$

By the Monotone convergence theorem, $\mathbb{E}$ preserves limits, so,

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$$

$\square$

2.4 Fatou's Lemma.

Theorem 2.5. *(Fatou's Lemma) Let Y be a random variable that satisfies $\mathbb{E}[|Y|] < \infty$. Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. Then the following holds:*

- If $Y \leq X_n$, for all n, then $\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[X_n\right]$.
- If $Y \geq X_n$, for all n, then $\mathbb{E}\left[\limsup_{n\to\infty} X_n\right] \geq \limsup_{n\to\infty} \mathbb{E}\left[X_n\right]$.

Proof. Firstly, if $X_n \geq Y$, that is, $(X_1, X_2, X_3, \ldots)$ is any sequence of random variables bounded below, analogous to a sequence of real numbers, the point-wise limit,$\liminf_{n\to\infty} X_n$ always exists and therefore lim inf random variable is defined. Similarly, if $X_n \leq Y$, that is, $(X_1, X_2, X_3, \ldots)$ is any sequence of random variables bounded above, then $\limsup_{n\to\infty} X_n$ always exists and therefore lim sup random variable is defined.

Fix some $n \in \mathbf{N}$. From the definition of infimum, we have:

$$\inf_{k\geq n} X_k - Y \leq X_m - Y, \quad \forall m \geq n$$

By the monotonicity property, it follows that:

$$\mathbb{E}\left[\inf_{k\geq n} X_k - Y\right] \leq \mathbb{E}\left[X_m - Y\right] \quad \forall m \geq n$$

The left-hand side is a constant real number. The right-hand side is indexed by m. So, this inequality holds for a sequence of real numbers (a_m), $m \geq n$, where $a_m = X_m(\omega) - Y(\omega)$.

Consider the set:

$$\{a_m, a_{m+1}, a_{m+2}, \ldots\}$$

This set is bounded below for all $m \geq n$. Hence, its infimum exists. I can take infimum with respect to m, on both sides. By the order limit theorem, we have:

$$\inf_{m\geq n} \mathbb{E}\left[\inf_{k\geq n} X_k - Y\right] \leq \inf_{m\geq n} \mathbb{E}\left[X_m - Y\right] \quad \forall m \geq n$$

Thus,

$$\mathbb{E}\left[\inf_{k\geq n} X_k - Y\right] \leq \inf_{m\geq n} \mathbb{E}\left[X_m - Y\right] \quad \forall m \geq n$$

Define $Z_n = \inf_{k\geq n} X_k - Y$ and $S_n = \inf_{m\geq n} \mathbb{E}\left[X_m - Y\right]$. So, we can write:

$$\mathbb{E}Z_n \leq S_n$$

Passing to the limit as $n \to \infty$, by the Order limit theorem, we have:

$$\lim_{n\to\infty} \mathbb{E}Z_n \leq \lim_{n\to\infty} S_n$$

Note that, $Z_n \geq 0$, since $X_k \geq Y$. And Z_n is a sequence of monotonically increasing random variables. Thus, $\lim Z_n$ exists. By the Monotone Convergence theorem,

$$\lim_{n\to\infty} \mathbb{E}Z_n = \mathbb{E}\left[\lim_{n\to\infty} Z_n\right] = \mathbb{E}\left[\liminf_{n\to\infty} X_n - Y\right] \leq \lim_{n\to\infty} S_n = \liminf_{n\to\infty} \mathbb{E}\left[X_n - Y\right]$$

and so, it follows that:

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}\left[X_n\right]$$

□

The DCT is an important result which asserts a sufficient condition under which we can interchange a limit and expectation.

Theorem 2.6. *(Dominated Convergence Theorem). Consider a sequence of random variables that converges almost surely to X. Suppose that there exists a random variable Y, such that $|X_n| \leq Y$ almost surely for all n and $\mathbb{E}[Y] < \infty$. Then, we have:*

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$$

Proof. Since $-Y \leq X_n \leq Y$ for all $n \in \mathbf{N}$, we can invoke both sides of Fatou's Lemma:

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}X_n$$

and

$$\mathbb{E}\left[\limsup_{n\to\infty} X_n\right] \geq \limsup_{n\to\infty} \mathbb{E}X_n$$

Thus,

$$\mathbb{E}X = \mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X_n \leq \limsup_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}\left[\limsup_{n\to\infty n} X_n\right] = \mathbb{E}X$$

This implies that:

$$\liminf_{n\to\infty} \mathbb{E}X_n = \limsup_{n\to\infty} EX_n$$

so

$$\lim_{n\to\infty} \mathbb{E}X_n$$

exists and further

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X$$

$\square$

3 Gaussian Processes.

3.1 Random Vectors.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We can define several random variables on Ω. A n-tuple of random variables on this space is called a random vector. For example, if $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, then the n-tuple $(X_1, X_2, \ldots, X_n)$ is a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$. The vector is said to be n-dimensional because it contains n-variables. We will sometimes denote a random vector by X.

A good point of view is to think of a random vector $X = (X_1, \ldots, X_n)$ as a random variable (point) in $\mathbf{R}^n$. In other words, for an outcome $\omega \in \Omega$, $X(\omega)$ is a point sampled in $\mathbf{R}^n$, where $X_j(\omega)$ represents the j-th coordinate of the point. The distribution of X, denoted μ_X is the probability on $\mathbf{R}^n$defined by the events related to the values of X:

$$\mathbb{P}\{X \in A\} = \mu_X(A) \quad \text{for a subset } A \text{ in } \mathbf{R}^n$$

In other words, $\mathbb{P}(X \in A) = \mu_X(A)$ is the probability that the random point X falls in A. The distribution of the vector X is also called the joint distribution of $(X_1, \ldots, X_n)$.

Definition 3.1. The **joint distribution function** of $\mathbf{X} = (X, Y)$ is the function $F : \mathbf{R}^2 \to [0, 1]$ given by:

$$F_{\mathbf{X}}(x, y) = \mathbb{P}(X \leq x, Y \leq y) \tag{3.1}$$

Definition 3.2. The joint **PDF** $f_{\mathbf{X}}(x_1, \ldots, x_n)$ of a random vector $\mathbf{X}$ is a function $f_{\mathbf{X}} : \mathbf{R}^n \to \mathbf{R}$ such that the probability that X falls in a subset A of $\mathbf{R}^n$ and is expressed as the multiple integral of $f(x_1, x_{2,} \ldots, x_n)$ over A:

$$\mathbb{P}(X \in A) = \int_A f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

Note that: we must have that the integral of f over the whole of $\mathbf{R}^n$ is 1.

If F is differentiable at the point (x, y), then we usually specify:

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) \tag{3.2}$$

Theorem 3.1. *Let (X, Y) be the random variables with joint density function $f_{X,Y}(x, y)$. The marginal density function $f_X(x)$ and $f_Y(y)$ of the random variables X and Y respectively is given by:*

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{+\infty} f_{(X,Y)}(x, y) dy \\ f_Y(y) &= \int_{-\infty}^{+\infty} f_{(X,Y)}(x, y) dx \end{aligned}$$

Proof. We have:

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \int_{-\infty}^{x} \int_{y=-\infty}^{y=+\infty} f(x, y) dy dx \end{aligned}$$

Differentiating both sides with respect to x,

$$f_X(x) = \int_{y=-\infty}^{y=+\infty} f(x, y) dy dx$$

□

Definition 3.3. For continuous random variables X and Y with the joint density function $f_{(X,Y)}$, the conditional density of Y given $X = x$ is:

$$f_{Y|X}(y|x) = \frac{f_{(X,Y)}(x, y)}{f_X(x)}$$

for all x with $f_X(x) > 0$. This is considered as a function of y for a fixed x. As a convention, in order to make $f_{Y|X}(y|x)$ well-defined for all real x, let $f_{Y|X}(y|x) = 0$ for all x with $f_X(x) = 0$.

We are essentially slicing the the joint density function of $f_{(X,Y)}(x, y)$ by a thin plane $X = x$. How can we speak of conditioning on $X = x$ for X being a continuous random variable, considering that this event has probability zero. Rigorously speaking, we are actually conditioning on the event that X falls within a small interval containing x, say $X \in (x - \epsilon, x + \epsilon)$ and then taking the limit as ϵ approaches zero from the right.

We can recover the joint PDF $f_{(X,Y)}$ if we have the conditional PDF $f_{Y|X}$ and the corresponding marginal f_X:

$$f_{(X,Y)}(x, y) = f_{Y|X}(y|x) \cdot f_X(x)$$

Theorem 3.2. *(Bayes rule and LOTP) Let (X, Y) be continuous random variables. We have the following continuous form of the Bayes rule:*

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{f_X(x)} \tag{3.3}$$

And we have the following continuous form of the law of total probability:

$$f_X(x) = \int_{y=-\infty}^{y=+\infty} f_{X|Y}(x|y) \cdot f_Y(y) dy$$

Proof. By the definition of conditional PDFs, we have:

$$f_{X|Y}(x|y) \cdot f_Y(y) = f_{(X,Y)}(x,y) = f_{Y|X}(y|x) \cdot f_X(x)$$

Dividing throughout by $f_X(x)$, we have:

$$f_{Y|X}(x) = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{f_X(x)} = \frac{f_{(X,Y)}(x,y)}{f_X(x)}$$

□

Example 3.1. (Sampling uniformly in the unit disc). Consider the random vector $\mathbf{X} = (X, Y)$ corresponding to a random point chosen uniformly in the unit disc $\{(x, y) : x^2 + y^2 \leq 1\}$. $\mathbf{X}$ is said to have uniform on the unit circle distribution. In this case the PDF is 0 outside the disc and $\frac{1}{\pi}$ inside the disc:

$$f(x, y) = \frac{1}{\pi} \quad \text{if } x^2 + y^2 \leq 1$$

The random point (X, Y) has x-coordinate X and Y coordinate Y. Each of these are random variables and their PDFs and CDFs can be computed. This is a valid PDF, because:

$$\begin{aligned}
\int\int_D f(x,y) dy dx &= \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy dx \\
&= \frac{1}{\pi} \int_{-1}^{1} [y]_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} dx \\
&= \frac{2}{\pi} \int_{-1}^{1} \sqrt{1-x^2} dx
\end{aligned}$$

Substituting $x = \sin\theta$, we have: $dx = \cos\theta d\theta$ and $\sqrt{1-x^2} = \cos\theta$. The limits of integration are $\theta = -\pi/2$ to $\theta = \pi/2$. Thus,

$$\int\int_D f(x,y)dydx = \frac{2}{\pi}\int_{-\pi/2}^{\pi/2}\cos^2\theta d\theta$$
$$= \frac{1}{\pi}\int_{-\pi/2}^{\pi/2}(1+\cos 2\theta)d\theta$$
$$= \frac{1}{\pi}\left[\theta + \frac{1}{2}\sin 2\theta\right]_{-\pi/2}^{\pi/2}$$
$$= \frac{1}{\pi}\cdot\pi$$
$$= 1$$

The CDF of X is given by:

$$F_X(a) = \int_{-1}^{a}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{\pi}dydx$$
$$= \frac{2}{\pi}\int_{-1}^{a}\sqrt{1-x^2}dx$$

I leave it in this integral form. The PDF of X is obtained by differentiating the CDF, so it is:

$$f_X(x) = \frac{2}{\pi}\sqrt{1-x^2} \tag{3.4}$$

Let's quickly plot the density of X over the domain of the definition $-1 \leq x \leq 1$.

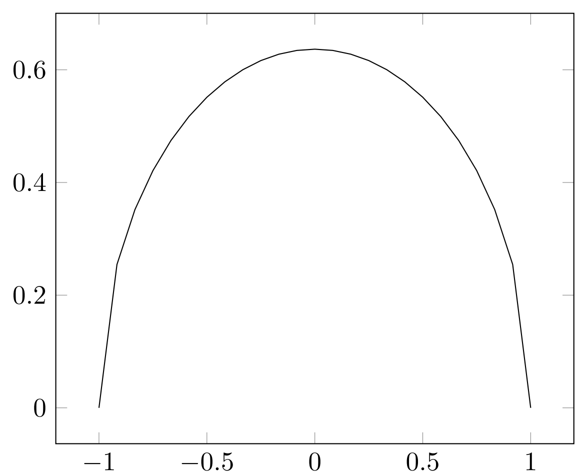


Figure. The PDF of the random variable X.

Not suprisingly the distribution of the x-coordinate is no longer uniform!

If $(X_1, X_2, \ldots, X_n)$ is a random vector, the distribution of a single coordinate, say X_1 is called the *marginal distribution*. In the example 3.1, the marginal distribution of X is determined by the PDF 3.4.

Random variables $X_1, X_2, \ldots, X_n$ defined on the same probability space are said to be independent if for any intervals $A_1, A_2, \ldots, A_n$ in $\mathbf{R}$, the probability factors:

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \times \mathbb{P}(X_2 \in A_2) \times \ldots \times \mathbb{P}(X_n \in A_n)$$

We say that the random variables are independent and identically distributed (IID) if they are independent and their marginal distributions are the same.

When the random vector $(X_1, X_2, \ldots, X_n)$ has a joint PDF $f(x_1, x_2, \ldots, x_n)$, the independence of random variables is equivalent to saying that the joint PDF is given by the product of the marginal PDFs:

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) \times f_2(x_2) \times \ldots \times f_n(x_n) \tag{3.5}$$

3.2 Basic Probabilistic Inequalities.

Inequalities are extremely useful tools in the theoretical development of probability theory.

3.2.1 Jensen's inequality.

Theorem 3.3. *If g is a convex function, and $a > 0$, $b > 0$, with $p \in [0, 1]$, it follows that:*

$$g(pa + (1-p)b) \le pg(a) + (1-p)g(b) \tag{3.6}$$

Proof. This directly follows from the definition of convex functions. □

3.2.2 Jensen's inequality for Random variables.

Theorem 3.4. *If g is a convex function, then it follows that:*

$$\mathbb{E}(g(X)) \ge g(\mathbb{E}X) \tag{3.7}$$

Proof. Another way to express the idea, that a function is convex is to observe that the tangent line at an arbitrary point $(t, g(t))$ always lies below the curve. Let $y = a + bx$ be the tangent to g at the point t. Then, it follows that:

$$\begin{aligned} a + bt &= g(t) \\ a + bx &\le g(x) \end{aligned}$$

for all x.

Thus, it follows that, for any point t, there exists b such that:

$$g(x) - g(t) \ge b(x - t)$$

for all x. Set $t = \mathbb{E}X$ and $x = X$. Then,

$$g(X) - g(\mathbb{E}X) \ge b(X - \mathbb{E}X)$$

Taking expectations on both sides and simplifying:

$$\begin{aligned} \mathbb{E}\left(g(X)\right) - g(\mathbb{E}X) &\ge b(\mathbb{E}X - \mathbb{E}X) = 0 \\ \mathbb{E}g(X) &\ge g(\mathbb{E}X) \end{aligned}$$

□

3.2.3 Young's Inequality.

Theorem 3.5. *If $a \geq 0$ and $b \geq 0$ are non-negative real numbers and if $p > 1$ and $q > 1$ are real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$, then:*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \tag{3.8}$$

Proof. Consider $g(x) = \log x$. Being a concave function, Jensen's inequality can be reversed. We have:

$$\begin{aligned}
g\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) &\geq \frac{1}{p}g(a^p) + \frac{1}{q}g(b^q) \\
\log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) &\geq \frac{1}{p}\log(a^p) + \frac{1}{q}\log(b^q) \\
\log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) &\geq \frac{1}{p}\cdot p\log(a) + \frac{1}{q}\cdot q\log(b) \\
\log\left(\frac{1}{p}a^p + \frac{1}{q}b^q\right) &\geq \log ab
\end{aligned}$$

By the Monotonicity of the $\log x$ function, it follows that :

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

□

3.2.4 Chebyshev's inequality.

One of the simplest and very useful probabilistic inequalities is a tail bound by expectation: the so called Chebyshev's inequality.

Theorem 3.6. *(Chebyshev's inequality) If X is a non-negative random variable, then for every $t \geq 0$:*

$$\mathbb{P}(X \geq t) \leq \frac{1}{t}\mathbb{E}X \tag{3.9}$$

Proof. We have:

$$t \cdot \mathbf{1}_{\{X \geq t\}} \leq X \cdot \mathbf{1}_{\{X \geq t\}}$$

By the monotonicity of expectations, we have:

$$\begin{aligned}
\mathbb{E}\mathbf{1}_{\{X \geq t\}} &\leq \frac{1}{t}\mathbb{E}X \\
\implies \mathbb{P}\{X \geq t\} &\leq \frac{1}{t}\mathbb{E}X
\end{aligned}$$

This closes the proof. □

There are several variants, easily deduced from Chebyshev's inequality using monotonicity of several functions. For a non-negative random variable X and $t > 0$, using the power function x^p, $p > 0$, we get:

$$\mathbb{P}(X \geq t) = \mathbb{P}(X^p \geq t^p) \leq \frac{1}{t^p}\mathbb{E}X^p \tag{3.10}$$

For a real valued random variable X, every $t \in \mathbf{R}$, using the square function x^2 and variance, we have:

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{1}{t^2}\mathbb{E}|X - \mathbb{E}X|^2 = \frac{1}{t^2}Var(X) \tag{3.11}$$

For a real-valued random variable X, every $t \in \mathbf{R}$ and $\lambda > 0$, using the exponential function $e^{\lambda x}$(which is monotonic), we have:

$$\mathbb{P}(X \geq t) = \mathbb{P}(\lambda X \geq \lambda t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{1}{e^{\lambda t}}\mathbb{E}e^{\lambda X} \tag{3.12}$$

Our next inequality, the so-called Holder's inequality is a very effective inequality to factor out the expectation of a product.

3.2.5 Holder's inequality.

Theorem 3.7. *Let $p, q \geq 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$, For random variables X and Y, we have:*

$$\mathbb{E}|XY| \leq (\mathbb{E}|X^p|)^{1/p} (\mathbb{E}|Y^q|)^{1/q}$$

Proof. From the Young's inequality, for any $a, b \in \mathbf{R}$, $p, q \geq 1$, we have:

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Setting $a = \frac{|X|}{(\mathbb{E}|X^p|)^{1/p}}$ and $b = \frac{|Y|}{(\mathbb{E}|Y^q|)^{1/q}}$, we get:

$$\frac{|XY|}{(\mathbb{E}|X^p|)^{1/p} (\mathbb{E}|Y^q|)^{1/q}} \leq \frac{1}{p} \cdot \frac{|X|^p}{\mathbb{E}|X^p|} + \frac{1}{q} \cdot \frac{|Y|^q}{\mathbb{E}|Y^q|}$$

Taking expectations on both sides, and using the monotonicity of expectation property, we get:

$$\frac{\mathbb{E}|XY|}{(\mathbb{E}|X^p|)^{1/p} (\mathbb{E}|Y^q|)^{1/q}} \leq \frac{1}{p} \cdot \frac{\mathbb{E}|X|^p}{\mathbb{E}|X^p|} + \frac{1}{q} \cdot \frac{\mathbb{E}|Y|^q}{\mathbb{E}|Y^q|} = \frac{1}{p} + \frac{1}{q} = 1$$

Consequently,

$$\mathbb{E}|XY| \leq (\mathbb{E}|X^p|)^{1/p} (\mathbb{E}|Y^q|)^{1/q}$$

Let $p = 2$ and $q = 2$. Then, we get the Cauchy-Schwarz inequality:

$$\mathbb{E}|XY| \leq \left[\mathbb{E}(X^2)\right]^{1/2} \left[\mathbb{E}(Y^2)\right]^{1/2}$$

In some ways, the p-th moment of a random variable can be thought of as it's length or p-norm.

Define:

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$$

□

3.2.6 Minkowski's Inequality.

Theorem 3.8. *For random variables X and Y, and for all* $p \geq 1$ *we have:*

$$\|X+Y\|_p \leq \|X\|_p + \|Y\|_p \tag{3.13}$$

Proof. The basic idea of the proof is to use Holder's inequality. Let $\frac{1}{q} = 1 - \frac{1}{p}$ or in other words, $q = \frac{p}{p-1}$. We have:

$$\begin{aligned} \mathbb{E}|X||X+Y|^{p-1} &\leq (\mathbb{E}|X|^p)^{1/p} \left(\mathbb{E}|X+Y|^{(p-1)q}\right)^{1/q} \quad (a) \\ \mathbb{E}|Y||X+Y|^{p-1} &\leq (\mathbb{E}|Y|^p)^{1/p} \left(\mathbb{E}|X+Y|^{(p-1)q}\right)^{1/q} \quad (b) \end{aligned}$$

Adding the above two equations, we get:

$$\begin{aligned} \mathbb{E}(|X+Y||X+Y|^{p-1}) \leq \mathbb{E}(|X|+|Y|)(|X+Y|^{p-1}) &\leq \left\{(\mathbb{E}|X|^p)^{1/p} + (\mathbb{E}|Y|^p)^{1/p}\right\} \left(\mathbb{E}|X+Y|^{(p-1)q}\right)^{1/q} \\ \mathbb{E}|X+Y|^p &\leq \left\{\|X\|_p + \|Y\|_p\right\} (\mathbb{E}|X+Y|^p)^{1/q} \\ (\mathbb{E}|X+Y|^p)^{1/p} &\leq \|X\|_p + \|Y\|_p \\ \|X+Y\|_p &\leq \|X\|_p + \|Y\|_p \end{aligned}$$

□

3.3 A quick refresher of linear algebra.

Many of the concepts in this chapter have very elegant interpretations, if we think of real-valued random variables on a probability space as vectors in a vector space. In particular, variance is related to the concept of norm and distance, while covariance is related to inner-products. These concepts can help unify some of the ideas in this chapter from a geometric point of view. Of course, real-valued random variables are simply measurable, real-valued functions on the abstract space Ω.

Definition 3.4. (Vector Space).

By a vector space, we mean a non-empty set V with two operations:

- Vector addition: $+ : (\mathbf{x}, \mathbf{y}) \to \mathbf{x} + \mathbf{y}$
- Scalar multiplication: $\cdot : (\alpha, \mathbf{x}) \to \alpha\mathbf{x}$

such that the following conditions are satisfied:

(A1) Commutativity. $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$ for all $\mathbf{x},\mathbf{y}\in V$

(A2) Associativity: $(\mathbf{x}+\mathbf{y})+\mathbf{z}=\mathbf{x}+(\mathbf{y}+\mathbf{z})$ for all $\mathbf{x},\mathbf{y},\mathbf{z}\in V$

(A3) Zero Element: There exists a zero element, denoted $\mathbf{0}$ in V, for all $\mathbf{x}\in V$, such that $\mathbf{x}+\mathbf{0}=\mathbf{x}$.

(A4) Additive Inverse: For all $\mathbf{x}\in V$, there exists an additive inverse(negative element) denoted $-\mathbf{x}$ in V, such that $\mathbf{x}+(-\mathbf{x})=\mathbf{0}$.

(M1) Scalar multiplication by identity element in F: For all $\mathbf{x}\in V$, $1\cdot\mathbf{x}=\mathbf{x}$, where 1 denotes the multiplicative identity in F.

(M2) Scalar multiplication and field multiplication mix well: For all $\alpha,\beta\in F$ and $\mathbf{v}\in V$, $(\alpha\beta)\mathbf{v}=\alpha(\beta\mathbf{v})$.

(D1) Distribution of scalar multiplication over vector addition: For all $\alpha\in F$, and $\mathbf{u},\mathbf{v}\in V$, $\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{u}+\alpha\mathbf{v}$.

(D2) Distribution of field addition over scalar multiplication: For all $\alpha,\beta\in F$, and $\mathbf{v}\in V$, $(\alpha+\beta)\mathbf{v}=\alpha\mathbf{v}+\beta\mathbf{v}$.

As usual, our starting point is a random experiment modeled by a probability space $(\Omega,\mathcal{F},\mathbb{P})$, so that Ω is the set of outcomes, $\mathcal{F}$ is the σ-algebra of events and $\mathbb{P}$ is the probability measure on the measurable space $(\Omega,\mathcal{F})$. Our basic vector space V consists of all real-valued random variables defined on $(\Omega,\mathcal{F},\mathbb{P})$. We define vector addition and scalar multiplication in the usual way point-wise.

- Vector addition: $(X+Y)(\omega)=X(\omega)+Y(\omega)$.
- Scalar multiplication: $(\alpha X)(\omega)=\alpha X(\omega)$

Clearly, any function g of a random variable $X(\omega)$ is also a random variable on the same probability space and any linear combination of random variables on $(\Omega,\mathcal{F},\mathbb{P})$ also define a new random variable on the same probability space. Thus, V is closed under vector addition and scalar-multiplication. Since vector-addition and scalar multiplication is defined point-wise, it is easy to see that - all the axioms of a vector space (A1)-(A4), (M1-M2), (D1), (D2) are satisfied. The constantly zero random variable $0(\omega)=0$ and the indicator random variable $I_\Omega(\omega)$ can be thought of as the zero and identity vectors in this vector space.

3.3.1 Inner Products.

In Euclidean geometry, the angle between two vectors is specified by their dot product, which is itself formalized by the abstract concept of inner products.

Definition 3.5. (Inner Product). An inner product on the real vector space V is a pairing that takes two vectors $\mathbf{v},\mathbf{w}\in V$ and produces a real number $\langle\mathbf{v},\mathbf{w}\rangle\in\mathbf{R}$. The inner product is required to satisfy the following three axioms for all $\mathbf{u},\mathbf{v},\mathbf{w}\in V$ and scalars $c,d\in\mathbf{R}$.

(i) Bilinearity:

$$\langle c\mathbf{u}+d\mathbf{v},\mathbf{w}\rangle=c\langle\mathbf{u},\mathbf{w}\rangle+d\langle\mathbf{v},\mathbf{w}\rangle \tag{3.14}$$

$$\langle\mathbf{u},c\mathbf{v}+d\mathbf{w}\rangle=c\langle\mathbf{u},\mathbf{v}\rangle+d\langle\mathbf{u},\mathbf{w}\rangle \tag{3.15}$$

(ii) Symmetry:

$$\langle\mathbf{v},\mathbf{w}\rangle=\langle\mathbf{w},\mathbf{v}\rangle \tag{3.16}$$

(iii) Positive Definiteness:

$$\langle\mathbf{v},\mathbf{v}\rangle>0\quad\text{whenever }\mathbf{v}\neq\mathbf{0} \tag{3.17}$$

$$\langle\mathbf{v},\mathbf{v}\rangle=0\quad\text{whenever }\mathbf{v}=\mathbf{0} \tag{3.18}$$

Definition 3.6. (Norm). A norm on a real vector space V is a function $\|\cdot\| : V \to \mathbf{R}$ satisfying :

(i) Positive Definiteness.

$$\|\mathbf{v}\| \geq 0 \tag{3.19}$$

and

$$\|\mathbf{v}\| = 0 \quad \text{if and only if } \mathbf{v} = \mathbf{0} \tag{3.20}$$

(ii) Scalar multiplication.

$$\|\alpha \mathbf{v}\| = |\alpha| \, \|\mathbf{v}\| \tag{3.21}$$

(iii) Triangle Inequality.

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \tag{3.22}$$

As mentioned earlier, we can define the p-norm of a random variable as :

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$$

(i) Positive semi-definiteness: Since $|X|$ is a non-negative random variable, $|X|^p \geq 0$ and the expectation of a non-negative random variable is also non-negative. Hence, $(\mathbb{E}|X|^p)^{1/p} \geq 0$. Moreover, $\|X\|_p = 0$ implies that $\mathbb{E}|X|^p = 0$. From property (iv) of expectations, $X = 0$.

(ii) Scalar-multiplication: We have:

$$\begin{aligned}
\|cX\|_p &= (\mathbb{E}|cX|^p)^{1/p} \\
&= (|c|^p)^{1/p} (\mathbb{E}|X|^p)^{1/p} \\
&= |c| \cdot \|X\|_p
\end{aligned}$$

(iii) Triangle Inequality. This followed from the Minkowski's inequality.

The space of all random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that $||X||_p < \infty$ is finite is called the L^p space.

3.3.2 Orthogonal Matrices.

Definition 3.7. (Orthogonal Matrix). Let A be an $n \times n$ square matrix. We say that the matrix A is orthogonal, if its transpose is equal to its inverse.

$$A' = A^{-1}$$

This may seem like an odd property to study, but the following theorem explains why it is so useful. Essentially, an orthogonal matrix rotates (or reflects) vectors without distorting angles or distances.

Proposition 3.1. *For an $n \times n$ square matrix A, the following are equivalent:*

(1) A is orthogonal. That is, $A'A = I$.

(2) A preserves norms. That is, for all $\mathbf{x}$,

$$\|A\mathbf{x}\| = \|\mathbf{x}\|$$

(3) A preserves inner products, that is, for every $\mathbf{x}$, $\mathbf{y} \in \mathbf{R}^n$*:*

$$(A\mathbf{x}) \cdot (A\mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$$

Proof. We have:

$$\begin{aligned} \|A\mathbf{x}\|^2 &= (A\mathbf{x})'(A\mathbf{x}) \\ &= \mathbf{x}'(A'A)\mathbf{x} \\ &= \mathbf{x}'I\mathbf{x} \\ &= \mathbf{x}'\mathbf{x} \\ &= ||\mathbf{x}||^2 \end{aligned}$$

Consequently, $||A\mathbf{x}|| = ||\mathbf{x}||$. The matrix A preserves norms. Thus, (1) implies (2).

Moreover, consider

$$\begin{aligned} ||A(\mathbf{x}+\mathbf{y})||^2 &= (A\mathbf{x} + A\mathbf{y}) \cdot (A\mathbf{x} + A\mathbf{y}) \\ &= (A\mathbf{x}) \cdot (A\mathbf{x}) + (A\mathbf{x}) \cdot (A\mathbf{y}) + (A\mathbf{y}) \cdot (A\mathbf{x}) + (A\mathbf{y}) \cdot (A\mathbf{y}) \\ &= ||A\mathbf{x}||^2 + 2(A\mathbf{x}) \cdot (A\mathbf{y}) + ||A\mathbf{y}||^2 && \{\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}\} \\ &= ||\mathbf{x}||^2 + 2(A\mathbf{x}) \cdot (A\mathbf{y}) + ||\mathbf{y}||^2 && \{A \text{ preserves norms}\} \end{aligned}$$

But, $||A(\mathbf{x}+\mathbf{y})||^2 = ||\mathbf{x}+\mathbf{y}||^2 = ||\mathbf{x}||^2 + 2\mathbf{x} \cdot \mathbf{y} + ||\mathbf{y}||^2$. Equating the two expressions, we have the desired result. Hence, (2) implies (3).

Lastly, if A preserves inner products, we may write:

$$\begin{aligned} \langle A\mathbf{x}, A\mathbf{x}\rangle &= \langle \mathbf{x}, \mathbf{x}\rangle \\ (A\mathbf{x})'(A\mathbf{x}) &= \mathbf{x}'\mathbf{x} \\ \mathbf{x}'A'A\mathbf{x} &= 0 \end{aligned}$$

Since $\mathbf{x} \neq \mathbf{0}$, it must be true that $\mathbf{x}'A'A - \mathbf{x}' = 0$. Again, since $\mathbf{x}' \neq \mathbf{0}$, it follows that $A'A - I = 0$. □

Theorem 3.9. *If $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k \in V$ be mutually orthogonal elements, such that $\mathbf{q}_i \neq \mathbf{0}$ for all i, then $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ are linearly independent.*

Proof. Let

$$c_1\mathbf{q}_1 + c_2\mathbf{q}_2 + \ldots + c_k\mathbf{q}_k = \mathbf{0}$$

Since $\langle \mathbf{q}_i, \mathbf{q}_i\rangle = 1$ and $\langle \mathbf{q}_i, \mathbf{q}_j\rangle = 0$ where $i \neq j$, we can take the inner product of the vector $(c_1\mathbf{q}_1 + c_2\mathbf{q}_2 + \ldots + c_i\mathbf{q}_i + \ldots + c_k\mathbf{q}_k)$ with $\mathbf{q}_i$ for each $i = 1, 2, 3, \ldots, k$. It results in $c_i||\mathbf{q}_i||^2 = 0$. Since $\mathbf{q}_i \neq \mathbf{0}$, $||\mathbf{q}_i||^2 > 0$. So, $c_i = 0$. We conclude that $c_1 = c_2 = \ldots = c_k = 0$. Consequently, $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ are linearly independent. □

Theorem 3.10. *Let* $Q = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \dots & \mathbf{q}_n \end{bmatrix}$ *be an* $n \times n$ *orthogonal matrix. Then,* $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ *form an orthonormal basis for* $\mathbf{R}^n$.

Proof. We have $Q\mathbf{e}_i = \mathbf{q}_i$. Consequently,

$$\begin{aligned}
\langle \mathbf{q}_i, \mathbf{q}_i \rangle &= \mathbf{q}_i'\mathbf{q}_i \\
&= (Q\mathbf{e}_i)'(Q\mathbf{e}_i) \\
&= \mathbf{e}_i'Q'Q\mathbf{e}_i \\
&= \mathbf{e}_i'I\mathbf{e}_i \\
&= \mathbf{e}_i'\mathbf{e}_i \\
&= 1
\end{aligned}$$

Assume that $i \neq j$. We have:

$$\begin{aligned}
\langle \mathbf{q}_i, \mathbf{q}_j \rangle &= \mathbf{q}_i'\mathbf{q}_j \\
&= \mathbf{e}_i'Q'Q\mathbf{e}_j \\
&= \mathbf{e}_i'\mathbf{e}_j \\
&= 0
\end{aligned}$$

From theorem (3.9), $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ are linearly independent and hence form an orthonormal basis for $\mathbf{R}^n$. □

3.3.3 Quadratic Forms.

An expression of the form:

$$\mathbf{x}'A\mathbf{x}$$

where $\mathbf{x}$ is a $n \times 1$ column vector and A is an $n \times n$ matrix is called a quadratic form in $\mathbf{x}$ and

$$\mathbf{x}'A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$$

If A and B are $n \times n$ and $\mathbf{x}, \mathbf{y}$ are n-vectors, then

$$\mathbf{x}'(A + B)\mathbf{y} = \mathbf{x}'A\mathbf{y} + \mathbf{x}'B\mathbf{y}$$

The quadratic form of the matrix A is called positive definite if:

$$\mathbf{x}'A\mathbf{x} > 0 \quad \text{whenever } \mathbf{x} \neq \mathbf{0}$$

and positive semidefinite if:

$$\mathbf{x}'A\mathbf{x} \geq 0 \quad \text{whenever } \mathbf{x} \neq \mathbf{0}$$

Letting $\mathbf{e}_i$ be the unit vector with it's ith coordinate vector 1, we have:

$$\mathbf{e}_i' A \mathbf{e}_i = [a_{i1} a_{i2} \dots a_{ii} \dots a_{in}] \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = a_{ii}$$

3.3.4 Eigenthingies and diagonalizability.

Let V and W be finite dimensional vector spaces with $dim(V) = n$ and $dim(W) = m$. A linear transformation $T : V \to W$, is defined by its action on the basis vectors. Suppose:

$$T(\mathbf{v}_j) = \sum_{i=1}^{n} a_{ij} \mathbf{w}_i$$

for all $1 \le i \le m$.

Then, the matrix $A = [T]_{\mathcal{B}_V}^{\mathcal{B}_W}$ of the linear transformation is defined as:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

Definition 3.8. A linear transformation $T : V \to V$ is **diagonalizable** if there exists an ordered basis $\mathcal{B} = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ for V so that the matrix for T with respect to $\mathcal{B}$ is diagonal. This means precisely that, for some scalars $\lambda_1, \lambda_2, \dots, \lambda_n$, we have:

$$\begin{aligned} T(\mathbf{v}_1) &= \lambda_1 \mathbf{v}_1 \\ T(\mathbf{v}_2) &= \lambda_2 \mathbf{v}_2 \\ &\vdots \\ T(\mathbf{v}_n) &= \lambda_n \mathbf{v}_n \end{aligned}$$

In other words, if $A = [T]_{\mathcal{B}}$, then we have:

$$A\mathbf{v}_i = \lambda_i \mathbf{v}_i$$

Thus, if we let P be the $n \times n$ matrix whose columns are the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ and Λ be the $n \times n$ diagonal matrix with diagonal entries $\lambda_1, \lambda_2, \dots, \lambda_n$, then we have:

$$A\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

$$AP = P\Lambda$$

$$A = P\Lambda P^{-1}$$

There exists a large class of diagonalizable matrices - the symmetric matrices. A square matrix A is symmetric, if $A = A'$.

Definition 3.9. Let $T : V \to V$ be a linear transformation. A **non-zero** vector $\mathbf{v} \in V$ is called the eigenvector of T, if there is a scalar λ so that $T(\mathbf{v}) = \lambda\mathbf{v}$. The scalar λ is called the eigenvalue of T.

Lemma 3.1. *Let A be an $n \times n$ matrix, and let λ be any scalar. Then,*

$$E(\lambda) = \{\mathbf{x} \in \mathbf{R}^n : A\mathbf{x} = \lambda\mathbf{x}\} = \ker(A - \lambda I)$$

is a subspace of $\mathbf{R}^n$. Moreover, if $E(\lambda) \neq \{\mathbf{0}\}$ if and only if λ is an eigenvalue, in which case we call $E(\lambda)$ the λ-eigenspace of the matrix A.

Proof. We know that, $E(\lambda)$ is a subset of $\mathbf{R}^n$. Moreover, if $\mathbf{u}, \mathbf{v} \in E(\lambda)$, then $A(c_1\mathbf{u} + c_2\mathbf{v}) = c_1 A\mathbf{u} + c_2 A\mathbf{v} = \lambda(c_1\mathbf{u} + c_2\mathbf{v})$. Consequently, $c_1\mathbf{u} + c_2\mathbf{v} \in E(\lambda)$. Thus, $E(\lambda)$ is a subspace of $\mathbf{R}^n$.

Moreover, by definition, λ is an eigenvalue of A precisely when $\mathbf{x} \neq \mathbf{0}$ vector in $E(\lambda)$. This closes the proof. ☐

Theorem 3.11. *Let A be a $n \times n$ square matrix. If A is a singular matrix, then* $\det A = 0$.

Proof. By definition, a square matrix is said to be non-singular, if it can be reduced to an upper triangular form with all non-zero elements on the diagonal - the pivots, by elementary row operations. A singular matrix is such that it's echelon form has a row of zeroes, and its row vectors are linearly dependent and $\det A = 0$. ☐

Theorem 3.12. *Let A be a $n \times n$ square matrix. Then, λ is an eigenvalue of A if and only if* $\det(A - \lambda I) = 0$.

Proof. λ is an eigenvalue of A, if and only, the homogenous system of linear equations $(A - \lambda I)\mathbf{x} = \mathbf{0}$ has non-trivial solutions. Consequently, the only possibility is that there are one more free variables (more variables than the number of equations). In other words, $(A - \lambda I)$ must be a singular matrix and $\det(A - \lambda I) = 0$. ☐

Example 3.2. Let's find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix}$$

We begin by computing

$$\begin{aligned} \det(A - \lambda I) &= \begin{bmatrix} 1-\lambda & 2 & 1 \\ 0 & 1-\lambda & 0 \\ 1 & 2 & 1-\lambda \end{bmatrix} \\ &= (1-\lambda)(1-\lambda)^2 - (1-\lambda) \\ &= (1-\lambda)[(1-\lambda)^2 - 1)] \\ &= -\lambda(1-\lambda)(2-\lambda) \end{aligned}$$

Thus, the eigenvalues of A are $\lambda = 0$, $\lambda = 1$ and $\lambda = 2$.

We find the respective eigenspaces:

1) Fix $\lambda = 0$. We see that:

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 0 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The augmented matrix $[A|b]$ is :

$$\begin{bmatrix} 1 & 2 & 1 & | & 0 \\ 0 & 1 & 0 & | & 0 \\ 1 & 3 & 1 & | & 0 \end{bmatrix}$$

$R_3 - R_1$, $R_3 - R_2$ and $R_1 - 2R_2$ leaves us with:

$$\begin{bmatrix} 1 & 0 & 1 & | & 0 \\ 0 & 1 & 0 & | & 0 \\ 0 & 0 & 0 & | & 0 \end{bmatrix}$$

So, $x_1 + x_3 = 0$ and $x_2 = 0$. Here, x_3 is a free variable. Thus,

$$E(0) = \{\alpha(1, 0, -1) | \alpha \in \mathbf{R}\}$$

2) Fix $\lambda = 1$. We see that:

$$\begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & 0 \\ 1 & 3 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Thus, $2x_2 + x_3 = 0$ and $x_1 + 3x_2 = 0$. Here x_3 is a free variable. Let $x_3 = -2\alpha$. Then, $x_2 = \alpha$ and $x_1 = -3\alpha$. Consequently,

$$E(1) = \{\alpha(-3, 1, -2) | \alpha \in \mathbf{R}\}$$

3) Fix $\lambda = 3$. We see that:

$$\begin{bmatrix} -1 & 2 & 1 \\ 0 & -1 & 0 \\ 1 & 3 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The augmented matrix $[A|b]$ is :

$$\begin{bmatrix} -1 & 2 & 1 & | & 0 \\ 0 & -1 & 0 & | & 0 \\ 1 & 3 & -1 & | & 0 \end{bmatrix}$$

$R_3 + R_1$, $R_3 + 5R_2$ followed by $R_1 + 2R_2$ gives:

$$\left[\begin{array}{ccc|c} -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Thus, $x_2 = 0$ and $x_1 - x_3 = 0$. Here x_3 is the free variable. Hence,

$$E(2) = \{\alpha(1, 0, 1) : \alpha \in \mathbf{R}\}$$

Clearly, there exists a basis $\mathcal{B} = \{(1, 0, -1), (-3, 1, -2), (1, 0, 1)\}$ with respect to which the matrix of T is diagonal. Hence, A is diagonalizable.

Judging from the previous example, it appears that when an $n \times n$ square matrix has n distinct eigen values, the corresponding eigenvectors form a linearly independent set and will therefore give a *diagonalizing basis.* Let's begin with a slightly stronger statement.

Theorem 3.13. *Let $T : V \to V$ be a linear transformation. Suppose $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are eigenvectors of T corresponding to the distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. Then, $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ is a linearly independent set of vectors.*

Proof. Let m be the largest number between 1 and k (inclusive) so that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is linearly independent. We proceed by contradiction. We want to see $m = k$. Assume that $m < k$. Then, we know that $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is linearly independent and $\{\mathbf{v}_1, \ldots, \mathbf{v}_m, \mathbf{v}_{m+1}\}$ is linearly dependent. Thus, $\mathbf{v}_{m+1} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_m\mathbf{v}_m$ such that atleast one of $c_1, c_2, \ldots, c_m$ are non-zero. Then, using repeatedly the fact that $T(\mathbf{v}_i) = \lambda_i\mathbf{v}_i$:

$$\begin{aligned} \mathbf{0} &= (T - \lambda_{m+1}I)\mathbf{v}_{m+1} = (T - \lambda_{m+1}I)(c_1\mathbf{v}_1 + \ldots + c_m\mathbf{v}_m) \\ &= c_1\left(T\mathbf{v}_1 - \lambda_{m+1}I\mathbf{v}_1\right) + c_2\left(T\mathbf{v}_2 - \lambda_{m+1}I\mathbf{v}_2\right) + \ldots + c_m\left(T\mathbf{v}_m - \lambda_{m+1}I\mathbf{v}_m\right) \\ &= c_1(\lambda_1 - \lambda_{m+1})\mathbf{v}_1 + c_2(\lambda_2 - \lambda_{m+1})\mathbf{v}_2 + \ldots + c_m(\lambda_m - \lambda_{m+1})\mathbf{v}_m \end{aligned}$$

Since $\lambda_i \neq \lambda_{m+1}$ for $i = 1, 2, 3, \ldots, m$ and since $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_m\}$ is linearly independent, the only other possibility is $c_1 = c_2 = \ldots = c_m = 0$. But, this contradicts the fact that $\mathbf{v}_{m+1}$ is an eigenvector since $\mathbf{v}_{m+1} \neq \mathbf{0}$. Thus, it cannot happen that $m < k$. Consequently, $m = k$. ☐

What is underlying this formal argument is the observation that: if $\mathbf{v} \in E(\lambda) \cap E(\mu)$, then $T\mathbf{v} = \lambda\mathbf{v}$ and $T\mathbf{v} = \mu\mathbf{v}$. Hence, if $\lambda \neq \mu$, then $\mathbf{v} = \mathbf{0}$. That is, if $\lambda \neq \mu$, we have $E(\lambda) \cap E(\mu) = \{\mathbf{0}\}$.

Corollary 3.1. *Suppose V is an n-dimensional vector space and $T : V \to V$ has n distinct eigenvalues. Then T is diagonalizable.*

Proof. The set of n corresponding eigenvectors must be linearly independent and hence form a basis for V. The matrix of T with respect to the eigenbasis is always diagonal. ☐

The converse of this statement is not true. There are many diagonalizable matrices with repeated eigen-values.

Definition 3.10. Let λ be an eigenvalue of a linear transformation. The algebraic multiplicity of λ is its multiplicity as a root of the characteristic polynomial $p(t)$ that is, the highest power of $t - \lambda$ dividing $p(t)$. The geometric multiplicity of λ is the dimension of the eigenspace $E(\lambda)$.

Proposition 3.2. *Let λ be an eigenvalue of algebraic multiplicity m and geometric multiplicity d. Then, the geometric multiplicity is always bounded by the algebraic multiplicity, and $1 \leq d \leq m$.*

Proof. Suppose λ is the eigenvalue of the linear transformation T. Then, $d = \dim E(\lambda) \geq 1$ by definition. Now, choose a basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d\}$ for $E(\lambda)$ and extend it to a basis $\mathcal{B} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ for V. Then, the matrix of T with respect to $\mathcal{B}$ is of the form

$$A = \begin{bmatrix} \lambda I_d & B \\ 0_{(n-d)\times d} & C \end{bmatrix}$$

The characteristic polynomial $p(t)$ of the matrix A is given by:

$$\begin{aligned} p(t) &= \det(A - tI) \\ &= \det((\lambda - t)I_d) \cdot \det(C - tI) \\ &= (\lambda - t)^d \cdot \det(C - tI) \end{aligned}$$

Since the characteristic polynomial does not depend on the choice of basis, the algebraic multiplicity of λ is atleast d. $\square$

Lemma 3.2. *(Lagrange Multipliers) Suppose $f, g : \mathbf{R}^n \to \mathbf{R}$ are scalar-valued C^1 functions - that is partial derivatives ∂_{x_i} in all variables are continuous. Let $S = \{\mathbf{x} \in \mathbf{R}^n | g(\mathbf{x}) = c\}$ denote the level set of g at height c. Then if $f|_S$ (the restriction of f to S) has an extremum point $\mathbf{x}_0$ in S such that $\nabla g(\mathbf{x}_0) \neq \mathbf{0}$, there exists a scalar λ such that*

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0) \tag{3.23}$$

Proof. Let's visualize the situation for the case $n = 3$, where the constraint equation $g(x, y, z) = c$ defines a surface S in $\mathbf{R}^3$.

Thus, suppose that $\mathbf{x}_0$ is an extremum of f restricted to S. We consider a further restriction of f - to a curve lying in S and passing through $\mathbf{x}_0$. Let $\mathbf{x}(t) = (x(t), y(t), z(t))$ be the parametric equation of one such arbitrary path $\mathbf{x} : I \subseteq \mathbf{R} \to \mathbf{R}^3$ lying in S with $\mathbf{x}(t_0) = \mathbf{x}_0$ for some $t_0 \in I$. Then, the restriction of f to $\mathbf{x}$ can be written as a function of a single variable t. That is:

$$F(t) := f(\mathbf{x}(t))$$

Because $\mathbf{x}_0$ is an extremum of f on the whole of S, it is also an extremum on the path $\mathbf{x}$. Since F is a differentiable function of t, by the interior-extremum theorem, it follows that $F'(t_0) = 0$. The chain rule implies that:

$$F'(t) = \nabla f(\mathbf{x}) \cdot \mathbf{x}'(t)$$

Evaluating at $t = t_0$, we have:

$$F'(t_0) = 0 = \nabla f(\mathbf{x}(t_0)) \cdot \mathbf{x}'(t_0)$$

Thus, $\nabla f(\mathbf{x}(t_0))$ is perpendicular to any curve in S passing through $\mathbf{x}_0$; that is $\nabla f(\mathbf{x}_0)$ is normal to S at $\mathbf{x}_0$. We've already seen previously that the gradient vector $\nabla g(\mathbf{x}_0)$ is also normal to S at $\mathbf{x}_0$. Since the normal direction to the level S is uniquely determined, we must conclude that $\nabla f(\mathbf{x}_0)$ and $\nabla g(\mathbf{x}_0)$ are parallel vectors. Therefore, there exists a scalar λ such that:

$$\nabla f(\mathbf{x}_0) = \lambda \nabla g(\mathbf{x}_0)$$

$\square$

3.3.5 The Gram-Schmidt Process.

The advantage of using an orthonormal basis is, that the coordinates of any vector are explicitly given as inner products. Let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be an orthonormal basis of $\mathbf{R}^n$. And let $\mathbf{v} = c_1\mathbf{u}_1 + \ldots + c_n\mathbf{u}_n$ be an arbitrary vector. Then we have:

$$c_i = \mathbf{v} \cdot \mathbf{u}_i$$

Moreover, the magnitude (norm) of the vector is given by the Pythagorean formula:

$$\begin{aligned}\|\mathbf{v}\|_2^2 &= \langle \mathbf{v}, \mathbf{v} \rangle \\ &= c_1^2 + c_2^2 + \ldots + c_n^2\end{aligned}$$

Once we are convinced of the utility of orthogonal and orthonormal bases, a natural question arises: how can we construct them? A practical algorithm was discovered Pierre-Simon Laplace in the eighteenth century. Today, the algorithm is known as the *Gram-Schmidt process,* after its rediscovery by Gram and twentieth century mathematician Schmidt.

Let W be a finite dimensional vector space, such that $\dim W = n$. We assume that, we already know some basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$ of W, where $n = \dim W$. Our goal is to use this information to construct an orthogonal basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

We will construct the orthogonal basis one-by-one. Since initially, we are not worrying about normality, there are no conditions on the first orthogonal basis element $\mathbf{v}_1$, so there is no harm in choosing :

$$\mathbf{v}_1 = \mathbf{w}_1$$

Note that, $\mathbf{v}_1 \neq \mathbf{0}$, since $\mathbf{w}_1$ appears in the original basis. Starting with $\mathbf{w}_2$, the second basis vector $\mathbf{v}_2$ must be orthogonal to the first: $\langle \mathbf{v}_2, \mathbf{v}_1 \rangle = 0$.

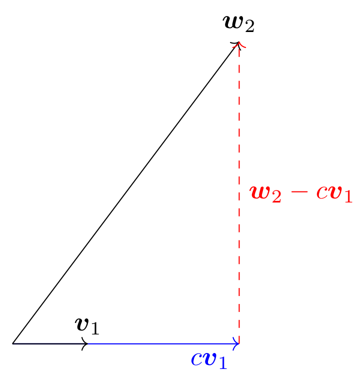


Figure. Resolving the vector $\mathbf{w}_2$ into two components (1) along $\mathbf{u}_1$ and (2) perpendicular to $\mathbf{u}_1$.

Let us try to arrange this, by subtracting a suitable multiple of $\mathbf{v}_1$, and set:

$$\mathbf{v}_2 = \mathbf{w}_2 - c\mathbf{v}_1$$

The orthogonality condition

$$
\begin{aligned}
0 &= \langle \mathbf{v}_2, \mathbf{v}_1 \rangle \\
&= (\mathbf{w}_2 - c\mathbf{v}_1) \cdot \mathbf{v}_1 \\
&= \mathbf{w}_2 \cdot \mathbf{v}_1 - c \left\| \mathbf{v}_1 \right\|^2 \\
c &= \frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\left\| \mathbf{v}_1 \right\|^2}
\end{aligned}
$$

and therefore

$$
\mathbf{v}_2 = \mathbf{w}_2 - \left(\frac{\mathbf{w}_2 \cdot \mathbf{v}_1}{\left\| \mathbf{v}_1 \right\|^2} \right) \mathbf{v}_1
$$

The linear independence of $\mathbf{v}_1 = \mathbf{w}_1$ and $\mathbf{w}_2$ ensures that $\mathbf{v}_2 \neq \mathbf{0}$.

Next, we construct:

$$
\mathbf{v}_3 = \mathbf{w}_3 - c_1 \mathbf{v}_1 - c_2 \mathbf{v}_2
$$

by subtracting suitable multiples of the first two orthogonal basis elements from $\mathbf{w}_3$. We want $\mathbf{v}_3$ to be orthogonal to both $\mathbf{v}_1$ and $\mathbf{v}_2$. Since we already arranged that $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, this requires:

$$
\begin{aligned}
0 = \mathbf{v}_3 \cdot \mathbf{v}_1 = (\mathbf{w}_3 \cdot \mathbf{v}_1) - c_1 \left\| \mathbf{v}_1 \right\|^2 \\
0 = \mathbf{v}_3 \cdot \mathbf{v}_2 = (\mathbf{w}_3 \cdot \mathbf{v}_2) - c_2 \left\| \mathbf{v}_2 \right\|^2
\end{aligned}
$$

And hence:

$$
\begin{aligned}
c_1 &= \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\left\| \mathbf{v}_1 \right\|^2} \\
c_2 &= \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\left\| \mathbf{v}_2 \right\|^2}
\end{aligned}
$$

Therefore the next orthogonal basis vector is given by the formula:

$$
\mathbf{v}_3 = \mathbf{w}_3 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_1}{\left\| \mathbf{v}_1 \right\|^2} \mathbf{v}_1 - \frac{\mathbf{w}_3 \cdot \mathbf{v}_2}{\left\| \mathbf{v}_2 \right\|^2} \mathbf{v}_2
$$

Since $\mathbf{v}_1$ and $\mathbf{v}_2$ are linear combinations of $\mathbf{w}_1$ and $\mathbf{w}_2$, we must have that $\mathbf{v}_3 \neq \mathbf{0}$, since otherwise this would imply that $\mathbf{w}_3$ can be written as a linear combination of $\mathbf{w}_1$ and $\mathbf{w}_2$ making them linearly dependent.

Continuing in the same manner, suppose we have already constructed the mutually orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$ as linear combinations of $\mathbf{w}_1, \ldots, \mathbf{w}_{k-1}$. The next orthogonal basis element $\mathbf{v}_k$ will be obtained from $\mathbf{w}_k$ by subtracting a suitable linear combination of the previous orthogonal basis elements. In this fashion we establish the general *Gram-Schmidt* formula -

$$
\mathbf{v}_k = \mathbf{w}_k - \sum_{j=1}^{k-1} \frac{\mathbf{w}_k \cdot \mathbf{v}_j}{\left\| \mathbf{v}_j \right\|^2} \mathbf{v}_j \tag{3.24}
$$

If we are after an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ we merely normalize the resulting orthogonal basis vectors, setting $\mathbf{u}_k = \frac{\mathbf{v}_k}{\|\mathbf{v}_k\|}$.

3.3.6 Modifications of the Gram-Schmidt process.

With the basic Gram-Schmidt algorithm now in hand, it is worth looking at a couple of reformulations that have both practical and theoretical advantages. The first can be used to construct orthonormal basis vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ directly from the basis $\mathbf{w}_1, \ldots, \mathbf{w}_n$.

We begin by replacing each orthogonal basis vector in the basic Gram-Schmidt formula (3.24) by its normalized version $\mathbf{u}_j = \mathbf{v}_j / \|\mathbf{v}_j\|$. The original basis vectors can be expressed in terms of the orthonormal basis via a triangular system.

$$
\begin{aligned}
\mathbf{w}_1 &= r_{11}\mathbf{u}_1 \\
\mathbf{w}_2 &= r_{12}\mathbf{u}_1 + r_{22}\mathbf{u}_2 \\
\mathbf{w}_3 &= r_{13}\mathbf{u}_1 + r_{23}\mathbf{u}_2 + r_{33}\mathbf{u}_3 \\
&\;\vdots \\
\mathbf{w}_n &= r_{1n}\mathbf{u}_1 + r_{2n}\mathbf{u}_2 + r_{3n}\mathbf{u}_3 + \ldots + r_{nn}\mathbf{u}_n
\end{aligned}
\tag{3.25}
$$

The coefficients r_{ij} can, in fact, be computed directly from these formulas. Indeed taking, the inner product of the equation for $\mathbf{w}_j$ with the orthonormal basis vector $\mathbf{u}_i$ for $i \leq j$, we obtain in view of the orthonormality constraints:

$$
\begin{aligned}
\mathbf{w}_j \cdot \mathbf{u}_i &= r_{1j}\mathbf{u}_1 \cdot \mathbf{u}_i + \ldots + r_{ij}\mathbf{u}_i \cdot \mathbf{u}_i + \ldots + r_{jj}\mathbf{u}_j \cdot \mathbf{u}_i \\
&= r_{ij}
\end{aligned}
$$

and hence:

$$
r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle \tag{3.26}
$$

On the other hand, we have:

$$
\begin{aligned}
\|\mathbf{w}_j\|^2 &= \|r_{1j}\mathbf{u}_1 + r_{2j}\mathbf{u}_2 + \ldots + r_{jj}\mathbf{u}_j\|^2 \\
&= r_{1j}^2 + r_{2j}^2 + \ldots + r_{jj}^2
\end{aligned}
\tag{3.27}
$$

The pair of equations (3.26) and (3.27) can be rearranged to devise a recursive procedure to compute the orthonormal basis. We begin by setting $r_{11} = \|\mathbf{w}_1\|$ and so $\mathbf{u}_1 = \mathbf{w}_1 / r_{11}$. At each subsequent stage, $j \geq 2$, we assume that we have already constructed $\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}$. We then compute

$$
r_{ij} = \langle \mathbf{w}_j, \mathbf{u}_i \rangle \quad \text{for each} \quad i = 1, 2, \ldots, j-1 \tag{3.28}
$$

We obtain next the orthonormal basis vector $\mathbf{u}_j$ by computing

$$
\begin{aligned}
r_{jj} &= \sqrt{\|\mathbf{w}_j\|^2 - r_{1j}^2 - r_{2j}^2 - \ldots - r_{j-1,j}^2} \\
\mathbf{u}_j &= \frac{\mathbf{w}_j - r_{1j}\mathbf{u}_1 - r_{2j}\mathbf{u}_2 - \ldots - r_{j-1,j}\mathbf{u}_{j-1}}{r_{jj}}
\end{aligned}
\tag{3.29}
$$

3.3.7 The QR Factorization.

The Gram-Schmidt procedure for orthonormalizing bases of $\mathbf{R}^n$ can be reinterpreted as a matrix factorization.

Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be a basis of $\mathbf{R}^n$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the corresponding orthonormal basis that results from any one of the implementations of the Gram-Schmidt process. We assemble both sets of column vectors to form non-singular $n \times n$ matrices:

$$A = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \ldots & \mathbf{w}_n \end{bmatrix}, \quad Q = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{bmatrix}$$

Since the $\mathbf{u}_i$ form an orthonormal basis, Q is an orthogonal matrix. In view of the matrix multiplication formula, the Gram-Schmidt equations (3.25) can be recast into an equivalent matrix form:

$$\begin{aligned} A &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & \ldots & r_{1n} \\ & r_{22} & r_{23} & \ldots & r_{2n} \\ & & r_{33} & \ldots & r_{3n} \\ & & & \ddots & \\ & & & \ldots & r_{nn} \end{bmatrix} \\ &= QR \end{aligned}$$

Since the Gram-Schmidt algorithm works on any basis, the only requirement on the matrix A is that it's columns are linearly-independent and form a basis of $\mathbf{R}^n$, and hence A can be any non-singular matrix. We have therefore established the celebrated QR-factorization of non-singular matrices.

Theorem 3.14. *Every non-singular matrix A can be factored, $A = QR$ into the product of an orthogonal matrix Q and an upper triangular matrix R.*

3.3.8 Numerically stable implementation of QR-Factorization.

We take a slightly different approach to generating orthogonal vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Define:

$$\mathbf{w}_1^{(1)} = r_{11}\mathbf{u}_1$$

and define the jth iterate of the procedure as:

$$\mathbf{w}_k^{(j)} = r_{1k}\mathbf{u}_1 + r_{2k}\mathbf{u}_2 + \ldots + r_{jk}\mathbf{u}_j, \quad j \leq k$$

Observe that:

$$\left\langle \mathbf{w}_k^{(j)}, \mathbf{u}_j \right\rangle = r_{jk}$$

We can treat all vectors simultaneously instead of sequentially and compute in the $j = 1$st iteration:

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{w}_1 / r_{11} \\ \mathbf{w}_2^{(2)} &= \left(\mathbf{w}_2^{(1)} - \left\langle \mathbf{w}_2^{(1)}, \mathbf{u}_1 \right\rangle \mathbf{u}_1 \right) \\ \mathbf{w}_3^{(2)} &= \left(\mathbf{w}_3^{(1)} - \left\langle \mathbf{w}_3^{(1)}, \mathbf{u}_1 \right\rangle \mathbf{u}_1 \right) \\ &\vdots \\ \mathbf{w}_n^{(2)} &= \left(\mathbf{w}_n^{(1)} - \left\langle \mathbf{w}_n^{(1)}, \mathbf{u}_1 \right\rangle \mathbf{u}_1 \right) \end{aligned}$$

Note that, the updated vectors $\mathbf{w}_2^{(2)}, \mathbf{w}_3^{(2)}, \ldots, \mathbf{w}_n^{(2)}$ are orthogonal to $\mathbf{u}_1$.

In the $j = 2$nd iteration, we compute:

$$
\begin{aligned}
\mathbf{u}_2 &= \mathbf{w}_2^{(2)} / r_{22} \\
\mathbf{w}_3^{(3)} &= \left(\mathbf{w}_3^{(2)} - \left\langle \mathbf{w}_3^{(2)}, \mathbf{u}_2 \right\rangle \mathbf{u}_2\right) \\
\mathbf{w}_4^{(3)} &= \left(\mathbf{w}_4^{(2)} - \left\langle \mathbf{w}_4^{(2)}, \mathbf{u}_2 \right\rangle \mathbf{u}_2\right) \\
&\vdots \\
\mathbf{w}_n^{(3)} &= \left(\mathbf{w}_n^{(2)} - \left\langle \mathbf{w}_n^{(2)}, \mathbf{u}_2 \right\rangle \mathbf{u}_2\right)
\end{aligned}
$$

Since $\mathbf{w}_2^{(2)}$ was orthogonal to $\mathbf{u}_1$, $\mathbf{u}_2$ must also be orthogonal to $\mathbf{u}_1$. Further, $\mathbf{w}_3^{(3)}, \ldots, \mathbf{w}_3^{(n)}$ are orthogonal to both $\mathbf{u}_1, \mathbf{u}_2$. In particular, in the jth iteration we compute:

$$
\begin{aligned}
\mathbf{u}_j &= \mathbf{w}_j^{(j)} / r_{jj} \\
\mathbf{w}_{j+1}^{(j+1)} &= \left(\mathbf{w}_{j+1}^{(j)} - \left\langle \mathbf{w}_{j+1}^{(j)}, \mathbf{u}_j \right\rangle \mathbf{u}_j\right) \\
\mathbf{w}_{j+2}^{(j+1)} &= \left(\mathbf{w}_{j+2}^{(j)} - \left\langle \mathbf{w}_{j+2}^{(j)}, \mathbf{u}_j \right\rangle \mathbf{u}_j\right) \\
&\vdots \\
\mathbf{w}_n^{(j+1)} &= \left(\mathbf{w}_n^{(j)} - \left\langle \mathbf{w}_n^{(j)}, \mathbf{u}_j \right\rangle \mathbf{u}_j\right)
\end{aligned}
$$

We can summarize the above steps as follows. We iterate $j = 1$ to n. For $j = 1$, we start with the initial basis $\mathbf{w}_k^{(1)} = \mathbf{w}_k$, and set $\mathbf{u}_1 = \mathbf{w}_1^{(1)} / r_{11}$.

In the jth iteration, we set $\mathbf{u}_j = \mathbf{w}_j^{(j)} / r_{jj}$ and for all $k = j+1$ to n, we let $\mathbf{w}_k^{(j+1)} = \mathbf{w}_k^{(j)} - \left\langle \mathbf{w}_k^{(j)}, \mathbf{u}_j \right\rangle \mathbf{u}_j$. Also, we set $r_{jk} = \left\langle \mathbf{w}_k^{(j)}, \mathbf{u}_j \right\rangle$.

Listing 1: QR Factorization

```cpp
#include <iostream>
#include <Eigen/Dense>
#include <tuple>

using namespace Eigen;

std::pair<MatrixXd,MatrixXd> qrFactorization(MatrixXd& A)
{
    int n = A.rows();

    // A = [w_1,w_2,...,w_n] = [u_1 u_2 ...u_n]R = QR

    MatrixXd Q = MatrixXd::Zero(n,n);
    MatrixXd R = MatrixXd::Zero(n,n);

    //We proceed iteratively and build the column vectors u_0, u_1, ... u_{n-1}
    for (int j{ 0 }; j < n; ++j)
    {
        // The scalar r_jj = ||w_j^(j)||
        for (int i{ 0 }; i < n; ++i)
            R(j, j) += A(i, j) * A(i, j);

        R(j, j) = sqrt(R(j, j));
```

```
        // u_j = w_j / r_jj
        for (int i{ 0 };i < n; ++i)
        {
            Q(i, j) = A(i,j) / R(j, j);
        }

        for (int k{ j + 1 }; k < n; ++k)
        {
            // Dot product of <w_k^(j),u_j>
            for (int i{ 0 }; i < n; ++i)
            {
                R(j,k) += A(i, k) * Q(i, j);
            }

            // w_k^(j+1) = w_k^(j) - <w_k^(j),u_j> u_j
            for (int i{ 0 }; i < n; ++i)
            {
                A(i, k) = A(i, k) - R(j,k) * Q(i, j);
            }
        }
    }
    return std::make_pair(Q, R);
}
```

3.3.9 Gram Matrices.

Symmetric matrices whose entries are given by the inner products of elements of an inner product space are called *Gram matrices,* after the Danish mathematician *Jorgen Gram.*

Definition 3.11. Let V be an inner product space, and let $\mathbf{v}_1, \ldots, \mathbf{v}_n \in V$. The associated *Gram matrix*

$$K = \begin{bmatrix} \langle \mathbf{v}_1, \mathbf{v}_1 \rangle & \langle \mathbf{v}_1, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_1, \mathbf{v}_n \rangle \\ \langle \mathbf{v}_2, \mathbf{v}_1 \rangle & \langle \mathbf{v}_2, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_2, \mathbf{v}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{v}_n, \mathbf{v}_1 \rangle & \langle \mathbf{v}_n, \mathbf{v}_2 \rangle & \ldots & \langle \mathbf{v}_n, \mathbf{v}_n \rangle \end{bmatrix}$$

is the $n \times n$ symmetric matrix whose entries are the inner-products between the selected vector space elements.

Theorem 3.15. *All Gram matrices are positive semi-definite.*

Proof. Let K be an arbitrary Gram matrix. To prove the positive semi-definiteness of K, we need to examine the associated quadratic form:

$$\begin{aligned} q(\mathbf{x}) &= \mathbf{x}' K \mathbf{x} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} x_i x_j \end{aligned}$$

But, $k_{ij} = \langle \mathbf{v}_i, \mathbf{v}_j \rangle$. Substituting the values for the matrix entries, we obtain:

$$q(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

For intuition, let's choose $n = 2$. The quadratic form becomes:

$$
\begin{aligned}
q(\mathbf{x}) &= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle x_1^2 + \langle \mathbf{v}_1, \mathbf{v}_2 \rangle x_1 x_2 + \langle \mathbf{v}_2, \mathbf{v}_1 \rangle x_2 x_1 + \langle \mathbf{v}_2, \mathbf{v}_1 \rangle x_2^2 \\
&= \langle x_1\mathbf{v}_1 + x_2\mathbf{v}_2, x_1\mathbf{v}_1 + x_2\mathbf{v}_1 \rangle && \{\text{Bi-linearity of inner products}\} \\
&= \|x_1\mathbf{v}_1 + x_2\mathbf{v}_2\|^2
\end{aligned}
$$

Therefore, we can write the original quadratic form as a single inner product:

$$
\begin{aligned}
q(\mathbf{x}) &= \left\langle \sum_{i=1}^{n} x_i\mathbf{v}_i, \sum_{j=1}^{n} x_j\mathbf{v}_j \right\rangle \\
&= \left\| \sum_{i=1}^{n} x_i\mathbf{v}_i \right\|^2 \\
&= \|\mathbf{v}\|^2 && \{\text{Norm } \|\cdot\| \text{ is positive semi-definite}\} \\
&\geq 0
\end{aligned}
$$

□

3.3.10 Positive Definiteness.

Gram matrices furnish us with an almost inexhaustible supply of positive semi-definite matrices. However, we still do not know how to test whether a given symmetric matrix is positive definite.

From elementary school, we recall the algebraic technique known as *completing the square*, first arising in the derivation of the formula for the solution to the quadratic equation

$$
q(x) = ax^2 + 2bx + c = 0 \tag{3.30}
$$

The idea is to combine the first two terms in the equation (3.30) to form a perfect square and thereby rewrite the quadratic function in the form :

$$
\begin{aligned}
q(x) &= a\left[x^2 + 2\frac{b}{a}x + \frac{c}{a}\right] \\
&= a\left[x^2 + 2x \cdot \frac{b}{a} + \left(\frac{b}{a}\right)^2 + \frac{c}{a} - \left(\frac{b}{a}\right)^2\right] \\
&= a\left[\left(x + \frac{b}{a}\right)^2 + \frac{ac - b^2}{a^2}\right]
\end{aligned}
$$

As a consequence,

$$
\left(x + \frac{b}{a}\right)^2 = \frac{b^2 - ac}{a^2}
$$

The familiar *quadratic formula*:

$$
x = \frac{-b \pm \sqrt{b^2 - ac}}{a}
$$

follows by taking the square root on both sides and then solving for x.

We can perform the same kind of manipulation on a homogenous quadratic form:

$$q(x_1, x_2) = ax_1^2 + 2bx_1x_2 + cx_2^2 \tag{3.31}$$

In this case, provided $a \neq 0$, completing the square amounts to writing:

$$\begin{aligned} q(x_1, x_2) &= ax_1^2 + 2bx_1x_2 + cx_2^2 \\ &= a\left[x_1^2 + 2x_1 \cdot \frac{b}{a}x_2 + \left(\frac{b}{a}x_2\right)^2 + \frac{c}{a}x_2^2 - \frac{b^2}{a^2}x_2^2\right] \\ &= a\left[\left(x_1 + \frac{b}{a}x_2\right)^2 + \frac{ac-b^2}{a^2}x_2^2\right] \\ &= ay_1^2 + \frac{ac-b^2}{a}y_2^2 \end{aligned} \tag{3.32}$$

The net result is to re-express $q(x_1, x_2)$ as a simpler sum of squares of the new variables:

$$y_1 = x_1 + \frac{b}{a}x_2, \quad y_2 = x_2 \tag{3.33}$$

It is not hard to see that the final expression in (3.32) is positive definite, as a function of y_1 and y_2 if and only if both coefficients are positive:

$$a > 0, \quad \frac{ac-b^2}{a} > 0 \tag{3.34}$$

Our goal is to adapt this simple idea to analyse the positive semi-definiteness of quadratic forms depending on more than two variables. To this end, let us write the quadratic form identity in the matrix form. The original quadratic form in (3.31) can be written as:

$$\begin{aligned} q(\mathbf{x}) &= \mathbf{x}'K\mathbf{x} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \end{aligned}$$

Similarly, the right hand side of (3.32) can be written as:

$$\hat{q}(\mathbf{y}) = \mathbf{y}'D\mathbf{y}, \quad \text{where} \quad D = \begin{bmatrix} a & 0 \\ 0 & \frac{ac-b^2}{a} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{3.35}$$

Anticipating the final result, the equations (3.33) connecting $\mathbf{x}$ and $\mathbf{y}$ can themselves be written in the matrix form as:

$$\mathbf{y} = L'\mathbf{x} \quad \text{or} \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 + \frac{b}{a}x_2 \\ x_2 \end{bmatrix}, \quad \text{where} \quad L' = \begin{bmatrix} 1 & 0 \\ b/a & 1 \end{bmatrix} \tag{3.36}$$

Substituting $\mathbf{y}$into (3.35), we obtain:

$$\mathbf{y}'D\mathbf{y} = (L'\mathbf{x})'D(L'\mathbf{x}) = \mathbf{x}'LDL'\mathbf{x} = \mathbf{x}'K\mathbf{x}, \quad \text{where} \quad K = LDL' \tag{3.37}$$

We are thus led to the realization that completing the square is the same as the LDL' factorization of a symmetric matrix K.

From basic algebra, we know that, if A is a non-singular matrix, with all it's pivot elements $a_{kk}^{(k)}$ non-zero in the Gaussian elimination process, then $A = LDU$ where L and U are lower and upper uni-triangular matrices and D is a diagonal matrix consisting of the pivots of A. If the matrix is symmetric, then it admits the unique factorization LDL'.

The identity (3.37) is therefore valid for all real symmetric matrices that are non-singular and can be reduced to an upper triangular matrix by performing elementary row operations (without row interchanges). It also shows how to write the associated quadratic form as a sum of squares:

$$q(\mathbf{x}) = \mathbf{x}'K\mathbf{x} = \mathbf{y}'D\mathbf{y} = d_1y_1^2 + d_2y_2^2 + \ldots + d_ny_n^2 \quad \text{where} \quad \mathbf{y} = L'\mathbf{x} \tag{3.38}$$

The coefficients d_i are the diagonal entries of D, which are the pivots of K. The diagonal quadratic form is positive definite, $\mathbf{y}'D\mathbf{y} > 0$ for all $\mathbf{y} \neq \mathbf{0}$ if and only if, when performing the Gaussian elimination process, all the pivots are positive. We can now add this to our list of standard results.

Theorem 3.16. *(Positive Definiteness) Let K be a $n \times n$ real symmetric positive definite (SPD) matrix. Then the following statements are equivalent.*

(i) K is non-singular and can be reduced to an upper triangular matrix by performing elementary row operations (without row permutations), and it has positive pivot elements when performing Gaussian elimination.

(ii) K admits a factorization $K = LDL'$, where $D = diag(d_1, \ldots, d_n)$ such that $d_i > 0$ for all $i = 1, 2, 3, \ldots, n$.

3.3.11 Cholesky Factorization.

The identity (3.37) shows us how to write an arbitrary regular quadratic form $q(\mathbf{x})$ as linear combination of squares. We can push this result slightly further in the positive definite case. Since each pivot d_i is positive, we can write the quadratic form as a sum of squares:

$$\begin{aligned} d_1y_1^2 + d_2y_2^2 + \ldots + d_ny_n^2 &= (\sqrt{d_1}y_1)^2 + (\sqrt{d_2}y_2)^2 + \ldots + (\sqrt{d_n}y_n)^2 \\ &= z_1^2 + z_2^2 + \ldots + z_n^2 \end{aligned}$$

where $z_i = \sqrt{d_i}y_i$. In the matrix form, we are writing:

$$\begin{aligned} \hat{q}(\mathbf{y}) &= \mathbf{y}'D\mathbf{y} \\ &= \mathbf{z}'\mathbf{z} \\ &= \|\mathbf{z}\|^2 \end{aligned}$$

where $\mathbf{z} = S\mathbf{y}$, with $S = diag(\sqrt{d_1}, \sqrt{d_2}, \ldots, \sqrt{d_n})$. Since $D = S^2$, the matrix S can be thought of as a square root of the diagonal matrix D. Substituting back into the equation $K = LDL'$, we deduce the *Cholesky factorization:*

$$\begin{aligned} K &= LDL' \\ &= LSS'L' \\ &= LS(LS)' \\ &= MM' \end{aligned}$$

of a positive definite matrix, first proposed by the early twentieth-century French geographer Andrew Louis Cholesky for solving problems in geodetic surveying. Note that, M is a lower triangular matrix with all positive diagonal entries, namely the square roots of the pivots: $m_{ii} = \sqrt{d_i}$.

Example 3.3. Let the matrix $K = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 6 & 0 \\ -1 & 0 & 9 \end{bmatrix}$. Let $KX = I$. We consider the augmented matrix $\begin{bmatrix} K & | & I \end{bmatrix}$. Performing Gaussian elimination, we have:

$$\begin{bmatrix} 1 & 2 & -1 & | & 1 & 0 & 0 \\ 2 & 6 & 0 & | & 0 & 1 & 0 \\ -1 & 0 & 9 & | & 0 & 0 & 1 \end{bmatrix}$$

The pivot element $a_{11}^{(1)} = 1$. Performing $R_2 = R_2 - 2R_1$and $R_3 = R_3 + R_1$, the above system is row-equivalent to:

$$\begin{bmatrix} 1 & 2 & -1 & | & 1 & 0 & 0 \\ 0 & 2 & 2 & | & -2 & 1 & 0 \\ 0 & 2 & 8 & | & 1 & 0 & 1 \end{bmatrix}$$

The pivot element $a_{22}^{(2)} = 2$. Performing $R_3 = R_3 - R_2$, the above system is row-equivalent to:

$$\begin{bmatrix} 1 & 2 & -1 & | & 1 & 0 & 0 \\ 0 & 2 & 2 & | & -2 & 1 & 0 \\ 0 & 0 & 6 & | & 3 & -1 & 1 \end{bmatrix}$$

The pivot element $a_{33}^{(3)} = 6$. We have now reduced the system to the form $\begin{bmatrix} DU & | & C \end{bmatrix}$, where U is an upper unitriangular matrix. Thus, Gaussian Elimination produces the factors:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix}, \quad L^T = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Thus,

$$M = LS = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{6} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & \sqrt{2} & 0 \\ -1 & \sqrt{2} & \sqrt{6} \end{bmatrix}$$

and $K = MM'$.

We conclude our discussion by observing the following:

Lemma 3.3. *If a square matrix K is SPD, it admits a Cholesky factorization of the form $K = MM^T$.*

Example 3.4. Prove that, if K is real SPD(symmetric positive definite matrix), then the diagonal elements of K are positive.

Proof. Since K is real SPD, K admits a factorization $K = LL^T$. Since the diagonal element (j, j) is the inner product of the j-th row of L and the j-th column of L^T, we have:

$$k_{jj} = \sum_{m=1}^{n} l_{jm} l'_{mj}$$

But, $l_{jm} = l'_{mj}$, since $L = \left(L^T\right)^T$. Hence, k_{jj} is a sum of squares. Further, since the diagonal elements of L, that is, all elements l_{jj} are strictly positive, the sum $k_{jj} = l_{j1}^2 + \ldots + l_{jj}^2 + \ldots + l_{jn}^2 > 0$. Consequently, the diagonal elements of K are positive. $\square$

3.3.12 Cholesky Factorization Algorithm.

We adopt the commonly used notation where Greek lower-case letters refer to scalars, lower-case letters refer to (column) vectors and upper case letters refer to matrices. The $\star$ refers to a part of A that is neither stored nor updated. By substituting these partitioned matrices into $A = LL'$ we find that:

$$\begin{bmatrix} \alpha_{11} & a_{21}^T \\ a_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 \\ l_{21} & L_{22} \end{bmatrix} \begin{bmatrix} \lambda_{11} & l_{21}^T \\ 0 & L_{22}^T \end{bmatrix} = \begin{bmatrix} \lambda_{11}^2 & \star \\ \lambda_{11} l_{21} & l_{21} l_{21}^T + L_{22} L_{22}^T \end{bmatrix}$$

so that :

$$\begin{array}{c|c} \alpha_{11} = \lambda_{11}^2 & \star \\ \hline a_{21} = \lambda_{11} l_{21} & A_{22} = l_{21} l_{21}^T + L_{22} L_{22}^T \end{array}$$

and hence:

$$\begin{array}{c|c} \lambda_{11} = \sqrt{a_{11}} & \star \\ \hline l_{21} = a_{21}/\lambda_{11} & L_{22} = \text{Cholesky}(A_{22} - l_{21} l_{21}^T) \end{array}$$

The last equality is clever. Essentially, if $A_{22} = l_{21} l_{21}^T - L_{22} L_{22}^T$, we must have: $L_{22} L_{22}^T = A_{22} - l_{21} l_{21}^T$. So, to find L_{22}, we recursively perform the cholesky factorization of the matrix $A_{22} - l_{21} l_{21}^T$. These equalities motivate the following block algorithm:

1. Partition $A = \begin{array}{c|c} \alpha_{11} & \star \\ \hline a_{21} & A_{22} \end{array}$.
2. Overwrite $\alpha_{11} := \lambda_{11} = \sqrt{\alpha_{11}}$.
3. Overwrite $a_{21} := l_{21} = a_{21}/\lambda_{11}$.
4. Overwrite $A_{22} := A_{22} - l_{21} l_{21}^T$.
5. Continue with $A = A_{22}$.

We can also implement a serial algorithm by multiplying out the matrices:

$$\begin{aligned} \begin{bmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} &= \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} & l_{41} \\ 0 & l_{22} & l_{32} & l_{42} \\ 0 & 0 & l_{33} & l_{43} \\ 0 & 0 & 0 & l_{44} \end{bmatrix} \\ &= \begin{bmatrix} l_{11}^2 & & & \\ l_{21} l_{11} & l_{21}^2 + l_{22}^2 & & \\ l_{31} l_{11} & l_{31} l_{21} + l_{32} l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 & \\ l_{41} l_{11} & l_{41} l_{31} + l_{42} l_{32} & l_{41} l_{31} + l_{42} l_{32} + l_{43} l_{33} & l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 \end{bmatrix} \end{aligned}$$

We can thus solve for the elements of the matrix L, column-by-column. The expressions for l_{jj} and l_{ij} in general, are given by:

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$$

$$l_{ij} = \frac{1}{l_{jj}} (a_{ij} - \sum_{k=1}^{j-1} l_{ik} \cdot l_{jk}), \quad \forall i > j$$

Listing 2: Cholesky Factorization

```
#include <iostream>
#include <Eigen/Dense>
#include <cmath>

using Eigen::MatrixXd;

// Cholesky-Crout algorithm starts from the upper-left corner of the matrix L and proceeds
// to calculate matrix column by column
MatrixXd choleskyDecomposition(const MatrixXd& A)
{
    MatrixXd L = MatrixXd::Zero(A.rows(), A.cols());

    for (int j{ 0 }; j < A.cols(); ++j)
    {
        double sum{ 0.0 };
        for (int k{ 0 }; k < j; ++k)
        {
            sum += L(j, k) * L(j, k);
        }
        L(j, j) = sqrt(A(j, j) - sum);

        for (int i{ j + 1 }; i < A.rows(); ++i)
        {
            double sum{ 0.0 };
            for (int k{ 0 }; k < j; ++k) {
                sum += L(i, k) * L(j, k);
            }

            L(i, j) = (A(i, j) - sum)/L(j,j);
        }
    }

    return L;
}

int main()
{
    MatrixXd K(3, 3);

    K <<    4, 12, -16,
            12, 37, -43,
            -16, -43, 98;

    MatrixXd L = choleskyDecomposition(K);

    std::cout << "The␣SPD(Symmetric␣Positive␣Definite)␣matrix␣K␣is␣:␣" << std::endl;
    std::cout << K << std::endl;
    std::cout << "The␣Cholesky␣Decomposition␣of␣K␣into␣K=LL\'␣yields␣L␣:" << std::endl;
    std::cout << L << std::endl;

    return 0;
}
```

3.3.13 Eigen-decomposition of real symmetric matrices.

We review couple of lemmas from basic algebra, which we shall need in the main result.

Lemma 3.4. *Every linearly independent sequence can be extended to a basis.*

Let V be a finite-dimensional vector space and let $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n$ be linearly independent. Then, there exists a basis of V containing $\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n$.

Proof. Let $\mathscr{L} = \mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n$. Since V is finite-dimensional, there exist elements $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ of V such that they span V.

Define a sequence of sequences of the elements of V as follows. Set $\mathscr{L}_0 = \mathscr{L}$ and for $i \geq 0$, define:

$$\mathscr{L}_{i+1} = \begin{cases} \mathscr{L}_i & \text{if } \mathbf{v}_i \in \text{span}(\mathscr{L}_i) \\ \mathscr{L}_i, \mathbf{v}_{i+1} & \text{otherwise} \end{cases}$$

Here, $\mathscr{L}_i, \mathbf{v}_{i+1}$ just means take the sequence $\mathscr{L}_i$ and add $\mathbf{v}_{i+1}$ on to the end.

Note that in either case, $\mathbf{v}_{i+1} \in \text{span}(\mathscr{L}_{i+1})$ and also that $\mathscr{L}_0 \subseteq \mathscr{L}_1 \subseteq \ldots \subseteq \mathscr{L}_m$.

By construction, each sequence $\mathscr{L}_i$ is linearly independent and in particular $\mathscr{L}_m$ is linearly independent. Furthermore, $\text{span}(\mathscr{L}_m)$ contains $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ and therefore contains $\text{span}(\mathscr{L}_m) = V$. Therefore, $\mathscr{L}_m$ is a basis for V containing $\mathscr{L}$. This completes the proof. □

Lemma 3.5. *(EMHE) Every matrix has an (atleast one) eigenvalue, and a corresponding eigenvector.*

Proof. This is just the Fundamental Theorem of Algebra(FTA), but it's still worth enumerating as a theorem.

Let $A \subseteq \mathbf{C}^{n \times n}$ and the scalar field $\mathbf{F} = \mathbf{C}$.

Let $\mathbf{v}$ be any non-zero vector in $\mathbf{C}^n$. Consider the list $\mathscr{L} = \mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \ldots, A^n\mathbf{v}$. There are $n+1$ vectors in the list, so they must be linearly dependent. There exists scalars $a_0, a_1, \ldots, a_n$ from $\mathbf{C}$ not all zero, such that:

$$a_0\mathbf{v} + a_1 A\mathbf{v} + a_2 A^2\mathbf{v} + \ldots + a_n A^n\mathbf{v} = \mathbf{0}$$

By FTA, the polynomial equation of degree n:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n = 0$$

has n linear factors

$$p(x) = (x - \lambda_1)(x - \lambda_2) \cdots (x - \lambda_n) = 0$$

where $\lambda_i \in \mathbf{C}$, $i = 1, 2, \ldots, n$.

Putting it all together,

$$\begin{aligned} p(A)\mathbf{v} = \mathbf{0} &= a_0\mathbf{v} + a_1 A\mathbf{v} + a_2 A^2\mathbf{v} + \ldots + a_n A^n\mathbf{v} \\ &= (a_0 + a_1 A + a_2 A^2 + \ldots + a_n A^n)\mathbf{v} \\ &= (A - \lambda_1 I)(A - \lambda_2 I) \cdots (A - \lambda_n I)\mathbf{v} \end{aligned}$$

This shows that the composition of the factors has a non-trivial nullspace. $\ker((A - \lambda_1 I)(A - \lambda_2 I) \cdots (A - \lambda_n I)) \neq \{\mathbf{0}\}$. So, atleast one of the factors must fail to be injective. There exists λ_i, such that $(A - \lambda_i)\mathbf{v} = \mathbf{0}$ such that $\mathbf{v} \neq \mathbf{0}$. Thus, A has atleast one eigenvalue and a corresponding eigenvector. □

Theorem 3.17. *(Spectral Theorem) Every real symmetric matrix is diagonalizable.*

Let A be a symmetric $n \times n$ real matrix. Then,

1) The eigenvalues of A are real.

2) There exists an orthonormal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ for $\mathbf{R}^n$ consisting of the eigenvectors of A. That is, there is an orthogonal matrix Q so that $Q^{-1}AQ = \Lambda$ is diagonal.

Proof. (I) Before we get to the proof, note that for any square matrix A, we have:

$$\begin{aligned}\langle A\mathbf{x}, \mathbf{y}\rangle &= \mathbf{x}'A'\mathbf{y} \\ &= \langle \mathbf{x}, A'\mathbf{y}\rangle\end{aligned}$$

Since for a symmetric matrix A, we have, $A = A'$, it follows that:

$$\langle A\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, A\mathbf{y}\rangle$$

Or using the dot-product notation, we could write:

$$(A\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (A\mathbf{y})$$

Suppose $\mathbf{v} \neq \mathbf{0}$ be a non-zero vector in $\mathbf{R}^n$ such that there exists a complex scalar λ, satisfying:

$$A\mathbf{v} = \lambda\mathbf{v} \tag{3.39}$$

We can therefore write:

$$(A\mathbf{v}) \cdot \mathbf{v} = (\lambda\mathbf{v}) \cdot \mathbf{v} = \lambda(\mathbf{v} \cdot \mathbf{v}) \tag{3.40}$$

Alternatively,

$$(A\mathbf{v}) \cdot \mathbf{v} = \mathbf{v} \cdot (A\mathbf{v}) \tag{3.41}$$

We can now take the complex conjugate of the (3.39) equation. Remember that A is a real matrix so $\overline{A} = A$. Thus, we have the conjugated version of the eigen-value equation:

$$\overline{(A\mathbf{v})} = \overline{A}\overline{\mathbf{v}} = A\overline{\mathbf{v}} = \overline{(\lambda\mathbf{v})} = \overline{\lambda}\overline{\mathbf{v}}$$

In equation (3.40), if we replace the second vector $\mathbf{v}$ with its conjugate, $\overline{\mathbf{v}}$, we get:

$$(A\mathbf{v}) \cdot \overline{\mathbf{v}} = \lambda(\mathbf{v} \cdot \overline{\mathbf{v}}) \tag{3.42}$$

In equation (3.41), if we replace the second vector $\mathbf{v}$ with its conjugate, $\overline{\mathbf{v}}$, we get:

$$(A\mathbf{v}) \cdot \overline{\mathbf{v}} = \mathbf{v} \cdot (A\overline{\mathbf{v}}) = \mathbf{v} \cdot (\overline{\lambda}\overline{\mathbf{v}}) = \overline{\lambda}(\mathbf{v} \cdot \overline{\mathbf{v}}) \tag{3.43}$$

Now, since $\mathbf{v}$ is an eigenvector, it cannot be the zero vector.

Without loss of generality, if $\mathbf{v} = (v_1, \ldots, v_n)$, then $\mathbf{v} \cdot \overline{\mathbf{v}} = |v_1|^2 + \ldots + |v_n|^2 \neq 0$, so $\mathbf{v} \cdot \overline{\mathbf{v}} \neq 0$.

The two expressions for $(A\mathbf{v}) \cdot \overline{\mathbf{v}}$ are equal, so $(\lambda - \overline{\lambda})(\mathbf{v} \cdot \overline{\mathbf{v}}) = 0$. But, $(\mathbf{v} \cdot \overline{\mathbf{v}}) \neq 0$, so $\lambda = \overline{\lambda}$. Therefore, $\lambda \in \mathbf{R}$.

(II) We proceed by mathematical induction on n.

For $n = 1$, any 1×1 symmetric matrix is already diagonal. Since A and $v \in V$ are both scalars, $Av = \lambda v$ where $\lambda = A$. Thus, we can pick any non-zero scalar v to form a basis of $\mathbf{R}$. And we can write, $A = P^{-1}\Lambda P$, where $P = I$ and $\Lambda = A$.

Induction hypothesis: Every $k \times k$ symmetric matrix is diagonalizable for $k = 1, 2, 3, \ldots, n-1$. If C is a real symmetric matrix of size $k \times k$, then there exists an orthogonal matrix R such that $R^{-1}CR$ is diagonal.

By lemma (3.5), the square matrix A has atleast one eigenvalue. Suppose λ_1 is an eigenvalue of the matrix A. By part (I), we know that $\lambda_1 \in \mathbf{R}$. Choose a unit vector $\mathbf{q}_1$ that is an eigenvector with eigenvalue λ_1. (Obviously, this is no problem. We can pick an eigenvector and then make it a unit vector by dividing by it's length.)

By lemma (3.4), we can extend this to a basis $\{\mathbf{q}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$ of V. By the Gram-Schmidt orthogonalization algorithm, given the basis $\{\mathbf{q}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$, we can find a corresponding orthonormal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ of V.

Now, we huddle these basis vectors together as column-vectors of a matrix and formulate the matrix P.

$$P = \begin{bmatrix} \mathbf{q_1} & \mathbf{q}_2 & \ldots & \mathbf{q}_n \end{bmatrix}$$

By definition, P is an orthogonal matrix.

Let

$$B = P^{-1}AP$$

We are interested to show that B is diagonal.

Step I. B is symmetric.

We have:

$$\begin{aligned}
B^T &= (P^{-1}AP)^T \\
&= (P^TAP)^T && \{P^{-1} = P^T\} \\
&= P^TA^T(P^T)^T \\
&= P^TA^TP \\
&= P^TAP && \{A \text{ is symmetric}\} \\
&= B
\end{aligned}$$

We are now going to try and write B in the block form to try to see the structure that this matrix must have and hope that it looks like, it is going to be diagonal.

Step II. The structure of B.

The way we do this, is to consider the matrix B post-multiplied by $\mathbf{e}_1$. Consider $B\mathbf{e}_1$. This should actually give us the first column of B. Now, we also know that $B = P^TAP$. So, we could actually say, well,

$$P^TAP\mathbf{e}_1 = P^TA\mathbf{q}_1$$

Now, remember that $\mathbf{q}_1$ is the normalized eigenvector corresponding to the eigenvalue λ_1. So, $A\mathbf{q}_1 = \lambda_1\mathbf{q}_1$. That means, this is equal to:

$$
\begin{aligned}
P^T A\mathbf{q}_1 &= P^T\lambda_1\mathbf{q}_1 \\
&= \lambda_1 P^t\mathbf{q}_1 \\
&= \lambda_1 \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \\ \mathbf{q}_n^T \end{bmatrix} \mathbf{q}_1 \\
&= \lambda_1 \begin{bmatrix} \mathbf{q}_1^T\mathbf{q}_1 \\ \mathbf{q}_2^T\mathbf{q}_1 \\ \vdots \\ \mathbf{q}_n^T\mathbf{q}_1 \end{bmatrix} \\
&= \lambda_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

This is the first column of the matrix B. Since $B = B^T$, the first row should also be

$$
\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \end{bmatrix}
$$

So, we can write the matrix B in the form:

$$
B = \begin{bmatrix} \lambda_1 & O \\ O & C \end{bmatrix}
$$

The first row and the first column are satisying the need to be diagonal.

Step III.

We know that C is a $n-1 \times n-1$ symmetric matrix. By the inductive hypothesis, C is diagonalizable and further there exists an orthogonal matrix R, such that $R^{-1}CR = D$ where D is diagonal.

Define the matrix Q as:

$$
Q := P \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R \end{bmatrix} \tag{3.44}
$$

Our claim is that Q is orthogonal and $Q^{-1}AQ$ is diagonal.

(i) We have:

$$
\begin{aligned}
Q^{-1} &= \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^{-1} \end{bmatrix} P^{-1} && \{\text{Reverse order law}\} \\
&= \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T \end{bmatrix} P^T && \{P \text{ and } R \text{ are orthogonal}\}
\end{aligned}
$$

But,

$$Q^T = \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T \end{bmatrix} P^T$$

So,

$$Q^T = Q^{-1}$$

Thus, Q is orthogonal.

(ii) Well, let's compute $Q^{-1}AQ$.

$$\begin{aligned}
Q^{-1}AQ &= Q^T AQ && \{Q \text{ is orthogonal}\} \\
&= \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T \end{bmatrix} P^T AP \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T \end{bmatrix} B \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T \end{bmatrix} \begin{bmatrix} \lambda_1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & C \end{bmatrix} \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T C \end{bmatrix} \begin{bmatrix} 1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1 & 0_{1\times n-1} \\ 0_{n-1\times 1} & R^T CR \end{bmatrix}
\end{aligned}$$

Since $R^T CR$ is diagonal, it follows that $Q^{-1}AQ$ is diagonal. This closes the proof. □

3.4 Covariance and MGF of random variables.

Definition 3.12. If (X, Y) is a random vector, then the covariance of (X, Y) is given by:

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y] \tag{3.45}$$

3.4.1 Expected value of a random matrix.

Suppose our random experiment is modeled by the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We can define the expected value of a random matrix in a component-wise manner.

Suppose that $\mathbf{X}$ is an $m \times n$ matrix of real-valued random variables, whose (i, j) entry is denoted by X_{ij}. Equivalently, $\mathbf{X}$ is a random $m \times n$ matrix. The expected value $\mathbb{E}(\mathbf{X})$ is defined to be the $m \times n$ matrix whose (i, j) entry is $\mathbb{E}X_{ij}$, the expected value of X_{ij}.

Many of the basic properties of expected value of random variables have analogous results for expected values of random matrices/vectors. If $\mathbf{X}$ and $\mathbf{Y}$are random $m \times n$ matrices, the linearity property holds: $\mathbb{E}(\mathbf{X} + \mathbf{Y}) = \mathbb{E}\mathbf{X} + \mathbb{E}\mathbf{Y}$. Similarly, if $\mathbf{X}$ is a $n \times p$ random matrix and $\mathbf{a}$ is a constant $m \times n$ matrix, th of the expectation. $\mathbb{E}[\mathbf{aX}] = \mathbf{a}\mathbb{E}[\mathbf{X}]$.

3.4.2 Covariance Matrices.

Definition 3.13. Suppose that $\mathbf{X}$ is a random vector in $\mathbf{R}^m$ and $\mathbf{Y}$ is a random vector in $\mathbf{R}^n$. The covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$ is the $m \times n$ matrix Cov$(\mathbf{X},\mathbf{Y})$ whose (i,j) entry is Cov(X_i, Y_j).

Definition 3.14. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector in $\mathbf{R}^n$. Then the covariance matrix of $\mathbf{X}$, denoted by Σ is the $n \times n$ matrix, whose (i,j) entry is Cov(X_i, X_j).

Theorem 3.18. *Let (X,Y) be random variables. Cov(X,Y) has the following properties:*

(i) $Cov(X,X) = Var(X)$

Theorem 3.19. *(ii)* $Cov(X,Y) = Cov(Y,X)$

(iii) $Cov(X,c) = 0$

(iv) Scaling property: $Cov(aX,Y) = aCov(X,Y)$

(v) Bi-linearity:

$Cov(aX+bY,Z) = aCov(X,Z) + bCov(Y,Z)$

$Cov(X,cY+dZ) = cCov(X,Y) + dCov(X,Z)$

(vi) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Since $Cov(X,-Y) = -Cov(X,Y)$*, it follows that* $Var(X-Y) = Var(X)+Var(-Y)+2Cov(X,-Y) = Var(X)+Var(Y) - 2Cov(X,Y)$

$Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^n Var(X_i) + \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j)$

Theorem 3.20. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *be a random vector with mean vector* $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ *and* $n \times n$ *covariance matrix* Σ*. Then,* Σ *is positive semi-definite.*

Proof. We have:

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \mathbb{E}(X_1-\mu_1)(X_1-\mu_1) & \mathbb{E}(X_1-\mu_1)(X_2-\mu_2) & \ldots & \mathbb{E}(X_1-\mu_1)(X_n-\mu_n) \\ \mathbb{E}(X_2-\mu_2)(X_1-\mu_1) & \mathbb{E}(X_2-\mu_2)(X_2-\mu_2) & \ldots & \mathbb{E}(X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n)(X_2-\mu_2) & \ldots & (X_n-\mu_n)(X_n-\mu_n) \end{bmatrix} \\
&= \mathbb{E}\left[\begin{bmatrix} X_1-\mu_1 \\ X_2-\mu_2 \\ \vdots \\ X_n-\mu_n \end{bmatrix} \begin{bmatrix} X_1-\mu_1 & X_2-\mu_2 & \ldots & X_n-\mu_n \end{bmatrix}\right] \\
&= \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right]
\end{aligned}$$

Let $\mathbf{a}$ be an arbitrary(not random) vector in $\mathbf{R}^n$. Then,

$$\begin{aligned}
\mathbf{a}'\Sigma\mathbf{a} &= \mathbf{a}'\mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right]\mathbf{a} \\
&= \mathbb{E}\left[\mathbf{a}'(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\mathbf{a}\right] \\
&= \mathbb{E}\left[((\mathbf{X}-\boldsymbol{\mu})'\mathbf{a})'\,((\mathbf{X}-\boldsymbol{\mu})'\mathbf{a})\right] \\
&= \mathbb{E}[(\mathbf{X}-\boldsymbol{\mu})'\mathbf{a}]^2 \\
&\geq 0
\end{aligned}$$

Consequently, Σ is a positive semi-definite matrix. □

Definition 3.15. The MGF of a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is the function on **R** defined by:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

Example 3.5. The MGF of a standard Gaussian random variable given by:

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\left[e^{tZ}\right] \\
&= \int_{-\infty}^{\infty} e^{tz}\phi(z)dz \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz}e^{-z^2/2}dz
\end{aligned}$$

We can complete the square in the exponent as follows:

$$\begin{aligned}
\exp\left(tz - \frac{z^2}{2}\right) &= \exp\left[-\frac{1}{2}\left(z^2 - 2tz + t^2 - t^2\right)\right] \\
&= \exp\left[-\frac{1}{2}(z-t)^2 + \frac{t^2}{2}\right]
\end{aligned}$$

So,

$$\begin{aligned}
M_Z(t) &= \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z-t)^2/2}dz \\
&= \frac{e^{t^2/2}}{\sqrt{2\pi}}\sqrt{2\pi} \\
&= e^{t^2/2}
\end{aligned}$$

Differentiating with respect to t, we have:

$$\begin{aligned}
M_Z'(t) &= te^{t^2/2} \\
M_Z''(t) &= e^{t^2/2} + t^2e^{t^2/2} \\
M_Z^{(3)}(t) &= 3te^{t^2/2} + t^3e^{t^2/2} \\
M_Z^{(4)}(t) &= 3e^{t^2/2} + 6t^2e^{t^2/2} + t^4e^{t^2/2}
\end{aligned}$$

So, the mean of the standard gaussian random variable is $M_Z'(0) = 0$, the second moment and variance of a standard gaussian random variable is $M_Z''(0) = 1$. The skewness of the standard gaussian random variable is $M_Z^{(3)}(0) = 0$, while the kurtosis of a standard gaussian random variable is $M_Z^{(4)}(t) = 3$.

Definition 3.16. (*Joint Moment Generating Function (MGF)*). The joint MGF of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is the function defined on $\mathbf{R}^n$ by:

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[\exp\left(\mathbf{t}^T\mathbf{X}\right)\right] = \mathbb{E}\left[\exp\left(t_1X_1 + t_2X_2 + \ldots + t_nX_n\right)\right] \tag{3.46}$$

The following result will be stated without proof. It will be useful when studying Gaussian vectors.

Proposition 3.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Two random vectors X and Y that have the same moment generating function have the same distribution.*

Example 3.6. Consider (X, Y) a random vector with value in $\mathbf{R}^2$ such that X and Y are IID with standard Gaussian distribution. Then, the joint PDF is:

$$f(x,y) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \times \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{\sqrt{2\pi}} e^{-(x^2+y^2)/2}$$

The moment generating function is obtained by independence:

$$\begin{aligned} M_{(X,Y)}(t_1, t_2) &= \mathbb{E}[e^{t_1 X + t_2 Y}] \\ &= \mathbb{E}\left[e^{t_1 X} \cdot e^{t_2 Y}\right] \\ &= \mathbb{E}\left[e^{t_1 X}\right] \cdot \mathbb{E}\left[e^{t_2 X}\right] \\ &= e^{t_1^2/2} \cdot e^{t_2^2/2} \\ &= e^{(t_1^2+t_2^2)/2} \end{aligned}$$

More generally, we can consider n IID random variables with standard Gaussian distribution. We then have the joint PDF:

$$f(x_1, x_2, \ldots, x_n) = \frac{e^{-(x_1^2+x_2^2+\ldots+x_n^2)/2}}{(2\pi)^{n/2}}$$

In order to work with random vectors, we frequently use the change-of-variables theorem from vector calculus.

Theorem 3.21. *(Change of Variables theorem for double integrals). If $f : \mathbf{R}^2 \to \mathbf{R}$ and $\mathbf{T}$ is a linear transformation such that $D = \mathbf{T}(D^*)$, and $\mathbf{x} = T\mathbf{y}$, then by the change of variables formula, we have:*

$$\int\int_D f(x_1, x_2) dx_1 dx_2 = \int\int_{D^*} f(\mathbf{T}(y_1, y_2)) \left|J(y_1, y_2)\right| dy_1 dy_2$$

where $J(y_1, y_2) = \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}$. In particular, if $T(\mathbf{y}) = A\mathbf{y}$ is a linear transformation which is one-to-one and onto, then it can be easily shown that any parallelogram with unit area in D^* is scaled by a factor $|\det A|$ in D. Hence, $J(y_1, y_2) = \frac{1}{|\det A|}$.

Corollary 3.2. *If X_1, X_2 have the joint density function f, and T is any linear transformation, then the pair $(Y_1, Y_2) = T(X_1, X_2)$ has the density function:*

$$f_{(Y_1,Y_2)}(y_1, y_2) = f(x_1(y_1, y_2), x_2(y_1, y_2)) \left|\frac{\partial(x_1, x_2)}{\partial(y_1, y_2)}\right|$$

Example 3.7. (*Computations with random vectors*). Let (X, Y) be two IID standard Gaussian random variables. We can think of (X, Y) as the random point in $\mathbf{R}^2$ with x-coordinate X and y-coordinate Y.

First off, let's compute the probability that the point (X, Y) is in the unit disc $D = \{(x, y)|x^2 + y^2 = 1\}$. The probability is given by the double integral:

$$\begin{aligned} P((X,Y) \in D) &= \int\int_D \frac{1}{2\pi} e^{-(x^2+y^2)/2} dx dy \\ &= \int_{-1}^{+1} \int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dx dy \end{aligned}$$

We apply the linear transformation:

$$
\begin{aligned}
r &= \sqrt{x^2 + y^2} \\
\tan\theta &= \frac{y}{x}
\end{aligned}
$$

The inverse map is:

$$
\begin{aligned}
x &= r\cos\theta \\
y &= r\sin\theta
\end{aligned}
$$

The Jacobian $\frac{\partial(x,y)}{\partial(r,\theta)}$ is given by:

$$
\begin{aligned}
\frac{\partial(x,y)}{\partial(r,\theta)} &= \det\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} \\
&= \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} \\
&= r(\cos^2\theta + \sin^2\theta) \\
&= r
\end{aligned}
$$

We need to identify the region D^* that T maps in a one-to-one fashion to D. We have:

$$
D^* = \{(\theta, r) | 0 \leq \theta \leq 2\pi, 0 \leq r \leq 1\}
$$

Thus, D^* is a rectangular region. We can write our double integral as:

$$
\begin{aligned}
P((X,Y) \in D) &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^1 e^{-r^2/2} r dr d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi}\int_0^{1/2} e^{-u} du d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} - \left[e^{-u}\right]_0^{1/2} d\theta \\
&= \frac{1}{2\pi}\int_0^{2\pi} (1 - e^{-1/2}) d\theta \\
&= (1 - e^{-1/2})\frac{1}{2\pi}\int_0^{2\pi} d\theta \\
&= (1 - e^{-1/2})
\end{aligned}
$$

We are given that X, Y are IID Gaussian random variables. Consider now the random variable $R = (X^2 + Y^2)$ giving the distance of the point to the origin. Let's compute $\mathbb{E}[R]$. Now, R is a function of the random variables (X, Y). Hence, by LOTUS, we must have:

$$
\begin{aligned}
\mathbb{E}\left[R\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+y^2)^{1/2}f_{(X,Y)}(x,y)dxdy \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+y^2)^{1/2}e^{-(x^2+y^2)/2}dxdy
\end{aligned}
$$

Again by transforming to the polar coordinates, we have:

$$
\begin{aligned}
\mathbb{E}\left[R\right] &= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} re^{-r^2/2}rdrd\theta \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} r^2e^{-r^2/2}drd\theta
\end{aligned}
$$

By the product rule, the inner integral can be simplified as follows:

u	dv
r	$re^{-r^2/2}dr$
1	$-e^{-r^2/2}$

We have:

$$
\begin{aligned}
\int_0^\infty udv &= uv|_0^\infty - \int_0^\infty vdu \\
&= -re^{-r^2/2}\Big|_0^\infty + \int_0^\infty e^{-r^2/2}dr \\
&= 0 + \frac{\sqrt{2\pi}}{2}
\end{aligned}
$$

So, the desired expectation is:

$$
\mathbb{E}[R] = \frac{\sqrt{2\pi}}{2}\cdot\frac{1}{2\pi}\int_0^{2\pi} d\theta = \sqrt{\frac{\pi}{2}}
$$

Consider the joint PDF $f_{(R,\Theta)}(r,\theta)$ of R, the distance to the origin and Θ, the angle made with the positive x-axis. By the change-of-variables theorem, this is:

$$
\begin{aligned}
f_{(R,\Theta)}(r,\theta) &= f_{X,Y}(x(r,\theta),y(r,\theta))\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right| \\
&= \frac{1}{2\pi}re^{-r^2/2}
\end{aligned}
$$

The joint CDF of (R,Θ) is :

$$
\begin{aligned}
F_{(R,\theta)}(r,\theta) &= \frac{1}{2\pi}\int_0^\theta \int_0^r re^{-r^2/2}drd\theta \\
&= \frac{1}{2\pi}\int_0^\theta \int_0^{r^2/2} e^{-u}dud\theta \\
&= \frac{1}{2\pi}\int_0^\theta \left[\frac{e^{-u}}{-1}\right]_0^{r^2/2} d\theta \\
&= \frac{1}{2\pi}\int_0^\theta \left[\frac{e^{-u}}{-1}\right]_0^{r^2/2} d\theta \\
&= \frac{1}{2\pi}\int_0^\theta (1-e^{-r^2/2})d\theta \\
&= \frac{\theta}{2\pi}(1-e^{-r^2/2})
\end{aligned}
$$

In particular, the variables (R, Θ) are independent since the joint PDF is the product of the marginals. Θ is uniformly distributed on $[0, 2\pi]$ and has PDF $f_\Theta(\theta) = \frac{1}{2\pi}$.

3.4.3 The Box-Mueller Method.

The above example gives an interesting method to generate a pair of IID standard Gaussian random variables. This is called the Box-Mueller method. Let U_1 and U_2be two independent uniform random variables on $[0, 1]$. Define the random variables (Z_1, Z_2) as follows:

$$
\begin{aligned}
Z_1 &= \sqrt{-2\log U_1}\cos(2\pi U_2) \\
Z_2 &= \sqrt{-2\log U_1}\sin(2\pi U_2)
\end{aligned}
$$

The CDF of the random variable R defined in the previous section is:

$$F_R(r) = 1 - e^{-r^2/2}$$

The inverse CDF is obtained by expressing r in terms of u:

$$
\begin{aligned}
e^{-r^2/2} &= 1-u \\
\frac{-r^2}{2} &= \log(1-u) \\
r^2 &= -2\log(1-u) \\
r &= \sqrt{-2\log(1-u)}
\end{aligned}
$$

By probability integral transform, we know that if U_1' is a Uniform$[0, 1]$ random variable, then the random variable $F_X^{-1}(U_1')$ has the CDF F_X. By symmetry, $U_1 := 1-U_1'$ is also uniformly distributed on $[0, 1]$. Thus, the random variable $\sqrt{-2\log U_1}$ has the same distribution as R.

The CDF of the random variable Θ defined above is:

$$F_\Theta(\theta) = \frac{\theta}{2\pi}$$

So, if U_2 is a uniform random variable, then the random variable $2\pi U_2$ has the same distribution as Θ in the discussion above.

As seen in the example above, if R and Θ are independent and their marginal CDFs are $F_R(r) = 1 - e^{-r^2/2}$ and $F_\Theta(\theta) = \frac{\theta}{2\pi}$, we know that the random variables defined by $X = R\cos\Theta$ and $Y = R\sin\Theta$ are IID standard normal random variables.

More formally, we are making the transformation $T : (R, \Theta) \mapsto (X, Y)$:

$$T(R, \Theta) = (R\cos\Theta, R\sin\Theta)$$

The inverse map is

$$\begin{aligned} R &= \sqrt{X^2 + Y^2} \\ \Theta &= \arctan\frac{Y}{X} \end{aligned}$$

So, the density function of the pair (X, Y) is given :

$$\begin{aligned} f_{(X,Y)}(x, y) &= f_{(R,\Theta)}(\sqrt{x^2+y^2}, \text{arctan}(y/x)) \cdot \begin{vmatrix} \frac{\partial r}{\partial x} & \frac{\partial r}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} \\ &= \frac{1}{2\pi} r e^{-r^2/2} \begin{vmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix} \\ &= \frac{1}{2\pi}\sqrt{x^2+y^2} \cdot e^{-(x^2+y^2)/2} \cdot \frac{1}{\sqrt{x^2+y^2}} \\ &= \frac{1}{2\pi} e^{-(x^2+y^2)/2} \end{aligned}$$

Hence, X and Y are IID standard Gaussian random variables.

Problem 3.1. Find the PDF of e^{-X} for $X \sim Expo(1)$.

Solution.

By change of variables, $Y = e^{-X}$, $X = -\log Y$. $\frac{\partial x}{\partial y} = -\frac{1}{y}$. The PDF of Y is:

$$\begin{aligned} f_Y(y) &= f_X(-\log y) \cdot \left|\frac{\partial x}{\partial y}\right| \\ &= y \times \frac{1}{y} \\ &= 1 \end{aligned}$$

Problem 3.2. Find the PDF of X^7 for $X \sim \text{Expo}(\lambda)$.

Solution.

By change of variables, $Y = X^7$, $X = Y^{1/7}$. The PDF of Y is:

$$
\begin{aligned}
f_Y(y) &= f_X(y^{1/7}) \cdot \left|\frac{\partial x}{\partial y}\right| \\
&= \lambda e^{-\lambda y^{1/7}} \times \frac{1}{7}\frac{1}{y^{6/7}} \\
&= \frac{\lambda e^{-\lambda y^{1/7}}}{7y^{6/7}}
\end{aligned}
$$

Problem 3.3. Find the PDF of Z^3 for $Z \sim \mathcal{N}(0,1)$.

Solution.

By change of variables, $Y = Z^3$, $Z = Y^{1/3}$. The PDF of Y is:

$$
\begin{aligned}
f_Y(y) &= f_Z(y^{1/3}) \cdot \left|\frac{\partial x}{\partial y}\right| \\
&= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{y^{2/3}}{2}\right] \times \frac{1}{3}y^{-2/3}
\end{aligned}
$$

3.5 Gaussian Vectors.

Definition 3.17. A n-dimensional random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is said to be jointly Gaussian if and only if for all real vectors $\mathbf{t} = (t_1, \ldots, t_n)$, the linear combination $\mathbf{t}^T\mathbf{X} = t_1X_1 + t_2X_2 + \ldots + t_nX_n$ of $(X_1, X_2, \ldots, X_n)$ is a Gaussian random variable.

As a simple consequence of the above definition, if $(X_1, \ldots, X_n)$ is Gaussian, then setting $t_i = 1$ and $t_j = 0$ for all $i \neq j$, we have that each X_i is also Gaussian.

An equivalent definition can also be stated in terms of the joint MGF since an MGF uniquely characterizes the distribution of a random variable. Before introducing the second definition, we first make two important observations about the mean and variance of a linear combination of random variables.

First, the mean of a linear combination of random variables is:

$$
\mathbb{E}[a_1X_1 + a_2X_2 + \ldots + a_nX_n] = a_1\mathbb{E}X_1 + \ldots + a_n\mathbb{E}X_n = \mathbf{a}^T\mathbb{E}\mathbf{X}
$$

where $\mathbb{E}\mathbf{X}$ is the mean vector. The variance is obtained with a short calculation using the linearity of expectations:

$$
\begin{aligned}
Var(a_1X_1 + \ldots + a_nX_n) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_jCov(X_i, X_j) \\
&= \mathbf{a}^T\Sigma\mathbf{a}
\end{aligned}
$$

where Σ is the covariance matrix of $\mathbf{X}$.

Proposition 3.4. *A random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is Gaussian if and only if the moment generating function of X is:*

$$
\mathbb{E}\left[\exp\left\{\mathbf{t}^T\mathbf{X}\right\}\right] = \exp\left[\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right] \tag{3.47}
$$

where $\boldsymbol{\mu}$ is the mean vector and Σ is the covariance matrix of $\mathbf{X}$.

Proof. By the definition of joint MGF:

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[\exp\left\{\mathbf{t}^T\mathbf{X}\right\}\right] = \mathbb{E}\left[\exp\left\{t_1X_1 + \ldots + t_nX_n\right\}\right] \tag{3.48}$$

But, we know that $t_1X_1 + \ldots + t_nX_n$ is a Gaussian random variable with mean $\mu = \mathbf{t}^T\boldsymbol{\mu}$ and variance $\sigma^2 = \mathbf{t}^T\Sigma\mathbf{t}$. The MGF of a univariate Gaussian random variable is :

$$M_X(s) = \mathbb{E}[\exp(sX)] = \exp(\mu s + \frac{\sigma^2 s^2}{2})$$

At $s = 1$, we have:

$$M_X(1) = \mathbb{E}[\exp(X)] = \exp(\mu + \frac{\sigma^2}{2}) \tag{3.49}$$

Thus, if $X = t_1X_1 + \ldots + t_nX_n$ then it follows that:

$$\mathbb{E}\left[\exp\left(t_1X_1 + \ldots + t_nX_n\right)\right] = \exp\left[\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right]$$

But from (3.48), this is the joint MGF of $\mathbf{X}$. This closes the proof. □

Proposition 3.5. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a Gaussian vector. Then, the covariance matrix is diagonal, if and only if the random variables are independent.*

Proof. ($\Longrightarrow$) direction.

We are given that the covariance matrix is diagonal. Our proposition is that the random variables are independent.

Remember, that if X_1 and X_2 are independent random variables, $Cov(X_1, X_2) = 0$. But, the converse is not true. We use the MGF of the random vector $\mathbf{X}$, to prove this claim.

We have:

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left[\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right]$$

Since $\Sigma = Diag(\sigma_1^2, \ldots, \sigma_n^2)$, we can express:

$$\mathbf{t}^T\Sigma\mathbf{t} = t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + \ldots + t_n^2\sigma_n^2$$

So:

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \exp\left[t_1\mu_1 + \frac{\sigma_1^2 t_1^2}{2}\right]\cdots\exp\left[t_n\mu_n + \frac{\sigma_n^2 t_n^2}{2}\right] \\ &= M_{X_1}(t_1)\cdots M_{X_n}(t_n) \end{aligned}$$

Consequently, the MGF can factored into a product of the MGFs of $X_1, \ldots, X_n$. Thus, $X_1, X_2, \ldots, X_n$ are independent random variables.

($\Longleftarrow$) direction.

This direction is trivial. We are given that the random variables are independent. Then, $Cov(X_i, X_j) = 0$ for all $i \neq j$. So, the covariance matrix is diagonal. □

Before writing the joint PDF of a Gaussian vector in terms of the mean vector and the covariance matrix, we need to introduce the important notion of degenerate vector. We say a Gaussian vector is *degenerate* if its covariance matrix Σ is singular, $\det \Sigma = 0$.

Example 3.8. Consider (Z_1, Z_2, Z_3) IID standard Gaussian random variables. We define $X = Z_1 + Z_2 + Z_3$, $Y = Z_1 + Z_2$ and $W = Z_3$. Clearly, (X, Y, W) is a Gaussian vector. It has 0 mean and covariance:

$$\begin{bmatrix} 3 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

It is easy to check that $\det \Sigma = 3 \cdot 2 - 2 \cdot 2 + 1 \cdot (-2) = 0$. Thus, (X, Y, W) is a degenerate Gaussian vector.

The above example is helpful to illustrate the notion. Note that we have the linear relation $X - Y - W = 0$ between the random variables. Therefore, the random variables are linearly dependent. In other words, one vector is redundant, say X, in the sense that its value can be recovered from others for any outcome. The relation between degeneracy and linear dependence is general.

Lemma 3.6. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *be a Gaussian vector. Then,* X *is degenerate if and only if the coordinates are linearly dependent. That is, there exists* $c_1, c_2, \ldots, c_n$*, not all zero, such that* $c_1 X_1 + c_2 X_2 + \ldots + c_n X_n = 0$ *with probability one.*

Proof. ($\Longrightarrow$) direction.

We are given that the vector X is degenerate. This implies that $\det \Sigma = 0$ and the columns of Σ are linearly dependent. Σ is non-singular.

TODO. $\square$

We are now ready to state the form of the PDF of Gaussian vectors.

Definition 3.18. (Joint PDF of Gaussian vectors). Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a non-degenerate Gaussian vector with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, written $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the joint density of X is given by the PDF:

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n |\det \Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \tag{3.50}$$

where $\mathbf{x} \in \mathbf{R}^n$ and Σ is PSD (Positive symmetric definite).

Example 3.9. Consider a Gaussian vector (X_1, X_2) of mean 0 and covariance matrix $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. The inverse of Σ can be found out as follows.

We consider the augmented matrix $[\Sigma | I]$.

$$\begin{bmatrix} 2 & 1 & | & 1 & 0 \\ 1 & 2 & | & 0 & 1 \end{bmatrix}$$

Performing $R_1 = 1/2 R_1$, the above system is row equivalent to:

$$\begin{bmatrix} 1 & \frac{1}{2} & | & \frac{1}{2} & 0 \\ 1 & 2 & | & 0 & 1 \end{bmatrix}$$

Performing $R_2 = R_2 - R_1$, the above system is row equivalent to:

$$\begin{bmatrix} 1 & \frac{1}{2} & | & \frac{1}{2} & 0 \\ 0 & \frac{3}{2} & | & -\frac{1}{2} & 1 \end{bmatrix}$$

Peforming $R_2 = \frac{2}{3}R_2$, the above system is row equivalent to:

$$\left[\begin{array}{cc|cc} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & -\frac{1}{3} & \frac{2}{3} \end{array}\right]$$

Performing $R_1 = R_1 - \frac{1}{2}R_2$, the above system is row equivalent to

$$\left[\begin{array}{cc|cc} 1 & 0 & \frac{2}{3} & -\frac{1}{3} \\ 0 & 1 & -\frac{1}{3} & \frac{2}{3} \end{array}\right]$$

So, $\Sigma^{-1} = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$ and $\det C = 3$. By doing matrix operations, the joint PDF of (X_1, X_2) is:

$$\begin{aligned} f_{(X_1,X_2)}(x_1,x_2) &= \frac{1}{\sqrt{(2\pi)^2 \cdot 3}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{1}{2}\begin{bmatrix} 2/3x_1 - 1/3x_2 & -1/3x_1 + 2/3x_2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi\sqrt{3}} \exp\left(-\frac{1}{3}x_1^2 + \frac{1}{3}x_1x_2 - \frac{1}{3}x_2^2\right) \end{aligned}$$

We will not prove proposition (3.18) yet. Instead, we will take a short detour and derive it from a powerful decomposition of Gaussian vectors as a linear combination of IID Gaussians. The decomposition is the generalization of making a random variable *standard*. Suppose X is Gaussian with mean 0 and variance σ^2.Then, we can write it as $X = \sigma Z$, where Z is a standard normal random variable. (This makes sense even when X is degenerate that is $\sigma^2 = 0$). If $\sigma^2 \neq 0$, then we can reverse the relation to get:

$$Z = \frac{X}{\sigma}$$

We generalize this procedure to Gaussian vectors.

Proposition 3.6. *(Decomposition into IID). Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *be a Gaussian vector of mean* 0 *and* $n \times n$ *covariance matrix* C. *If* $\mathbf{X}$ *is non-degenerate, there exists* n *IID gaussian random variables* $Z_1, Z_2, \ldots, Z_n$ *and an invertible* $n \times n$ *matrix* A *such that:*

$$\mathbf{X} = AZ, \quad Z = A^{-1}\mathbf{X} \tag{3.51}$$

The choice of Zs and thus the matrix A is generally not unique as the following simple example shows:

Example 3.10. Consider the Gaussian vector (X_1, X_2) given by:

$$\begin{aligned} X_1 &= Z_1 + Z_2 \\ X_2 &= Z_1 - Z_2 \end{aligned}$$

where Z_1, Z_2 are IID standard gaussians.

The matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. The covariance matrix of $\mathbf{X}$ is $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$. Since, the covariance matrix is diagonal, by proposition (3.5), the random variables X_1 and X_2 are independent.

Another choice of decomposition is simply $W_1 = X_1/\sqrt{2}$ and $W_2 = X_2/\sqrt{2}$.

Proof of proposition 3.6.

Proof. This is done using the same Gram-Schmidt procedure as for $\mathbf{R}^n$. The idea is to take the variables one-by-one and subtract the components in the directions of the previous ones using covariance. The lemma (3.6) ensures that no random variables are linear combinations of the others.

To start, we take $Z_1 = \frac{X_1}{\sqrt{C_{11}}}$. Clearly, Z_1 is a standard normal random variable.

Then, we define Z_2' as:

$$Z_2' = X_2 - \mathbb{E}\left[X_2 Z_1\right] Z_1$$

And let

$$Z_2 = \frac{Z_2'}{\sqrt{Var(Z_2')}}$$

Firstly, since (X_1, X_2) is Gaussian, any linear combination of X_1 and X_2 must be a Gaussian random variable. It follows that Z_2' is also a Gaussian random variable. Moreover, $\mathbb{E}[Z_2] = \frac{1}{\sqrt{Var(Z_2')}} \cdot \mathbb{E}[X_2] = 0$ and $Var(Z_2) = 1$. Further:

$$\begin{aligned}
Cov(Z_1, Z_2') &= Cov(Z_1, X_2 - \mathbb{E}\left[X_2 Z_1\right] Z_1) \\
&= Cov(Z_1, X_2) - \mathbb{E}\left[X_2 Z_1\right] Var(Z_1) \\
&= \mathbb{E}\left[X_2 Z_1\right] - \mathbb{E}\left[X_2 Z_1\right](1) \\
&= 0
\end{aligned}$$

Thus, Z_2 is independent Gaussian with mean 0 and variance 1.

In the same way, we take Z_3 to be:

$$Z_3' = X_3 - \mathbb{E}(X_3 Z_2) Z_2 - E(X_3 Z_1) Z_1$$

and

$$Z_3 = \frac{Z_3'}{\sqrt{Var(Z_3')}}$$

Again, its easy to check that Z_3' is independent of Z_2 and Z_1. As above, we define Z_3 to be Z_3' divided by the square root of variance.

Also, Z_2 is a linear combination of $X_2$and Z_1 and in turn, Z_1 is a linear combination of X_1. So, effectively, Z_3' is a linear combination of X_1, X_2, X_3. Since (X_1, X_2, X_3) is a Gaussian vector, every linear combination is Gaussian. So, Z_3' is Gaussian.

This procedure is carried on until we run out variables. Not that since C is non-degenerate, none of the variances of the Z_i' will be zero, and therefore they can be standardized. ☐

The covariance matrix C of the Gaussian vector $\mathbf{X}$ with mean vector $\boldsymbol{\mu} = \mathbf{0}$ can be written in terms of A. Write $A = (a_{ij})$ for the (i, j)th entry of the matrix A. By the relation $X = AZ$, we have:

$$
\begin{aligned}
Cov(X_i, X_j) &= \mathbb{E}(X_i X_j) \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{n} a_{ik} Z_k\right)\left(\sum_{l=1}^{n} a_{jl} Z_l\right)\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{n} a_{ik} a_{jk} Z_k^2 + \sum_{k \neq l} a_{ik} a_{jl} Z_k Z_l\right]
\end{aligned}
$$

Now, we know that:

$$
\mathbb{E}(Z_k \cdot Z_l) = \begin{cases} 1 & \text{if } k = l \\ 0 & \text{otherwise} \end{cases}
$$

So, the expectation simplifies to:

$$
\begin{aligned}
Cov(X_i, X_j) &= \mathbb{E}\left[\sum_{k=1}^{n} a_{ik} a_{jk} Z_k^2\right] \\
&= \sum_{k=1}^{n} a_{ik} a_{jk} \\
&= (AA^T)_{ij}
\end{aligned}
$$

Thus, we have:

$$
C = AA^T
$$

Thus, the covariance matrix C of a Gaussian vector $\mathbf{X}$ admits a Cholesky Factorization of the form, $C = AA^T$ and therefore, C is SPD(symmetric positive definite). For applications and numerical simulations, it is important to get the matrix A from the covariance matrix C. This decomposition is an exact analogue of the decomposition of a vector in $\mathbf{R}^3$ written as a sum of orthonormal basis vectors. In particular, the condition of being non-degenerate is equivalent to linear independence.

Proof of proposition 3.18

Proof. Without the loss of generality assume that $\boldsymbol{\mu} = (0, \ldots, 0)$. Otherwise, we just need to subtract it from $\mathbf{X}$. We use the decomposition in proposition **(3.6).** First note that, since $C = AA^T$, the determinant of C is:

$$
C = AA^T
$$

so the determinant of C is:

$$
\begin{aligned}
\det C &= \det A \cdot \det A^T \\
&= \det A \cdot \det A \\
&= (\det A)^2
\end{aligned}
$$

In particular, since $\mathbf{X}$ is non-degenerate, we have that $\det C \neq 0$, so $\det A \neq 0$. Thus, A is invertible. We also have by the decomposition that there exist IID Gausian random variables Z such that $\mathbf{X} = AZ$. Now, the event $\{\mathbf{X} \in B\} = \{AZ \in B\} = \{Z \in A^{-1}B\}$. So,

$$P(X \in B) = P(Z \in A^{-1}B)$$

But we know the joint density of n IID standard normal random variables $Z_1, Z_2, \ldots, Z_n$. Consequently, we have:

$$P(X \in B) = \int \cdots \int_{A^{-1}B} \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\mathbf{z}^T\mathbf{z}\right] dz_1 \cdots dz_n$$

because $\mathbf{z} = (z_1, \ldots, z_n)$ and $\mathbf{z}^T\mathbf{z} = z_1^2 + \ldots + z_n^2$. It remains to do the change of variable $\mathbf{x} = A\mathbf{z}$.

Let us define the map T as:

$$\mathbf{x} = A\mathbf{z}$$

Then, the inverse map T^{-1} is:

$$\mathbf{z} = A^{-1}\mathbf{x}$$

Since $\mathbf{X}$ is non-degenerate, A^{-1} exists and the right-hand side vector is well-defined. The Jacobian $\frac{\partial(z_1,\ldots,z_n)}{\partial(x_1,\ldots,x_n)}$ is:

$$\frac{\partial(z_1, \ldots, z_n)}{\partial(x_1, \ldots, x_n)} = |\det(A^{-1})| = \frac{1}{|\det A|} = \frac{1}{\sqrt{|\det C|}}$$

Moreover, $\mathbf{z} \in A^{-1}B$ is equivalent to $\mathbf{x} \in B$. Further, $\mathbf{z}^T\mathbf{z} = (A^{-1}\mathbf{x})^T(A^{-1}\mathbf{x}) = \mathbf{x}^T(A^{-1})^T(A^{-1})\mathbf{x}$. Now, note that if $C = AA^T$, by the reverse order law, $C^{-1} = (A^T)^{-1}(A^{-1}) = (A^{-1})^T(A^{-1})$. Consequently, $\mathbf{z}^T\mathbf{z} = \mathbf{x}^T C^{-1}\mathbf{x}$.

$$P(X \in B) = \int \cdots \int_{B} \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\left[-\frac{1}{2}\mathbf{x}^T C^{-1}\mathbf{x}\right] dx_1 \cdots dx_n$$

Consequently, the joint density function of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\left[-\frac{1}{2}\mathbf{x}^T C^{-1}\mathbf{x}\right]$$

If $\mathbf{X}$ has a non-zero mean vector $\boldsymbol{\mu}$, then $\mathbf{X}' = \mathbf{X} - \boldsymbol{\mu}$ has a mean vector zero. Thus, the joint density function becomes:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\det C|}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T C^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

$\square$

We now explore three ways to find the matrix A in the decomposition of Gaussian vectors of proposition (3.6). We proceed by example:

Example 3.11. (Cholesky by Gram-Schmidt). This is the method suggested by the proof of proposition **3.6.** It suffices to successively go through the X's by subtracting the projection of a given X_i onto the previous random variables. Consider the random vector $\mathbf{X} = (X_1, X_2)$ with mean 0 and covariance matrix :

$$C = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

It is easy to check that X is non-degenerate, $\det C = 3$. Take:

$$Z_1 = \frac{X_1}{\sqrt{2}}$$

This is obviously a standard Gaussian random variable. For Z_2, first consider:

$$Z_2' = X_2 - \mathbb{E}(X_2 Z_1) Z_1$$

It is straightforward to check that Z_1 and Z_2' are jointly Gaussian. Z_2' is a linear combination of Z_1 and X_2, so Z_2' is Gaussian. Since all linear combinations of Z_2' and Z_1 are Gaussian, by definition, (Z_1, Z_2') is jointly Gaussian. They are also independent, because:

$$\begin{aligned}
\mathbb{E}(Z_1 Z_2') &= \mathbb{E}[Z_1(X_2 - \mathbb{E}(X_2 Z_1)Z_1)] \\
&= \mathbb{E}[Z_1 X_2] - \mathbb{E}(X_2 Z_1)\mathbb{E}(Z_1^2) \\
&= \mathbb{E}[Z_1 X_2] - \mathbb{E}(X_2 Z_1) \cdot 1 \\
&= 0
\end{aligned}$$

Note that:

$$\begin{aligned}
Z_2' &= X_2 - \mathbb{E}[X_2 Z_1] Z_1 \\
&= X_2 - \mathbb{E}\left[X_2 \frac{X_1}{\sqrt{2}}\right] \frac{X_1}{\sqrt{2}} \\
&= X_2 - \frac{1}{2}\mathbb{E}[X_1 X_2] X_1 \\
&= X_2 - \frac{1}{2} X_1
\end{aligned}$$

In particular, we have by linearity of expectations:

$$\begin{aligned}
\mathbb{E}[(Z_2')^2] &= \mathbb{E}[X_2^2 - X_1 X_2 + \frac{1}{4} X_1^2] \\
&= \mathbb{E}(X_2^2) - \mathbb{E}(X_1 X_2) + \frac{1}{4}\mathbb{E}(X_1^2) \\
&= 2 - 1 + \frac{1}{4} \cdot 2 \\
&= \frac{3}{2}
\end{aligned}$$

To get a random variable of variance 1, that is a multiple of Z_2', we take $Z_2 = \frac{Z_2'}{\sqrt{3/2}} = \sqrt{\frac{2}{3}} Z_2 = \sqrt{\frac{2}{3}} X_2 - \frac{1}{\sqrt{6}} X_1$. Altogether, we get :

$$
\begin{aligned}
Z_1 &= \frac{X_1}{\sqrt{2}} \\
Z_2 &= -\frac{1}{\sqrt{6}} X_1 + \sqrt{\frac{2}{3}} X_2
\end{aligned}
$$

We thus constructed two standard IID Gaussians from (X_1, X_2). In particular we have:

$$
A^{-1} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & \sqrt{\frac{2}{3}} \end{bmatrix}, \quad A = \begin{bmatrix} \sqrt{2} & 0 \\ \frac{1}{\sqrt{2}} & \sqrt{\frac{3}{2}} \end{bmatrix}
$$

We can check that :

$$
\begin{aligned}
AA^T &= \begin{bmatrix} \sqrt{2} & 0 \\ \frac{1}{\sqrt{2}} & \sqrt{\frac{3}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{2} & \frac{1}{\sqrt{2}} \\ 0 & \sqrt{\frac{3}{2}} \end{bmatrix} \\
&= \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \\
&= C
\end{aligned}
$$

The probability of the event $P(X_1 > 2, X_2 < 3)$ can be computed as follows:

$$
P(X_1 > 2, X_2 < 3) = P\left(\sqrt{2} Z_1 > 2, \frac{1}{\sqrt{2}} Z_1 + \sqrt{\frac{3}{2}} Z_2 < 3\right)
$$

Example 3.12. (Cholesky by solving a system of equations). Consider the same example as above. Write $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then the relation $C = AA^T$ yields:

$$
\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & c^2 + d^2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}
$$

and so we have the three equations:

$$
\begin{aligned}
a^2 + b^2 &= 2 \\
ac + bd &= 1 \\
c^2 + d^2 &= 2
\end{aligned}
$$

There are several solutions. One of them is $a = \sqrt{2}$, $b = 0$, $c = \frac{1}{\sqrt{2}}$ and $d = \sqrt{\frac{3}{2}}$.

Example 3.13. (*Cholesky by diagonalization*) This method takes advantage of the symmetry of the covariance matrix. From the spectral theorem, we know that, if C is a symmetric matrix, it is diagonalizable, it admits a factorization of the form $Q\Lambda Q^{-1}$ where Q is an orthogonal matrix. The entries of Λ are the eigenvalues of C. Furthermore, the eigenvectors are orthogonal to each other.

Since $C = AA^T$, we get:

$$\begin{aligned} C &= Q\Lambda Q^T \\ \iff AA^T &= Q\Lambda Q^T \end{aligned}$$

It suffices to take:

$$A = Q\Lambda^{1/2}$$

where Q is the matrix with the columns given by the eigenvectors of C and $\Lambda^{1/2}$ is the diagonal matrix with the square root of the eigenvalues on the diagonal.

Example 3.14. (IID Decomposition). Let $X = (X_1, X_2, X_3)$ be a Gaussian vector with mean 0 and covariance matrix:

$$C = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

Let's find a matrix A such that $X = AZ$ for $Z = (Z_1, Z_2, Z_3)$ IID standard gaussians. The vector is not degenerate since $\det C = 1 \cdot (2-1) = 1$. If we do a Gram-Schmidt procedure, we get:

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= (X_2 - X_1) \\ Z_3 &= X_3 - (X_2 - X_1) - X_1 \\ &= X_3 - X_2 \end{aligned}$$

Consequently, $X_1 = Z_1$, $X_2 = Z_1 + Z_2$ and $X_3 = Z_1 + Z_2 + Z_3$. So, the matrix A is:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

As we will see in the next section, this random vector corresponds to the position of the Brownian motion at time 1, 2 and 3.

3.6 Gaussian Processes.

In general, a *stochastic process* is an infinite collection of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The collection can be countable or uncountable. We are mostly interested in the case, where the variables are indexed by time; for example

$$X = (X_t, t \in \mathcal{J})$$

where $\mathcal{J}$ can be a finite set, **N** or some uncountable set such as the closed interval $[0, T]$ or $[0, \infty)$.

In the case where $\mathcal{J} = [0, \infty)$ or $[0, T]$, the realization of the process $X(\omega)$ can be thought of as a function of time for each outcome ω. This function $t \mapsto X_t(\omega)$ is sometimes called a *path* or the *trajectory* of the process. With this in mind, we can think of the process $(X)_{t\geq 0}$ as a function-valued random variable as each outcome ω produces a function.

(a) For each t, $X(t, \cdot)$ is a random variable.

(b) For each ω, $X(\cdot, \omega)$ is a *function* (called a *sample path*)

For convenience, the random variable $X(t, \cdot)$ will be written as $X(t)$ or X_t. Thus a stochastic process $X(t, \omega)$ can also be expressed as $(X(t))_{t\geq 0}$ or simply $X(t)$.

How can we compute the probabilities for a stochastic process? In other words, what object captures it's distribution? The most common way (there are others) is to use finite dimensional distributions. The idea here is to describe the probabilities related to any finite set of time. More precisely, the finite-dimensional distributions are given by:

$$\mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2, \ldots, X_{t_n} \in B_n)$$

for any $n \in \mathbf{N}$, any choice of $t_1, \ldots, t_n \in \mathcal{J}$, and any events $B_1, \ldots, B_n$ in **R**. Of course, for any fixed choice of t's $(X_{t_1}, \ldots, X_{t_n})$ is a random vector as seen in the previous section. The fact that we can control the probabilities for the whole random function comes from the fact that we have the distributions of these vectors for any n and any choice of t's.

Some important types of stochastic processes include Markov processes, martingales and Gaussian processes. We will encounter them along the way. Let's start with Gaussian processes.

Definition 3.19. A *Gaussian process* $(X_t)_{t\geq 0}$ is a stochastic process whose finite dimensional distributions are jointly Gaussian. In other words, for any $n \in \mathbf{N}$ and any choice of $t_1 < \ldots < t_n$ we have that $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ is a Gaussian vector. In particular, its distribution is defined by the mean function $m(t) = \mathbb{E}(X_t)$ and the covariance function $C(s, t) = Cov(X_t, X_s)$.

As before, linear combinations of Gaussian processes remain Gaussian.

Lemma 3.7. *Let $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$ be m Gaussian processes on $[0, \infty)$ defined on the same probability space. Then, any process constructed by taking linear combinations is also a Gaussian process:*

$$a_1 X^{(1)} + \ldots + a_m X^{(m)} = \left(a_1 X_t^{(1)} + \ldots + a_m X_t^{(m)}, t \geq 0\right)$$

Proof. It suffices to take a finite set of times $t_1 < t_2 < \ldots < t_n$. Consider the vector:

$$\mathbf{Y} = \begin{bmatrix} a_1 X_{t_1}^{(1)} + \ldots + a_m X_{t_1}^{(m)} \\ a_1 X_{t_2}^{(1)} + \ldots + a_m X_{t_2}^{(m)} \\ \vdots \\ a_1 X_{t_n}^{(1)} + \ldots + a_m X_{t_n}^{(m)} \end{bmatrix}$$

Consider any linear combination of these random variables:

$$\begin{aligned} \beta_1 Y_1 + \ldots \beta_n Y_n &= \beta_1 \sum_{j=1}^{m} a_j X_{t_1}^{(j)} + \ldots + \beta_n \sum_{j=1}^{m} a_j X_{t_n}^{(j)} \\ &= a_1 \sum_{i=1}^{n} \beta_i X_{t_i}^{(1)} + \ldots + a_n \sum_{i=1}^{n} \beta_i X_{t_i}^{(m)} \end{aligned}$$

Since $(X_{t_1}^{(j)}, \ldots, X_{t_n}^{(j)})$ is a Gaussian vector, any linear combination $\sum \beta_i X_{t_i}^{(j)}$ is Gaussian. Moreover, since $\square$

The most important example of a Gaussian process is a Brownian motion.

Definition 3.20. (*Standard Brownian motion or Wiener process*). A stochastic process $B(t,\omega)$ is called a Brownian motion if it satisfies the following conditions:

- $\mathbb{P}\{\omega : B(0,\omega) = 0\} = 1$.
- For any $0 \leq s < t$, the random variable $B(t) - B(s)$ is normally distributed with mean 0 and variance $t - s$. That is for any $a < b$:

$$\mathbb{P}\{a \leq B(t) - B(s) \leq b\} = \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b e^{-\frac{x^2}{2(t-s)}} dx$$

- $B(t,\omega)$ has independent increments, i.e. for any $0 \leq t_1 < t_2 < \ldots < t_n$, the random variables :

$$B(t_1), B(t_2) - B(t_1), B(t_3) - B(t_2), \ldots, B(t_n) - B(t_{n-1})$$

  are independent.
- Almost all sample paths of $B(t,\omega)$ are continuous functions, that is,

$$P(\omega \quad | \quad B(\cdot,\omega) \text{ is continuous}) = 1$$

Proposition 3.7. *The standard Brownian motion* $(B_t, t \geq 0)$ *is a Gaussian process.*

Proof.

Consider any finite set of times $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$. Our claim is that the vector $(B_{t_1}, \ldots, B_{t_n})$ is Gaussian.

First consider $n = 1$. $B(t_1)$ is a Gaussian random variable centered at 0. Any scalar multiple $\alpha_1 B_{t_1}$ is also Gaussian.

Next, consider $n = 2$. We have:

$$\alpha_1 B_{t_1} + \alpha_2 B_{t_2} = (\alpha_1 + \alpha_2) B_{t_1} + \alpha_2 (B_{t_2} - B_{t_1})$$

Now, B_{t_1} is a Gaussian random variable and it is $\mathcal{F}_{t_1}$ measurable. $(B_{t_2} - B_{t_1})$ is also Gaussian and independent of $\mathcal{F}_{t_1}$. Hence, $\alpha_1 B_{t_1} + \alpha_2 B_{t_2}$ is Gaussian.

Assume that $\alpha_1 B_{t_1} + \ldots + \alpha_k B_{t_k}$ is Gaussian for all $\alpha_1, \ldots, \alpha_k \in \mathbf{R}$. Our claim is that $\alpha_1 B_{t_1} + \ldots + \alpha_k B_{t_k} + \alpha_{k+1} B_{t_{k+1}}$ is Gaussian.

We have:

$$\alpha_1 B_{t_1} + \ldots + \alpha_k B_{t_k} + \alpha_{k+1} B_{t_{k+1}} = (\alpha_1 B_{t_1} + \ldots + (\alpha_k + \alpha_{k+1}) B_{t_k}) + \alpha_{k+1} (B_{t_{k+1}} - B_{t_k})$$

Now, from our inductive assumption $\alpha_1 B_{t_1} + \ldots + (\alpha_k + \alpha_{k+1}) B_{t_k}$ is Gaussian and $\mathcal{F}_{t_k}$ measurable. Also, $(B_{t_{k+1}} - B_{t_k})$ is Gaussian and independent of $\mathcal{F}_{t_k}$.

Example 3.15. (Sampling a Gaussian process using Cholesky decomposition). The IID decomposition of proposition 3.6 is useful for generating a sample of the Gaussian process. $(X_t)_{t \in [0,T]}$. First, we need to fix the discretization or step-size. Take for example, a step size of 0.01, meaning we approximate the process by evaluating the position at every 0.01. This is given by the Gaussian vector:

$$(X_{\frac{j}{100}}, j = 1, 2, 3, \ldots, 100T)$$

This Gaussian vector has covariance matrix C and a matrix A from the IID decomposition. Note that, we start with the vector at 0.01 and not 0. This is because in some cases (like the standard Brownian motion) the value at time 0 is 0. Including it in the covariance matrix would result in a degenerate covariance matrix. You can always add position 0 at time 0 after performing the cholesky decomposition. It then suffices to sample $100T$ IID standard Gaussian random variable $Z = (Z_1, Z_2, \dots, Z_{100T})$ and to apply the deterministic matrix A to the sample vector to get :

$$(X_{\frac{j}{100}}, j = 1, 2, \dots, 100T) = AZ$$

Example 3.16. Simulating Brownian Motion. The goal of this project is to simulate 100 paths of Brownian motion on $[0, 1]$ using a step-size of 0.01 using the Cholesky decomposition.

(a) Construct the covariance matrix of $(B_{j/100})_{1 \le j \le 100}$ using a for-loop. Recall that for a Brownian motion $C(s, t) = s \wedge t$ with mean 0.

(b) The command `numpy.linalg.cholesky` in Python gives the Cholesky decomposition of the covariance matrix C. Use this to find the matrix A.

(c) Define a function whose output is a sample of N standard Gaussian random variables and whose input is N.

(d) Use the above to plot $n = 100$ paths of the Brownian motion on $[0, 1]$ with a step size of 0.01. Do not forget B_0!

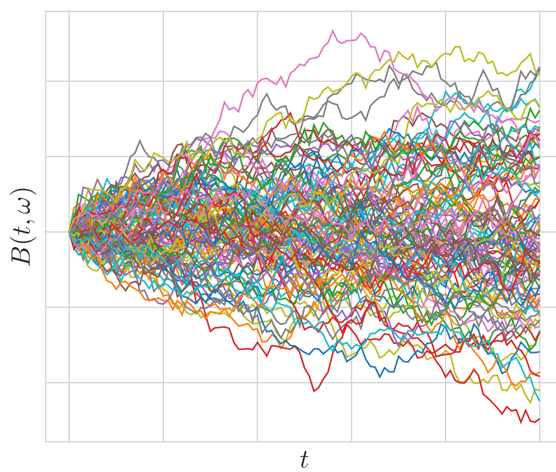
Solution.

Listing 3: Generating 100 paths of a standard brownian motion

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

sns.set_style("whitegrid")

# A generator for N standard gaussian random variables
def standardNormalGenerator(N):
    return np.random.standard_normal(N)

# Produces 1 sample (path) of a gaussian process
# N : Number of time-steps
# A : The transformation that maps IID gaussians (Z_1,Z_2,...,Z_N) to a gaussian vector (X_1,X_2
    ,...,X_N)
# with covariance matrix C = AA'
def sampleGaussianProcess(A,N):

    Z = standardNormalGenerator(N)
    X = np.matmul(A,Z)
    return X

# Produces `numOfPaths` paths of a standard brownian motion
# N : Number of time-steps, 1/N : step-size
def standardBrownianMotion(numOfPaths,N):
    C = np.zeros((N,N))

    for i in range(N):
        for j in range(N):
            s = (i+1)/N
            t = (j+1)/N

            C[i][j] = np.min([s,t])

    A = np.linalg.cholesky(C)

    B = []
    for i in range(numOfPaths):
        X = sampleGaussianProcess(A,N)
        X = np.concatenate(([0], X), axis=0)
```

```
        B.append(X)

    return B

if __name__ == "__main__":

    T = 1.0
    N = 100

    C = covarMatrix(N)
    B = standardBrownianMotion(numOfPaths=100,covarianceMatrix=C,N=100)

    plt.xlabel(r'$t$')
    plt.ylabel(r'$B(t,\omega)$')
    plt.grid(True)
    plt.title(r'$100$ sample paths of a standard brownian motion')

    t = np.linspace(start=0,stop=1.0,num=101)
    for n in range(100):
        plt.plot(t,B[n])


    plt.show()
```

100 sample paths of a standard brownian motion



Example 3.17. (*Brownian motion with a drift.*) For $\sigma > 0$ (called the volatility or diffusion coefficient) and $\mu \in \mathbf{R}$ (called the drift), we define the process:

$$X_t = \sigma B_t + \mu t$$

where $(B_t)_{t\geq 0}$ is a standard brownian motion. This is a Gaussian process, because it is a linear transformation of a brownian motion, which is itself a Gaussian process by lemma (3.7).

A straightfoward computation shows that:

$$\begin{aligned}\mathbb{E}[X_t] &= \sigma\mathbb{E}[B_t] + \mathbb{E}[\mu t] \\ &= \mu t\end{aligned}$$

and if $0 \leq s \leq t$,

$$
\begin{aligned}
\mathbb{E}[X_s X_t] &= \mathbb{E}[(\sigma B_s + \mu s)(\sigma B_t + \mu t)] \\
&= \mathbb{E}[\sigma^2 B_s B_t + \mu t B_s + \mu s \sigma B_t + \mu^2 st] \\
&= \sigma^2 s + \mu^2 st
\end{aligned}
$$

so,

$$Cov(X_s, X_t) = \sigma^2 s$$

Listing 4: Brownian motion with a drift

```
# Given a standard brownian motion, this function produces a
# brownian motion with drift = mu and diffusion coefficient=sigma
def brownianMotionWithDrift(mu,sigma,B_t):
    numOfPaths = len(B_t)
    N = len(B_t[0])
    t = np.linspace(start=0,stop=1.0,num=N)

    Y = []
    for omega_i in range(numOfPaths):
        X_t = sigma * B_t[omega_i] + mu * t
        Y.append(X_t)
```

Brownian motion with drift $\mu = 1.0$, diffusion coeff $\sigma = 1.0$

Example 3.18. (*Brownian Bridge*). The Brownian bridge is a Gaussian process $(Z_t)_{t\in[0,1]}$ defined by the mean $\mathbb{E}[Z_t] = 0$ and covariance $Cov(Z_t, Z_s) = s(1-t)$ if $0 \leq s \leq t$. Note that by construction, $Z_0 = Z_1 = 0$. It turns out that if $(B_t)_{t\in[0,1]}$ is a standard brownian motion on $[0, 1]$, then the process

$$Z_t = B_t - tB_1, \quad t \in [0, 1]$$

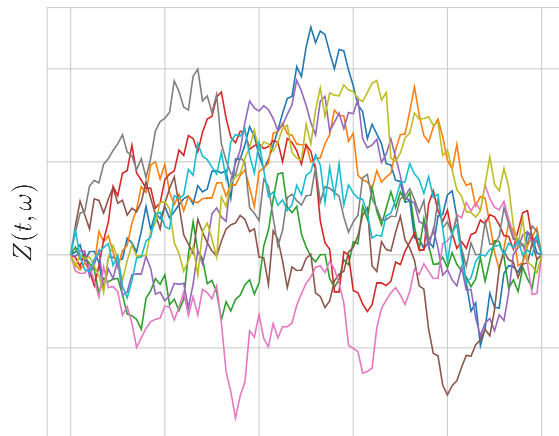
has the distribution of a Brownian bridge.

Listing 5: Brownian bridge

```
def brownianBridge(B_t):
    numOfPaths = len(B_t)
```

```
    N = len(B_t[0])

    t = np.linspace(start=0, stop=1.0, num=N)

    Z = []
    for omega_i in range(numOfPaths):
        X = B_t[omega_i] - B_t[omega_i][N-1]* t
        Z.append(X)

    return Z
```

10 sample paths of Brownian Bridge on $[0, 1]$



Example 3.19. (*Fractional Brownian Motion*). The fractional Brownian motion $(B_t^{(H)})_{t\geq 0}$ with index $0 < H < 1$ (called the Hurst Index), is the Gaussian process with mean 0 and covariance

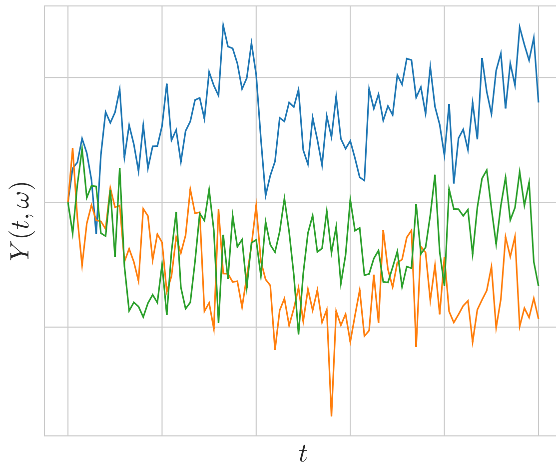
$$Cov(Y_s, Y_t) = \mathbb{E}[B_t^{(H)}, B_s^{(H)}] = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H})$$

The case of $H = 1/2$ corresponds to the Brownian motion.

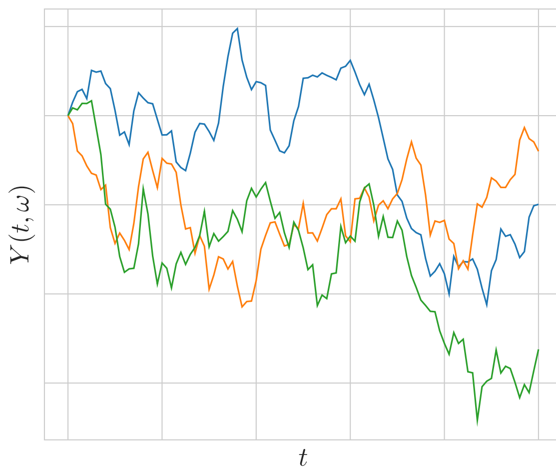
Listing 6: Fractional Brownian Motion

```
def fBM(H,numOfPaths,N):
    # Initialize the covariance matrix
    C = np.zeros((N,N))

    for i in range(N):
        for j in range(N):
            s = (i+1)/N
            t = (j+1)/N

            C[i][j] = 0.50 * (s**(2*H) + t**(2*H) - (np.abs(t - s))**(2*H))

    A = np.linalg.cholesky(C)

    Y = []
    for i in range(numOfPaths):
        X = sampleGaussianProcess(A, N)
        X = np.concatenate(([0], X), axis=0)
        Y.append(X)

    return Y
```

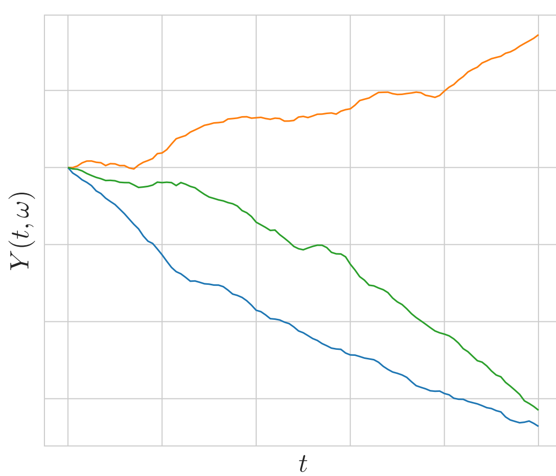
3 paths of fractional brownian motion with $H = 0.1$ on $[0, 1]$



3 paths of fractional brownian motion with $H = 0.5$ on $[0, 1]$



3 paths of fractional brownian motion with $H = 0.9$ on $[0, 1]$



Example 3.20. (Ornstein-Uhlenbeck process). The Ornstein-Uhlenbeck process $(Y_t)_{t\geq 0}$ starting at $Y_0 = 0$ is the Gaussian process with mean $\mathbb{E}[Y_t] = 0$ and covariance:

$$Cov(Y_s, Y_t) = \frac{e^{-2(t-s)}}{2}(1 - e^{-2s}), \quad \text{for } s \leq t$$

If the starting point Y_0 is random, specifically Gaussian with mean 0 and variance $1/2$, then we have: $\mathbb{E}Y_t = 0$ and

$$Cov(Y_s, Y_t) = \frac{e^{-2(t-s)}}{2}, \quad \text{for } s \leq t$$

The covariance only depends on the difference of the time! This means that the process $(Y_t)_{t\geq 0}$ has the same distribution if we shift time by an amount a for any $a \geq 0$: $(Y_{t+a})_{t\geq 0}$. Processes with this property are called *stationary*. As can be observed from the figure below, the statistics of stationary processes do not change over time.

3.7 A Geometric Point of View.

Before turning to Gaussian processes in more detail, it is worthwhile to spend some time to further explore the analogy between random variables in $L^2(\Omega)$ and vectors in$\mathbf{R}^n$. We've already seen earlier, how the space of all random variables form a vector space. We shall now observe that L^2 is a subspace of this vector space.

Definition 3.21. For a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ consists of all random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$ such that:

$$\left[\mathbb{E}(X^2)\right] < \infty$$

Such random variables are called square integrable.

In the same spirit, the space of integrable random variables is denoted by $L^1(\Omega, \mathcal{F}, \mathbb{P})$. We will see that any square-integrable random variable must be integrable. In other words, $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a subset of $L^1(\Omega, \mathcal{F}, \mathbb{P})$. In particular, square integrable random variables have a well-defined expectation. This means that, we can think of L^2 as the set of all random variables on a given probability space with finite variance. Clearly, random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with the Gaussian distribution are in L^2.

The space L^2 is a vector space.

1) If X and Y are two random variables in L^2, then the linear combination $aX + bY$ is also a random variable in L^2. If $u, v \in \mathbf{R}$, we know that :

$$\begin{aligned} (u-v)^2 &\geq 0 \\ u^2 - 2uv + v^2 &\geq 0 \\ 2uv &\leq u^2 + v^2 \end{aligned}$$

Setting $u = aX$ and $v = bY$, we get:

$$\begin{aligned} 2abXY &\leq a^2X^2 + b^2Y^2 \\ 2ab\mathbb{E}(XY) &\leq a^2\mathbb{E}X^2 + b^2\mathbb{E}Y^2 \end{aligned}$$

Having established this upper bound for $2ab\mathbb{E}(XY)$, we now proceed to show that $aX + bY$ belongs to L^2. We have:

$$
\begin{aligned}
\mathbb{E}[(aX+bY)^2] &= \mathbb{E}(a^2X^2+b^2Y^2+2abXY) \\
&= a^2\mathbb{E}X^2+b^2\mathbb{E}Y^2+2ab\mathbb{E}(XY) \\
&\leq a^2\mathbb{E}X^2+b^2\mathbb{E}Y^2+a^2\mathbb{E}X^2+b^2\mathbb{E}Y^2 \\
&= 2a^2\mathbb{E}X^2+2b^2\mathbb{E}Y^2
\end{aligned}
$$

Since $X, Y \in L^2$, $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$. Hence, $\mathbb{E}[(aX+bY)^2]$ is bounded.

2) The zero element of the linear space L^2 is the constantly zero random variable $X = 0$ (with probability one).

Example 3.21. Consider $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$ where $\Omega = \{0,1\} \times \{0,1\}$, $\mathbb{P}$ is the equiprobability and $\mathcal{P}(\Omega)$ is the power set of Ω i.e. all the subsets of Ω. An example of a random variable is $X = 2\mathbf{1}_{\{(0,0)\}}$ where $\mathbf{1}_{\{(0,0)\}}$ is the indicator function of the event $\{(0,0)\}$. In other words, X takes the value 2 on the outcome $(0,0)$ and 0 for the other outcomes. Of course, we can generalize this construction by taking a linear combination of multiples of indicator random functions. Namely, consider the random variable:

$$
X = a\mathbf{1}_{\{(0,0)\}} + b\mathbf{1}_{\{(1,0)\}} + c\mathbf{1}_{\{(0,1)\}} + d\mathbf{1}_{\{(1,1)\}}
$$

for some fixed $a, b, c, d \in \mathbf{R}$. Clearly, any random variable on this probability space can be written in this form. Moreover, any random variable of this form will have a finite variance. Therefore, the space L^2 in this example consists of random variables of the above form. This linear space has dimension 4, since we can write any random variables as the finite linear combination of the four indicator functions. In general, if Ω is finite, the space L^2 is finite dimensional as a linear space, if Ω is infinite, the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ might be infinite dimensional.

3.7.1 Norm in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Similar to $\mathbf{R}^n$, the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ has a norm or a length: for a random variable X in L^2, its norm $\|X\|_2$ is given by:

$$
\|X\|_2 = \left[\mathbb{E}X^2\right]^{1/2} \tag{3.52}
$$

Note that this is very close in spirit to the length for the vector $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \ldots + x_n^2}$ in $\mathbf{R}^n$, since the expectation is heuristically a sum over the outcomes. We have already seen that this definition satisfies the properties of a norm.

1) *Positive Semi-Definite*:

If X is a random variable, then $X^2 \geq 0$. By monotonicity of expectations, $\mathbb{E}X^2 \geq 0$. Moreover, if $\mathbb{E}X^2 = 0$, then since X^2 is a non-negative random variable, $X^2 = 0$ almost surely. It implies that $X = 0$ a.s.

2) *Scalar multiplication*.

If X is a random variable in L^2, we have:

$$
\begin{aligned}
\|aX\|_2 &= \left(\mathbb{E}[(aX)^2]\right)^{1/2} \\
&= \left(\mathbb{E}a^2X^2\right)^{1/2} \\
&= |a| \left(\mathbb{E}X^2\right)^{1/2} \\
&= |a| \, \|X\|_2
\end{aligned}
$$

3) *Triangle Inequality*.

We have:

$$
\begin{aligned}
\|X+Y\|_2^2 &= \mathbb{E}[(X+Y)^2] \\
&= \mathbb{E}X^2 + \mathbb{E}Y^2 + 2\mathbb{E}XY \\
&= \|X\|_2^2 + \|Y\|_2^2 + 2\mathbb{E}XY \\
&\leq \|X\|_2^2 + \|Y\|_2^2 + 2\mathbb{E}|XY| && \{\because XY \leq |XY|\} \\
&\leq \|X\|_2^2 + \|Y\|_2^2 + 2\left(\mathbb{E}X^2\right)^{1/2}\left(\mathbb{E}Y^2\right)^{1/2} && \{\text{Cauchy-Schwarz inequality}\} \\
&= \|X\|_2^2 + \|Y\|_2^2 + 2\|X\|_2\|Y\|_2 \\
&= (\|X\|_2 + \|Y\|_2)^2
\end{aligned}
$$

Thus, $\|X+Y\|_2 \leq \|X\|_2 + \|Y\|_2$.

3.7.2 Inner-product in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Like $\mathbf{R}^n$, the space L^2 has a dot-product or scalar product between two elements X, Y of the space. It is given by:

$$\langle X, Y\rangle = \mathbb{E}(XY)$$

More generally, this operation is called the inner-product. It has the same properties as the dot product in $\mathbf{R}^n$.

1) *Symmetric*:

We have:

$$\mathbb{E}(XY) = \mathbb{E}(YX)$$

2) *Linearity*:

$$\mathbb{E}((aX+bY)Z) = a\mathbb{E}(XZ) + b\mathbb{E}(YZ)$$

3) *Positive semi-definite*:

$$\langle X, X\rangle = \mathbb{E}X^2$$

Since X^2 is a non-negative random variable, $X^2 \geq 0$ and by the monotonicity of expectations $\mathbb{E}X^2 \geq 0$.

Let $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ and define $\hat{X} = X - \mathbb{E}X$, $\hat{Y} = Y - \mathbb{E}Y$

$$
\begin{aligned}
|\mathbb{E}(\hat{X}\hat{Y})| \leq \mathbb{E}|\hat{X}\hat{Y}| &\leq \left(\mathbb{E}\hat{X}^2\right)^{1/2}\left(\mathbb{E}\hat{Y}^2\right)^{1/2} \\
|\mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)| &\leq \left[\mathbb{E}(X-EX)^2\right]^{1/2}\left[\mathbb{E}(Y-EY)^2\right]^{1/2} \\
|Cov(X,Y)| &\leq \sqrt{Var(X)} \cdot \sqrt{Var(Y)} \\
|Corr(X,Y)| &\leq 1
\end{aligned}
$$

3.7.3 Projection of a random variable X on Y.

Consider the random variable

$$\begin{aligned} X^{\perp} &= X - \frac{\langle X, Y\rangle}{\|Y\|_2^2} Y \\ &= X - \frac{\mathbb{E}(XY)}{\mathbb{E}Y^2} Y \end{aligned}$$

This random variable is uncorrelated to Y or orthogonal to Y, in the sense that it's inner product with Y is zero. We have:

$$\begin{aligned} \left\langle X^{\perp}, Y \right\rangle &= \mathbb{E}(X^{\perp} Y) \\ &= \mathbb{E}\left[XY - \frac{\mathbb{E}(XY)}{\mathbb{E}Y^2} Y^2\right] \\ &= \mathbb{E}XY - \frac{\mathbb{E}(XY)}{\mathbb{E}Y^2} \cdot \mathbb{E}Y^2 \\ &= 0 \end{aligned}$$

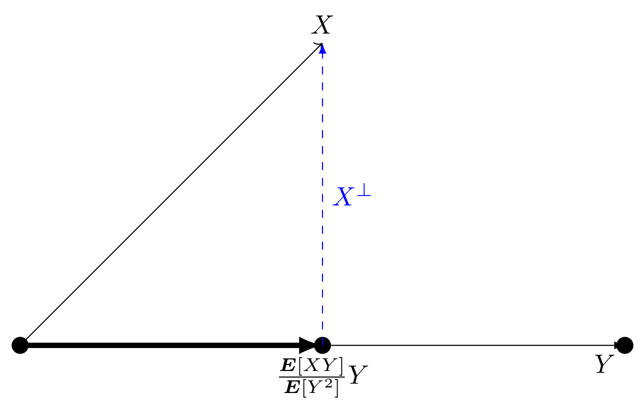


Figure. A representation of the decomposition of the random variable X in terms of its projection on Y and the component $X^{\perp}$ orthogonal to Y.

The random variable $\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y$ is the random variable of the form tY, $t \in \mathbf{R}$, that is closest to X in the L^2-sense. We will make this more precise when we define the conditional expectation of a random variable shortly ahead. For now, we simply note that these considerations imply the decomposition

$$X = X^{\perp} + \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]} Y$$

The random variable $X^{\perp} = X - \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y$ is the component of X *orthogonal* to Y. The random variable:

$$\mathrm{Proj}_Y(X) = \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]} Y$$

is called the orthogonal projection of the random variable X onto Y. Put another way, this is the component of X in the direction of Y. This is the equivalent of the orthogonal projection of $\mathbf{R}^n$ of a vector $\mathbf{w}$ in the direction of $\mathbf{v}$, given by $\frac{\langle \mathbf{w}, \mathbf{v}\rangle}{\|\mathbf{v}\|^2}\mathbf{v}$.

Example 3.22. Going back to example (3.21), let's define the random variables $Y = 2\mathbf{1}_{\{(0,0)\}} + \mathbf{1}_{\{(1,0)\}}$and $W = \mathbf{1}_{\{(0,0)\}}$. Then the orthogonal projection of Y onto W is:

$$\begin{aligned}\frac{\mathbb{E}(YW)}{\mathbb{E}W^2}W &= \frac{2\mathbb{P}(\{0,0\}}{\mathbb{P}(\{0,0\})}W \\ &= \frac{2\cdot\frac{1}{4}}{\frac{1}{4}}W \\ &= 2W\end{aligned}$$

The orthogonal decomposition of Y is simply :

$$Y = 2W + (Y - 2W)$$

The notion of norm induces a notion of distance between the random variables in L^2 given by $\|X - Y\|_2 = \mathbb{E}[(X - Y)^2]^{1/2}$. In particular, we see that the orthogonal projection of X onto Y is the closest point from X amongst all multiples of Y. This is what the proof of Cauchy-Schwarz inequality does. The L^2 distance also gives rise to a notion of convergence.

3.7.4 Borel-Cantelli Lemmas.

Lemma 3.8. *(Borel-Cantelli Lemmas)*

(a) (First Borel-Cantelli Lemma) Let $\{A_n\}$ be a sequence of events such that the series $\sum_n \mathbb{P}(A_n)$ converges to a finite value L. Then, almost surely, only finitely many A_n's will occur.

(b) (Second Borel-Cantelli Lemma) Let $\{A_n\}$ be a sequence of independent events such that $\sum_n \mathbb{P}(A_n)$ diverges to ∞. Then, almost surely, infinitely many A_n's will occur.

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $A_1, A_2, A_3, \ldots$ be an infinite sequence of events belonging to $\mathcal{F}$. We shall often be interested in finding out how many of the A_n occur.

The event "A_n occurs infinitely often (A_n $i.o.$) is the set of all ω that belong to infinitely many A_n's.

Imagine that an infinite number of A_n's occur. That is, $(\forall n)(\exists m \geq n)(\text{s.t.} A_m \text{ occurs})$. In other words:

$$\{A_n \text{ infinitely often }\} \triangleq \bigcap_{n=1}^{\infty} \underbrace{\bigcup_{m=n}^{\infty} A_m}_{B_n} \tag{3.53}$$

Here, B_n is the event that atleast one of $A_n, A_{n+1}, \ldots$ occur. For that reason, B_n is sometimes referred to as the n-th tail event. $\{A_n$ infinitely often $\}$ is the intersection of all the B_n's, so it is the event that all the B_n's occur. Therefore, no matter how far I go, no matter how big my n_0 is, beyond that n_0, atleast one of $A_{n_0}, A_{n_0+1}, \ldots$ occurs.

Taking the complement of both sides in (3.53), we get the expression for the event that A_n occurs finitely often.

$$\{A_n \text{ finitely often }\} \triangleq \bigcup_{n=1}^{\infty} \underbrace{\bigcap_{m=n}^{\infty} A_m^C}_{} \tag{3.54}$$

It means there exists an n, such that each of the further A_i's fail to occur.

In order to prove the Borel-Cantelli lemmas, we require the following lemma.

Lemma 3.9. *If $\sum_{i=1}^{\infty} p_i = \infty$, then $\lim_{n\to\infty} \prod_{i=1}^{n}(1 - p_i) = 0$.*

Proof. We know that:

$$\ln(1+x) \leq x$$

So,

$$\begin{aligned}
\ln(1-p_i) &\leq -p_i \\
\sum_{i=1}^{n} \ln(1-p_i) &\leq -\sum_{i=1}^{n} p_i \\
0 \leq \prod_{i=1}^{n}(1-p_i) &\leq e^{-\sum_{i=1}^{n} p_i}
\end{aligned}$$

Passing to the limit on both sides, as $n \to \infty$, we have:

$$0 \leq \lim_{n\to\infty} \prod_{i=1}^{n}(1-p_i) \leq \lim e^{-\sum_{i=1}^{n} p_i} = 0$$

By the squeeze theorem, the limit $\lim_{n\to\infty} \prod_{i=1}^{n}(1-p_i)$ exists and is equal to 0.

Consequently, the product series $\prod_{i=1}^{n}(1-p_i)$ converges to 0. ☐

Proof. (**First Borel-Cantelli Lemma**)

Our claim is that $\mathbb{P}(\bigcap_{n=1}^{\infty} B_n) = 0$.

Whenever we see something like $\bigcap_{n=1}^{\infty} B_n$, we can think of invoking continuity of probability. It turns out that, $B_n = \bigcup_{m\geq n} A_m$. So, $B_1 \supseteq B_2 \supseteq B_3 \supseteq \ldots$, that is the B_n's are nested decreasing sequence of sets. So, $\lim B_n = \bigcap_{n=1}^{\infty} B_n$. So, by continuity of probability measure:

$$\begin{aligned}
\mathbb{P}(\bigcap_{n=1}^{\infty} B_n) &= \lim_{n\to\infty} \mathbb{P}(B_n) \\
&= \lim_{n\to\infty} \mathbb{P}(\bigcup_{m\geq n}^{\infty} A_m) \\
&\leq \lim_{n\to\infty} [\mathbb{P}(A_n) + \mathbb{P}(A_{n+1}) + \ldots] \\
&\quad \{\text{ Union bound on probability}\} \\
&= \lim_{n\to\infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i)
\end{aligned}$$

We know that, $\sum_{n=1}^{\infty} \mathbb{P}(A_n)$ converges to some finite value L, and the above expression is the tail sum of a convergent series. The sequence of tail sums of a convergent series always converges to 0. Thus,

$$0 \leq \mathbb{P}(\bigcap_{n=1}^{\infty} B_n) \leq 0$$

so it follows that

$$\mathbb{P}\{A_n \; i.o.\} = 0$$

(**Second Borel-Cantelli Lemma**)

Our claim is $\mathbb{P}\{A_n \; i.o.\} = 1$. We must therefore prove that:

$$\begin{aligned} \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) &= 1 \\ \iff \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n^C\right) &= 0 \end{aligned}$$

We have:

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n^C\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n^C)$$

I want to prove that the above sum is zero. It means that each of these B_n^C events should have 0 probability. We will show that $\mathbb{P}(B_n^C) = 0$ for all $n \geq 1$.

Indeed fix n and $k \geq n$. Consider $\mathbb{P}\left(\bigcap_{i=n}^{k} A_i^C\right)$. That is I am taking finite intersection of A_i^C. I want to prove that B_n^C has probability zero.

If you look at A_i^C, these are independent events. So, $\mathbb{P}(\bigcap_{i=n}^{k} A_i^C) = \prod_{i=n}^{k} \mathbb{P}(A_i^C) = \prod_{i=n}^{k} [1 - \mathbb{P}(A_i)]$. Passing to the limit as $k \to \infty$,

$$\begin{aligned} \lim_{k\to\infty} \mathbb{P}(\bigcap_{i=n}^{k} A_i^C) &= \lim_{k\to\infty} \prod_{i=n}^{k} [1 - \mathbb{P}(A_i)] \\ \mathbb{P}\left(\bigcap_{i=n}^{\infty} A_i^C\right) &= \lim_{k\to\infty} \prod_{i=n}^{k} [1 - \mathbb{P}(A_i)] \\ \mathbb{P}(B_n^C) &= \lim_{k\to\infty} \prod_{i=n}^{k} [1 - \mathbb{P}(A_i)] \end{aligned}$$

Since, $\sum_{i=n}^{\infty} \mathbb{P}(A_i)$ diverges to ∞, it follows from lemma (3.9), that $\prod_{i=n}^{\infty} [1 - \mathbb{P}(A_i)] = 0$. Hence, $\mathbb{P}(B_n^C) = 0$ for all $n \in \mathbf{N}$. So, $0 \leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n^C\right) \leq 0$. Therefore, $\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n^C\right) = 0$, or equivalently, $\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = 1$. The event $\{A_n \; i.o.\}$ occurs almost surely. □

3.8 Convergence of random variables.

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ once for all. On this probability space, we will have a sequence $(X_1, X_2, X_3, \ldots)$ of random variables defined on it.

Definition 3.22. (*Point-wise convergence.*) A sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge point-wise to X, if and if, for all $\epsilon > 0$, and for all $\omega \in \Omega$, there exists $N(\epsilon, \omega) \in \mathbf{N}$ such that for all $n \geq N$, we have $|X_n(\omega) - X(\omega)| < \epsilon$.

It would be natural to say, that, for all $\omega \in \Omega$, $X_n(\omega) \to X(\omega)$. But, this is too demanding. So, we will weaken this convergence.

Definition 3.23. (*Almost-sure convergence.*) A sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge almost-surely to X, written $X_n \xrightarrow{a.s.} X$, if and only if, there exists a set $A \in \mathcal{F}$, such that $\mathbb{P}[A] = 1$ and for all $\omega \in A$, $X_n(\omega) \to X(\omega)$.

Theorem 3.22. *(Sufficient condition for almost-sure convergence.) If* $(\forall \epsilon > 0)$, $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty$*, then* $X_n \stackrel{a.s.}{\to} X$.

Remark. If you notice just the object $\mathbb{P}(|X_n - X| > \epsilon)$; if this term goes to zero, then it is convergence in probability. So, if the term $\mathbb{P}(|X_n - X| > \epsilon)$ goes to zero, we have convergence in probability. The condition $\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| > \epsilon) < \infty$ is a little bit stronger, it says a little bit more. As n tends to infinity, not only do the terms $a_n = \mathbb{P}(|X_n - X| > \epsilon)$ go to zero, for every ϵ; it goes to zero fast enough that the sum converges. For example, if this probability $\mathbb{P}(|X_n - X| > \epsilon)$ were to go to zero, as $\frac{1}{n}$, then you have convergence in probability, but $\sum \frac{1}{n}$ diverges. So, if the term a_n goes to zero fast enough to keep the summation finite, then we have convergence almost surely. For instance, if $a_n \approx \frac{1}{n^2}$, then we would have almost sure convergence.

This is just a sufficient condition. If it holds, we are guaranteed almost sure convergence, but even if it doesn't hold, sometimes we may have almost sure convergence.

Proof. Let $A_n(\epsilon)$ be the event $\{|X_n - X| > \epsilon\}$. We are given that, for all $\epsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\epsilon)) < \infty$. Using BCL1 (3.8), we see that that, for any $\epsilon > 0$, only finitely many $A_n(\epsilon)$ occur with probability 1. Thus, there exists an n_0, such that for all $n \geq n_0$, $A_n^C(\epsilon) = \{|X_n - X| \leq \epsilon\}$ occurs with probability 1. So, X_n converges to X with probability 1. $\square$

Remark. If we plot the distance between the random variables, $X_n - X$, there may be some excursions. But, essentially, BCL1 says that, with probability 1, there must be an n_0, beyond which the sequence X_n settles within an ϵ-band of X, and these excursions never occur. Because, only finitely many excursions occur. And this is true for every $\epsilon > 0$. So, with probability 1, $X_n \to X$. Thus, $X_n \stackrel{a.s.}{\to} X$.

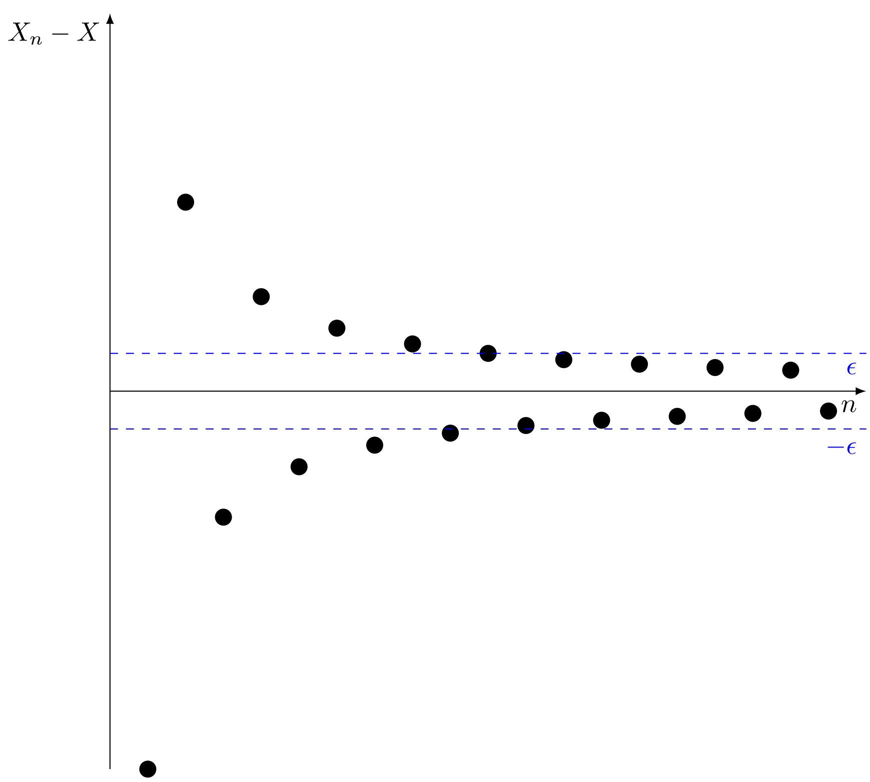


Figure. Convergence of X_n to X.

Example 3.23. The converse of the theorem (3.22) does not hold. Consider the sequence of random variables:

$$X_n = \begin{cases} 1 & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}$$

Then, for all $\epsilon > 0$, $\mathbb{P}(|X_n| < \epsilon) = 1 - \frac{1}{n}$. So, for all $\epsilon > 0$, $\lim \mathbb{P}(|X_n| < \epsilon) = 1$. Thus, the sequence (X_n) converges 0 with probability 1. So, $X_n \stackrel{a.s.}{\to} 0$. However, $\sum \mathbb{P}(|X_n| > \epsilon) = \sum \frac{1}{n} = \infty$.

Theorem 3.23. *(Necessary and sufficient condition for almost-sure convergence.) Let $A_n(\epsilon)$ be the event that the excursion $\{|X_n - X| > \epsilon\}$ happens and define:*

$$B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon)$$

Then,

$$X_n \stackrel{a.s.}{\to} X \quad \textit{if and only if} \quad \lim \mathbb{P}(B_m(\epsilon)) = 0 \quad \forall \epsilon > 0$$

Remark. Note that, $\mathbb{P}(A_n(\epsilon))$ going to 0 is convergence in probability. I am saying a little more. In words, $A_n(\epsilon)$ is the event that an excursion occurs at the nth term. In words, $B_m(\epsilon)$ is the event that atleast one of A_m, A_{m+1}, A_{m+2}, $\ldots$ occurs, which means that atleast one excursion occurs m or after. What this theorem says is, if the probability of this event goes to 0, then you have almost sure convergence. In other words, if you find some m; this m can be very large, but if you find some m beyond which no excursions ever occur, then you have almost sure convergence (and vice versa).

Proof. ($\Longrightarrow$ direction.)

We are given that $X_n \stackrel{a.s.}{\to} X$. Our claim is $\lim_{m \to \infty} \mathbb{P}(B_m(\epsilon)) = 0$.

Now, if $X_n \to X$ almost surely, then clearly, $(\forall \epsilon > 0)$, the event

$$\bigcup_{m \geq 1} \bigcap_{n \geq m} \{|X_n - X| < \epsilon\}$$

occurs with probability 1.

So, for all $\epsilon > 0$, the event

$$\bigcap_{m \geq 1} \bigcup_{n \geq m} \{|X_n - X| \geq \epsilon\} = \bigcap_{m \geq 1} B_m$$

occurs with probability 0. An excursion happens only finitely many times.

Now, the sequence events $B_1(\epsilon)$, $B_2(\epsilon)$, $B_3(\epsilon)$, $\ldots$ are nested decreasing. They are like Russian dolls. By continuity of probability measure, $\mathbb{P}\left(\bigcap_{m=1}^{\infty} B_m\right) = \lim_{m \to \infty} \mathbb{P}(B_m)$. Consequently, it follows that $\lim_{m \to \infty} \mathbb{P}(B_m) = 0$.

($\Longleftarrow$direction.)

We are given that, for all $\epsilon > 0$, $\lim_{m \to \infty} \mathbb{P}\left(B_m(\epsilon)\right) = 0$. We are interested to prove that $X_n \stackrel{a.s.}{\to} X$.

Let C be the event:

$$C = \{\omega | X_n(\omega) \to X(\omega)\}$$

Define the event:

$$A(\epsilon) = \bigcap_{m\geq 1} \bigcup_{n\geq m} \{|X_n - X| \geq \epsilon\}$$

I would like to prove that $\mathbb{P}(C) = 1$. What we will prove is that $\mathbb{P}(C^C) = 0$.

C^C is the event that, no matter how big an n you look at, there is some excursion. That is there are infinitely many excursions.

This means, there must be some $\epsilon_0 > 0$ for which $A(\epsilon)$ occurs.

$$\begin{aligned}
\mathbb{P}(C^C) &= \mathbb{P}\left(\bigcup_{\epsilon>0} A(\epsilon)\right) \\
&= \mathbb{P}\left(\bigcup_{k=0}^{\infty} A(\frac{1}{k})\right) \\
&\leq \sum_{k=0}^{\infty} \mathbb{P}\left(A(\frac{1}{k})\right)
\end{aligned}$$

Now,

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}(B_m(\frac{1}{k})) &= \mathbb{P}\left(\bigcap_{m=1}^{\infty} B_m(\frac{1}{k})\right) \\
&\quad \{\text{ Continuity of probability measure }\} \\
&= \mathbb{P}\left(A(\frac{1}{k})\right)
\end{aligned}$$

So, $\mathbb{P}\left(A(1/k)\right) = 0$. Consequently, $\mathbb{P}(C^C) = 0$ and $\mathbb{P}(C) = 1$. Thus, $X_n \xrightarrow{a.s.} X$. $\square$

Remark. Therein, lies the difference between convergence in probability and convergence almost surely. Convergence in probability just says the probability of $A_n(\epsilon)$ (an excursion occurs at n) goes to zero. It just looks at one n; it forgets about the rest of the sequence.

For convergence almost surely, you are not looking at a particular n. You fix a particular m and you're saying that the probability that beyond m an excursion occurs goes to zero.

This should convince you intuitively, that almost sure convergence implies convergence in probability.

Definition 3.24. (*Convergence in Probability.*) A sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge in probability to X, written $X_n \xrightarrow{P.} X$, if and only if

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

Definition 3.25. (*Convergence in L^p*) A sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge in the pth mean to X, if and only if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0$$

Definition 3.26. (*Convergence in Distribution.*) A sequence of random variables $(X_n)_{n=1}^{\infty}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to converge in distribution to X, if and only if:

$$\mathbb{P}(X_n \leq X) \to \mathbb{P}(X \leq x)$$

For example, let $\Omega = \{1, 2\}$ and $\mathbb{P}(1) = \mathbb{P}(2) = \frac{1}{2}$, $X_n(1) = \frac{-1}{n}$ and $X_n(2) = \frac{1}{n}$. We have:

1) $X_n \xrightarrow{a.s.} 0$ because $X_n(\omega) \to 0$ for all $\omega \in \Omega$.

2) $X_n \xrightarrow{L^2} 0$ because $\mathbb{E}(X_n^2) = \frac{1}{n^2} \to 0$.

3) $X_n \xrightarrow{P} 0$ because $P(|X_n| > \epsilon) = \mathbb{P}\left(\frac{1}{n} > \epsilon\right) = 0$.

Theorem 3.24. *(Hierarchy of Convergence) The following implications hold:*

$$\begin{array}{ccc}
(X_n \xrightarrow{a.s.} X) & & \\
\Downarrow & & \\
(X_n \xrightarrow{P} X) & \Longrightarrow & (X_n \xrightarrow{D} X) \\
\Uparrow & & \\
(X_n \xrightarrow{L^p} X) & &
\end{array}$$

Proof. (i) *Claim.* $X_n \overset{L^p}{\to} X$ implies $X_n \overset{P}{\to} X$.

This is a very easy proposition to prove. By definition, we have:

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0$$

By Markov's inequality:

$$\begin{aligned}
0 \le \mathbb{P}(|X_n - X| > \epsilon) &= \mathbb{P}(|X_n - X|^p > \epsilon^p) \\
&\le \frac{1}{\epsilon^p}\mathbb{E}\left[|X_n - X|^p\right]
\end{aligned}$$

Passing to the limit on both sides, as $n \to \infty$, by the squeeze limit theorem,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

(ii) *Claim.* $X_n \overset{P}{\to} X$ implies that $X_n \overset{D}{\to} X$.

Fix an $\epsilon > 0$.

We have:

$$\begin{aligned}
F_{X_n}(x) &= \mathbb{P}(X_n \le x) \\
&= \mathbb{P}(X_n \le x, X \le x + \epsilon) \\
&+ \mathbb{P}(X_n \le x, X > x + \epsilon) \\
&\le \mathbb{P}(X \le x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\
&\quad \because \{X_n \le x, X \le x + \epsilon\} \subseteq \{X \le x + \epsilon\} \\
&= F_X(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon)
\end{aligned}$$

Similarly,

$$
\begin{aligned}
F_X(x-\epsilon) &= \mathbb{P}(X \le x-\epsilon) \\
&= \mathbb{P}(X \le x-\epsilon, X_n \le x) + \mathbb{P}(X \le x-\epsilon, X_n > x) \\
&\le \mathbb{P}(X_n \le x) + \mathbb{P}(|X_n - x| > \epsilon) \\
&= F_{X_n}(x) + \mathbb{P}(|X_n - X| > \epsilon)
\end{aligned}
$$

Thus, we have the inequality:

$$\forall \epsilon > 0, \quad F_X(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \le F_{X_n}(x) \le F_X(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon)$$

We assume that F_X is continous for all x. Pick $\epsilon = \frac{1}{n}$ and passing to the limit as $n \to \infty$, we have:

$$
\begin{gathered}
\lim F_X(x-\epsilon) \le \lim F_{X_n}(x) \le \lim F_X(x+\epsilon) \\
F_X(x) \le \lim F_{X_n}(x) \le F_X(x)
\end{gathered}
$$

By the Squeeze Theorem, the limit $F_{X_n}(x)$ exists and $\lim F_{X_n}(x) = F_X(x)$.

(iii) **Counterexample**. (Convergence in distribution does not imply convergence in probability).

Convergence in distribution simply means that only the CDFs are converging; it doesn't mean that X_n and X are getting closer in any sense.

Let $X_1, X_2, X_3, \ldots$ be such that $X_i = X$ for all $i \ge 1$ and

$$X = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases}$$

The entire sequence is just $(X, X, X, \ldots)$. Let $Y = 1 - X$. By definition, $Y \sim Bernoulli(1/2)$. We have: $X_n \xrightarrow{D} Y$ in distribution, but $|X_n - Y| = 1$, we could choose $\epsilon_0 = \frac{1}{2}$ and we get:

$$\mathbb{P}(|X_n - Y| > \frac{1}{2}) = 1$$

so (X_n) does not converge to Y in probability.

(iv) **Counterexample**. (Convergence in probability does not imply convergence in the mean square sense).

Consider

$$X_n = \begin{cases} n^2 & \text{with probability } n^{-2} \\ 0 & \text{with probability } 1 - n^{-2} \end{cases}$$

Then,

$$\mathbb{P}(|X_n| > \epsilon) = \frac{1}{n^2}$$

and as $n \to \infty$, $\frac{1}{n^2} \to 0$. So, $X_n \xrightarrow{P} 0$.

But,

$$\mathbb{E}(X_n^2) = n^2$$

and as $n \to \infty$, $n^2 \to \infty$. Hence, $X_n \overset{L^2}{\not\to} 0$.

(v) **Counterexample**. (Convergence in probability does not imply convergence almost surely).

Consider

$$X_n = \begin{cases} 1 & \text{with probability} \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{1}{n} \end{cases}$$

and X_i's are independent. The larger the value of n, the more likely that X_n takes the value 0. We have:

$$\mathbb{P}(|X_n| > \epsilon) = \frac{1}{n}$$

So,

$$\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = \lim \frac{1}{n} = 0$$

Our claim is that X_n does not converge to 0 almost surely.

Let A_n be the event that $\{X_n = 1\}$. Then, A_n's are independent. We have:

$$\sum_{i=1}^{\infty} \mathbb{P}(A_n) = 0 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$$

The harmonic series $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges to ∞.

By the BCL2 (Borell-Cantelli Lemma 2) (3.8), it follows that, with probability 1, infinitely many A_n's will occur.

$$\mathbb{P}\{X_n = 1 \; i.o.\} = 1$$

So, $X_n \overset{a.s.}{\not\to} 0$.

Imagine a coin-tossing experiment, where X_n represents the outcome of the nth coin-toss, and the probability of the nth coin toss falling heads is $\frac{1}{n}$. Then, no matter how far out you go in the sequence, BCL2 says that, there will some occasional head $(X_n = 1)$ popping off at some-time. Which means that X_n does not converge to zero.

(vi) **Claim**. $X_n \overset{a.s.}{\to} X$ implies $X_n \overset{P}{\to} X$.

By the necessary and sufficient condition of almost sure convergence, $X_n \overset{a.s.}{\to} X$ is equivalent to saying that:

$$\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0$$

But, $B_m(\epsilon) = \bigcup_{n \geq m} A_n(\epsilon)$. Thus, $A_m(\epsilon) \subseteq B_m(\epsilon)$. So, $(\forall \epsilon > 0), 0 \leq \mathbb{P}(A_m(\epsilon)) \leq \mathbb{P}(B_m(\epsilon))$. Passing to the limit on both sides, $0 \leq \mathbb{P}(A_m(\epsilon)) \leq \lim \mathbb{P}(B_m(\epsilon)) = 0$. By the squeeze theorem, $\lim \mathbb{P}(A_m(\epsilon)) = 0$. Consequently, $X_n \overset{P}{\to} X$.

(vii) **Counterexample**. (Convergence almost surely does not imply convergence in mean square)

Let

$$X_n(\omega) = \begin{cases} n & \omega \in [0, \frac{1}{n}] \\ 0 & \text{otherwise} \end{cases}$$

Let $A_n(\epsilon)$ be the event that an excursion occurs at n, $\{|X_n| > \epsilon\}$. Now, $\mathbb{P}(B_m(\epsilon)) = \mathbb{P}(\bigcup_{n \geq m} A_n(\epsilon)) \leq \mathbb{P}(A_m(\epsilon)) = \frac{1}{m}$. So, $\lim_{m\to\infty} \mathbb{P}(B_m(\epsilon)) = 0$. Thus, $X_n \xrightarrow{a.s.} 0$. But, $\mathbb{E}[X_n^2] = n$. Thus, $X_n \overset{L^2}{\not\to} 0$.

(viii) **Counterexample.** (Convergence in mean square does not imply convergence almost surely)

Let

$$X_n = \begin{cases} 1 & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 1/n \end{cases}$$

where the X_n's are independent.

Now, $\mathbb{E}[X_n^2] = \frac{1}{n}$ so $\lim \mathbb{E}[X_n^2] = 0$. Thus, $X_n \xrightarrow{L^2} 0$. Define $A_n = \{|X_n| \geq \epsilon\}$. But, by BCL2, $\sum_n \mathbb{P}(A_n) = \infty$ and the events A_n are independent. Then, A_n occurs infinitely often. In other words, X_n does not converge to 0 almost surely. ☐

Theorem 3.25. *If a sequence (X_n) of random variables converges in probability to X, then there exists a subsequence $(X_{n_k})_k$ which converges to X almost surely.*

Proof. Since $X_n \xrightarrow{P} X$ it follows that:

$$\forall \epsilon > 0 \quad \lim \mathbb{P}(|X_n - X| > \epsilon) = 0$$

Let $\epsilon = 1/m$. In words, for all $m > 0$ and for all $k > 0$, there exists $N(m, k) \in \mathbf{N}$, such that for all $n \geq N$, the following condition is satisfied:

$$\mathbb{P}\left(|X_n - X| > \frac{1}{m}\right) < \frac{1}{k}$$

By the definition of the limit of a sequence, there exists n_1 such that:

$$\mathbb{P}\left(|X_{n_1} - X| > 1\right) < 1$$

There exists $n_2 \geq n_1$ such that:

$$\mathbb{P}\left(|X_{n_2} - X| > \frac{1}{2}\right) < \frac{1}{2^2}$$

There exists $n_3 \geq n_2$ such that:

$$\mathbb{P}\left(|X_{n_3} - X| > \frac{1}{3}\right) < \frac{1}{3^2}$$

In general, there exists a positive integer $n_i \geq n_{i-1}$ such that:

$$\mathbb{P}\left(|X_{n_i} - X| > \frac{1}{i}\right) < \frac{1}{i^2}$$

By the sufficient condition for almost sure convergence (3.22), $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \sum_{n=1}^{\infty} \frac{1}{n^2}$ which converges to a finite value. Hence, $X_{n_i} \stackrel{a.s.}{\to} X$. ☐

Consider a sequence of random variables (X_n), such that $X_n \stackrel{L^2}{\to} X$. it turns out that the limit random variable X of the convergent sequence in L^2 is guaranteed to be in L^2. This is because L^2 is complete. We will prove this very important result further ahead. This property is crucial for the construction of the Ito integral.

Example 3.24. (A version of the weak law of large numbers.) Consider a sequence of random variables $X_1, X_2, \ldots, X_n$ in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = \sigma^2 < \infty$ for all $i \geq 1$ and that they are orthogonal to each other, that is, $\mathbb{E}[X_i X_j] = 0$ for all $i \neq j$. We show that the empirical mean

$$\frac{1}{n}S_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

converges to zero in the L^2 sense.

Clearly,

$$\begin{aligned}
\mathbb{E}\left[\frac{S_n^2}{n^2}\right] &= \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}[X_i^2] \\
&= \lim_{n\to\infty} \frac{1}{n^2} \cdot n\sigma^2 \\
&= \lim_{n\to\infty} \frac{\sigma^2}{n} \\
&= 0
\end{aligned}$$

Consequently, the empirical mean $\frac{S_n}{n}$ converges to 0 in the mean square sense.

Exercise 3.1. (**Ornstein-Uhlenbeck Process**.) Generate 100 paths with step size = 0.01 of the following processes on $[0, 1]$:

(a) Ornstein Uhlenbeck process: $C(s,t) = \frac{e^{-2(t-s)}}{2}(1 - e^{-2s})$ for $s \leq t$. with mean 0 (so that $Y_0 = 0$).

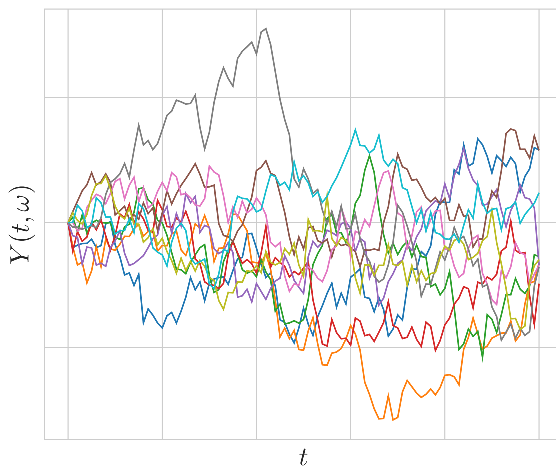
(b) Stationary Ornstein-Uhlenbeck process: $C(s,t) = \frac{e^{-2(t-s)}}{2}$ for $s \leq t$ with mean 0 (so Y_0 is a Gaussian random variable of mean 0 and variance $1/2$).

Solution.

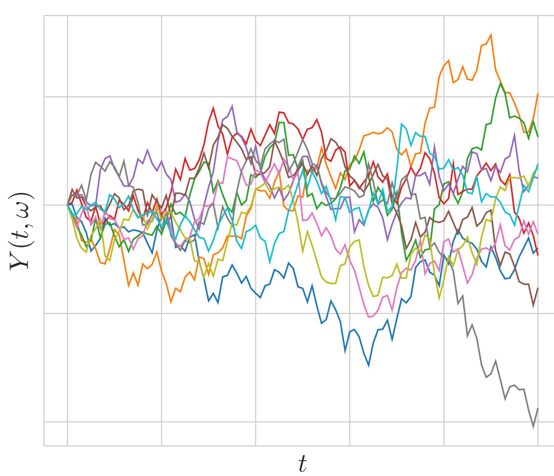
Listing 7: Orsntein-Uhlenbech(OU) process

```
def ornsteinUhlenbeck(numOfPaths,N):
    C = np.zeros((N,N))

    for i in range(N):
        for j in range(N):
            s = (i + 1)/N
            t = (j + 1)/N

            if s > t:
                s,t = t,s
            C[i][j] = np.exp(-2*(t-s))/2 * (1 - np.exp(-2*s))

    A = np.linalg.cholesky(C)

    Y =[]
    for i in range(numOfPaths):
        X = sampleGaussianProcess(A,N)
        X = np.concatenate(([0],X),axis=0)
        Y.append(X)

    return Y
```

10 paths of Ornstein Uhlenbeck process on $[0, 1]$



10 sample paths of stationary OU process on $[0, 1]$



4 Properties of Brownian Motion.

4.1 Properties of Brownian Motion.

Let $B(t)$ be a fixed Brownian motion. We give below some simple properties that follow directly from the definition of the Brownian Motion.

Proposition 4.1. *For any $t \geq 0$, $B(t)$ is normally distributed with mean 0 and variance t. For any $s, t \geq 0$ we have $\mathbb{E}(B_s B_t) = \min\{s, t\}$.*

Proof. From condition (1), we have that $B_0 = 0$. From condition (2), $B_t - B_0 = B_t$ is normally distributed with mean 0 and variance t.

Assume that $s < t$.

We have:

$$\begin{aligned}
\mathbb{E}(B_s B_t) &= \mathbb{E}\left[B_s(B_t - B_s + B_s)\right] && \{\text{Write } B_t = B_t - B_s + B_s\} \\
&= \mathbb{E}[B_s(B_t - B_s)] + \mathbb{E}[B_s^2] && \{\text{Linearity of expectations}\} \\
&= \mathbb{E}[B_s]\mathbb{E}(B_t - B_s) + s && \{B_s, (B_t - B_s) \text{ are independent}\} \\
&= 0 \cdot 0 + s \\
&= s
\end{aligned}$$

This closes the proof. ☐

Proposition 4.2. *(Translation Invariance) For fixed $t_0 \geq 0$, the stochastic process $\tilde{B}(t) = B(t + t_0) - B(t_0)$ is also a Brownian motion.*

Proof. Firstly, the stochastic process $\tilde{B}(t)$ is such that:

(1) $\tilde{B}(0) = B(t_0) - B(t_0) = 0$. Hence, it satisfies condition (1).

(2) Let $s < t$. We have: $\tilde{B}(t) - \tilde{B}(s) = B(t + t_0) - B(s + t_0)$ which a Gaussian random variable with mean 0 and variance $t - s$. Hence, for $a \leq b$,

$$\mathbb{P}\{a \leq \tilde{B}(t) \leq b\} = \frac{1}{\sqrt{2\pi(t-s)}} \int_a^b e^{-\frac{x^2}{2(t-s)}} dx$$

Hence, it satisfies condition (2).

(3) To check condition (3) for $\tilde{B}(t)$, we may assume $t_0 > 0$. Then, for any $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$, we have:

$$0 < t_0 \leq t_0 + t_1 \leq t_0 + t_2 \leq \ldots \leq t_0 + t_n$$

So, $B(t_1 + t_0) - B(t_0)$, $B(t_2 + t_0) - B(t_1 + t_0)$, $\ldots$, $B(t_k + t_0) - B(t_{k-1} + t_0)$, $\ldots$, $B(t_n + t_0) - B(t_{n-1} + t_0)$ are independent random variables. Consequently, $\tilde{B}(t)$ satisfies condition (3).

This closes the proof. ☐

The above translation invariance property says that a Brownian motion starts afresh at any moment as a new Brownian motion.

Proposition 4.3. *(Scaling Invariance) For any real number $\lambda > 0$, the stochastic process $\tilde{B}(t) = B(\lambda t)/\sqrt{\lambda}$ is also a Brownian motion.*

Proof. The scaled stochastic process $\tilde{B}(t)$ is such that:

(1) $\tilde{B}(0) = 0$. Hence it satisfies condition (1).

(2) Let $s < t$. Then, $\lambda s < \lambda t$. We have:

$$\tilde{B}(t) - \tilde{B}(s) = \frac{1}{\sqrt{\lambda}}(B(\lambda t) - B(\lambda s))$$

Now, $B(\lambda t) - B(\lambda s)$ is a Gaussian random variable with mean 0 and variance $\lambda(t-s)$. We know that, if X is a random variable with mean μ and variance σ^2, $Z = \left(\frac{X-\mu}{\sigma}\right)$ has mean 0 and variance 1. Consequently, $\frac{B(\lambda t)-B(\lambda s)}{\sqrt{\lambda}}$ is a Gaussian random variable with mean 0 and variance $(t-s)$.

Hence, $\tilde{B}(t) - \tilde{B}(s)$ is normal distributed with mean 0 and variance $t-s$ and it satisfies condition (2).

(3) To check condition (3) for $\tilde{B}(t)$, we may assume $t_0 > 0$. Then, for any $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$, we have:

$$0 \leq \lambda t_1 \leq \lambda t_2 \leq \ldots \leq \lambda t_n$$

Consequently, the random variables $B(\lambda t_k) - B(\lambda t_{k-1})$, $k = 1, 2, 3, \ldots, n$ are independent. Hence it follows that $\frac{1}{\sqrt{\lambda}}[B(\lambda t_k) - B(\lambda t_{k-1})]$ for $k = 1, 2, \ldots, n$ are also independent random variables.

This closes the proof. $\square$

It follows from the scaling invariance property that for any $\lambda > 0$ and $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$, the random vectors:

$$(B(\lambda t_1), B(\lambda t_2), \ldots, B(\lambda t_n)) \quad (\sqrt{\lambda}B(t_1), \sqrt{\lambda}B(t_1), \ldots, \sqrt{\lambda}B(t_n))$$

have the same distribution.

The scaling property shows that Brownian motion is *self-similar*, much like a fractal. To see this, suppose we zoom into a Brownian motion path very close to zero, say on the interval $[0, 10^{-6}]$. If the Brownian motion path were smooth and differentiable, the closer we zoom in around the origin, the flatter the function will look. In the limit, we would essentially see a straight line given by the derivative at 0. However, what we see with the Brownian motion is very different. The scaling property means that for $a = 10^{-6}$,

$$(B_{10^{-6}t,} t \in [0,1]) \overset{\text{distrib.}}{=} (10^{-3}B_t, t \in [0,1])$$

where $\overset{\text{distrib.}}{=}$ means equality of the distribution of the two processes. In other words, Brownian motion on $[0, 10^{-6}]$ looks like a Browian motion on $[0, 1]$, but with its amplitude multiplied by a factor of 10^{-3}. In particular, it will remain rugged as we zoom in, unlike a smooth function.

Proposition 4.4. *(Reflection at time s) The process $(-B_t, t \geq 0)$ is a Brownian motion. More generally, for any $s \geq 0$, the process $(\tilde{B}(t), t \geq 0)$ defined by:*

$$\tilde{B}(t) = \begin{cases} B_t & \textit{if } t \leq s \\ B_s - (B_t - B_s) & \textit{if } t > s \end{cases} \tag{4.1}$$

is a Brownian motion.

Proof. (a) Consider the process $\tilde{B}(t) = (-B_t, t \geq 0)$.

(1) $\tilde{B}(0) = 0$.

(2) If X is a Gaussian random variable with mean 0 and variance $t-s$, $-X$ is also Gaussian with mean 0 and variance $t-s$. Thus, $\tilde{B}(t) - \tilde{B}(s) = -(B(t) - B(s))$ is also Gaussian with mean 0 and variance $(t-s)$. Hence condition (2) is satisfied.

(3) Assume that $0 \leq t_0 \leq t_1 \leq \ldots \leq t_n$. Then, the random variables $-(B(t_k) - B(t_{k-1}))$ are independent for $k = 1, 2, 3, \ldots, n$. Hence, condition (3) is satisfied.

(b) Consider the process $\tilde{B}(t)$ as defined in (4.1).

Fix an $s \geq 0$.

(1) Let $t = 0$. Then, $t \leq s$. $\tilde{B}(t) = \tilde{B}(0) = B(0) = 0$.

(2) Let $t_1 < t_2 \leq s$. Then, $\tilde{B}(t_2) - \tilde{B}(t_1) = B(t_2) - B(t_1)$. This is a Gaussian random variable with mean 0 and variance $t_2 - t_1$.

Let $t_1 < s < t_2$. Then, $\tilde{B}(t_2) - \tilde{B}(t_1) = B(s) - (B(t_2) - B(s)) - B(t_1) = (B(s) - B(t_1)) - (B(t_2) - B(s))$. Since, $B(s) - B(t_1)$ and $B(t_2) - B(s)$ are independent Gaussian random variables, any linear combination of these is Gaussian. Moreover, its mean is zero. The variance is given by:

$$
\begin{aligned}
Var[\tilde{B}(t_2) - \tilde{B}(t_1)] &= Var[B(s) - B(t_1)] + Var[B(t_2) - B(s)] \\
&= (s - t_1) + (t_2 - s) \\
&= t_2 - t_1
\end{aligned}
$$

Let $s < t_1 < t_2$. Then,

$$
\begin{aligned}
\tilde{B}(t_2) - \tilde{B}(t_1) &= B_s - (B_{t_2} - B_s) - (B_s - (B_{t_1} - B_s)) \\
&= \cancel{B_s} - (B_{t_2} - \cancel{B_s}) - (\cancel{B_s} - (B_{t_1} - \cancel{B_s})) \\
&= -(B_{t_2} - B_{t_1})
\end{aligned}
$$

Hence, $\tilde{B}(t_2) - \tilde{B}(t_1)$ is again a Gaussian random variable with mean 0 and variance $t_2 - t_1$. Hence, condition (3) is satisfied.

(3) Assume that $0 \leq t_1 \leq \ldots \leq t_{k-1} \leq s \leq t_k \leq \ldots \leq t_n$. From the above discussion, the increments $\tilde{B}(t_2) - \tilde{B}(t_1)$, $\ldots$, $\tilde{B}(s) - \tilde{B}(t_{k-1})$, $\tilde{B}(t_k) - \tilde{B}(s)$, $\ldots$, $\tilde{B}(t_n) - \tilde{B}(t_{n-1})$ are independent increments. The increment $\tilde{B}(t_k) - \tilde{B}(t_{k-1})$ only depends on the random variables $\tilde{B}(s) - \tilde{B}(t_{k-1})$ and $\tilde{B}(t_k) - \tilde{B}(s)$. Thus, $\tilde{B}(t_2) - \tilde{B}(t_1), \ldots, \tilde{B}(t_k) - \tilde{B}(t_{k-1})$, $\ldots$, $\tilde{B}(t_n) - \tilde{B}(t_{n-1})$ are independent. $\square$

Proposition 4.5. *(Time Reversal). Let $(B_t, t \geq 0)$ be a Brownian motion. Show that the process $(B_1 - B_{1-t}, t \in [0,1])$ has the distribution of a standard brownian motion on $[0,1]$.*

Proof. (1) At $t = 0$, $B(1) - B(1-t) = B(1) - B(1) = 0$.

(2) Let $s < t$. Then, $1 - t < 1 - s$. So, the increment :

$$
(B(1) - B(1-t)) - (B(1) - B(1-s)) = B(1-s) - B(1-t)
$$

has a Gaussian distribution. It's mean is 0 and variance is $(1-s) - (1-t) = t - s$.

(3) Let $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$. Then:

$$
1 - t_n \leq \ldots \leq 1 - t_k \leq 1 - t_{k-1} \leq \ldots \leq 1 - t_2 \leq 1 - t_1
$$

Consider the increments of the process for $k = 1, 2, \ldots, n$:

$$(B(1) - B(1 - t_k)) - (B(1) - B(1 - t_{k-1})) = B(1 - t_{k-1}) - B(1 - t_k)$$

They are independent random variables. Hence, condition (3) is satisfied. $\square$

Example 4.1. (Evaluating Brownian Probabilities). Let's compute the probability that $B_1 > 0$ and $B_2 > 0$. We know from the definition that (B_1, B_2) is a Gaussian vector with mean 0 and covariance matrix:

$$C = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

The determinant of C is 1. By performing row operations on the augmented matrix $[C|I]$ we find that:

$$C^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

Thus, the probability $\mathbb{P}(B_1 > 0, B_2 > 0)$ can be expressed as:

$$\mathbb{P}(B_1 > 0, B_2 > 0) = \frac{1}{\sqrt{(2\pi)^2}} \int_0^\infty \int_0^\infty \exp\left[-\frac{1}{2}(2x_1^2 - 2x_1x_2 + x_2^2\right] dx_2 dx_1$$

This integral can be evaluated using a calculator or software and is equal to $3/8$. The probability can also be computed using the independence of increments. The increments $(B_1, B_2 - B_1)$ are IID standard Gaussians. We know their joint PDF. It remains to integrate over the correct region of $\mathbf{R}^2$ which in this case will be:

$$D^* = \{(z_1, z_2) : (z_1 > 0, z_1 + z_2 > 0)\}$$

We have:

$$\mathbb{P}(B_1 > 0, B_2 > 0) = \frac{1}{2\pi} \int_0^\infty \int_{z_2=-z_1}^{z_2=\infty} e^{-(z_1^2+z_2^2)/2} dz_2 dz_1$$

It turns out that this integral can be evaluated exactly. Indeed by writing $B_1 = Z_1$ and $Z_2 = B_2 - B_1$ and splitting the probability on the event $\{Z_2 \geq 0\}$ and its complement, we have that $\mathbb{P}(B_1 \geq 0, B_2 \geq 0)$ equals:

$$\begin{aligned}
\mathbb{P}(B_1 \geq 0, B_2 \geq 0) &= \mathbb{P}(Z_1 \geq 0, Z_1 + Z_2 > 0, Z_2 \geq 0) + \mathbb{P}(Z_1 \geq 0, Z_1 + Z_2 > 0, Z_2 < 0) \\
&= \mathbb{P}(Z_1 \geq 0, Z_2 \geq 0) + \mathbb{P}(Z_1 \geq 0, Z_1 > -Z_2, -Z_2 > 0) \\
&= \mathbb{P}(Z_1 \geq 0, Z_2 \geq 0) + \mathbb{P}(Z_1 \geq 0, Z_1 > Z_2, Z_2 > 0) \\
&= \frac{1}{4} + \frac{1}{8} \\
&= \frac{3}{8}
\end{aligned}$$

Note that, by symmetry, $\mathbb{P}(Z_1 \geq 0, Z_1 > Z_2, Z_2 > 0) = \mathbb{P}(Z_1 \geq 0, Z_1 \leq Z_2, Z_2 > 0) = \frac{1}{8}$.

Example 4.2. (Another look at Ornstein Uhlenbeck process.) Consider the process $(X_t, t \in \mathbf{R})$ defined by :

$$X_t = \frac{e^{-2t}}{\sqrt{2}} B(e^{4t}), \quad t \in \mathbf{R}$$

Here the process $(B_{e^{4t}}, t \geq 0)$ is called a time change of Brownian motion, since the time is now quantitfied by an increasing function of t namely e^{4t}. The example $(B(\lambda t), t \geq 0)$ in the scaling property is another example of time change.

It turns out that $(X_t, t \in \mathbf{R})$ is a stationary Ornstein-Uhlenbeck process. (Here the index of time is $\mathbf{R}$ instead of $[0, \infty)$, but the definition also applies as the process is stationary. Since the original brownian motion $B(t)$ is a Gaussian process, any finite dimensional vector $(B(t_1), \ldots, B(t_n))$ is Gaussian. It follows that:

$$(B(T_1), \ldots, B(T_n)) = \frac{1}{\sqrt{2}}(e^{-2t_1} B(e^{4t_1}), \ldots, e^{-2t_n} B(e^{4t_n}))$$

is also a Gaussian vector. (Note, once we fix $t_1, t_2, \ldots, t_n$, $e^{-4t_1}, \ldots, e^{-4t_n}$ are constants.) Hence, $(X_t, t \in \mathbf{R})$ is a Gaussian process.

The mean of $(X_t, t \in \mathbf{R})$ is:

$$\mathbb{E}[X_t] = \frac{e^{-2t}}{\sqrt{2}} \mathbb{E}[B(e^{4t})] = 0$$

And if $s < t$,

$$\begin{aligned}
\mathbb{E}[X_s X_t] &= \frac{e^{-2(s+t)}}{2} \mathbb{E}[B(e^{4s}) B(e^{4t})] \\
&= \frac{e^{-2(s+t)}}{2} e^{4s} \\
&= \frac{e^{-2(t-s)}}{2}
\end{aligned}$$

Two Gaussian processes having the same mean and covariance have the same distribution. Hence, it proves the claim that (X_t) is a stationary OU process.

4.2 Properties of the paths.

First we review the definitions of the Riemann integral and the Riemann-Stieljtes integral in Calculus.

Definition 4.1. A partition P of $[a, b]$ is a *finite* set of points from $[a, b]$ that includes both $[a, b]$.The notational convention is to always list the points of a partition $P = \{a = x_0, x_1, x_2, \ldots, x_n = b\}$ in increasing order. Thus:

$$a = x_0 < x_1 < \ldots < x_{k-1} < x_k < \ldots < x_n = b$$

For each subinterval $[x_{k-1}, x_k]$ of P, let

$$\begin{aligned}
m_k &= \inf\{f(x) : x \in [x_{k-1}, x_k]\} \\
M_k &= \sup\{f(x) : x \in [x_{k-1}, x_k]\}
\end{aligned}$$

The lower sum of f with respect to P is given by :

$$L(f,P) = \sum_{k=1}^{n} m_k(x_k - x_{k-1})$$

The upper sum of f with respect to P is given by:

$$U(f,P) = \sum_{k=1}^{n} M_k(x_k - x_{k-1})$$

For a particular partition P, it is clear that $U(f,P) \geq L(f,P)$ because $M_k \geq m_k$ for all $k = 0, 1, 2, \ldots, n$.

Definition 4.2. A partition Q is called a *refinement* of P if Q contains all of the points of P; that is $Q \subseteq P$.

Lemma 4.1. *If $P \subseteq Q$, then $L(f,P) \leq L(f,Q)$ and $U(f,Q) \leq U(f,P)$.*

Proof. Consider what happens when we refine P by adding a single point z to some subinterval $[x_{k-1}, x_k]$ of P. We have:

$$\begin{aligned} m_k(x_k - x_{k-1}) &= m_k(x_k - z) + m_k(z - x_{k-1}) \\ &\leq m_k'(x_k - z) + m_k''(z - x_{k-1}) \end{aligned}$$

where

$$\begin{aligned} m_k' &= \inf\{f(x) : x \in [z, x_k]\} \\ m_k'' &= \inf\{f(x) : x \in [x_{k-1}, z]\} \end{aligned}$$

By induction we have:

$$\begin{aligned} L(f,P) &\leq L(f,Q) \\ U(f,Q) &\leq U(f,P) \end{aligned}$$

□

Lemma 4.2. *If P_1 and P_2 are any two partitions of $[a,b]$, then $L(f,P_1) \leq U(f,P_2)$.*

Proof. Let $Q = P_1 \cup P_2$. Then, $P_1 \subseteq Q$ and $P_2 \subseteq Q$. Thus, $L(f,P_1) \leq L(f,Q) \leq U(f,Q) \leq L(f,P_2)$. □

Definition 4.3. Let $\mathcal{P}$ be the collection of all possible partitions of the interval $[a,b]$. The upper integral of f is defined to be:

$$U(f) = \inf\{U(f,P) : P \in \mathcal{P}\}$$

The lower integral of f is defined by:

$$L(f) = \sup\{L(f,P) : P \in \mathcal{P}\}$$

Consider the set of all upper sums of f - $\{U(f, P) : P \in \mathcal{P}\}$. Take an arbitrary partition $P' \in \mathcal{P}$. Since $L(f, P') \leq U(f, P)$ for all $P \in \mathcal{P}$, by the Axiom of Completeness(AoC), $\inf\{U(f, P) : P \in \mathcal{P}\}$ exists.We can similarly argue for the supremum of all lower Riemann sums.

Lemma 4.3. *For any bounded function f on $[a, b]$, it is always the case that $U(f) \geq L(f)$.*

Proof. By the properties of the infimum of a set, $(\forall \epsilon > 0)$, $\exists P(\epsilon)$ such that $U(f) < U(f, P(\epsilon)) < U(f) + \epsilon$. Pick $\epsilon = 1, \frac{1}{2}, \frac{1}{3} \ldots, \frac{1}{n}, \ldots$. Thus, we can produce a sequence of partitions P_n such that :

$$U(f) < \ldots < U(f, P_n) < U(f) + \frac{1}{n}$$

Consequently, $\lim U(f, P_n) = U(f)$. Similarly, we can produce a sequence of partitions (Q_m) such that :

$$L(f) - \frac{1}{m} < \ldots < L(f, Q_m) < L(f)$$

We know that:

$$L(f, Q_m) \leq U(f, P_n)$$

Keeping m fixed and passing to the limit, as $n \to \infty$ on both sides, we have:

$$\begin{aligned} \lim_{n\to\infty} L(f, Q_m) &\leq \lim_{n\to\infty} U(f, P_n) \quad \{\text{Order Limit Theorem}\} \\ L(f, Q_m) &\leq U(f) \end{aligned}$$

Now, passing to the limit, as $m \to \infty$ on both sides, we have:

$$\begin{aligned} \lim_{m\to\infty} L(f, Q_m) &\leq \lim_{m\to\infty} U(f) \quad \{\text{Order Limit Theorem}\} \\ L(f) &\leq U(f) \end{aligned}$$

□

Definition 4.4. (Riemann Integrability). A bounded function f on the interval $[a, b]$ is said to be Riemann integrable if $U(f) = L(f)$. In this case, we define $\int_a^b f$ or $\int_a^b f(x)dx$ to be the common value:

$$\int_a^b f(x)dx = U(f) = L(f)$$

Theorem 4.1. *(Integrability Criterion) A bounded function f is integrable on $[a, b]$ if and only if, for every $\epsilon > 0$, there exists a partition P_ϵ of $[a, b]$ such that:*

$$U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon$$

Proof. ($\Longleftarrow$ direction.) Let $\epsilon > 0$. If such a partition P_ϵ exists, then:

$$U(f) - L(f) \leq U(f, P_\epsilon) - L(f, P_\epsilon) < \epsilon$$

Because ϵ is arbitrary, it follows that $U(f) = L(f)$ and hence f is Riemann integrable.

($\Longrightarrow$ direction.) Let f be a bounded function on $[a, b]$ such that f is Riemann integrable.

Pick an arbitrary $\epsilon > 0$.

Then, since $U(f) = \inf\{U(f, P) : P \in \mathcal{P}\}$, there exists $P_\epsilon \in \mathcal{P}$, such that $U(f) < U(f, P_\epsilon) < U(f) + \frac{\epsilon}{2}$. Since $L(f) = \sup\{L(f, P) : P \in \mathcal{P}\}$, there exists $P_\epsilon \in \mathcal{P}$, such that $L(f) - \frac{\epsilon}{2} < L(f, P_\epsilon) < L(f)$. Consequently,

$$\begin{aligned} U(f, P_\epsilon) - L(f, P_\epsilon) &< U(f) + \frac{\epsilon}{2} - \left(L(f) - \frac{\epsilon}{2}\right) \\ &= U(f) - L(f) + \epsilon \\ &= \epsilon \end{aligned}$$

□

4.2.1 Functions considered in Stochastic Calculus.

Definition 4.5. A point c is called a discontinuity of the first kind or jump point if both limits $g(c+) = \lim_{t \uparrow c} g(t)$ and $g(c-) = \lim_{t \downarrow c} g(t)$ exist and are not equal. The jump at c is defined as $\Delta g(c) = g(c+) - g(c-)$. Any other discontinuity is said to be of the second kind.

Example 4.3. Consider the function

$$f(x) = \sin\left(\frac{1}{x}\right)$$

Let $x_n = \frac{1}{2n\pi}$. Then, $f(x_n) = (0, 0, 0, \ldots)$. Next, consider $y_n = \frac{1}{\pi/2 + 2n\pi}$. Then, $f(y_n) = (1, 1, 1, \ldots)$. Consequently, f is not continuous at 0. Hence, limits from the left or right don't exist. Consequently, this is a discontinuity of the second kind.

Functions in stochastic calculus are functions without discontinuities of the second kind, that is functions have both left and right hand limits at any point of the domain and have one-sided limits at the boundary. These functions are called *regular* functions. It is often agreed to identify functions if they have the same right and left limits at any point.

The class $D = D[0, T]$ of right-continuous functions on $[0, T]$ with left limits has a special name, *cadlag* functions (which is the abbreviation of right continuous with left limits in French). Sometimes these processes are called R.R.C. for regular right continuous. Notice that this class of processes includes C, the class of continuous functions.

Let $g \in D$ be a cadlag function, then, by definition, all the discontinuities of g are jumps. An important result in analysis is that, a function can have no more than a countable number of discontinuities.

4.2.2 Variation of a function.

If g is a function of a real variable, its variation over the interval $[a, b]$ is defined as:

$$V_g([a, b]) = \sup\left\{\sum_{i=1}^{n} |g(t_i) - g(t_{i-1})|\right\} \tag{4.2}$$

where the supremum is taken over all partitions $P \in \mathcal{P}$.

Clearly, by the Triangle Inequality, the sums in (4.2) increase as new points are added to the partitions. Therefore, the variation of g is:

$$V_g([a,b]) = \lim_{||\Delta_n|| \to 0} \sum_{i=1}^{n} |g(t_i) - g(t_{i-1})|$$

where $||\Delta_n|| = \max_{1 \le i \le n}(t_i - t_{i-1})$. If $V_g([a,b])$ is finite, then g is said to be a function of finite variation on $[a,b]$. If g is a function of $t \ge 0$, then the variation of g as a function of t is defined by:

$$V_g(t) = V_g([0,t])$$

Clearly, $V_g(t)$ is an increasing function of t.

Definition 4.6. g is a function of finite variation if $V_g(t) < \infty$ for all $t \in [0,\infty)$. g is of bounded variation if $\sup_t V_g(t) < \infty$, in other words there exists C, for all t, such that $V_g(t) < C$. Here C is independent of t.

Example 4.4. (1) If $g(t)$ is increasing then for any i, $g(t_i) \ge g(t_{i-1})$, resulting in a telescopic sum, where all terms excluding the first and the last cancel out, leaving

$$V_g(t) = g(t) - g(0)$$

(2) If $g(t)$ is decreasing, then similarly,

$$V_g(t) = g(0) - g(t)$$

Example 4.5. If $g(t)$ is differentiable with continuous derivative $g'(t)$, $g(t) = \int_0^t g'(s)ds$ then

$$V_g(t) = \int_0^t |g'(s)|ds$$

Proof. By definition,

$$V_g(t) = \lim_{||\Delta_n \to 0||} \sum_{i=1}^{n} |g(t_i) - g(t_{i-1})|$$

Since g is continuous and differentiable on $[t_{i-1}, t_i]$, there exists $z_i \in (t_{i-1}, t_i)$ such, that $g(t_i) - g(t_{i-1}) = g'(z_i)(t_i - t_{i-1})$. Therefore, we can write:

$$\begin{aligned} V_g(t) &= \lim_{||\Delta_n \to 0||} \sum_{i=1}^{n} |g'(z_i)|(t_i - t_{i-1}) \\ &= \int_0^t |g'(s)|ds \end{aligned}$$

□

Theorem 4.2. *If g is continuous, g' exists and $\int_0^t |g'(s)|ds$ is finite, then g is of finite variation.*

Example 4.6. The function $g(t) = t\sin(1/t)$ for $t > 0$ and $g(0) = 0$ is continuous on $[0, 1]$ and differentiable at all points except zero, but is not of bounded variation on any interval that includes 0. Consider the partition $\{x_n\} = \left\{\frac{1}{\pi/2+n\pi}\right\}$. Thus,

$$\sin(\frac{1}{x_n}) = \begin{cases} 1 & \text{if } n \text{ is even} \\ -1 & \text{if } n \text{ is odd} \end{cases}$$

Thus,

$$f(x_n) = \begin{cases} x_n & n \text{ is even} \\ -x_n & n \text{ is odd} \end{cases}$$

Therefore,

$$\begin{aligned}
\sum_{n=1}^{m} |f(x_n) - f(x_{n-1})| &= \sum_{n=1}^{m} (x_n + x_{n-1}) \\
&= x_0 + x_n + 2\sum_{n=1}^{m-1} x_n \\
&\geq \sum_{n=1}^{m-1} x_n
\end{aligned}$$

This is the lower bound on the variation of g on the partition $\{0, x_m, \ldots, x_1, x_0, 1\}$. Now, passing to the limit as m approaches infinity, $\sum \frac{1}{\pi/2+n\pi}$ is a divergent series. Consequently, $V_g([0, 1])$ has unbounded variation.

4.2.3 Jordan Decomposition.

Theorem 4.3. *Any function $g : [0, \infty) \to \mathbf{R}$ is of bounded variation if and only if it can be expressed as the difference of two increasing functions:*

$$g(t) = a(t) - b(t)$$

Proof. ($\Longrightarrow$direction). If g is of finite variation, $V_g(t) < \infty$ for all t, and we can write:

$$g(t) = V_g(t) - (V_g(t) - g(t))$$

Let $a(t) = V_g(t)$ and $b(t) = V_g(t) - g(t)$. Clearly, both $a(t)$ and $b(t)$ are increasing functions.

($\Longleftarrow$direction). Suppose a function g can be expressed as a difference of two bounded increasing functions. Then,

$$\begin{aligned}
V_g(t) &= \lim_{||\Delta_n|| \to 0} \sum_{i=1}^{n} |(a(t_i) - b(t_i)) - (a(t_{i-1}) - b(t_{i-1})| \\
&\{ \text{ Telescoping sum } \} \\
&= a(t) - b(t) - (a(0) - b(0))
\end{aligned}$$

Since both $a(t)$ and $b(t)$ are bounded, g has bounded variation. $\square$

4.2.4 Riemann-Stieltjes Integral.

Let g be a montonically increasing function on a finite closed interval $[a, b]$. A bounded function f defined on $[a, b]$ is said to *Riemann-Stieltjes integrable* with respect to g if the following limit exists:

$$\int_a^b f(t)dg(t) = \lim_{||\Delta_n|| \to 0} \sum_{i=1}^{n} f(\tau_i)(g(t_i) - g(t_{i-1})) \tag{4.3}$$

where τ_i is an evaluation point in the interval $[t_{i-1}, t_i]$. It is a well-known fact that continuous functions are Riemann integrable and Riemann-Stieltjes integrable with respect to any monotonically increasing function on $[a, b]$.

We ask the following question. For any continuous functions f and g on $[a, b]$, can we define the integral $\int_a^b f(t)dg(t)$ by Equation (4.3)?

Consider the special case $f = g$, namely, the integral:

$$\int_a^b f(t)df(t)$$

Let $\Delta_n = \{a = t_0, t_1, \ldots, t_n = b\}$ be a partition of $[a, b]$. Let L_n and R_n denote the corresponding Riemann sums with the evaluation points $\tau_i = t_{i-1}$ and $\tau_i = t_i$, respectively, namely,

$$L_n = \sum_{i=1}^{n} f(t_{i-1})(f(t_i) - f(t_{i-1})) \tag{4.4}$$

$$R_n = \sum_{i=1}^{n} f(t_i)(f(t_i) - f(t_{i-1})) \tag{4.5}$$

Is it true that, $\lim L_n = \lim R_n$ as $||\Delta_n|| \to 0$? Observe that:

$$R_n - L_n = \sum_{i=1}^{n} (f(t_i) - f(t_{i-1}))^2 \tag{4.6}$$

$$R_n + L_n = \sum_{i=1}^{n} (f(t_i)^2 - f(t_{i-1})^2) = f(b)^2 - f(a)^2 \tag{4.7}$$

Therefore, R_n and L_n are given by:

$$R_n = \frac{1}{2}\left(f(b)^2 - f(a)^2 + \sum_{i=1}^{n} (f(t_i) - f(t_{i-1}))^2\right) \tag{4.8}$$

$$L_n = \frac{1}{2}\left(f(b)^2 - f(a)^2 - \sum_{i=1}^{n} (f(t_i) - f(t_{i-1}))^2\right) \tag{4.9}$$

The limit of the right-hand side of equation (4.6) is called the *quadratic variation* of the function f on $[a, b]$. Obviously, $\lim_{||\Delta_n|| \to 0} R_n \neq \lim_{||\Delta_n|| \to 0} L_n$ if and only the quadratic variation of the function f is non-zero.

Example 4.7. Let f be a C^1-function that is $f'(t)$ is a continuous function. Then, by the mean value theorem:

$$
\begin{aligned}
|R_n - L_n| &= \sum_{i=1}^{n} (f(t_i) - f(t_{i-1}))^2 \\
&= \sum_{i=1}^{n} (f'(t_i^*)(t_i - t_{i-1}))^2 \\
&\quad \{\text{Mean Value Theorem}\} \\
&\leq \sum_{i=1}^{n} \|f'\|_\infty^2 (t_i - t_{i-1})^2 \\
&\quad \{ \text{ Interior Extremum Theorem } \} \\
&\leq \|f'\|_\infty^2 \|\Delta_n\| \sum_{i=1}^{n} (t_i - t_{i-1}) \\
&= \|f'\|_\infty^2 \|\Delta_n\| (b-a)
\end{aligned}
$$

where $\|f'\|_\infty = \sup_{x \in [a,b]} f(x)$. Thus, the limit as $\|\Delta_n\| \to 0$ of the distance $|R_n - L_n|$ also approaches zero. Thus, $\lim L_n = \lim R_n$ as $\|\Delta_n\| \to 0$ and the Riemann-Stieltjes integral exists. By equation (4.7), we have:

$$
\lim_{\|\Delta_n\| \to 0} L_n = \lim_{\|\Delta_n\| \to 0} R_n = \frac{1}{2}(f(b)^2 - f(a)^2) \tag{4.10}
$$

On the other hand, for such a C^1-function f, we may simply define the integral $\int_a^b f(t) df(t)$ by:

$$
\int_a^b f(t) df(t) = \int_a^b f(t) f'(t) dt
$$

Then, by the fundamental theorem of Calculus:

$$
\int_a^b f(t) df(t) = \int_a^b f(t) f'(t) dt = \frac{1}{2} f(t)^2 |_a^b = \frac{1}{2}(f(b)^2 - f(a)^2)
$$

Remark. There is a very close relationship between functions with bounded variation and functions for which the classical integral makes sense. For the Ito integral, the quadratic variation plays a similar role. The quadratic variation of a smooth fuction $f \in C^1([0,t])$ is zero.

Example 4.8. Suppose f is a continuous function satisfying the condition

$$
|f(t) - f(s)| \leq C|t-s|^{1/2}
$$

where $0 < C < 1$.

In this case we have:

$$
0 \leq |R_n - L_n| \leq C^2 \sum_{i=1}^{n} (t_i - t_{i-1}) = C^2(b-a)
$$

Hence, $\lim R_n \neq \lim L_n$ as $\|\Delta_n\| \to 0$ when $a \neq b$. Consequently, the integral $\int_a^b f(t) df(t)$ cannot be defined for such a function f. Observe that the quandratic variation of the function is $b - a$ (non-zero).

We see from the above examples, that definining the integral $\int_a^b f(t)dg(t)$ even when $f = g$ is a non-trivial problem. Consider the question posed earlier - if f and g are continuous functions on $[a, b]$, can we define the integral $\int_a^b f(t)dg(t)$? There is no simple answer to this question. But then in view of example (4.8), we can ask another question:

Question. Are there continuous functions f satisfying the condition

$$|f(t) - f(s)| \leq C|t - s|^{1/2}$$

4.2.5 Brownian motion as the limit of a symmetric random walk.

Consider a random walk starting at 0 with jumps h and $-h$ equally at times $\delta, 2\delta, \ldots$ where h and δ are positive numbers. More precisely, let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent and identically distributed random variables with :

$$\mathbb{P}\{X_j = h\} = \mathbb{P}\{X_j = -h\} = \frac{1}{2}$$

Let $Y_{\delta,h}(0) = 0$ and put:

$$Y_{\delta,h}(n\delta) = X_1 + X_2 + \ldots + X_n$$

For $t > 0$, define $Y_{\delta,h}(t)$ by linearization that is, for $n\delta < t < (n+1)\delta$, define:

$$Y_{\delta,h}(t) = \frac{(n+1)\delta - t}{\delta} Y_{\delta,h}(n\delta) + \frac{t - n\delta}{\delta} Y_{\delta,h}((n+1)\delta)$$

We can think of $Y_{\delta,h}(t)$ as the position of the random walk at time t. In particular, $X_1 + X_2 + \ldots + X_n$ is the position of this random walk at time $n\delta$.

Question. What is the limit of the random walk $Y_{\delta,h}$ as $\delta, h \to 0$?

Recall that the characteristic function of a random variable X is $\phi_X(\lambda) = \mathbb{E}\exp[i\lambda X]$. In order to find out the answer, let us compute the following limit of the characteristic function of $Y_{\delta,h}(t)$:

$$\lim_{\delta,h\to 0} \mathbb{E}\exp\left[i\lambda Y_{\delta,h}(t)\right]$$

where $\lambda \in$ **R**is fixed. For heuristic derivation, let $t = n\delta$ and so $n = t/\delta$. Then we have:

$$\begin{aligned}
\mathbb{E}\exp\left[i\lambda Y_{\delta,h}(t)\right] &= \prod_{j=1}^{n} \mathbb{E}e^{i\lambda X_j} \\
&= \prod_{j=1}^{n} \left(\frac{1}{2}e^{i\lambda h} + \frac{1}{2}e^{-i\lambda h}\right) \\
&= \left(\frac{1}{2}e^{i\lambda h} + \frac{1}{2}e^{-i\lambda h}\right)^n \\
&= (\cos \lambda h)^n \\
&= (\cos \lambda h)^{t/\delta}
\end{aligned}$$

For fixed t and λ, when δ and h independently approach 0, the limit of $\mathbb{E}\exp\left[i\lambda Y_{\delta,h}(t)\right]$ may not exist. For example, holding h constant, letting $\delta \to 0$, since $-1 \leq \cos\theta \leq 1$, the function $(\cos\lambda h)^{t/\delta} \to 0$. Holding δ constant, letting $h \to 0$, the function $(\cos\lambda h)^{t/\delta} \to 1$. In order for the limit to exist, we impose a certain relationship between δ and h. However, depending on the relationship, we may obtain different limits.

Let $u = \cos(\lambda h)^{1/\delta}$. Then $\ln u = \frac{1}{\delta}\ln\cos(\lambda h)$. Note that:

$$\cos(\lambda h) \approx 1 - \frac{1}{2}\lambda^2 h^2$$

And $\ln(1+x) \approx x$. Hence,

$$\ln\cos(\lambda h) \approx \ln\left(1 - \frac{1}{2}\lambda^2 h^2\right) \approx -\frac{1}{2}\lambda^2 h^2$$

Therefore for small λ and h, we have $\ln u \approx -\frac{1}{2\delta}\lambda^2 h^2$ and so:

$$u \approx \exp\left[-\frac{1}{2\delta}\lambda^2 h^2\right]$$

In particular, if δ and h are related by $h^2 = \delta$, then

$$\lim_{\delta\to 0}\mathbb{E}\exp\left[i\lambda Y_{\delta,h}(t)\right] = e^{-\frac{1}{2}\lambda^2 t}$$

But, $e^{-\frac{1}{2}\lambda^2 t}$ is the characteristic function of a Gaussian random variable with mean 0 and variance t. Thus, we have derived the following theorem about the limit of the random walk $Y_{\delta,h}$ as $\delta, h \to 0$ in such a way that $h^2 = \delta$.

Theorem 4.4. *Let $Y_{\delta,h}(t)$ be the random walk starting at 0 with jumps h and $-h$ equally likely at times δ, 2δ, 3δ, ... Assume that $h^2 = \delta$. Then, for each $t \geq 0$, the limit:*

$$\lim_{\delta\to 0} Y_{\delta,h}(t) = B(t)$$

exists in distribution. Moreover, we have:

$$\mathbb{E}e^{i\lambda B(t)} = e^{-\frac{1}{2}\lambda^2 t}$$

Theorem 4.5. *(Quadratic Variation of a Brownian motion). Let $(B_t, t \geq 0)$ be a standard brownian motion. Then, for any sequence of partitions $(t_j, j \leq n)$ of $[0,t]$ we have:*

$$\langle B\rangle_t = \sum_{j=1}^{n}(B_{t_{j+1}} - B_{t_j})^2 \xrightarrow{L^2} t$$

where the convergence is in the L^2 sense.

Remark. It is reasonable to have some sort of convergence as we are dealing with a sum of independent random variables. However, the conclusion would not hold if the increments were not squared. So there is something more at play here.

Proof. We have:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2-t\right)^2\right] &= \mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2-\sum_{j=0}^{n-1}(t_{j+1}-t_j)\right)^2\right]\\
&= \mathbb{E}\left[\left(\sum_{j=0}^{n-1}\left\{(B(t_{j+1})-B(t_j))^2-(t_{j+1}-t_j)\right\}\right)^2\right]
\end{aligned}$$

For simplicity, we define the variables $X_j = (B(t_{j+1})-B(t_j))^2-(t_{j+1}-t_j)$. Then, we may write:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2-t\right)^2\right] &= \mathbb{E}\left[\left(\sum_{j=0}^{n-1}X_j\right)^2\right]\\
&= \mathbb{E}\left[\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}X_iX_j\right]\\
&= \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}[X_iX_j]
\end{aligned}$$

Now, the random variables X_j are independent.

The expectation of X_j is $\mathbb{E}[X_j] = \mathbb{E}(B(t_{j+1})-B(t_j))^2-(t_{j+1}-t_j) = 0$.

Since, X_i and X_j are independent, for $i \neq j$, $\mathbb{E}[X_iX_j] = \mathbb{E}X_i \cdot \mathbb{E}X_j = 0$.

Hence, we have:

$$\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2-t\right)^2\right] = \sum_{i=0}^{n-1}\mathbb{E}[X_i^2]$$

We now develop the expectation of the square of X_i. We have:

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

The MGF of the random variable $B(t_{i+1})-B(t_i)$ is :

$$
\begin{aligned}
\phi(\lambda) &= \exp\left[\frac{\lambda^2(t_{i+1}-t_i)}{2}\right] \\
\phi'(\lambda) &= \lambda(t_{i+1}-t_i)\exp\left[\frac{\lambda^2(t_{i+1}-t_i)}{2}\right] \\
\phi''(\lambda) &= \left[(t_{i+1}-t_i)+\lambda^2(t_{i+1}-t_i)^2\right]\exp\left[\frac{\lambda^2(t_{i+1}-t_i)}{2}\right] \\
\phi^{(3)}(\lambda) &= \left[3\lambda(t_{i+1}-t_i)^2+\lambda^3(t_{i+1}-t_i)^3\right]\exp\left[\frac{\lambda^2(t_{i+1}-t_i)}{2}\right] \\
\phi^{(4)}(\lambda) &= \left[3(t_{i+1}-t_i)^2+6\lambda^2(t_{i+1}-t_i)^3+\lambda^4(t_{i+1}-t_i)^4\right]\exp\left[\frac{\lambda^2(t_{i+1}-t_i)}{2}\right]
\end{aligned}
$$

Thus, $\mathbb{E}[(B(t_{i+1})-B(t_i))^4] = 3(t_{i+1}-t_i)^2$. Consequently,

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

Putting all this together, we finally have that:

$$
\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2 - t\right)^2\right] &= 2\sum_{i=0}^{n-1}(t_{i+1}-t_i)^2 \\
&\leq 2\left\|\Delta_n\right\|\sum_{i=0}^{n-1}(t_{i+1}-t_i) \\
&= 2\left\|\Delta_n\right\|\cdot t
\end{aligned}
\tag{4.11}
$$

As $n \to \infty$, $\|\Delta_n\| \to 0$. Hence,

$$
\lim_{n\to\infty}\mathbb{E}\left[\left(\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2 - t\right)^2\right] = 0
$$

Hence, the sequence of random variables

$$
\sum_{j=0}^{n-1}(B(t_{j+1})-B(t_j))^2 \xrightarrow{L^2} t
$$

□

Corollary 4.1. *(Quadratic Variation of a Brownian Motion Path). Let* $(B_s, s \geq 0)$ *be a Brownian motion. For every* $n \in \mathbf{N}$*, consider the dyadic partition* $(t_j, j \leq 2^n)$ *of* $[0,t]$ *where* $t_j = \frac{j}{2^n}t$*. Then we have that:*

$$
\langle B\rangle_t = \sum_{j=1}^{2^n-1}(B_{t_{j+1}} - B_{t_j})^2 \xrightarrow{a.s.} t
$$

Proof. We have $(t_{i+1} - t_i) = \frac{t}{2^n}$. Borrowing equation (4.11) from the proof of theorem (4.5), we have that:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{j=0}^{2^n-1}(B(t_{j+1}) - B(t_j))^2 - t\right)^2\right] &= 2\sum_{i=0}^{2^n-1}\left(\frac{t}{2^n}\right)^2 \\ &= 2\cdot(2^n)\cdot\frac{t^2}{2^{2n}} \\ &= \frac{2t^2}{2^n}\end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{j=0}^{2^n-1}(B(t_{j+1}) - B(t_j))^2 - t\right| > \epsilon\right) &\leq \frac{1}{\epsilon^2}\mathbb{E}\left[\left(\sum_{j=0}^{2^n-1}(B(t_{j+1}) - B(t_j))^2 - t\right)^2\right] \\ &\leq \frac{1}{\epsilon^2}\cdot\frac{2t^2}{2^n}\end{aligned}$$

Define $A_n := \left\{\left|\sum_{j=0}^{2^n-1}(B(t_{j+1}) - B(t_j))^2 - t\right| > \epsilon\right\}$. Since, $\sum \frac{1}{2^n}$ is a convergent series, any multiple of it, $(2t^2/\epsilon^2)\sum \frac{1}{2^n}$ also converges. Now, $0 \leq \mathbb{P}(A_n) \leq \frac{(2t^2/\epsilon^2)}{2^n}$. By the comparison test, $\sum \mathbb{P}(A_n)$ converges to a finite value. By Theorem (3.22),

$$\sum_{j=0}^{2^n-1}(B(t_{j+1}) - B(t_j))^2 \overset{a.s.}{\to} t$$

□

We are now ready to show that every Brownian motion path has infinite variation.

If g is a C^1 function,

$$\begin{aligned}\int_0^t |g'(t)|dt &= \int_0^t \sqrt{g'(t)^2}dt \\ &\leq \int_0^t \sqrt{1 + g'(t)^2}dt \\ &= l_g(t)\end{aligned}$$

where $l_g(t)$ is the arclength of the function g between $[0, t]$. So, $V_g(t) \leq l_g(t)$ and further:

$$\begin{aligned}l_g(t) &= \int_0^t \sqrt{1 + g'(t)^2}dt \\ &\leq \int_0^t \left(1 + \sqrt{g'(t)^2}\right)dt \\ &\leq t + V_g(t)\end{aligned}$$

Consequently,

$$V_g(t) \le l_g(t) \le t + V_g(t)$$

The total variation of the function is finite if and only if it's arclength is.

Hence, intuitively, our claim is that a Brownian motion path on $[0, T]$ has infinite arc-length. Since $g \in C^1([a, b]) \implies (V_g(t) < \infty)$, it follows that $(V_g(t) \to \infty) \implies g \notin C^1$.

Corollary 4.2. *(Brownian Motion paths have unbounded total variation.) Let* $(B_s, s \ge 0)$ *be a Brownian motion. Then, the random functions* $B(s, \omega)$ *on the interval* $[0, t]$ *have unbounded variation almost surely.*

Proof. Take the sequence of dyadic partitions of $[0, t]$: $t_j = \frac{j}{2^n} t$, $n \in \mathbf{N}$, $j \le 2^n$. By pulling out the worst increment, we have the trivial bound for every ω:

$$\sum_{j=0}^{2^n-1} (B_{t_{j+1}}(\omega) - B_{t_j}(\omega))^2 \le \max_{0 \le j \le 2^n} \left| B_{t_{j+1}}(\omega) - B_{t_j}(\omega) \right| \cdot \sum_{j=0}^{2^n-1} (B_{t_{j+1}}(\omega) - B_{t_j}(\omega)) \tag{4.12}$$

We proceed by contradiction. Let A' be the set of all ω, for which the Brownian motion paths have bounded total variation. Let A be event that the Brownian motion paths have unbounded variation.

By the definition of total variation, that would imply, $\exists M \in \mathbf{N}$:

$$(\forall \omega \in A') \quad \lim_{n \to \infty} \sum_{j=0}^{2^n-1} \left| (B_{t_{j+1}}(\omega) - B_{t_j}(\omega)) \right| < M$$

Since Brownian Motion paths are continuous on the compact set $[\frac{j}{2^n} t, \frac{j+1}{2^n} t]$, they are uniformly continuous. So, as $n \to \infty$, $|t_{j+1} - t_j| \to 0$ and therefore $|B_{t_{j+1}}(\omega) - B_{t_j}(\omega)| \to 0$. And consequently, $\max_{0 \le j \le 2^n} \left| B_{t_{j+1}}(\omega) - B_{t_j}(\omega) \right| \to 0$.

Thus, for every $\omega \in A'$, the right hand side of the inequality (4.12), converges to 0 and therefore the left hand side converges to 0. But, this contradicts the fact that $\langle B \rangle_t \xrightarrow{a.s.} t$. So, A' is a null set, and $\mathbb{P}(A') = 0$ and $\mathbb{P}(A) = 1$. This closes the proof. □

4.3 What exactly is $(\Omega, \mathcal{F}, \mathbb{P})$ in mathematical finance?

If we make the simplifying assumption that the process paths are continuous, we obtain the set of all continuous functions on $[0, T]$, denoted by $C[0, T]$. This is a very rich space. In a more general model, it is assumed that the process paths are right continuous with left limits (regular right-continuous RRC, cadlag) functions.

Let the sample space $\Omega = D[0, T]$ be the set of all RRC functions on $[0, T]$. An element of this set is a RRC function from $[0, T]$ into **R**. First we must decide what kind of sets of these functions are measurable? The simplest set for which we would like to calculate the probabilities are sets of the form $\{a \le S(t_1) \le b\}$ for some t_1. If $S(t)$ represents the price of a stock at time t, then the probability of such a set gives the probability that the stock price at time t_1 is between a and b. We are also interested in how the price of the stock at time t_1 affects the price at another time t_2. Thus, we need to talk about the joint distribution of stock prices $S(t_1)$ and $S(t_2)$. This means that we need to define probability on the sets of the form $\{S(t_1) \in B_1, S(t_2) \in B_2\}$ where B_1 and B_2 are intervals on the line. More generally, we would like to have all the finite-dimensional distributions of the process $S(t)$, that is, the probabilities of the sets: $\{S(t_1) \in B_1, S(t_2) \in B_2, \ldots, S(t_n) \in B_n\}$ for any choice of $0 \le t_1 \le \ldots \le t_n \le T$.

The sets of the form $A = \{\omega(\cdot) \in D[0, T] : \omega(t_1) \in B_1, \ldots, \omega(t_n) \in B_n\}$, where B_i's are borel subsets of **R**, are called cylinder sets or finite-dimensional rectangles.

The stochastic process $S(t)$ is just a (function-valued) random variable on this sample space, which takes some value $\omega(t)$ - the value of the function ω at t.

Let $\mathcal{R}$ be the colllection of all cylindrical subsets of $D[0,1]$. Obviously $\mathcal{R}$ is not a σ-field.

Probability is first defined by on the elements of $\mathcal{R}$. Let $A \subseteq \mathcal{R}$.

$$\mathbb{P}(A) = \int_{B_1} \cdots \int_{B_n} \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)(t_i - t_{i-1})}} \exp\left[-\frac{(u_i - u_{i-1})^2}{2(t_i - t_{i-1})}\right] du_1 \cdots du_n$$

and then extended to the σ-field generated by taking unions, complements and intersections of cylinders. We take the smallest σ-algebra containing all the cylindrical subsets of $D[0,1]$. Thus, $\mathcal{F} = \mathcal{B}(D[0,1])$.

Hence, $(\Omega, \mathcal{F}, \mathbb{P}) = (D[0,1], \mathcal{B}(D[0,1]), \mathbb{P})$ is a probability space. It is called the *Wiener space* and $\mathbb{P}$ here is called the *Wiener measure*.

4.4 Continuity and Regularity of paths.

As discussed in the previous section, a stochastic process is determined by its finite-dimensional distribution. In studying stochastic processes, it is often natural to think of them as function-valued random variables in t. Let $S(t)$ be defined for $0 \leq t \leq T$, then for a fixed ω, it is a function in t, called the sample path or a realization of S. Finite-dimensional distributions do not determine the continuity property of sample paths. The following example illustrates this.

Example 4.9. Let $X(t) = 0$ for all t, $0 \leq t \leq 1$ and τ be a uniformly distributed random variable on $[0,1]$. Let $Y(t) = 0$ for $t \neq \tau$ and $Y(t) = 1$ if $t = \tau$. Then, for any fixed t, $\mathbb{P}(Y(t) \neq 0) = \mathbb{P}(\tau = t) = 0$, and hence $\mathbb{P}(Y(t) = 0) = 1$. So, that all one-dimensional distributions of $X(t)$ and $Y(t)$ are the same. Similarly, all finite-dimensional distributions of X and Y are the same. However, the sample paths of the process X, that is, the functions $X(t)_{0 \leq t \leq 1}$ are continuous in t, whereas every sample path $Y(t)_{0 \leq t \leq 1}$ has a jump at the (random) point τ. Notice that, $\mathbb{P}(X(t) = Y(t)) = 1$ for all t, $0 \leq t \leq 1$.

Definition 4.7. Two stochastic processes are called *versions* (modifications) of one another if

$$\mathbb{P}(X(t) = Y(t)) = 1 \quad \text{for all } 0 \leq t \leq T$$

Thus, the two processes in the example (4.9) are versions of one another, one has continuous sample paths, the other does not. If we agree to pick any version of the process we want, then we can pick the continous version when it exists. In general, we choose the smoothest possible version of the process.

For two processes, X and Y, denote by $N_t = \{X(t) \neq Y(t)\}, 0 \leq t \leq T$. In the above example, $\mathbb{P}(N_t) = \mathbb{P}(\tau = t) = 0$ for any t, $0 \leq t \leq 1$. However, $\mathbb{P}(\bigcup_{0 \leq t \leq 1} N_t) = \mathbb{P}(\tau = t \text{ for some } t \text{ in } [0,1]) = 1$. Although, each of N_t is a $\mathbb{P}$-null set, the union $N = \bigcup_{0 \leq t \leq 1} N_t$ contains uncountably many null sets, and in this particular case it is a set of of probability one.

If it happens that $\mathbb{P}(N) = 0$, then N is called an *evanescent set*, and the processes X and Y are called *indistinguishable*. Note that in this case, $\mathbb{P}(\{\omega : \exists t : X(t) \neq Y(t)\}) = \mathbb{P}(\bigcup_{0 \leq t \leq 1}\{X(t) \neq Y(t)) = 0$ and $\mathbb{P}(\bigcap_{0 \leq t \leq 1}\{X(t) = Y(t)\}) = 1$. It is clear, that if the time is discrete, then any two versions of the process are indistinguishable. It is also not hard to see, that if $X(t)$ and $Y(t)$ are versions of one another and they are both right-continuous, they are indistinguishable.

Theorem 4.6. *(Paul Levy's construction of Brownian Motion). Standard Brownian motion exists.*

Proof. I reproduce the standard proof as present in *Brownian Motion* by Morters and Peres. I added some remarks for greater clarity.

Let

$$\mathcal{D}_n = \left\{\frac{k}{2^n} : k = 0, 1, 2, \ldots, 2^n\right\}$$

be a finite set of dyadic points.

Let

$$\mathcal{D} = \bigcup_{n=0}^{\infty} \mathcal{D}_n$$

Let $\{Z_t : t \in \mathcal{D}\}$ be a collection of independent, standard normally distributed random variables. This is a countable set of random variables.

Let $B(0) := 0$ and $B(1) := Z_1$.

For each $n \in \mathbf{N}$, we define the random variables $B(d)$, $d \in \mathcal{D}_n$ such that, the following invariant holds:

(1) for all $r < s < t$ in $\mathcal{D}_n$ the random variable $B(t) - B(s)$ is normally distributed with mean zero and variance $t - s$ and is independent of $B(s) - B(r)$.

(2) the vectors $(B(d) : d \in \mathcal{D}_n)$ and $(Z_t : t \in \mathcal{D} \setminus \mathcal{D}_n)$ are independent.

Note that we have already done this for $\mathcal{D}_0 = \{0, 1\}$. Proceeding inductively, let's assume that the above holds for some $n - 1$. We are interested to prove that the invariant also holds for n.

We define $B(d)$ for $d \in \mathcal{D}_n \backslash \mathcal{D}_{n-1}$ by:

$$B(d) = \frac{B(d - 2^{-n}) + B(d + 2^{-n})}{2} + \frac{Z_d}{2^{(n+1)/2}}$$

Note that, the points $0, \frac{1}{2^{n-1}}, \ldots, \frac{k}{2^{n-1}}, \frac{k+1}{2^{n-1}}, \ldots, 1$ belong to $\mathcal{D}_{n-1}$. The first summand is the linear interpolation of the values of B at the neighbouring points of d in $\mathcal{D}_{n-1}$. That is,

$$B\left(\frac{2k+1}{2^n}\right) = \frac{B\left(\frac{k}{2^{n-1}}\right) + B\left(\frac{k+1}{2^{n-1}}\right)}{2} + \frac{Z_d}{2^{(n+1)/2}}$$

Since $P(n-1)$ holds, $B(d - 2^{-n})$ and $B(d + 2^{-n})$ are have no dependence on $(Z_t : t \in \mathcal{D} \setminus \mathcal{D}_{n-1})$. Consequently, $B(d)$ has no dependence on $(Z_t : t \in \mathcal{D} \setminus \mathcal{D}_n)$ and the second property is fulfilled.

Moreover, as $\frac{1}{2}[B(d + 2^{-n}) - B(d - 2^{-n})]$ depends only on $(Z_t : t \in \mathcal{D}_{n-1})$, it is independent of $\frac{Z_d}{2^{(n+1)/2}}$. By our induction assumptions, they are both nromally distributed with mean 0 and variance $\frac{1}{2^{(n+1)}}$.

So, their sum and difference random variables

$$\begin{aligned}
B(d) - B(d - 2^{-n}) &= \frac{B(d + 2^{-n}) - B(d - 2^{-n})}{2} + \frac{Z_d}{2^{(n+1)/2}} \\
B(d + 2^{-n}) - B(d) &= \frac{B(d + 2^{-n}) - B(d - 2^{-n})}{2} - \frac{Z_d}{2^{(n+1)/2}}
\end{aligned}$$

are also independent, with mean 0 and variance $\frac{1}{2^n}$ (the variance of independent random variables is the sum of the variances).

Indeed all increments $B(d) - B(d - 2^{-n})$ for $d \in \mathcal{D}_n \setminus \{0\}$ are independent. To see this, it suffices to show that they are pairwise independent. We have seen in the previous paragraph that the pairs $B(d) - B(d - 2^{-n})$ and $B(d + 2^{-n}) - B(d)$ with $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$ are independent. The other possibility is that the increments are over the intervals separated by some $d \in \mathcal{D}_{n-1}$. For concreteness, if n were 3, then the increments, $B_{7/8} - B_{6/8}$ and $B_{5/8} - B_{4/8}$ are seperated by $d = \frac{3}{4} \in \mathcal{D}_2$. Choose $d \in \mathcal{D}_j$ with this property and minimal j, so, the two intervals are contained in $[d - 2^{-j}, d]$ and

$[d, d+2^{-j}]$ respectively. By induction, the increments over these two intervals of length 2^{-j} are independent and the increments over the intervals of length 2^{-n} are constructed from the independent increments $B(d) - B(d-2^{-j})$ and $B(d+2^{-j}) - B(d)$ using a disjoint set of variables $(Z_t : t \in \mathcal{D}_n)$. Hence, they are independent and this implies pairwise independence. This implies the first property. Consequently, the vector of increments $(B(d) - B(d-2^{-n})$ for all $d \in \mathcal{D}_n$ is Gaussian.

Having thus chosen the value of the process on all the dyadic points, we interpolate between them. Formally, we define:

$$F_0(t) = \begin{cases} Z_1 & \text{for } t = 1 \\ 0 & \text{for } t = 0 \\ \text{linear in between} & \end{cases}$$

and for each $n \geq 1$,

$$F_n(t) = \begin{cases} \frac{Z_t}{2^{(n+1)/2}} & \text{for } t \in \mathcal{D} \setminus \mathcal{D}_{n-1} \\ 0 & \text{for } t \in \mathcal{D}_{n-1} \\ \text{linear between consecutive points in } \mathcal{D}_n & \end{cases}$$

These functions are continuous on $[0, 1]$ and for all n and $d \in \mathcal{D}_n$, we have:

$$B(d) = \sum_{i=0}^{n} F_i(d) = \sum_{i=0}^{\infty} F_i(d) \tag{4.13}$$

To see this, assume that above equation holds for all $d \in \mathcal{D}_{n-1}$.

Let's consider the point $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$.

$$\begin{aligned} B(d) &= \frac{B(d-2^{-n}) + B(d+2^{-n})}{2} + \frac{Z_d}{2^{(n+1)/2}} \\ &= \sum_{i=0}^{n-1} \frac{F_i(d-2^{-n}) + F_i(d+2^{-n})}{2} + \frac{Z_d}{2^{(n+1)/2}} \end{aligned} \tag{4.14}$$

Now, $d - 2^{-n}$ and $d + 2^{-n}$ belong to $\mathcal{D}_{n-1}$ and are not in $\bigcup_{i<n-1} \mathcal{D}_i$. Therefore, for $i = 0, 1, \ldots, n-2$, the points $(d - 2^{-n}, F_i(d - 2^{-n}))$ and $(d + 2^{-n}, F_i(d + 2^{-n})$ lie on some straight line and have $(d, F_i(d))$ as their midpoint. Moreover, $d - 2^{-n}$ and $d + 2^{-n}$ are vertices in $\mathcal{D}_{n-1}$. So, by definition of $F_{n-1}(d)$, we have $F_{n-1}(d) = [F_{n-1}(d - 2^{-n}) + F_{n-1}(d + 2^{-n})]/2$.

To summarize, the first term on the right hand side of expression (4.14) is equal to $\sum_{i=0}^{n-1} F_i(d)$. By mathematical induction, it follows that the claim (4.13) is true for all $n \in \mathbf{N}$.

It's extremely easy to find an upper bound on the probability contained in the Gaussian tails. Suppose $X \sim N(0, 1)$ and let $x > 0$. We are interested in the tail probability $\mathbb{P}(X > x)$. We have:

$$\mathbb{P}(X > x) = \int_x^\infty e^{-x^2/2} dx = \int_x^\infty \frac{x e^{-x^2/2} dx}{x}$$

Let $u = \frac{1}{x}$ and $dv = x e^{-x^2/2} dx$. We have:

$$\begin{array}{c|c} u = \frac{1}{x} & dv = xe^{-x^2/2}dx \\ \hline du = -\frac{1}{x^2}dx & v = -e^{-x^2/2} \end{array}$$

Thus,

$$\begin{aligned}
\mathbb{P}(X > x) &= -\left.\frac{1}{x}e^{-x^2/2}\right|_x^\infty - \int_x^\infty \frac{e^{-x^2/2}}{x^2}dx \\
&= \frac{e^{-x^2/2}}{x} - \int_x^\infty \frac{e^{-x^2/2}}{x^2}dx \\
&\left\{I(x) = \int_x^\infty \frac{e^{-x^2/2}}{x^2} \geq 0\right\} \\
&\leq \frac{e^{-x^2/2}}{x}
\end{aligned}$$

Thus, for $c > 1$ and large n, we have:

$$\mathbb{P}(|Z_d| \geq c\sqrt{n}) \leq \frac{1}{c\sqrt{n}}e^{-c^2n/2} \leq \exp\left(-\frac{c^2n}{2}\right)$$

So, the series:

$$\begin{aligned}
\sum_{n=0}^\infty \mathbb{P}\left\{\text{There exists atleast one } d \in \mathcal{D}_n \text{ with } |Z_d| \geq c\sqrt{n}\right\} &\leq \sum_{n=0}^\infty \sum_{d\in\mathcal{D}_n} \mathbb{P}\left\{|Z_d| \geq c\sqrt{n}\right\} \\
&\leq \sum_{n=0}^\infty (2^n+1)\exp\left(-\frac{c^2n}{2}\right)
\end{aligned}$$

Now, the series (a_n) given by, $a_n := (2^n+1)e^{-c^2n/2}$ has the ratio between successive terms:

$$\begin{aligned}
\lim\left|\frac{a_{n+1}}{a_n}\right| &= \lim_{n\to\infty} \frac{2^{n+1}+1}{2^n+1} \cdot \frac{e^{(c^2n)/2}}{e^{c^2(n+1)/2}} \\
&= \lim_{n\to\infty} \frac{\frac{1}{2}+\frac{1}{2^n}}{1+\frac{1}{2^n}} \cdot \frac{1}{e^{c^2/2}} \\
&= \frac{1}{2e^{c^2/2}}
\end{aligned}$$

If this ratio is less than unity, that is $c > \sqrt{2\log 2}$, than by the ratio test, $\sum(2^n+1)e^{-c^2n/2}$ converges to a finite value. Fix such a c.

By BCL1(Borel-Cantelli Lemma), if $A_n := \{\text{There exists atleast one } d \in \mathcal{D}_n \text{ with } |Z_d| \geq c\sqrt{n}\}$ and $\sum_{n=0}^\infty \mathbb{P}(A_n)$ converges to a finite value, then the event A_n occurs finitely many times with probability 1. There exists $N \in \mathbf{N}$, such that for all $n \geq N$, A_n fails to occur with probability 1. Thus, for all $n \geq N$, $\{Z_d \leq c\sqrt{n}\}$ occurs with probability 1. It follows that:

$$\sup_{t\in[0,1]} F_n(t) \leq \frac{c\sqrt{n}}{2^{(n+1)/2}}$$

Define

$$M_n = \frac{c\sqrt{n}}{2^{(n+1)/2}}$$

Since $\sum M_n$ converges, by the Weierstrass M-test, the infinite series of functions $\sum_{n=0}^{\infty} F_n(t)$ converges uniformly on $[0,1]$. Since, each $F_n(t)$ is piecewise linear and continuous, by the Term-by-Term continuity theorem, $\sum_{n=0}^{\infty} F_n(t)$ is continuous on $[0,1]$. □

4.5 A point of comparison: The Poisson Process.

Like the Brownian motion, the Poisson process is defined as a process with stationary and independent increments.

Definition 4.8. A process $(N_t, t \geq 0)$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ has the distribution of the Poisson process with rate $\lambda > 0$, if and only if the following hold:

(1) $N_0 = 0$.

(2) For any $s < t$, the increment $N_t - N_s$ is a Poisson random variable with parameter $\lambda(t-s)$.

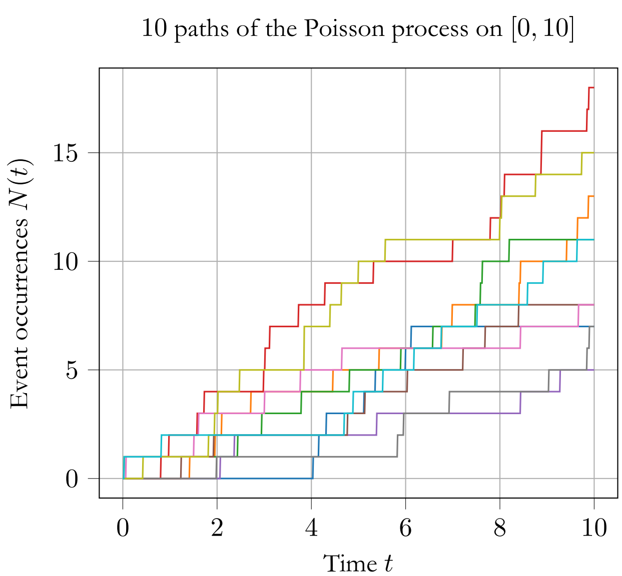
(3) For any $n \in \mathbf{N}$ and any choice $0 < t_1 < t_2 < \ldots < t_n < \infty$, the increments $N_{t_2} - N_{t_1}, N_{t_3} - N_{t_2}, \ldots, N_{t_n} - N_{t_{n-1}}$ are independent.

Poisson paths can be sampled using this definition. By construction, it is not hard to see that the paths of Poisson processes are piecewise, constant, integer-valued and non-decreasing. In particular, the paths of Poisson processes have finite variation. Poisson paths are much simpler than the ones of Brownian motion in many ways!

Example 4.10. (Simulating the Poisson Process.) Use the definition (4.8) to generate 10 paths of the Poisson process with rate 1 on the interval $[0, 10]$ with step-size 0.01.

Listing 8: Generating 10 paths of a Poisson process

```
def generatePoissonProcess(lam,T,stepSize):
    N = int(T/stepSize)
    x = np.random.poisson(lam=lam,size=N)
    y = np.cumsum(x)
    y = np.concatenate([[0.0],y])
    return y
```



We can construct a Poisson process as follows. Consider $(\tau_j, j \in \mathbf{N})$ IID exponential random variables with parameter $1/\lambda$. One should think of τ_j as the waiting time from the $(j-1)$st to the jth jump. Then, one defines :

$$\begin{aligned}
N_t &= \#\{k : \tau_1 + \tau_2 + \ldots + \tau_k \leq t\} \\
&= \text{Number of jumps upto and including time } t
\end{aligned}$$

Now, here is an idea! What about defining a new process with stationary and independent increments using a given distribution other than Poisson and Gaussian? Is this even possible? The answer is yes, but only if the distribution satisfies the property of being *infinitely divisible*. To see this, consider the value of the process at time 1, N_1. Then, no matter how many subintervals we chop the interval $[0, 1]$ into, we must have the increments add up to N_1. In other words, we must be able to write N_1 as a sum of n IID random variables for every possible n. This is certainly true for Poisson random variables and Gaussian random variables. Another example is the Cauchy distribution. In general, processes that can be constructed using independent, stationary increments are called Levy processes.

Example 4.11. Time Inversion. Let $(B_t, t \geq 0)$ be a standard brownian motion. We consider the process:

$$X_t = tB_{1/t} \quad \text{for } t > 0$$

This property relates the behavior of t *large to the behavior of* t *small.*

(a) Show that $(X_t, t > 0)$ has the distribution of Brownian motion on $t > 0$.

Proof.

Like $B(t)$, it is an easy exercise to prove that $X(t)$ is also a Gaussian process.

We have, $\mathbb{E}[X_s] = 0$.

Let $s < t$. We have:

$$\begin{aligned}
Cov(X_s, X_t) &= \mathbb{E}[sB(1/s) \cdot tB(1/t)] \\
&= st\mathbb{E}[B(1/s) \cdot B(1/t)] \\
&= st \cdot \frac{1}{t} \\
&\quad \left\{\because \frac{1}{t} < \frac{1}{s}\right\} \\
&= s
\end{aligned}$$

Consequently, $X(t)$ has the distribution of a Brownian motion.

(b) Argue that $X(t)$ converges to 0 as $t \to 0$ in the sense of L^2-convergence. It is possible to show convergence almost surely so that $(X_t, t \geq 0)$ is really a Brownian motion for $t \geq 0$.

Solution.

Let (t_n) be any arbitrary sequence of positive real numbers approaching 0 and consider the sequence of random variables $(X(t_n))_{n=1}^{\infty}$. We have:

$$\begin{aligned}
\mathbb{E}\left[X(t_n)^2\right] &= \mathbb{E}\left[t_n^2 B(1/t_n)^2\right] \\
&= t_n^2 \mathbb{E}\left[B(1/t_n)^2\right] \\
&= t_n^2 \cdot \frac{1}{t_n} \\
&= t_n
\end{aligned}$$

Hence,

$$\lim \mathbb{E}\left[X(t_n)^2\right] = \lim t_n = 0$$

Since (t_n) was an arbitrary sequence, it follows that $\lim_{t\to 0} \mathbb{E}[(X(t))^2] = 0$.

(c) Use this property of Brownian motion to show the law of large numbers for Brownian motion:

$$\lim_{t\to\infty} \frac{X(t)}{t} = 0 \quad \text{almost surely}$$

Solution.

What we need to do is to show that $X(t) \to 0$ as $t \to 0$ almost surely. That would show that $\frac{B(1/t)}{1/t} \to 0$ as $t \to 0$ almost surely, which is the same as showing $\frac{B(t)}{t} \to 0$ as $t \to \infty$, which is the law of large numbers for Brownian motion.

What we have done in part (b), is to prove the claim that $\mathbb{E}[X(t)^2] \to 0$ as $t \to 0$, which shows convergence in the L^2 sense and hence convergence in probability. This is infact the weak law of large numbers. $\frac{B(t)}{t} \xrightarrow{\mathbf{P}} 0$ as $t \to \infty$.

For $t > 0$, continuity is clear. However, it is the proof that as $t \to 0$, $X(t) \to 0$ almost surely which we have not done.

Note that, the limit $X(t) \to 0$ as $t \to 0$ if and only if $(\forall n \geq 1)$, $(\exists m \geq 1)$, such that $\forall r \in \mathbb{Q} \cap (0, \frac{1}{m}]$, we have $|X(r)| = \left|rB\left(\frac{1}{r}\right)\right| \leq \frac{1}{n}$.

To understand the above, we just recall the $\epsilon - \delta$ definition of continuity. Note that $\frac{1}{n}$ plays the role of ϵ and $\frac{1}{m}$ works as δ.

That is,

$$\Omega^X := \left\{\lim_{t\to 0} X(t) = 0\right\} = \bigcap_{n\geq 1} \bigcup_{m\geq 1} \bigcap_{r\in\mathbb{Q}\cap(0,\frac{1}{m}]} \left\{|X(r)| \leq \frac{1}{n}\right\}$$

Also, note that $X(t)$ is continuous on all $[a, 1]$ for all $a > 0$, thus, uniformly continuous on $[a, 1]$, and hence uniformly continuous on $\mathbb{Q} \cap (0, 1]$. So, there exists a continuous extension of $X(t)$ on $[0, 1]$. We already know from part (a), that $(X(t))_{t>0}$ and $(B(t))_{t>0}$ have the same finite dimensional distributions. Therefore, the RHS event has the same probability as $\Omega^B := \bigcap_{n\geq 1} \bigcup_{m\geq 1} \bigcap_{r\in\mathbb{Q}\cap(0,\frac{1}{m}]} \left\{|B(r)| \leq \frac{1}{n}\right\}$. Since $B(t) \to 0$ as $t \to 0$ almost surely, the event Ω^B has probability 1. Thus, $\mathbb{P}\left\{\lim_{t\to 0} X(t) = 0\right\} = 1$.

This actually shows that $X(t)$ is a bonafide standard brownian motion, as we have established continuity as well.

5 Martingales.

5.1 Elementary conditional expectation.

In elementary probability, the conditional expectation of a variable Y given another random variable X refers to the expectation of Y given the conditional distribution $f_{Y|X}(y|x)$ of Y given X. To illustrate this, let's go through a simple example. Consider $\mathcal{B}_1, \mathcal{B}_2$ to be two independent Bernoulli-distributed random variables with $p = 1/2$. Then, construct:

$$X = \mathcal{B}_1, \quad Y = \mathcal{B}_1 + \mathcal{B}_2$$

It is easy to compute $\mathbb{E}[Y|X = 0]$ and $\mathbb{E}[Y|X = 1]$. By definition, it is given by:

$$
\begin{aligned}
\mathbb{E}[Y|X=0] &= \sum_{j=0}^{2} j\mathbb{P}(Y=j|X=0) \\
&= \sum_{j=0}^{2} j \cdot \frac{\mathbb{P}(Y=j, X=0)}{P(X=0)} \\
&= 0 + 1 \cdot \frac{(1/4)}{(1/2)} + 2 \cdot \frac{0}{(1/2)} \\
&= \frac{1}{2}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[Y|X=1] &= \sum_{j=0}^{2} j\mathbb{P}(Y=j|X=1) \\
&= \sum_{j=0}^{2} j \cdot \frac{\mathbb{P}(Y=j, X=1)}{P(X=1)} \\
&= 0 + 1 \cdot \frac{(1/4)}{(1/2)} + 2 \cdot \frac{(1/4)}{(1/2)} \\
&= \frac{3}{2}
\end{aligned}
$$

With this point of view, the conditional expectation is computed given the information that the event $\{X = 0\}$ occurred or the event $\{X = 1\}$ occurred. It is possible to regroup both conditional expectations in a single object, if we think of the conditional expectation as a random variable and denote it by $\mathbb{E}[Y|X]$. Namely, we take:

$$
\mathbb{E}[Y|X](\omega) = \begin{cases} \frac{1}{2} & \text{if } X(\omega) = 0 \\ \frac{3}{2} & \text{if } X(\omega) = 1 \end{cases} \tag{5.1}
$$

This random variable is called the *conditional expectation* of Y given X. We make two important observations:

(i) If the value of X is known, then the value of $\mathbb{E}[Y|X]$ is determined.

(ii) If we have another random variable $g(X)$ constructed from X, then we have:

$$
\mathbb{E}[g(X)Y] = \mathbb{E}[g(X)\mathbb{E}[Y|X]]
$$

In other words, as far as X is concerned, the conditional expectation $\mathbb{E}[Y|X]$ is a proxy for Y in the expectation. We sometimes say that $\mathbb{E}[Y|X]$ is the best estimate of Y given the information of X.

The last observation is easy to verify since:

$$
\begin{aligned}
\mathbb{E}[g(X)Y] &= \sum_{i=0}^{1}\sum_{j=0}^{2} g(i) \cdot j \cdot \mathbb{P}(X=i, Y=j) \\
&= \sum_{i=0}^{1} \mathbb{P}(X=i)g(i) \left\{ \sum_{j=0}^{2} j \cdot \frac{\mathbb{P}(X=i, Y=j)}{\mathbb{P}(X=i)} \right\} \\
&= \mathbb{E}[g(X)\mathbb{E}[Y|X]]
\end{aligned}
$$

Example 5.1. (Elementary Definitions of Conditional Expectation).

(1) (X, Y) discrete. The treatment is similar to the above. If a random variable X takes values $(x_i, i \geq 1)$ and Y takes values $(y_j, j \geq 1)$, we have by definition that the conditional expectation as a random variable is:

$$\mathbb{E}[Y|X](\omega) = \sum_{j\geq 1} y_j \mathbb{P}(Y = y_j | X = x_i) \quad \text{for } \omega \text{ such that } X(\omega) = x_i$$

(2) (X, Y) continuous with joint PDF $f_{X,Y}(x, y)$: In this case, the conditional expectation is the random variable given by

$$\mathbb{E}[Y|X] = h(X)$$

where

$$h(x) = \int_{\mathbf{R}} y f_{Y|X}(y|x) dy = \int_{\mathbf{R}} y \frac{f_{X,Y}(x, y)}{f_X(x)} dy = \frac{\int_{\mathbf{R}} y f_{X,Y}(x, y) dy}{\int_{\mathbf{R}} f_{X,Y}(x, y) dy}$$

In the two examples above, the expectation of the random variable $\mathbb{E}[Y|X]$ is equal to $\mathbb{E}[Y]$. Indeed in the discrete case, we have:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \sum_{i=0}^{1} P(X = x_i) \cdot \sum_{j=0}^{2} y_j \mathbb{P}(Y = y_j | X = x_i) \\
&= \sum_{i=0}^{1} \sum_{j=0}^{2} y_j \mathbb{P}(Y = y_j, X = x_i) \\
&= \sum_{j=0}^{2} y_j \mathbb{P}(Y = y_j) \\
&= \mathbb{E}[Y]
\end{aligned}$$

Example 5.2. (Conditional Probability vs Conditional expectation). The conditional probability of the event A given B can be recast in terms of conditional expectation using indicator functions. If $0 < \mathbb{P}(B) < 1$, it is not hard to check that: $\mathbb{P}(A|B) = \mathbb{E}[\mathbf{1}_A | \mathbf{1}_B = 1]$ and $\mathbb{P}(A|B^C) = \mathbb{E}[\mathbf{1}_A | 1_B = 0]$. Indeed the random variables $\mathbf{1}_A$ and $\mathbf{1}_B$ are discrete. If we proceed as in the discrete case above, we have:

$$\begin{aligned}
\mathbb{E}[\mathbf{1}_A | \mathbf{1}_B = 1] &= 1 \cdot \mathbb{P}(\mathbf{1}_A = 1 | \mathbf{1}_B = 1) \\
&= \frac{\mathbb{P}(\mathbf{1}_A = 1, \mathbf{1}_B = 1)}{\mathbb{P}(\mathbf{1}_B = 1)} \\
&= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= \mathbb{P}(A|B)
\end{aligned}$$

A similar calculation gives $\mathbb{P}(A|B^C)$. In particular, the formula for total probability for A is a rewriting of the expectation of the random variable $\mathbb{E}[\mathbf{1}_A | \mathbf{1}_B]$:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[\mathbf{1}_A | \mathbf{1}_B]] &= \mathbb{E}[\mathbf{1}_A | \mathbf{1}_B = 1] \mathbb{P}(\mathbf{1}_B = 1) + \mathbb{E}[\mathbf{1}_A | \mathbf{1}_B = 0] \mathbb{P}(\mathbf{1}_B = 0) \\
&= \mathbb{P}(A|B) \cdot \mathbb{P}(B) + \mathbb{P}(A|B^C) \cdot \mathbb{P}(B^C) \\
&= \mathbb{P}(A)
\end{aligned}$$

5.2 Conditional Expectation as a projection.

Conditioning on one variable. We start by giving the definition of conditional expectation given a single variable. This relates to the two observations (A) and (B) made previously. We assume that the random variable is integrable for the expectations to be well-defined.

Definition 5.1. Let X and Y be integrable random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. The conditional expectation of Y given X is the random variable denoted by $\mathbb{E}[Y|X]$ with the following two properties:

(A) There exists a function $h : \mathbf{R} \to \mathbf{R}$ such that $\mathbb{E}[Y|X] = h(X)$.

(B) For any bounded random variable of the form $g(X)$ for some function g,

$$\mathbb{E}[g(X)Y] = \mathbb{E}[g(X)\mathbb{E}[Y|X]] \tag{5.2}$$

We can intepret the second property as follows. The conditional expectation $\mathbb{E}[Y|X]$ serves as a proxy for Y as far as X is concerned. Note that in equation (5.2), the expectation on the left can be seen as an average over the joint values of (X, Y), whereas the one on the right is an average over the values of X only! Another way to see this property is to write is as:

$$\mathbb{E}[g(X)(Y - \mathbb{E}[Y|X])] = 0 \tag{5.3}$$

In other words, the *random variable* $Y - \mathbb{E}[Y|X]$ *is orthogonal to any random variable constructed from* X.

Finally, it is important to notice that if we take $g(X) = 1$, then the second property implies :

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

In other words, the expectation of the conditional expectation of Y is simply the expectation of Y.

The existence of the conditional expectation $\mathbb{E}[Y|X]$ is not obvious. We know, it exists in particular cases given in example (5.1). We will show more generally, that it exists, it is unique whenever Y is in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ (In fact, it can be shown to exist whenever Y is integrable). Before doing so, let's warm up by looking at the case of Gaussian vectors.

Example 5.3. (Conditional expectation of Gaussian vectors - I). Let (X, Y) be a Gaussian vector of mean 0. Then:

$$\mathbb{E}[Y|X] = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} X \tag{5.4}$$

This candidate satisfies the two defining properties of conditional expectation : (A) It is clearly a function of X; in fact it is a simple multiple of X. (B) We have that the random variable $\left(Y - \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} X\right)$ is orthogonal and thus independent to X. This is a consequence of the proposition (3.5), since:

$$\begin{aligned}\mathbb{E}\left[X\left(Y - \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}X\right)\right] &= \mathbb{E}XY - \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}\mathbb{E}X^2 \\ &= \mathbb{E}XY - \frac{\mathbb{E}[XY]}{\cancel{\mathbb{E}[X^2]}}\cancel{\mathbb{E}X^2} \\ &= 0\end{aligned}$$

Therefore, we have for any bounded function $g(X)$ of X:

$$\mathbb{E}[g(X)(Y - \mathbb{E}(Y|X))] = \mathbb{E}[g(X)]\mathbb{E}[Y - \mathbb{E}[Y|X]] = 0$$

Example 5.4. (Brownian conditioning-I) Let $(B_t, t \geq 0)$ be a standard Brownian motion. Consider the Gaussian vector $(B_{1/2}, B_1)$. Its covariance matrix is:

$$C = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1 \end{bmatrix}$$

Let's compute $\mathbb{E}[B_1|B_{1/2}]$ and $\mathbb{E}[B_{1/2}|B_1]$. This is easy using the equation (5.4). We have:

$$\begin{aligned}\mathbb{E}[B_1|B_{1/2}] &= \frac{\mathbb{E}[B_1 B_{1/2}]}{\mathbb{E}[B_{1/2}^2]} B_{1/2} \\ &= \frac{(1/2)}{(1/2)} B_{1/2} \\ &= B_{1/2}\end{aligned}$$

In other words, the best approximation of B_1 given the information of $B_{1/2}$ is $B_{1/2}$. There is no problem in computing $\mathbb{E}[B_{1/2}|B_1]$, even though we are conditioning on a future position. Indeed the same formula gives

$$\mathbb{E}[B_{1/2}|B_1] = \frac{\mathbb{E}[B_1 B_{1/2}]}{\mathbb{E}[B_1^2]} B_1 = \frac{1}{2} B_1$$

This means that the best approximation of $B_{1/2}$ given the position at time 1, is $\frac{1}{2}B_1$ which makes a whole lot of sense!

In example (5.4) for the Gaussian vector (X, Y), the conditional expectation was equal to the *orthogonal projection* of Y onto X in L^2. In particular, the conditional expectation was a multiple of X. Is this always the case? Unfortunately, it is not. For example, in the equation (5.1), the conditional expectation is clearly not a multiple of the random variable X. However, it is a function of X, as is always the case by definition (5.1).

The idea to construct the conditional expectation $\mathbb{E}[Y|X]$ in general is to *project* Y *on the space of all random variables that can be constructed from* X*. To make this precise, consider the following subspace of* $L^2(\Omega, \mathcal{F}, \mathbb{P})$ *:*

Definition 5.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable defined on it. The space $L^2(\Omega, \sigma(X), \mathbb{P})$ is the linear subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ consisting of the square-integrable random variables of the form $g(X)$ for some function $g : \mathbf{R} \to \mathbf{R}$.

This is a linear subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$: It contains the random variable 0, and any linear combination of random variables of this kind is also a function of X and must have a finite second moment. We note the following:

Remark. $L^2(\Omega, \sigma(X), \mathbb{P})$ is a subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, very much how a plane or line (going through the origin) is a subspace of $\mathbf{R}^3$.

In particular, as in the case of a line or a plane, we can project an element of Y of $L^2(\Omega, \mathcal{F}, \mathbb{P})$ onto $L^2(\Omega, \sigma(X), \mathbb{P})$. The resulting projection is an element of $L^2(\Omega, \sigma(X), \mathbb{P})$, a square-integrable random-variable that is a function of X. For a subspace $\mathcal{S}$ of $\mathbf{R}^3$ (e.g. a line or a plane), the projection of the vector $\mathbf{v} \in \mathbf{R}^3$ onto the subspace $\mathcal{S}$, denoted $\text{Proj}_{\mathcal{S}}(\mathbf{v})$ is the closest point to $\mathbf{v}$ lying in the subspace $\mathcal{S}$. Moreover, $\mathbf{v} - \text{Proj}_{\mathcal{S}}(\mathbf{v})$ is orthogonal to the subspace. This picture of orthogonal projection also holds in L^2. Let Y be a random variable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ and let $L^2(\Omega, \sigma(X), \mathbb{P})$ be the subspace of those random variables that are functions of X. We write $Y^\star$ for the random variable in $L^2(\Omega, \sigma(X), \mathbb{P})$ that is *closest* to Y. In other words, we have (using the definition of the L^2-distance square):

$$\inf_{Z \in L^2(\Omega, \sigma(X), \mathbb{P})} \mathbb{E}[(Y - Z)^2] = \mathbb{E}[(Y - Y^\star)^2] \tag{5.5}$$

It turns out that $Y^\star$ is the right candidate for the conditional expectation.

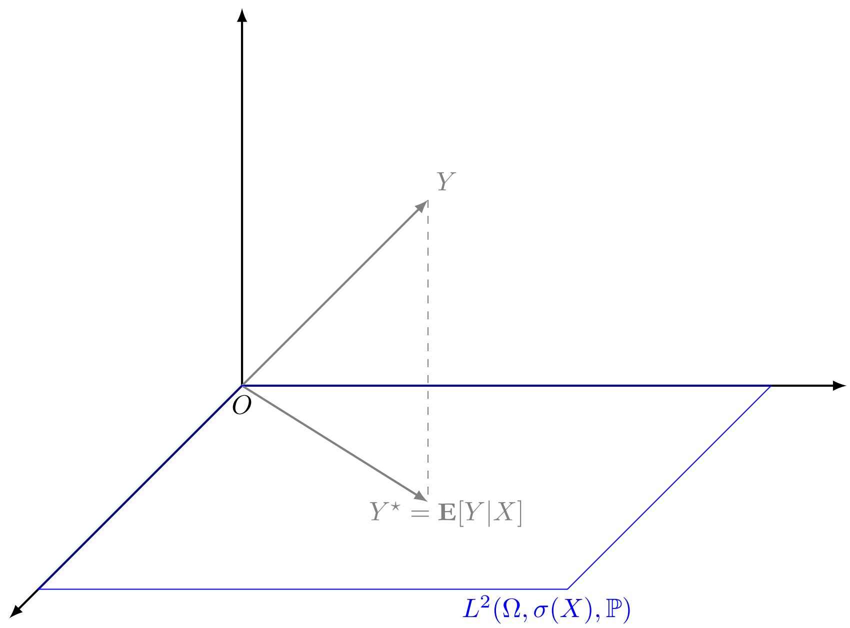


Figure. An illustration of the conditional expectation $\mathbb{E}[Y|X]$ as an orthogonal projection of Y onto the subspace $L^2(\Omega, \sigma(X), \mathbb{P})$.

Theorem 5.1. *(Existence and uniqueness of the conditional expectation) Let X be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Let Y be a random variable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then the conditional expectation $\mathbb{E}[Y|X]$ is the random variable $Y^\star$ given in the equation (5.5). Namely, it is the random variable in $L^2(\Omega, \sigma(X), \mathbb{P})$ that is closest to Y in the L^2-distance.*

In particular we have the following:

1) It is the orthogonal projection of Y onto $L^2(\Omega, \sigma(X), \mathbb{P})$, that is $Y - Y^\star$ is orthogonal to any random variables in the subspace $L^2(\Omega, \sigma(X), \mathbb{P})$.

2) It is unique.

Remark. This result reinforces the meaning of the conditional expectation $\mathbb{E}[Y|X]$ as the best estimation of Y given the information of X: it is the closest random variable to Y among all the functions of X in the sense of L^2.

Proof. We write for short $L^2(X)$ for the subspace $L^2(\Omega, \sigma(X), \mathbb{P})$. Let $Y^\star$ be as in equation (5.5). We show successively that (1) $Y - Y^\star$ is orthogonal to any element of $L^2(X)$, so it is the orthogonal projection (2) $Y^\star$ has the properties of conditional expectation in definition (5.2) (3) $Y^\star$ is unique.

(1) Let $W = g(X)$ be a random variable in $L^2(X)$. We show that W is orthogonal to $Y - Y^\star$; that is $\mathbb{E}[(Y - Y^\star)W] = 0$. This should be intuitively clear from figure above. On the one hand, we have by developing the square:

$$\begin{aligned}\mathbb{E}[(W - (Y - Y^\star))^2] &= \mathbb{E}[W^2 - 2W(Y - Y^\star) + (Y - Y^\star)^2] \\ &= \mathbb{E}[W^2] - 2\mathbb{E}[W(Y - Y^\star)] + \mathbb{E}(Y - Y^\star)^2]\end{aligned} \tag{5.6}$$

On the other hand, $Y^\star + W$ is an arbitrary vector in $L^2(X)$(it is a linear combination of the elements in $L^2(X)$), we must have from equation (5.5):

$$\begin{aligned}\mathbb{E}[(W - (Y - Y^\star))^2] &= \mathbb{E}[(Y - (Y^\star + W))^2] \\ &\geq \inf_{Z \in L^2(X)} \mathbb{E}[(Y - Z)^2] \\ &= \mathbb{E}[(Y - Y^\star)^2]\end{aligned} \tag{5.7}$$

Putting the last two equations (5.6), (5.7) together, we get that for any $W \in L^2(X)$:

$$\mathbb{E}[W^2] - 2\mathbb{E}[W(Y - Y^\star)] \geq 0$$

In particular, this also holds for aW, in which case we get:

$$\begin{aligned} &a^2\mathbb{E}[W^2] - 2a\mathbb{E}[W(Y - Y^\star)] \geq 0 \\ \implies &a\left\{a\mathbb{E}[W^2] - 2\mathbb{E}[W(Y - Y^\star)]\right\} \geq 0 \end{aligned}$$

If $a > 0$, then:

$$a\mathbb{E}[W^2] - 2\mathbb{E}[W(Y - Y^\star)] \geq 0 \tag{5.8}$$

whereas if $a < 0$, then the sign changes upon dividing throughout by a, and we have:

$$a\mathbb{E}[W^2] - 2\mathbb{E}[W(Y - Y^\star)] \leq 0 \tag{5.9}$$

Rearranging (5.8) yields:

$$\mathbb{E}[W(Y - Y^\star)] \leq a\mathbb{E}[W^2]/2 \tag{5.10}$$

Rearranging (5.9) yields:

$$\mathbb{E}[W(Y - Y^\star)] \geq a\mathbb{E}[W^2]/2 \tag{5.11}$$

Since (5.10) holds for all $a > 0$, the stronger inequality, $\mathbb{E}[W(Y - Y^\star)] \leq 0$ must hold. Since, (5.11) holds for all $a < 0$, the stronger inequality $\mathbb{E}[W(Y - Y^\star)] \geq 0$ must hold. Consequently,

$$\mathbb{E}[W(Y - Y^\star)] = 0 \tag{5.12}$$

(2) It is clear that $Y^\star$ is a function of X by construction, since it is in $L^2(X)$. Moreover, for any $W \in L^2(X)$, we have from (1) that:

$$\mathbb{E}[W(Y - Y^\star)] = 0$$

which is the second defining property of conditional expectations.

(3) Lastly, suppose there is another element Y' that is in $L^2(X)$ that minimizes the distance to Y. Then we would get:

$$\begin{aligned} \mathbb{E}[(Y - Y')^2] &= \mathbb{E}[(Y - Y^\star + Y^\star - Y')^2] \\ &= \mathbb{E}[(Y - Y^\star)^2] + 2\mathbb{E}[(Y - Y^\star)(Y^\star - Y')] + \mathbb{E}[(Y^\star - Y')^2] \\ &= \mathbb{E}[(Y - Y^\star)^2] + 0 + \mathbb{E}[(Y^\star - Y')^2] \\ &\quad \left\{(Y^\star - Y') \in L^2(X) \perp (Y - Y^\star)\right\} \end{aligned}$$

where we used the fact, that $Y^\star - Y'$ is a vector in $L^2(X)$ and the orthogonality of $Y - Y^\star$ with $L^2(X)$ as in (1). But, this implies that:

$$\cancel{\mathbb{E}[(Y - Y')^2]} = \cancel{\mathbb{E}[(Y - Y^\star)^2]} + \mathbb{E}[(Y^\star - Y')^2]$$
$$\mathbb{E}[(Y^\star - Y')^2] = 0$$

So, $Y^\star = Y'$ almost surely. □

Example 5.5. [Arguin-4.1] **Conditional Expectation of continuous random variables.** Let (X, Y) be two random variables with joint density $f_{X,Y}(x, y)$ on $\mathbf{R}^2$. Suppose for simplicity, that $\int_{\mathbf{R}} f(x, y)dx > 0$ for every y belonging to $\mathbf{R}$. Show that the conditional expectation $\mathbf{E}[Y|X]$ equals $h(X)$ where h is the function:

$$h(x) = \frac{\int_{\mathbf{R}} y f_{X,Y}(x, y)dy}{\int_{\mathbf{R}} f_{X,Y}(x, y)dy} \tag{5.13}$$

In particular, verify that $\mathbf{E}[\mathbf{E}[Y|X]] = \mathbf{E}[Y]$.

Hint: To prove this, verify that the above formula satisfies both the properties of conditional expectations; then invoke uniqueness to finish it off.

Solution. (i) The density function $f_{X,Y}(x, y)$ is a map $f : \mathbf{R}^2 \to \mathbf{R}$. The integral $\int_{y=-\infty}^{y=+\infty} y f_{X,Y}(x_0, y)dy$ is the area under the curve $yf(x, y)$ at the point $x = x_0$. Let's call it $A(x_0)$. If instead, we have an arbitrary x, $\int_{y=-\infty}^{y=+\infty} y f_{X,Y}(x, y)dy$ represents the area $A(x)$ of an arbitrary slice of the surface $y f_{X,Y}$ at the point x. Hence, it is a function of x. The denominator $\int_{\mathbf{R}} f_{X,Y}(x, y)dy = f_X(x)$, the density of X, which is a function of x. Hence, the ratio is a function of x.

(ii) Let $g(X)$ is a bounded random variable. We have:

$$\begin{aligned}
\mathbf{E}[g(X)(Y - h(X))] &= \mathbf{E}[Y g(X)] - \mathbf{E}[g(X)h(X)] \\
&= \int\int_{\mathbf{R}^2} y g(x) f_{X,Y}(x, y)dydx - \int_{\mathbf{R}} g(x)h(x)f(x)dx \\
&= \int\int_{\mathbf{R}^2} y g(x) f_{X,Y}(x, y)dydx \\
&\quad - \int_{\mathbf{R}} g(x) \cdot \frac{\int_{\mathbf{R}} y f_{X,Y}(x,y)dy}{\int_{\mathbf{R}} f_{X,Y}(x,y)dy} \cdot \int_{\mathbf{R}} f_{X,Y}(x, y)dy\, dx \\
&= \int\int_{\mathbf{R}^2} y g(x) f_{X,Y}(x, y)dydx \\
&\quad - \int_{\mathbf{R}} g(x) \cdot \frac{\int_{\mathbf{R}} y f_{X,Y}(x,y)dy}{\cancel{\int_{\mathbf{R}} f_{X,Y}(x,y)dy}} \cdot \cancel{\int_{\mathbf{R}} f_{X,Y}(x, y)dy}\, dx \\
&= \int\int_{\mathbf{R}^2} y g(x) f_{X,Y}(x, y)dydx - \int_{\mathbf{R}^2} y g(x) f_{X,Y}(x, y) \cdot dx \cdot dy \\
&= 0
\end{aligned}$$

Thus, $h(X)$ is a valid candidate for the conditional expectation $\mathbf{E}[Y|X]$. Moreover, by the existence and uniqueness theorem (5.1), $\mathbf{E}[Y|X]$ is unique and equals $h(X)$.

Conditioning on several random variables. We would like to generalize the conditional expectation to the case when we condition on the information of more than one random variable. Taking the L^2 point of view, we should expect that the conditional expectation is the orthogonal projection of the given random variable on the subspace generated by square integrable functions of all the variables on which we condition.

It is now useful to study sigma-fields, an object that was defined in chapter 1.

Definition 5.3. (Sigma-Field) A sigma-field or sigma-algebra $\mathcal{F}$ of a sample space Ω is a collection of all measurable events with the following properties:

(1) Ω is in $\mathcal{F}$.

(2) Closure under complement. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.

(3) Closure under countable unions. If $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.

Such objects play a fundamental role in the rigorous study of probability and real analysis in general. We will focus on the intuition behind them. First let's mention some examples of sigma-fields of a given sample space Ω to get acquainted with the concept.

Example 5.6. (Examples of sigma-fields).

(1) *The trivial sigma-field.* Note that the collection of events $\{\emptyset, \Omega\}$ is a sigma-field of Ω. We generally denote it by $\mathcal{F}_0$.

(2) *The σ-field generated by an event A.* Let A be an event that is not $\emptyset$ and not the entire Ω. Then the smallest sigma-field containing A ought to be:

$$\mathcal{F}_1 = \{\emptyset, A, A^C, \Omega\}$$

This sigma-field is denoted by $\sigma(A)$.

(3) The *sigma-field generated by a random variable X.*

We now define the $\mathcal{F}_X$ as follows:

$$\mathcal{F}_X = X^{-1}(\mathcal{B}) := \{\omega : X(\omega) \in B\}, \forall B \in \mathcal{B}(\mathbf{R})$$

where $\mathcal{B}$ is the Borel σ-algebra on $\mathbf{R}$. $\mathcal{F}_X$ is sometimes denoted as $\sigma(X)$. $\mathcal{F}_X$is the set of all events pertaining to X. It is a sigma-algebra because:

(i) $\Omega \in \sigma(X)$ because $\Omega = \{\omega : X(\omega) \in \mathbf{R}\}$ and $\mathbf{R} \in \mathcal{B}(\mathbf{R})$.

(ii) Let any event $C \in \sigma(X)$. We need to show that $\Omega \setminus C \in \sigma(X)$.

Since $C \in \sigma(X)$, there exists $A \in \mathcal{B}(\mathbf{R})$, such that:

$$C = \{\omega \in \Omega : X(\omega) \in A\}$$

Now, we calculate:

$$\Omega \setminus C = \{\omega \in \Omega : X(\omega) \in \mathbf{R} \setminus A\}$$

Since $\mathcal{B}(\mathbf{R})$ is a sigma-algebra, it is closed under complementation. Hence, if $A \in \mathcal{B}(\mathbf{R})$, it implies that $\mathbf{R} \setminus A \in \mathcal{B}(\mathbf{R})$. So, $\Omega \setminus C \in \sigma(X)$.

(iii) Consider a sequence of events $C_1, C_2, \ldots, C_n, \ldots \in \sigma(X)$. We need to prove that $\bigcup_{n=1}^{\infty} C_n \in \sigma(X)$.

Since $C_n \in \sigma(X)$, there exists $A_n \in \mathcal{B}(\mathbf{R})$ such that:

$$C_n = \{\omega \in \Omega : X(\omega) \in A_n\}$$

Now, we calculuate:

$$\bigcup_{n=1} C_n = \{\omega \in \Omega : X(\omega) \in \bigcup_{n=1}^{\infty} A_n\}$$

But, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{B}(\mathbf{R})$. So, $\bigcup_{n=1}^{\infty} C_n \in \sigma(X)$.

Consequently, $\sigma(X)$ is indeed a σ-algebra.

Intuitively, we think of $\sigma(X)$ as containing all information about X.

(4) *The sigma-field generated by a stochastic process* $(X_s, s \leq t)$. Let $(X_s, s \geq 0)$ be a stochastic process. Consider the process restricted to $[0, t]$, $(X_s, s \leq t)$. We consider the smallest sigma-field containing all events pertaining to the random variables $X_s, s \leq t$. We denote it by $\sigma(X_s, s \leq t)$ or $\mathcal{F}_t$.

The sigma-fields on Ω have a natural (partial) ordering: two sigma-fields $\mathcal{G}$ and $\mathcal{F}$ of Ω are such that $\mathcal{G} \subseteq \mathcal{F}$ if all the events in $\mathcal{G}$ are in $\mathcal{F}$. For example, the trivial σ-field $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is contained in all the σ-fields of Ω. Clearly, the σ-field $\mathcal{F}_t = \sigma(X_s, s \leq t)$ is contained in $\mathcal{F}_{t'}$ if $t \leq t'$.

If all the events pertaining to a random variable X are in the σ-field $\mathcal{G}$ (and thus we can compute $\mu(X^{-1}((a, b]))$), we will say that X is $\mathcal{G}$-measurable. This means that all information about X is contained in $\mathcal{G}$.

Definition 5.4. Let X be a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Consider another $\mathcal{G} \subseteq \mathcal{F}$. Then X is said to be $\mathcal{G}$-measurable, if and only if:

$$\{\omega : X(\omega) \in (a, b]\} \in \mathcal{G} \text{ for all intervals } (a, b] \in \mathbf{R}$$

Example 5.7. ($\mathcal{F}_0$-measurable random variables). Consider the trivial sigma-field $\mathcal{F}_0 = \{\emptyset, \Omega\}$. A random variable that is $\mathcal{F}_0$-measurable must be a constant. Indeed, we have that for any interval $(a, b]$, $\{\omega : X(\omega) \in (a, b]\} = \emptyset$ or $\{\omega : X(\omega) \in (a, b]\} = \Omega$. This can only hold if X takes a single value.

Example 5.8. ($\sigma(X)$-measurable random variables). Let X be a given random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Roughly speaking, a $\sigma(X)$-measurable random variable is determined by the information of X only. Here is the simplest example of a $\sigma(X)$-measurable random variable. Take the indicator function $Y = \mathbf{1}_{\{X \in B\}}$ for some event $\{X \in B\}$ pertaining to X. Then the pre-images $\{\omega : Y(\omega) \in (a, b]\}$ are either $\emptyset$, $\{X \in B\}$, $\{X \in B^C\}$ or Ω depending on whether $0, 1$ are in $(a, b]$ or not. All of these events are in $\sigma(X)$. More generally, one can construct a $\sigma(X)$-measurable random variable by taking linear combinations of indicator functions of events of the form $\{X \in B\}$.

It turns out that any (Borel measurable) function of X can be approximated by taking limits of such simple functions.

Concretely, this translates to the following statement:

$$\text{If } Y \text{ is } \sigma(\text{X})\text{-measurable, then Y=g(X) for some function g} \tag{5.14}$$

In the same way, if Z is $\sigma(X, Y)$-measurable, then $Z = h(X, Y)$ for some h. These facts can be proved rigorously using measure theory.

We are ready to give the general definition of conditional expectation.

Example 5.9. (Coin-Tossing Space). Suppose a coin is tossed infinitely many times. Let Ω be the set of all infinite sequences of Hs and Ts. A generic element of Ω is denoted by $\omega_1\omega_2\ldots$, where ω_n indicates the result of the nth coin toss. Ω is an uncountable sample space. The trivial sigma-field $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Assume that we don't know anything about the outcome of the experiement. Even without any information, we know that the true ω belongs to Ω and does not belong to $\emptyset$. It is the information learned at time 0.

Next, assume that we know the outcome of the first coin toss. Define $A_H = \{\omega : \omega_1 = H\}$=set of all sequences beginning with H and $A_T = \{\omega : \omega_1 = T\}$=set of all sequences beginning with T. The four sets resolved by the first coin-toss form the the σ-field $\mathcal{F}_1 = \{\emptyset, A_H, A_T, \Omega\}$. We shall think of this σ-field as containing the information learned by knowing the outcome of the first coin toss. More precisely, if instead of being told about the first coin toss, we are told for each set in $\mathcal{F}_1$, whether or not the true ω belongs to that set, then we know the outcome of the first coin toss and nothing more.

If we are told the first two coin tosses, we obtain a finer resolution. In particular, the four sets:

$$\begin{aligned} A_{HH} &= \{\omega : \omega_1 = H, \omega_2 = H\} \\ A_{HT} &= \{\omega : \omega_1 = H, \omega_2 = T\} \\ A_{TH} &= \{\omega : \omega_1 = T, \omega_2 = H\} \\ A_{TT} &= \{\omega : \omega_1 = T, \omega_2 = T\} \end{aligned}$$

are resolved. Of course, the sets in $\mathcal{F}_1$ are resolved. Whenever a set is resolved, so is its complement, which means that A_{HH}^C, A_{HT}^C, A_{TH}^C and A_{TT}^C are resolved, so is their union which means that $A_{HH} \cup A_{TH}$, $A_{HH} \cup A_{TT}$, $A_{HT} \cup A_{TH}$ and $A_{HT} \cup A_{TT}$ are resolved. The other two pair-wise unions $A_{HH} \cup A_{HT} = A_H$ and $A_{TH} \cup A_{TT} = A_T$ are already resolved. Finally, the triple unions are also resolved, because $A_{HH} \cup A_{HT} \cup A_{TH} = A_{TT}^C$ and so forth. Hence, the information pertaining to the second coin-toss is contained in:

$$\begin{aligned} \mathcal{F}_2 = \{&\emptyset, \Omega, \\ &A_H, A_T, \\ &A_{HH}, A_{HT}, A_{TH}, A_{TT}, \\ &A_{HH}^C, A_{HT}^C, A_{TH}^C, A_{TT}^C, \\ &A_{HH} \cup A_{TH}, A_{HH} \cup A_{TT}, A_{HT} \cup A_{TH}, A_{HT} \cup A_{TT}\} \end{aligned}$$

Hence, if the outcome of the first two coin tosses is known, all of the events in $\mathcal{F}_2$ are resolved - we exactly know, if each event has ocurred or not. $\mathcal{F}_2$ is the information learned by observing the first two coin tosses.

Exercise 5.1. [Arguin-4.2] (**Exercises on sigma-fields**).

(a) Let A, B be two proper subsets of Ω such that $A \cap B \neq \emptyset$ and $A \cup B \neq \Omega$. Write down $\sigma(\{A, B\})$, the smallest sigma-field containing A and B explicitly. What if $A \cap B = \emptyset$?

(b) The Borel sigma-field is the smallest sigma-field containing intervals of the form $(a, b]$ in **R**. Show that all singletons $\{b\}$ are in $\mathcal{B}(\mathbf{R})$ by writing $\{b\}$ as a countable intersection of intervals $(a, b]$. Conclude that all open intervals (a, b) and all closed intervals $[a, b]$ are in $\mathcal{B}(\mathbf{R})$. Is the subset **Q** of rational numbers a Borel set?

Proof. (a) The sigma-field generated by the two events A, B is given by:

$$\begin{aligned} \sigma(\{A, B\}) = \{&\emptyset, \Omega, \\ &A, B, A^C, B^C, \\ &A \cup B, A \cap B, \\ &A \cup B^C, A^C \cup B, A^C \cup B^C, \\ &A \cap B^C, A^C \cap B, A^C \cap B^C, \\ &(A \cup B) \cap (A \cap B)^C, \\ &(A \cup B)^C \cup (A \cap B)\} \end{aligned}$$

(b) Firstly, recall that:

$$\mathcal{B}(\mathbf{R}) = \bigcap_{\alpha \in \Lambda} \mathcal{F}_\alpha = \bigcap \sigma(\{I : I \text{ is an interval } (a, b] \subseteq \mathbf{R}\})$$

We can write:

$$\{b\} = \bigcap_{n=1}^{\infty} \left(b - \frac{1}{n}, b\right]$$

As $\mathcal{B}(\mathbf{R})$ is a sigma-field, it is closed under countable intersections. Hence, the singleton set $\{b\}$is a Borel set.

Similarly, we can write, any open interval as the countable union:

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right]$$

We can convince ourselves, that equality indeed holds. Let $x \in (a, b)$ and choose N, such that $\frac{1}{N} < |b - x|$. Then, for all $n \geq N$, $x \in (a, b - 1/n]$. Thus, it belongs to the RHS. In the reverse direction, let x belong to $\bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right]$. So, x belongs to atleast one of these sets. Therefore, $x \in (a, b)$ is trivially true. So, the two sets are equal.

Hence, open intervals are Borel sets.

Similarly, we may write:

$$[a, b] = \bigcap_{n=1}^{\infty} \left(a - \frac{1}{n}, b + \frac{1}{n}\right)$$

Consequently, closed intervals are Borel sets. Since $\mathbf{Q}$ is countable, it is a Borel set. Moreover, the empty set $\emptyset$ and $\mathbf{R}$ are Borel sets. So, $\mathbf{R} \backslash \mathbf{Q}$ is also a Borel set. □

Exercise 5.2. [Arguin-4.4] Let (X, Y) be a Gaussian vector with mean 0 and covariance matrix

$$C = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

for $\rho \in (-1, 1)$. We verify that the example (5.3) and exercise (5.5) yield the same conditional expectation.

(a) Use equation (5.4) to show that $\mathbf{E}[Y|X] = \rho X$.

(b) Write down the joint PDF $f(x, y)$ of (X, Y).

(c) Show that $\int_{\mathbf{R}} y f(x, y) dy = \rho x$ and that $\int_{\mathbf{R}} f(x, y) dy = 1$.

(d) Deduce that $\mathbf{E}[Y|X] = \rho X$ using the equation (5.13).

Proof. (a) Since (X, Y) have mean 0 and variance 1, it follows that:

$$\begin{aligned}
\mathbf{E}[(X - EX)(Y - EY)] &= \mathbf{E}(XY) \\
\sqrt{(\mathbf{E}[X^2] - (\mathbf{E}X)^2)} \cdot \sqrt{(\mathbf{E}[Y^2] - (\mathbf{E}Y)^2)} &= \sqrt{(1 - 0)(1 - 0)} \\
&= 1
\end{aligned}$$

and therefore,

$$\rho = \frac{\mathbf{E}(XY)}{1} = \frac{\mathbf{E}[XY]}{\mathbf{E}[X^2]}$$

Since (X, Y) is a Gaussian vector, using (5.4), we have:

$$\mathbf{E}[Y|X] = \frac{\mathbf{E}[XY]}{\mathbf{E}[X^2]} X = \rho X$$

(b) Consider the augmented matrix $[C|I]$. We have:

$$[C|I] = \left[\begin{array}{cc|cc} 1 & \rho & 1 & 0 \\ \rho & 1 & 0 & 1 \end{array}\right]$$

Performing $R_2 = R_2 - \rho R_1$, the above system is row-equivalent to:

$$\left[\begin{array}{cc|cc} 1 & \rho & 1 & 0 \\ 0 & 1-\rho^2 & -\rho & 1 \end{array}\right]$$

Performing $R_2 = \frac{1}{1-\rho^2} R_2$, the above system is row-equivalent to:

$$\left[\begin{array}{cc|cc} 1 & \rho & 1 & 0 \\ 0 & 1 & \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{array}\right]$$

Performing $R_1 = R_1 - \rho R_2$, we have:

$$\left[\begin{array}{cc|cc} 1 & 0 & \frac{1}{1-\rho^2} & -\frac{\rho}{1-\rho^2} \\ 0 & 1 & \frac{-\rho}{1-\rho^2} & \frac{1}{1-\rho^2} \end{array}\right]$$

Thus,

$$C^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

Moreover, $\det C = 1 - \rho^2$.

Therefore, the joint density of (X, Y) is given by:

$$\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}\right] \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \begin{bmatrix} x - \rho y & -\rho x + y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}\right] \\
& \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} (x^2 - 2\rho xy + y^2)\right]
\end{aligned}$$

(c) Claim I. $\int_{\mathbf{R}} y f(x,y) dy = \rho x$.

Completing the square, we have:

$$(x^2 - 2\rho xy + y^2) = (y - \rho x)^2 + x^2(1 - \rho^2)$$

Thus, we can write:

$$\int_{\mathbf{R}} yf(x,y)dy = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}x^2} \int_{\mathbf{R}} ye^{-\frac{1}{2}\frac{(y-\rho x)^2}{(1-\rho^2)}} dy$$

Let's substitute

$$\begin{aligned} z &= \frac{(y-\rho x)}{\sqrt{1-\rho^2}} \\ dz &= \frac{dy}{\sqrt{1-\rho^2}} \end{aligned}$$

Therefore,

$$\begin{aligned} \int_{\mathbf{R}} ye^{-\frac{1}{2}\frac{(y-\rho x)^2}{(1-\rho^2)}} dy &= \sqrt{1-\rho^2} \int_{\mathbf{R}} (\rho x + \sqrt{1-\rho^2} z) e^{-\frac{z^2}{2}} dz \\ &= \rho x \cdot \sqrt{1-\rho^2} \int_{\mathbf{R}} e^{-\frac{z^2}{2}} dz + (1-\rho^2) \int_{\mathbf{R}} ze^{-\frac{z^2}{2}} dz \\ &= \rho x \cdot \sqrt{1-\rho^2} \cdot \sqrt{2\pi} + (1-\rho^2) \cdot 0 \\ &= \rho x \cdot \sqrt{1-\rho^2} \cdot \sqrt{2\pi} \end{aligned}$$

Consequently,

$$\begin{aligned} \int_{\mathbf{R}} yf(x,y)dy &= \frac{1}{2\pi\cancel{\sqrt{1-\rho^2}}} e^{-\frac{1}{2}x^2} \rho x \cdot \cancel{\sqrt{1-\rho^2}} \cdot \sqrt{2\pi} \\ &= \rho x \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \\ &= \rho x \cdot f_X(x) \\ \frac{\int_{\mathbf{R}} yf(x,y)dy}{f_X(x)} &= \frac{\int_{\mathbf{R}} yf(x,y)dy}{\int_{\mathbf{R}} f(x,y)} = \rho x \end{aligned}$$

(d) For a Gaussian vector (X, Y), the conditional expectation $\mathbf{E}[Y|X] = h(X)$. Hence, $\mathbf{E}[Y|X] = \rho X$. $\square$

Definition 5.5. (Conditional Expectation) Let Y be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a sigma-field of Ω. The conditional expectation of Y given $\mathcal{G}$ is the random variable denoted by $\mathbb{E}[Y|\mathcal{G}]$ such that the following hold:

(a) $\mathbb{E}[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable.

In other words, all events pertaining to the random variable $\mathbb{E}[Y|\mathcal{G}]$ are in $\mathcal{G}$.

(b) For any (bounded) random variable W, that is $\mathcal{G}$-measurable,

$$\mathbb{E}[WY] = \mathbb{E}[W\mathbb{E}[Y|\mathcal{G}]]$$

In other words, $\mathbf{E}[Y|\mathcal{G}]$ is a proxy for Y as far as the events in $\mathcal{G}$ are concerned.

Note that, by taking $W = 1$ in the property (B), we recover:

$$\mathbf{E}[\mathbf{E}[Y|\mathcal{G}]] = \mathbf{E}[Y]$$

Remark. Beware of the notation! If $\mathcal{G} = \sigma(X)$, then the conditional expectation $\mathbf{E}[Y|\sigma(X)]$ is usually denoted by $\mathbf{E}[Y|X]$ for short. However, one should always keep in mind that conditioning on X is in fact projecting on the linear subspace *generated by all variables constructed from* X and not on the linear space generated by generated by X alone. In the same way, the conditional expectation $\mathbf{E}[Z|\sigma(X,Y)]$ is often written $\mathbf{E}[Z|X,Y]$ for short.

As expected, if Y is in $L^2(\Omega,\mathcal{F},\mathbb{P})$, then $\mathbf{E}[Y|\mathcal{G}]$ is given by the orthogonal projection of Y onto the subspace $L^2(\Omega,\mathcal{G},\mathbb{P})$, the subspace of square integrable random variables that are $\mathcal{G}$-measurable. We write $Y^\star$ for the random variable in $L^2(\Omega,\mathcal{G},\mathbb{P})$ that is closest to Y that is:

$$\min_{Z\in L^2(\Omega,\mathcal{G},\mathbb{P})} \mathbf{E}[(Y-Z)^2] = \mathbf{E}[(Y-Y^\star)^2] \tag{5.15}$$

Theorem 5.2. *(Existence and Uniqueness of Conditional Expectations) Let $\mathcal{G} \subset \mathcal{F}$ be a sigma-field of Ω. Let Y be a random variable in $L^2(\Omega,\mathcal{F},\mathbb{P})$. Then, the conditional expectation $\mathbf{E}[Y|\mathcal{G}]$ is the random variable $Y^\star$ given in the equation (5.15). Namely, it is the random variable in $L^2(\Omega,\mathcal{G},\mathbb{P})$ that is closest to Y in the L^2-distance. In particular we have the following:*

- It is the orthogonal projection of Y onto $L^2(\Omega,\mathcal{G},\mathbb{P})$, that is, $Y - Y^\star$ is orthogonal to the random variables in $L^2(\Omega,\mathcal{G},\mathbb{P})$.
- It is unique.

Again, the result should be interpreted as follows: The conditional expectation $\mathbf{E}[Y|\mathcal{G}]$ is the best approximation of Y given the information included in $\mathcal{G}$.

Remark. The conditional expectation in fact exists and is unique for any integrable random variable Y (i.e. $Y \in L^1(\Omega,\mathcal{F},\mathbb{P})$ as the definition suggests. However, there is no orthogonal projection in L^1, so the intuitive geometric picture is lost.

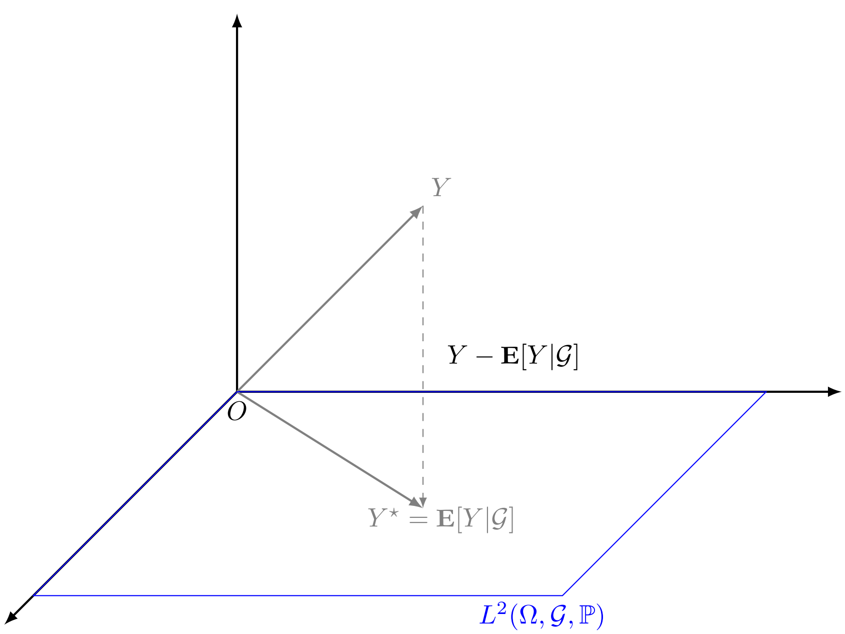


Figure. An illustration of the conditional expectation $\mathbb{E}[Y|\mathcal{G}]$ as an orthogonal projection of Y onto the subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Example 5.10. (Conditional Expectation for Gaussian Vectors. II.) Consider the Gaussian vector $(X_1, \ldots, X_n)$. Without loss of generality, suppose it has mean 0 and is non-degenerate. What is the best approximation of X_n given the information $X_1, \ldots, X_{n-1}$? In other words, what is:

$$\mathbf{E}[X_n|\sigma(X_1, \ldots, X_{n-1})$$

With example (5.8) in mind, let's write $\mathbf{E}[X_n|X_1 \ldots X_{n-1}]$ for short. From example (5.5), we know that if (X, Y) is a Gaussian vector with mean 0, then $\mathbf{E}[Y|X]$ is a multiple of X. Thus, we expect, that $\mathbf{E}[X_n|X_1 X_2 \ldots X_{n-1}]$ is a linear combination of $X_1, X_2, \ldots, X_{n-1}$. That is, there exists $a_1, \ldots, a_{n-1}$ such that:

$$\mathbf{E}[X_n|X_1 X_2 \ldots X_{n-1}] = a_1 X_1 + a_2 X_2 + \ldots + a_{n-1} X_{n-1}$$

In particular, since the conditional expectation is a linear combination of the X's, it is itself a Gaussian random variable. The best way to find the coefficient a's is to go back to IID decomposition of Gaussian vectors.

Let $(Z_1, Z_2, \ldots, Z_{n-1})$ be IID standard Gaussians constructed from the linear combination of $(X_1, X_2, \ldots, X_{n-1})$. Then, we have:

$$\mathbf{E}[X_n|X_1 X_2 \ldots X_{n-1}] = b_1 Z_1 + \ldots + b_{n-1} Z_{n-1}$$

Now, recall, that we construct the random variables Z_1, Z_2, ..., Z_n using Gram-Schmidt orthogonalization:

$$\begin{aligned}
\tilde{Z}_1 &= X_1, & Z_1 &= \frac{\tilde{Z}_1}{\mathbf{E}(\tilde{Z}_1^2)} \\
\tilde{Z}_2 &= X_2 - \mathbf{E}(X_2 Z_1) Z_1 & Z_2 &= \frac{\tilde{Z}_2}{\mathbf{E}(\tilde{Z}_2^2)} \\
\tilde{Z}_3 &= X_3 - \sum_{i=1}^{2} \mathbf{E}(X_3 Z_i) Z_i & Z_3 &= \frac{\tilde{Z}_3}{\mathbf{E}(\tilde{Z}_3^2)} \\
&\vdots
\end{aligned}$$

The simple case for $n = 2$ random variables.

We have already seen before:

$$\begin{aligned}
\mathbf{E}[X_1(X_2 - \mathbf{E}(X_2 Z_1) Z_1)] &= \mathbf{E}[\tilde{Z}_1(X_2 - \mathbf{E}(X_2 Z_1) Z_1)] \\
&= \frac{\mathbf{E}[\tilde{Z}_1^2]}{\mathbf{E}[\tilde{Z}_1^2]} \times \mathbf{E}\left[\tilde{Z}_1(X_2 - \mathbf{E}(X_2 Z_1) Z_1)\right] \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \mathbf{E}\left[\frac{\tilde{Z}_1}{\mathbf{E}[\tilde{Z}_1^2]}(X_2 - \mathbf{E}(X_2 Z_1) Z_1)\right] \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \mathbf{E}[Z_1(X_2 - \mathbf{E}(X_2 Z_1) Z_1)] \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \left(\mathbf{E}[Z_1 X_2] - \mathbf{E}(X_2 Z_1)\mathbf{E}[Z_1^2]\right) \\
&= 0
\end{aligned}$$

So,$X_2 - \mathbf{E}(X_2 Z_1) Z_1$ is orthogonal to X_1.

Moreover, $\mathbf{E}(X_2 Z_1) Z_1$ is a function of X_1. Thus, both the properties of conditional expectation are satisfied. Since conditional expectations are unique, we must have, $\mathbf{E}[X_2 | X_1] = \mathbf{E}(X_2 Z_1) Z_1$.

The case for $n = 3$ random variables.

We have seen that:

$$\begin{aligned}
\mathbf{E}[X_1(X_3 - \mathbf{E}(X_3 Z_1)Z_1 - \mathbf{E}(X_3 Z_2)Z_2)] &= \frac{\mathbf{E}[\tilde{Z}_1^2]}{\mathbf{E}[\tilde{Z}_1^2]} \times \mathbf{E}[\tilde{Z}_1(X_3 - \mathbf{E}(X_3 Z_1)Z_1 - \mathbf{E}(X_3 Z_2)Z_2)] \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \mathbf{E}\left\{ \frac{\tilde{Z}_1}{\mathbf{E}[\tilde{Z}_1^2]} (X_3 - \mathbf{E}(X_3 Z_1)Z_1 - \mathbf{E}(X_3 Z_2)Z_2) \right\} \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \mathbf{E}\left\{ Z_1(X_3 - \mathbf{E}(X_3 Z_1)Z_1 - \mathbf{E}(X_3 Z_2)Z_2) \right\} \\
&= \mathbf{E}[\tilde{Z}_1^2] \times \mathbf{E}[X_3 Z_1] - \mathbf{E}[X_3 Z_1]\mathbf{E}[Z_1^2] - \mathbf{E}[X_3 Z_2]\mathbf{E}[Z_1 Z_2] \\
&= 0
\end{aligned}$$

It is an easy exercise to show that it is orthogonal to X_2.

Hence, $X_3 - \mathbf{E}(X_3 Z_1)Z_1 - \mathbf{E}(X_3 Z_2)Z_2$ is orthogonal to X_1 and X_2. Moreover, $\mathbf{E}(X_3 Z_1)Z_1 + \mathbf{E}(X_3 Z_2)Z_2$ is a function of X_1, X_2. Thus, we must have:

$$\mathbf{E}[X_3 | X_1 X_2] = \mathbf{E}(X_3 Z_1)Z_1 + \mathbf{E}(X_3 Z_2)Z_2$$

In general, $X_n - \sum_{i=1}^{n-1} \mathbf{E}(X_n Z_i) Z_i$ is orthogonal to X_1, X_2, $\ldots$, X_{n-1}. Hence,

$$\mathbf{E}[X_n | X_1 X_2 \ldots X_{n-1}] = \sum_{i=1}^{n-1} \mathbf{E}(X_n Z_i) Z_i$$

5.2.1 Properties of Conditional Expectation.

We now list the properties of conditional expectation that follow from the two defining properties (A), (B) in the definition. They are extremely useful, when doing explicit computations on martingales. A good way to remember them is to understand how they relate to the interpretation of conditional expectation as an orthogonal projection onto a subspace or, equivalently, as the best approximation of the variable given the information available.

Proposition 5.1. *Let Y be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be another sigma-field of Ω. Then, the conditional expectation $\mathbf{E}[Y|\mathcal{G}]$ has the following properties:*

(1) If Y is $\mathcal{G}$-measurable, then :

$$\mathbf{E}[Y|\mathcal{G}] = Y$$

(2) Taking out what is known. More generally, if Y is $\mathcal{G}-$measurable and X is another integrable random variable (with XY also integrable), then :

$$\mathbf{E}[XY|\mathcal{G}] = Y\mathbf{E}[X|\mathcal{G}]$$

This makes sense, since Y is determined by $\mathcal{G}$, so we can take out what is known; it can be treated as a constant for the conditional expectation.

(3) Independence. If Y is independent of $\mathcal{G}$, that is, for any events $\{Y \in (a, b]\}$ and $A \in \mathcal{G}$:

$$\mathbb{P}(\{Y \in I\} \cap A) = \mathbb{P}(\{Y \in I\}) \cdot \mathbb{P}(A)$$

then

$$\mathbf{E}[Y|\mathcal{G}] = \mathbf{E}[Y]$$

In other words, if you have no information on Y, your best guess for its value is simply plain expectation.

(4) Linearity of conditional expectations. Let X be another integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then,

$$\mathbf{E}[aX + bY|\mathcal{G}] = a\mathbf{E}[X|\mathcal{G}] + b\mathbf{E}[Y|\mathcal{G}], \quad \textit{for any } a, b \in \mathbf{R}$$

The linearity justifies the cumbersom choice of notation $\mathbf{E}[Y|\mathcal{G}]$ for the random variable.

(5) Tower Property : If $\mathcal{H} \subseteq \mathcal{G}$ is another sigma-field of Ω, then:

$$\mathbf{E}[Y|\mathcal{H}] = \mathbf{E}[\mathbf{E}[Y|\mathcal{G}]|\mathcal{H}]$$

Think in terms of two successive projections: first on a plane, then on a line in the plane.

(6) Pythagoras Theorem. We have:

$$\mathbf{E}[Y^2] = \mathbf{E}\left[(\mathbf{E}[Y|\mathcal{G}])^2\right] + \mathbf{E}\left[(Y - \mathbf{E}[Y|\mathcal{G}])^2\right]$$

In particular:

$$\mathbf{E}\left[(\mathbf{E}\left[Y|\mathcal{G}\right])^2\right] \leq \mathbf{E}[Y^2]$$

In words, the L^2 norm of $\mathbf{E}[X|\mathcal{G}]$ is smaller than the one of X, which is clear if you think in terms of orthogonal projection.

(7) Expectation of the conditional expectation.

$$\mathbf{E}\left[\mathbf{E}[Y|\mathcal{G}]\right] = \mathbf{E}[Y]$$

Proof.

The uniqueness property of conditional expectations in theorem (5.2) might appear to be an academic curiosity. On the contrary, it is very practical, since it ensures, that if we find a candidate for the conditional expectation that has the two properties in Definition (5.1), then it must be *the* conditional expectation. To see this, let's prove property (1).

Claim 5.1. If Y is $\mathcal{G}$-measurable, then $\mathbf{E}[Y|\mathcal{G}] = Y$.

It suffices to show that Y has the two defining properties of conditional expectation.

(1) We are given that, Y is $\mathcal{G}$-measurable. So, property (A) is satisfied.

(2) For any bounded random variable W that is $\mathcal{G}$-measurable, we have:

$$\mathbf{E}[W(Y - Y)] = \mathbf{E}[0] = 0$$

So, property (B) is also a triviality.

Claim 5.2. (Taking out what is known.) If Y is $\mathcal{G}$-measurable and X is another integrable random variable, then:

$$\mathbf{E}[XY|\mathcal{G}] = Y\mathbf{E}[X|\mathcal{G}]$$

In a similar vein, it suffices to show that, $Y\mathbf{E}[X|\mathcal{G}]$ has the two defining properties of conditional expectation.

(1) We are given that Y is $\mathcal{G}$-measurable; from property (1), $\mathbf{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable. It follows that, $Y\mathbf{E}[X|\mathcal{G}]$ is $\mathcal{G}$-measurable.

(2) From theorem (5.2), $X - \mathbf{E}[X|\mathcal{G}]$ is orthogonal to the random variables $L^2(\Omega, \mathcal{G}, \mathbb{P})$. So, if W is any bounded $\mathcal{G}$-measurable random variable, it follows that:

$$\begin{aligned}\mathbf{E}[WY(X - \mathbf{E}[X|\mathcal{G}])] &= 0 \\ \implies \mathbf{E}[W \cdot XY] &= \mathbf{E}[WY\mathbf{E}[X|\mathcal{G}]]\end{aligned}$$

This closes the proof.

Claim 5.3. (Independence.) If Y is independent of $\mathcal{G}$, that is, for all events $\{Y \in (a, b]\}$ and $A \in \mathcal{G}$,

$$\mathbb{P}\{Y \in (a, b] \cap A\} = \mathbb{P}\{Y \in (a, b]\} \cdot \mathbb{P}(A)$$

then

$$\mathbf{E}[Y|\mathcal{G}] = \mathbf{E}[Y]$$

Let us show that $\mathbf{E}[Y]$ has the two defining properties of conditional expectations.

(1) $\mathbf{E}[Y]$ is a constant and so it is $\mathcal{F}_0$ measurable. Hence, it is $\mathcal{G}$ measurable.

(2) If W is another $\mathcal{G}$-measurable random variable,

$$\mathbf{E}[WY] = \mathbf{E}[W] \cdot \mathbf{E}[Y]$$

since Y is independent of $\mathcal{G}$ and therefore it is independent of Y. Hence,

$$\mathbf{E}[W(Y - \mathbf{E}[Y])] = 0$$

Consequently, $\mathbf{E}[Y|\mathcal{G}] = \mathbf{E}[Y]$.

Claim 5.4. (Linearity of conditional expectations) Let X be another integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. Then,

$$\mathbf{E}[aX + bY|\mathcal{G}] = a\mathbf{E}[X|\mathcal{G}] + b\mathbf{E}[Y|\mathcal{G}], \quad \text{for any } a, b \in \mathbf{R}$$

Since $\mathbf{E}[X|\mathcal{G}]$ and $\mathbf{E}[Y|\mathcal{G}]$ are $\mathcal{G}-$measurable, any linear combination of these two random variables is also $\mathcal{G}$-measurable. Also, if W is any bounded $\mathcal{G}-$measurable random variable, we have:

$$\begin{aligned}\mathbf{E}[W(aX + bY - (a\mathbf{E}[X|\mathcal{G}] + b\mathbf{E}[Y|\mathcal{G}]))] &= a\mathbf{E}[W(X - \mathbf{E}[X|\mathcal{G}])] \\ &\quad + b\mathbf{E}[W(Y - \mathbf{E}[Y|\mathcal{G}])]\end{aligned}$$

By definition, $X - \mathbf{E}(X|\mathcal{G})$ is orthogonal t o the subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$ and hence to all $\mathcal{G}$-measurable random-variables. Hence, the two expectations on the right hand side of the above expression are 0. Since, conditional expectations are unique, we have the desired result.

Claim 5.5. If $\mathcal{H} \subseteq \mathcal{G}$ is another sigma-field of Ω, then

$$\mathbf{E}[Y|\mathcal{H}] = \mathbf{E}[\mathbf{E}[Y|\mathcal{G}]|\mathcal{H}]$$

Define $U := \mathbf{E}[Y|\mathcal{G}]$. By definition, $\mathbf{E}[U|\mathcal{H}]$ is $\mathcal{H}$-measurable.

Let W be any bounded $\mathcal{H}$-measurable random variable. We have:

$$\mathbf{E}[W\{\mathbf{E}(Y|\mathcal{G}) - \mathbf{E}(\mathbf{E}(Y|\mathcal{G})|\mathcal{H})\}] = \mathbf{E}[W(U - \mathbf{E}(U|\mathcal{H})]$$

But, by definition $U - \mathbf{E}(U|\mathcal{H})$ is always orthogonal to the subspace $L^2(\Omega, \mathcal{H}, \mathbb{P})$ and hence, $\mathbf{E}[W(U - \mathbf{E}(U|\mathcal{H})] = 0$. Since, conditional expectations are unique, we have the desired result.

Claim 5.6. **Pythagoras's theorem**. We have:

$$\mathbf{E}[Y^2] = \mathbf{E}[(\mathbf{E}[Y|\mathcal{G}])^2] + \mathbf{E}[(Y - \mathbf{E}(Y|\mathcal{G}))^2]$$

In particular,

$$\mathbf{E}[(\mathbf{E}[Y|\mathcal{G}])^2] \leq \mathbf{E}[Y^2]$$

Consider the orthogonal decomposition:

$$Y = \mathbf{E}[Y|\mathcal{G}] + (Y - \mathbf{E}[Y|\mathcal{G}])$$

Squaring on both sides and taking expectations, we have:

$$\mathbf{E}[Y^2] = \mathbf{E}[(\mathbf{E}(Y|\mathcal{G}))^2] + \mathbf{E}[(Y - \mathbf{E}[Y|\mathcal{G}])^2] + 2\mathbf{E}\left[\mathbf{E}[Y|\mathcal{G}](Y - \mathbf{E}[Y|\mathcal{G}])\right]$$

By definition of conditional expectation, $(Y - \mathbf{E}[Y|\mathcal{G}])$ is orthogonal to the subspace $L^2(\Omega, \mathcal{G}, \mathbb{P})$. By the properties of conditional expectation, $\mathbf{E}[Y|\mathcal{G}]$ is $\mathcal{G}-$measurable, so it belongs to $L^2(\Omega, \mathcal{G}, \mathbb{P})$. Hence, the dot-product on the right-hand side is 0. Consequently, we have the desired result.

Moreover, since $(Y - \mathbf{E}[Y|\mathcal{G}])^2$ is a non-negative random variable, $\mathbf{E}[(Y - \mathbf{E}[Y|\mathcal{G}])^2] \geq 0$. It follows that: $\mathbf{E}[Y^2] \geq \mathbf{E}[(\mathbf{E}(Y|\mathcal{G}))^2]$.

Claim 5.7. Our claim is:

$$\mathbf{E}\left[\mathbf{E}[Y|\mathcal{G}]\right] = \mathbf{E}[Y]$$

We know that, if W is any bounded $\mathcal{G}$-measurable random variable:

$$\mathbf{E}\left[WY\right] = \mathbf{E}[W\mathbf{E}[Y|\mathcal{G}]]$$

Taking $W = 1$, we have:

$$\mathbf{E}\left[Y\right] = \mathbf{E}[\mathbf{E}[Y|\mathcal{G}]]$$

Example 5.11. (Brownian Conditioning II). We continue the example (5.4). Let's now compute the conditional expectations $\mathbf{E}[e^{aB_1}|B_{1/2}]$ and $\mathbf{E}[e^{aB_{1/2}}|B_1]$ for some parameter a. We shall need the properties of conditional expectation in proposition (5.1). For the first one we use the fact that $B_{1/2}$ is independent of $B_1 - B_{1/2}$ to get:

$$\begin{aligned}
\mathbf{E}[e^{aB_1}|B_{1/2}] &= \mathbf{E}[e^{a((B_1-B_{1/2})+B_{1/2})}|B_{1/2}] \\
&= \mathbf{E}[e^{a(B_1-B_2)} \cdot e^{aB_{1/2}}|B_{1/2}] \\
&\quad \{\text{Taking out what is known}\} \\
&= e^{aB_{1/2}}\mathbf{E}[e^{a(B_1-B_{1/2})}|B_{1/2}] \\
&= e^{aB_{1/2}} \cdot \mathbf{E}[e^{a(B_1-B_{1/2})}] \\
&\quad \{\text{Independence}\}
\end{aligned}$$

We know that, $a(B_1 - B_{1/2})$ is a gaussian random variable with mean 0 and variance $a^2/2$. We also know that, $\mathbf{E}[e^{tZ}] = e^{t^2/2}$. So, $\mathbf{E}[e^{a(B_1-B_{1/2})}] = e^{a^2/4}$. Consequently, $\mathbf{E}[e^{aB_1}|B_{1/2}] = e^{aB_{1/2}+a^2/4}$.

The result itself has the form of the MGF of a Gaussian with mean $B_{1/2}$ and variance $1/2$. (The MGF of $X = \mu + \sigma Z$, $Z = N(0,1)$ is $M_X(a) = \exp\left[\mu + \frac{1}{2}\sigma^2 a^2\right]$.) In fact, this shows that the conditional distribution of B_1 given $B_{1/2}$ is Gaussian of mean $B_{1/2}$ and variance $1/2$.

For the other expectation, note that $B_{1/2} - \frac{1}{2}B_1$ is independent of B_1. We have:

$$\begin{aligned}
\mathbf{E}\left[\left(B_{1/2} - \frac{1}{2}B_1\right)B_1\right] &= \mathbf{E}(B_{1/2}B_1) - \frac{1}{2}\mathbf{E}[B_1^2] \\
&= \frac{1}{2} - \frac{1}{2} \cdot 1 \\
&= 0
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\mathbf{E}[e^{aB_{1/2}}|B_1] &= \mathbf{E}[e^{a(B_{1/2}-\frac{1}{2}B_1)+\frac{a}{2}B_1}|B_1] \\
&= \mathbf{E}[e^{a(B_{1/2}-\frac{1}{2}B_1)} \cdot e^{\frac{a}{2}B_1}|B_1] \\
&= e^{\frac{a}{2}B_1}\mathbf{E}[e^{a(B_{1/2}-\frac{1}{2}B_1)}|B_1] \\
&\quad \{\text{Taking out what is known }\} \\
&= e^{\frac{a}{2}B_1}\mathbf{E}[e^{a(B_{1/2}-\frac{1}{2}B_1)}] \\
&\quad \{\text{Independence}\}
\end{aligned}$$

Now, $a(B_{1/2} - \frac{1}{2}B_1)$ is a random variable with mean 0 and variance $a^2(\frac{1}{2} - \frac{1}{4}) = \frac{a^2}{4}$. Consequently, $\mathbf{E}[e^{(a/2)Z}] = e^{\frac{a^2}{8}}$. Thus, $\mathbf{E}[e^{aB_{1/2}}|B_1] = e^{\frac{a}{2}B_1 + \frac{a^2}{8}}$.

Example 5.12. (Brownian bridge is conditioned Brownian motion). We know that the Brownian bridge $M_t = B_t - tB_1$, $t \in [0,1]$ is independent of B_1. We use this to show that the conditional distribution of the Brownian motion given the value at the end-point B_1 is the one of a Brownian bridge shifted by the straight line going from 0 to B_1. To see this, we compute the conditional MGF of $(B_{t_1}, B_{t_2}, \ldots, B_{t_n})$ given B_1 for some arbitrary choices of $t_1, t_2, \ldots, t_n$ in $[0,1]$. We get the following by adding and subtracting $t_j B_1$:

$$
\begin{aligned}
\mathbf{E}[e^{a_1 B_{t_1} + \ldots + a_n B_{t_n}}|B_1] &= \mathbf{E}[e^{a_1(B_{t_1} - t_1 B_1) + \ldots + a_n(B_{t_n} - t_n B_1)} \cdot e^{(a_1 t_1 B_1 + \ldots + a_n t_n B_1)}|B_1] \\
&= e^{(a_1 t_1 B_1 + \ldots + a_n t_n B_1)}\mathbf{E}[e^{a_1 M_{t_1} + \ldots + a_n M_{t_n}}|B_1] \\
&\quad \{\text{Taking out what is known}\} \\
&= e^{(a_1 t_1 B_1 + \ldots + a_n t_n B_1)}\mathbf{E}[e^{a_1 M_{t_1} + \ldots + a_n M_{t_n}}] \\
&\quad \{\text{Independence}\}
\end{aligned}
$$

The right side is exactly the MGF of the process $M_t + tB_1, t \in [0,1]$ (for a fixed value B_1), where $(M_t, t \in [0,1])$ is a Brownian bridge. This proves the claim.

Lemma 5.1. *(Conditional Jensen's Inequality) If c is a convex function on $\mathbf{R}$ and X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, then:*

$$\mathbf{E}[c(X)] \geq c(\mathbf{E}[X])$$

More generally, if $\mathcal{G} \subseteq \mathcal{F}$ is a sigma-field, then:

$$\mathbf{E}[c(X)|\mathcal{G}] \geq c(\mathbf{E}[X|\mathcal{G}]) \tag{5.16}$$

Proof. We know that, if $c(x)$ is a convex function, the tangent to the curve c at any point lies below the curve. The tangent to the cuve at this point, is a straight-line of the form:

$$c(t) = y = mt + c$$

where $m(t) = c'(t)$. This holds for all $t \in \mathbf{R}$. At an arbitrary point x we have:

$$c(x) \geq y = mx + c$$

Therefore, we have:

$$c(x) - c(t) \geq m(t)(x - t)$$

for any x and any point of tangency t.

$$c(X) - c(Y) \geq m(Y)(X - Y)$$

Substituting $Y = \mathbf{E}[X|\mathcal{G}]$, we get:

$$c(X) - c(\mathbf{E}[X|\mathcal{G}]) \geq m(\mathbf{E}[X|\mathcal{G}])(X - \mathbf{E}[X|\mathcal{G}])$$

Taking expectations on both sides, we get:

$$\mathbf{E}[(c(X) - c(\mathbf{E}[X|\mathcal{G}]))|\mathcal{G}] \geq \mathbf{E}[m(\mathbf{E}[X|\mathcal{G}])(X - \mathbf{E}[X|\mathcal{G}])|\mathcal{G}]$$

The left-hand side simplifies as:

$$\begin{aligned}
\mathbf{E}[(c(X) - c(\mathbf{E}[X|\mathcal{G}]))|\mathcal{G}] &= \mathbf{E}[c(X)|\mathcal{G}] - \mathbf{E}[c(\mathbf{E}[X|\mathcal{G}]))|\mathcal{G}] \\
&\quad \{\text{Linearity}\} \\
&= \mathbf{E}[c(X)|\mathcal{G}] - c(\mathbf{E}[X|\mathcal{G}]) \\
&\quad \{\text{c}(\mathbf{E}[\text{X}|\mathcal{G}])\text{) is } \mathcal{G}\text{-measurable}\}
\end{aligned}$$

On the right hand side, we have:

$$\begin{aligned}
\mathbf{E}[m(\mathbf{E}[X|\mathcal{G}])(X - \mathbf{E}[X|\mathcal{G}])|\mathcal{G}] &= \mathbf{E}[m(\mathbf{E}[X|\mathcal{G}]) \cdot X|\mathcal{G}] - \mathbf{E}[m(\mathbf{E}[X|\mathcal{G}]) \cdot \mathbf{E}[X|\mathcal{G}]|\mathcal{G}] \\
&= \mathbf{E}[X|\mathcal{G}]m(\mathbf{E}[X|\mathcal{G}]) - m(\mathbf{E}[X|\mathcal{G}]) \cdot \mathbf{E}[X|\mathcal{G}] \\
&= 0
\end{aligned}$$

Consequently, it follows that $\mathbf{E}[c(X)|\mathcal{G}] \geq c(\mathbf{E}[X|\mathcal{G}])$. □

Example 5.13. (Embeddings of L^p spaces) Square-integrable random variables are in fact integrable. In other words, there is always the inclusion $L^2(\Omega, \mathcal{F}, \mathbb{P}) \subseteq L^1(\Omega, \mathcal{F}, \mathbb{P})$. In particular, square integrable random variables always have a well-defined variance. This embedding is a simple consequence of Jensen's inequality since:

$$|\mathbf{E}[X]|^2 \leq \mathbf{E}[|X|^2]$$

as $f(x) = |x|^2$ is convex. By taking the square root on both sides, we get:

$$\|X\|_1 \leq \|X\|_2$$

More generally, for any $1 < p < \infty$, we can define $L^p(\Omega, \mathcal{F}, \mathbb{P})$ to be the linear space of random variables such that $\mathbf{E}[|X|^p] < \infty$. Then for $p < q$, since $x^{q/p}$ is convex, we get by Jensen's inequality :

$$\mathbf{E}[|X|^q] = \mathbf{E}[(|X|^p)^{\frac{q}{p}}] \geq (\mathbf{E}[|X|^p])^{\frac{q}{p}}$$

Taking the q-th root on both sides:

$$\mathbf{E}[|X|^p]^{1/p} \leq \mathbf{E}[|X|^q]^{1/q}$$

So, if $X \in L^q$, then it must also be in L^p. Concretely, this means that any random variable with a finite q-moment will also have a finite p-moment, for $q > p$.

5.3 Martingales.

We now have all the tools to define martingales.

Definition 5.6. (Filtration). A *filtration* $(\mathcal{F}_t : t \geq 0)$ of Ω is an increasing sequence of σ-fields of Ω. That is,

$$\mathcal{F}_s \subseteq \mathcal{F}_t, \quad \forall s \leq t$$

We will usually take $\mathcal{F}_0 = \{\emptyset, \Omega\}$. The canonical example of a filtration is the natural filtration of a given process $(M_s : s \geq 0)$. This is the filtration given by $\mathcal{F}_t = \sigma(M_s, s \leq t)$. The inclusions of the σ-fields are then clear. For a given Brownian motion $(B_t, t \geq 0)$, the filtration $\mathcal{F}_t = \sigma(B_s, s \leq t)$ is sometimes called the *Brownian filtration*. We think of the filtration as the *flow of information of the process*.

Definition 5.7. A stochastic process $(X_t : t \geq 0)$ is said to be adapted to $(\mathcal{F}_t : t \geq 0)$, if for each t, the random variable X_t is $\mathcal{F}_t-$measurable.

Definition 5.8. (Martingale). A process $(M_t : t \geq 0)$ is a martingale for the filtration $(\mathcal{F}_t : t \geq 0)$ if the following hold:

(1) The process is *adapted*, that is M_t is $\mathcal{F}_t-$measurable for all $t \geq 0$.

(2) $\mathbf{E}[|M_t|] < \infty$ for all $t \geq 0$. (This ensures that the conditional expectation is well defined.)

(3) *Martingale property:*

$$\mathbf{E}[M_t|\mathcal{F}_s] = M_s \quad \forall s \leq t$$

Roughly, speaking this means that the best approximation of a process at a future time t is its value at the present.

In particular, the martingale property implies that:

$$\begin{aligned} \mathbf{E}[M_t|\mathcal{F}_0] &= M_0 \\ \mathbf{E}[\mathbf{E}[M_t|\mathcal{F}_0]] &= \mathbf{E}[M_0] \\ \mathbf{E}[M_t] &= \mathbf{E}[M_0] \\ &\{\text{Tower Property}\} \end{aligned} \tag{5.17}$$

Usually, we take $\mathcal{F}_0$ to be the trivial sigma-field $\{\emptyset, \Omega\}$. A random variable that is $\mathcal{F}_0$-measurable must be a constant, so M_0 is a constant. In this case, $\mathbf{E}[M_t] = M_0$ for all t. If properties (1) and (2) are satisfied, but the best approximation is larger, $\mathbf{E}[M_t|\mathcal{F}_s] \geq M_s$, the process is called a *submartingale*. If it is smaller on average, $\mathbf{E}[M_t|\mathcal{F}_s] \leq \mathbf{E}[M_s]$, we say it is a supermartingale.

We will be mostly interested in martingales that are continuous and square-integrable. Continuous martingales are martingales whose paths $t \mapsto M_t(\omega)$ are continuous almost surely. Square-integrable martingales are such that $\mathbf{E}[|M_t|^2] < \infty$ for all t's. This condition is stronger than $\mathbf{E}[|M_t|] < \infty$ due to Jensen's inequality.

Remark. (Martingales in Discrete-time). Martingales can be defined the same way if the index set of the process is discrete. For example, the filtration $(\mathcal{F}_n : n \in \mathbf{N})$ is a countable set and the martingale property is then replaced by $\mathbf{E}[M_{n+1}|\mathcal{F}_n] = M_n$ as expected. The tower-property then yields the martingale property $\mathbf{E}[M_{n+k}|\mathcal{F}_n] = M_n$ for $k \geq 1$.

Remark. (Continuous Filtrations). Filtrations with continuous time can be tricky to handle rigorously. For example, one has to make sense of what it means for $\mathcal{F}_s$ as s approaches t from the left. Is it equal to $\mathcal{F}_t$? Or is there actually less information in $\lim_{s \to t^-} \mathcal{F}_s$ than in $\mathcal{F}_t$? This is a bit of headache when dealing with processes with jumps, like the Poisson process. However, if the paths are continuous, the technical problems are not as heavy.

Let's look at some of the important examples of martingales constructed from Brownian Motion.

Example 5.14. (Examples of Brownian Martingales)

(i) *Standard Brownian Motion.* Let $(B_t : t \geq 0)$ be a standard Brownian motion and let $(\mathcal{F}_t : t \geq 0)$ be a *Brownian filtration*. Then $(B_t : t \geq 0)$ is a square integrable martingale for the filtration $(\mathcal{F}_t : t \geq 0)$. Property (1) is obvious, because all the sets in $\mathcal{F}_t$ are resolved, upon observing the outcome of B_t. Similarly, $\mathbf{E}[|B_t|] = 0$. As for the martingale property, note that, by the properties of conditional expectation in proposition (5.1), we have:

$$\begin{aligned}
\mathbf{E}[B_t|\mathcal{F}_s] &= \mathbf{E}[B_t|B_s] \\
&= \mathbf{E}[B_t - B_s + B_s|B_s] \\
&= \mathbf{E}[B_t - B_s|B_s] + \mathbf{E}[B_s|B_s] \\
&\quad \{\text{Linearity}\} \\
&= \mathbf{E}[B_t - B_s] + B_s \\
&\quad \{\text{Independence}\} \\
&= B_s
\end{aligned}$$

(ii) *Geometric Brownian Motion.* Let $(B_t, t \geq 0)$ be a standard brownian motion, and $\mathcal{F}_t = \sigma(B_s, s \leq t)$. A *geometric brownian motion* is a process $(S_t, t \geq 0)$ defined by:

$$S_t = S_0 \exp\left(\sigma B_t + \mu t\right)$$

for some parameter $\sigma > 0$ and $\mu \in \mathbf{R}$. This is simply the exponential of the Brownian motion with drift. This is not a martingale for most choices of μ! In fact, one must take

$$\mu = -\frac{1}{2}\sigma^2$$

for the process to be a martingale for the Brownian filtration. Let's verify this. Property (1) is obvious since S_t is a function of B_t for each t. So, it is $\mathcal{F}_t$ measurable. Moreover, property (2) is clear: $\mathbf{E}[\exp(\sigma B_t + \mu t)] = \mathbf{E}[\exp(\sigma\sqrt{t}Z + \mu t)] = \exp(\mu t + \frac{1}{2}\sigma^2 t)$. So, its a finite quantity. As for the martingale property, note that by the properties of conditional expectation, and the MGF of Gaussians, we have for $s \leq t$:

$$\begin{aligned}
\mathbf{E}[S_t|\mathcal{F}_s] &= \mathbf{E}\left[S_0 \exp\left(\sigma B_t - \frac{1}{2}\sigma^2 t\right)|\mathcal{F}_s\right] \\
&= S_0 \exp(-\frac{1}{2}\sigma^2 t)\mathbf{E}[\exp(\sigma(B_t - B_s + B_s))|\mathcal{F}_s] \\
&= S_0 \exp(-\frac{1}{2}\sigma^2 t)\exp(\sigma B_s)\mathbf{E}[\exp(\sigma(B_t - B_s))|\mathcal{F}_s] \\
&\quad \{\text{Taking out what is known}\} \\
&= S_0 \exp\left(\sigma B_s - \frac{1}{2}\sigma^2 t\right)\mathbf{E}\left[\exp\left(\sigma(B_t - B_s)\right)\right] \\
&\quad \{\text{Independence}\} \\
&= S_0 \exp\left(\sigma B_s - \frac{1}{2}\sigma^2 t + \frac{1}{2}\sigma^2(t-s)\right) \\
&= S_0 \exp(\sigma B_s - \frac{1}{2}\sigma^2 s) \\
&= S_s
\end{aligned}$$

We will sometimes abuse terminology and refer to the martingale case of geometric brownian motion simply as geometric Brownian Motion when the context is clear.

(iii) *The square of the Brownian motion, compensated.* It is easy to check $(B_t^2, t \geq 0)$ is a submartingale by direct computation using increments or by Jensen's inequality: $\mathbf{E}[B_t^2|\mathcal{F}_s] > (\mathbf{E}[B_t|\mathcal{F}_s])^2 = B_s^2$, $s < t$. It is nevertheless possible to compensate to get a martingale:

$$M_t = B_t^2 - t$$

It is an easy exercise to verify that $(M_t : t \geq 0)$ is a martingale for the Brownian filtration $(\mathcal{F}_t : t \geq 0)$.

$$\begin{aligned}
\mathbf{E}[M_t|\mathcal{F}_s] &= \mathbf{E}[B_t^2 - t|\mathcal{F}_s] \\
&= \mathbf{E}[B_t^2|\mathcal{F}_s] - t \\
&= \mathbf{E}[(B_t - B_s + B_s)^2|\mathcal{F}_s] - t \\
&= \mathbf{E}[(B_t - B_s)^2|\mathcal{F}_s] + 2\mathbf{E}[(B_t - B_s)B_s|\mathcal{F}_s] + \mathbf{E}[B_s^2|\mathcal{F}_s] - t \\
&= \mathbf{E}[(B_t - B_s)^2] + 2B_s\mathbf{E}[(B_t - B_s)|\mathcal{F}_s] + B_s^2 - t \\
&= \mathbf{E}[(B_t - B_s)^2] + 2B_s\mathbf{E}[(B_t - B_s)] + B_s^2 - t \\
&\left\{ \begin{array}{c} (B_t - B_s) \text{ is independent of } \mathcal{F}_s \\ \text{Also, } B_s \text{ is known at time } s \end{array} \right\} \\
&= (t - s) + 2B_s \cdot 0 + B_s^2 - t \\
&= B_s^2 - s \\
&= M_s
\end{aligned}$$

Example 5.15. (Other important martingales).

(1) *Symmetric random walks.* This is an example of a martingale in discrete time. Take $(X_i : i \in \mathbf{N})$ to be IID random variables with $\mathbf{E}[X_i] = 0$ and $\mathbf{E}[|X_i|] < \infty$. Take $\mathcal{F}_n = \sigma(X_i, i \leq n)$ and

$$S_n = X_1 + X_2 + \ldots + X_n, \quad S_0 = 0$$

Firstly, the information learned by observing the outcomes of $X_1, \ldots, X_n$ is enough to completely determine S_n. Hence, S_n is $\mathcal{F}_n-$measurable.

Next,

$$\begin{aligned}
|S_n| &= \left| \sum_{i=1}^{n} X_i \right| \\
&\leq \sum_{i=1}^{n} |X_i|
\end{aligned}$$

Consequently, by the montonocity of expectations, we have:

$$\mathbf{E}[|S_n|] \leq \sum_{i=1}^{n} \mathbf{E}[|X_i|] < \infty$$

The martingale property is also satisfied. We have:

$$
\begin{aligned}
\mathbf{E}[S_{n+1}|\mathcal{F}_n] &= \mathbf{E}[S_n + X_{n+1}|\mathcal{F}_n] \\
&= \mathbf{E}[S_n|\mathcal{F}_n] + \mathbf{E}[X_{n+1}|\mathcal{F}_n] \\
&= S_n + \mathbf{E}[X_{n+1}] \\
&\left\{ \begin{array}{c} S_n \text{ is } \mathcal{F}_n\text{-measurable} \\ X_{n+1} \text{ is independent of } \mathcal{F}_n \end{array} \right\} \\
&= S_n + 0 \\
&= S_n
\end{aligned}
$$

(2) *Compensated Poisson process.* Let $(N_t : t \geq 0)$ be a Poisson process with rate λ and $\mathcal{F}_t = \sigma(N_s, s \leq t)$. Then, N_t is a submartingale for its natural filtration. Again, properties (1) and (2) are easily checked. N_t is $\mathcal{F}_t$ measurable. Moreover, $\mathbf{E}[|N_t|] = \mathbf{E}[N_t] = \frac{1}{\lambda t} < \infty$. The submartingale property follows by the independence of increments : for $s \leq t$,

$$
\begin{aligned}
\mathbf{E}[N_t|\mathcal{F}_s] &= \mathbf{E}[N_t - N_s + N_s|\mathcal{F}_s] \\
&= \mathbf{E}[N_t - N_s|\mathcal{F}_s] + \mathbf{E}[N_s|\mathcal{F}_s] \\
&= \mathbf{E}[N_t - N_s] + N_s \\
&= \lambda(t-s) + N_s \\
&\{\because \mathbf{E}[N_t] = \lambda t\}
\end{aligned}
$$

More importantly, we get a martingale by slightly modifying the process. Indeed, if we subtract λt, we have that the process :

$$
M_t = N_t - \lambda t
$$

is a martingale. We have:

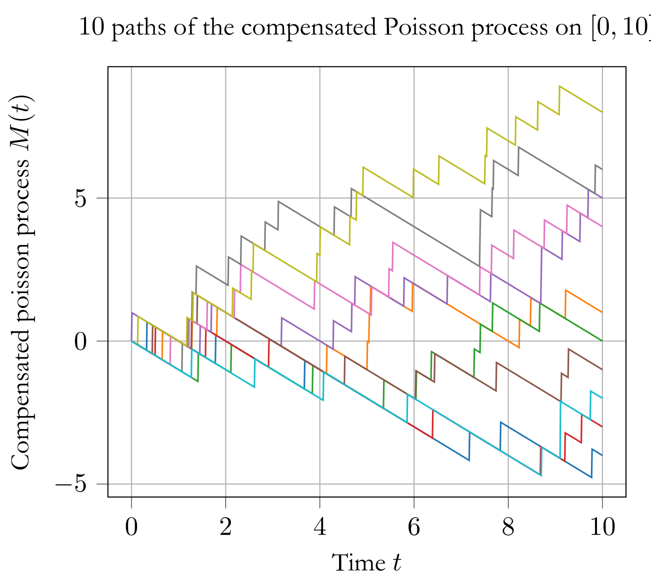
$$
\begin{aligned}
\mathbf{E}[M_t|\mathcal{F}_s] &= \mathbf{E}[N_t - \lambda t|\mathcal{F}_s] \\
&= \lambda t - \lambda s + N_s - \lambda t \\
&= N_s - \lambda s \\
&= M_s
\end{aligned}
$$

This is called the *compensated Poisson process.* Let us simulate 10 paths of the compensated poisson process on $[0, 10]$.

Listing 9: Generating 10 paths of a compensated Poisson process

```
import numpy as np
import matplotlib.pyplot as plt

# Generates a sample path of a compensated poisson process
# with rate : `lambda_` per unit time
# on the interval [0,T], and subintervals of size `stepSize`.

def generateCompensatedPoissonPath(lambda_,T,stepSize):
    N = int(T/stepSize)

    poissonParam = lambda_ * stepSize

    x = np.random.poisson(lam=poissonParam,size=N)
    x = np.concatenate([[0.0], x])
```

```
    N_t = np.cumsum(x)
    t = np.linspace(start=0.0,stop=10.0,num=1001)

    M_t = np.subtract(N_t,lambda_ * t)
    return M_t


t = np.linspace(0,10,1001)
plt.grid(True)

plt.xlabel(r'Time␣$t$')
plt.ylabel(r'Compensated␣poisson␣process␣$M(t)$')
plt.grid(True)
plt.title(r'$10$␣paths␣of␣the␣compensated␣Poisson␣process␣on␣$[0,10]$')

for i in range(10):
    # Generate a poisson path with rate 1 /sec = 0.01 /millisec
    n_t = generateCompensatedPoissonPath(lambda_=1.0, T=10, stepSize=0.01)
    plt.plot(t, n_t)


plt.show()
plt.close()
```



We saw in the two examples, that, even though a process is not itself a martingale, we can sometimes *compensate* to obtain a martingale! Ito Calculus will greatly extend this perspective. We will have systematic rules that show when a function of Brownian motion is a martingale and if not, how to modify it to get one.

For now, we observe that a convex function of a martingale is always a submartingale by Jensen's inequality.

Corollary 5.1. *If c is a convex function on* $\mathbf{R}$ *and* $(M_t : t \geq 0)$ *is a martingale for* $(\mathcal{F}_t : t \geq 0)$*, then the process* $(c(M_t) : t \geq 0)$ *is a submartingale for the same filtration, granted that* $\mathbf{E}[|c(M_t)|] < \infty$.

Proof. The fact that $c(M_t)$ is adapted to the filtration is clear since it is an explicit function of M_t. The integrability is by assumption. The submartingale property is checked as follows:

$$\mathbf{E}[c(M_t)|\mathcal{F}_s] \geq c(\mathbf{E}[M_t|\mathcal{F}_s]) = c(M_s)$$

□

Remark. (The Doob-Meyer Decomposition Theorem). Let $(X_n : n \in \mathbf{N})$ be a submartingale with respect to a filtration $(\mathcal{F}_n : n \in \mathbf{N})$. Define a sequence of random variables $(A_n : n \in \mathbf{N})$ by $A_0 = 0$ and

$$A_n = \sum_{i=1}^{n} (\mathbf{E}[X_i|\mathcal{F}_{i-1}] - X_{i-1}), \quad n \geq 1$$

Note that A_n is $\mathcal{F}_{n-1}$-measurable. Moreover, since $(X_n : n \in \mathbf{N})$ is a submartingale, we have $\mathbf{E}[X_i|\mathcal{F}_{i-1}] - X_{i-1} \geq 0$ almost surely. Hence, $(A_n : n \in \mathbf{N})$ is an increasing sequence almost surely. Let $M_n = X_n - A_n$.

We have:

$$\begin{aligned}
\mathbf{E}[M_n|\mathcal{F}_{n-1}] &= \mathbf{E}[X_n - A_n|\mathcal{F}_{n-1}] \\
&= \mathbf{E}[X_n|\mathcal{F}_{n-1}] - \mathbf{E}[A_n|\mathcal{F}_{n-1}] \\
&= \mathbf{E}[X_n|\mathcal{F}_{n-1}] - \mathbf{E}\left[\mathbf{E}[X_n|\mathcal{F}_{n-1}] - X_{n-1} + A_{n-1}|\,\mathcal{F}_{n-1}\right] \\
&= \mathbf{E}[X_n|\mathcal{F}_{n-1}] - \mathbf{E}[X_n|\mathcal{F}_{n-1}] + \mathbf{E}[X_{n-1}|\mathcal{F}_{n-1}] - \mathbf{E}[A_{n-1}|\mathcal{F}_{n-1}] \\
&= \cancel{\mathbf{E}[X_n|\mathcal{F}_{n-1}]} - \cancel{\mathbf{E}[X_n|\mathcal{F}_{n-1}]} + X_{n-1} - A_{n-1} \\
&= M_{n-1}
\end{aligned}$$

Thus, $(M_n : n \in \mathbf{N})$ is a martingale. Thus, we have obtained the Doob decomposition:

$$X_n = M_n + A_n \tag{5.18}$$

This decomposition of a submartingale as a sum of a martingale and an adapted increasing sequence is unique, if we require that $A_0 = 0$ and that A_n is $\mathcal{F}_{n-1}$-measurable.

For the continuous-time case, the situation is much more complicated. The analogue of equation (5.18) is called the *Doob-Meyer decomposition.* We briefly describe this decomposition and avoid the technical details. All stochastic processes $X(t)$ are assumed to be right-continuous with left-hand limits $X(t-)$.

Let $X(t)$, $a \leq t \leq b$ be a submartingale with respect to a right-continuous filtration $(\mathcal{F}_t : a \leq t \leq b)$. If $X(t)$ satisfies certain conditions, then it can be uniquely decomposed as:

$$X(t) = M(t) + C(t), \quad a \leq t \leq b$$

where $M(t)$, $a \leq t \leq b$ is a martingale with respect to $(\mathcal{F}_t; a \leq t \leq b)$, $C(t)$ is right-continuous and increasing almost surely with $\mathbf{E}[C(t)] < \infty$.

Example 5.16. (Square of a Poisson Process). Let $(N_t : t \geq 0)$ be a Poisson process with rate λ. We consider the compensated process $M_t = N_t - \lambda t$. By (5.1), the process $(M_t^2 : t \geq 0)$ is a submartingale for the filtration $(\mathcal{F}_t : t \geq 0)$ of the Poisson process. How should we compensated M_t^2 to get a martingale? A direct computation using the properties of conditional expectation yields:

$$\begin{aligned}
\mathbf{E}[M_t^2|\mathcal{F}_s] &= \mathbf{E}[(M_t - M_s + M_s)^2|\mathcal{F}_s] \\
&= \mathbf{E}[(M_t - M_s)^2 + 2(M_t - M_s)M_s + M_s^2|\mathcal{F}_s] \\
&= \mathbf{E}[(M_t - M_s)^2|\mathcal{F}_s] + 2\mathbf{E}[(M_t - M_s)M_s|\mathcal{F}_s] + \mathbf{E}[M_s^2|\mathcal{F}_s] \\
&= \mathbf{E}[(M_t - M_s)^2] + 2M_s \underbrace{\mathbf{E}[M_t - M_s]}_{\text{equals } 0} + M_s^2 \\
&= \mathbf{E}[(M_t - M_s)^2] + M_s^2
\end{aligned}$$

Now, if $X \sim \text{Poisson}(\lambda t)$, then $\mathbf{E}[X] = \lambda t$ and $\mathbf{E}[X^2] = \lambda t(\lambda t + 1)$.

$$\begin{aligned}
\mathbf{E}[(M_t - M_s)^2] &= \mathbf{E}\left[\{(N_t - N_s) - \lambda(t-s)\}^2\right] \\
&= \mathbf{E}\left[(N_t - N_s)^2\right] - 2\lambda(t-s)\mathbf{E}\left[(N_t - N_s)\right] + \lambda^2(t-s)^2 \\
&= \lambda^2(t-s)^2 + \lambda(t-s) - 2\lambda(t-s)\cdot\lambda(t-s) + \lambda^2(t-s)^2 \\
&= \lambda(t-s)
\end{aligned}$$

Thus,

$$\mathbf{E}[M_t^2 - \lambda t | \mathcal{F}_s] = M_s^2 - \lambda s$$

We conclude that the process $(M_t^2 - \lambda t : t \geq 0)$ is a martingale. The Doob-Meyer decomposition of the submartingale M_t^2 is then:

$$M_t^2 = (M_t^2 - \lambda t) + \lambda t$$

Example 5.17. Consider a Brownian motion $B(t)$. The quadratic variation of the process $(B(t) : t \geq 0)$ over the interval $[0, t]$ is given by $[B]_t = t$. On the other hand, we saw, that the square of Brownian motion compensated, $(B_t^2 - t : t \geq 0)$ is a martingale. Hence, the Doob-Meyer decomposition of $B(t)^2$ is given by:

$$B(t)^2 = (B(t)^2 - t) + t$$

5.4 Computations with Martingales.

Martingales are not only conceptually interesting, they are also formidable tools to compute probabilities and expectations of processes. For example, in this section, we will solve the *gambler's ruin* problem for Brownian motion. For convenience, we introduce the notion of *stopping time* before doing so.

Definition 5.9. A random variable $\tau : \Omega \to \mathbf{N} \cup \{+\infty\}$ is said to be a *stopping time* for the filtration $(\mathcal{F}_t : t \geq 0)$ if and only if:

$$\{\omega : \tau(\omega) \leq t\} \in \mathcal{F}_t, \quad \forall t \geq 0$$

Note that since $\mathcal{F}_t$ is a sigma-field, if τ is a stopping time, then we must also have that $\{\omega : \tau(\omega) > t\} \in \mathcal{F}_t$.

In other words, τ is a stopping time, if we can decide if the events $\{\tau \leq t\}$ occurred or not based on the information available at time t.

The term *stopping time* comes from gambling: a gambler can decide to stop playing at a random time (depending for example on previous gains or losses), but when he or she decides to stop, his/her decision is based solely upon the knowledge of what happened before, and does not depend on future outcomes. In other words, the stopping policy/strategy can only depend on past outcomes. Otherwise, it would mean that he/she has a crystall ball.

Example 5.18. (Examples of stopping times).

(i) *First passage time.* This is the first time when a process reaches a certain value. To be precise, let $X = (X_t : t \geq 0)$ be a process and $(\mathcal{F}_t : t \geq 0)$ be its natural filtration. For $a > 0$, we define the first passage time at a to be:

$$\tau(\omega) = \inf\{s \geq 0 : X_s(\omega) \geq a\}$$

If the path ω never reaches a, we set $\tau(\omega) = \infty$. Now, for t fixed and for a given path $X(\omega)$, it is possible to know if $\{\tau(\omega) \leq t\}$ (the path has reached a before time t) or $\{\tau(\omega) > t\}$ (the path has not reached a before time t) with the information available at time t, since we are looking at the first time the process reaches a. Hence, we conclude that τ is a stopping time.

(ii) *Hitting time.* More generally, we can consider the first time (if ever) that the path of a process $(X_t : t \geq 0)$ enters or hits a subset B of **R**:

$$\tau(\omega) = \min\{s \geq 0 : X_s(\omega) \in B\}$$

The first passage time is the particular case in which $B = [a, \infty)$.

(iii) *Minimum of two stopping times.* If τ and τ' are two stopping times for the same filtration $(\mathcal{F}_t : t \geq 0)$, then so is the minimum $\tau \wedge \tau'$ between the two, where

$$(\tau \wedge \tau')(\omega) = \min\{\tau(\omega), \tau'(\omega)\}$$

This is because for any $t \geq 0$:

$$\{\omega : (\tau \wedge \tau')(\omega) \leq t\} = \{\omega : \tau(\omega) \leq t\} \cup \{\omega : \tau'(\omega) \leq t\}$$

Since the right hand side is the union of two events in $\mathcal{F}_t$, it must also be in $\mathcal{F}_t$ by the properties of a sigma-field. We conclude that $\tau \wedge \tau'$ is a stopping time. Is it also the case that the maximum $\tau \vee \tau'$ is a stopping time?

For any fixed $t \geq 0$, we have:

$$\{\omega : (\tau \vee \tau')(\omega) \leq t\} = \{\omega : \tau(\omega) \leq t\} \cap \{\omega : \tau'(\omega) \leq t\}$$

Since the right hand side is the intersection of two events in $\mathcal{F}_t$, it must also be in $\mathcal{F}_t$ by the properties of a sigma-field. We conclude that $\tau \vee \tau'$ is a stopping time.

Example 5.19. (Last passage time is not a stopping time). What if we look at the last time the process reaches a, that is:

$$\rho(\omega) = \sup\{t \geq 0 : X_t(\omega) \geq a\}$$

This is a well-defined random variable, but it is not a stopping time. Based on the information available at time t, we are not able to decide whether or not $\{\rho(\omega) \leq t\}$ occurred or not, as the path can always reach a one more time after t.

It turns out that a martingale that is stopped when the stopping time is attained remains a martingale.

Proposition 5.2. *(Stopped Martingale). If $(M_t : t \geq 0)$ is a continuous martingale for the filtration $(\mathcal{F}_t : t \geq 0)$ and τ is a stopping time for the same filtration, then the stopped process defined by*

$$M_{t\wedge\tau} = \begin{cases} M_t & t \leq \tau \\ M_\tau & t > \tau \end{cases}$$

is also a continuous martingale for the same filtration.

Theorem 5.3. *(Doob's Optional sampling theorem). If $(M_t : t \geq 0)$ is a continuous martingale for the filtration $(\mathcal{F}_t : t \geq 0)$ and τ is a stopping time such that $\tau < \infty$ and the stopped process $(M_{t\wedge\tau} : t \geq 0)$ is bounded, then:*

$$\mathbf{E}[M_\tau] = M_0$$

Proof. Since $(M_{\tau\wedge t} : t \geq 0)$ is a martingale, we always have:

$$\mathbf{E}[M_{\tau\wedge t}] = M_0$$

Now, since $\tau(\omega) < \infty$, we must

have that $\lim_{t\to\infty} M_{\tau\wedge t} = M_\tau$ almost surely. In particular, we have:

$$\mathbf{E}[M_\tau] = \mathbf{E}\left[\lim_{t\to\infty} M_{\tau\wedge t}\right] = \lim_{t\to\infty} \mathbf{E}[M_{\tau\wedge t}] = \lim_{t\to\infty} M_0$$

where we passed to the limit, using the dominated convergence theorem (2.6). □

Example 5.20. (Gambler's ruin with Brownian motion). The *gambler's ruin problem* is known in different forms. Roughly speaking, it refers to the problem of computing the probability of a gambler making a series of bets reaching a certain amount before going broke. In terms of Brownian motion (and stochastic processes in general), it translates to the following questions: Let $(B_t : t \geq 0)$ be a standard brownian motion starting at $B_0 = 0$ and $a, b > 0$.

(1) What is the probability that a Brownian path reaches a before $-b$?

(2) What is the expected waiting time for the path to reach a or $-b$?

For the first question, it is a simple computation using stopping time and martingale properties. Define the hitting time:

$$\tau(\omega) = \inf\{t \geq 0 : B_t(\omega) \geq a \text{ or } B_t(\omega) \leq -b\}$$

Note that τ is the minimum between the first passage time at a and the one at $-b$.

We first show that $\tau < \infty$ almost surely. In other words, all Brownian paths reach a or $-b$ eventually. To see this, consider the event E_n that the n-th increment exceeds $a + b$

$$E_n := \{|B_n - B_{n-1}| > a + b\}$$

Note that, if E_n occurs, then we must have that the Brownian motion path exits the interval $[-b, a]$. Moreover, we have $\mathbb{P}(E_n) = \mathbb{P}(E_1)$ for all n. Call this probability p.

Since the events E_n are independent, we have:

$$\mathbb{P}(E_1^C \cap E_2^C \cap \ldots \cap E_n^C) = (1 - p)^n$$

As $n \to \infty$ we have:

$$\lim_{n\to\infty} \mathbb{P}(E_1^C \cap E_2^C \cap \ldots \cap E_n^C) = 0$$

The sequence of events (F_n) where $F_n = E_1^C \cap E_2^C \cap \ldots \cap E_n^C$ is a decreasing sequence of events. By the continuity of probability measure lemma (1.8), we conclude that:

$$\lim_{n\to\infty} \mathbb{P}(F_n) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} F_n\right) = 0$$

Therefore, it must be the case $\mathbb{P}(\cup_{n=1}^{\infty} E_n) = 1$. So, E_n must occur for some n, so all brownian motion paths reach a or $-b$ almost surely.

Since $\tau < \infty$ with probability one, the random variable B_τ is well-defined : $B_\tau(\omega) = B_t(\omega)$ if $\tau(\omega) = t$. It can only take two values: a or $-b$. Question (1) above translates into computing $\mathbb{P}(B_\tau = a)$. On one hand, we have:

$$\mathbf{E}[B_\tau] = a\mathbb{P}(B_\tau = a) + (-b)(1 - \mathbb{P}(B_\tau = a))$$

On the other hand, by corollary (5.3), we have $\mathbf{E}[B_\tau] = \mathbf{E}[B_0] = 0$. (Note that the stopped process $(B_{t\wedge\tau} : t \geq 0)$ is bounded above by a and by $-b$ below). Putting these two observations together, we get:

$$\mathbb{P}(B_\tau = a) = \frac{b}{a+b}$$

A very simple and elegant answer!

We will revisit this problem again and again. In particular, we will answer the question above for Brownian motion with a drift at length further ahead.

Example 5.21. (Expected Waiting Time). Let τ be as in the last example. We now answer question (2) of the gambler's ruin problem:

$$\mathbf{E}[\tau] = ab$$

Note that the expected waiting time is consistent with the rough heuristic that Brownian motion travels a distance $\sqrt{t}$ by time t. We now use the martingale $M_t = B_t^2 - t$. On the one hand, if we apply optional stopping in corollary (5.3), we get:

$$\mathbf{E}[M_\tau] = M_0 = 0$$

Moreover, we know the distribution of B_τ, thanks to the probability calculated in the last example. We can therefore compute $\mathbf{E}[M_\tau]$ directly:

$$\begin{aligned}
0 &= \mathbf{E}[M_\tau] \\
&= \mathbf{E}[B_\tau^2 - \tau] \\
&= \mathbf{E}[B_\tau^2] - \mathbf{E}[\tau] \\
&= a^2 \cdot \frac{b}{a+b} + b^2 \cdot \frac{a}{a+b} - \mathbf{E}[\tau] \\
\mathbf{E}[\tau] &= \frac{a^2 b + b^2 a}{a+b} \\
&= \frac{ab\cancel{(a+b)}}{\cancel{(a+b)}} = ab
\end{aligned}$$

Why can we apply optional stopping here? The random variable τ is finite with probability 1 as before. However, the stopped martingale is not necessarily bounded as before: $B_{\tau\wedge t}$ is bounded but τ is not. However, the conclusion of optional stopping still holds. Indeed, we have:

$$\mathbf{E}[M_{t\wedge\tau}] = \mathbf{E}[B_{t\wedge\tau}^2] - \mathbf{E}[t\wedge\tau]$$

By the bounded convergence theorem, we get $\lim_{t\to\infty}\mathbf{E}[B_{t\wedge\tau}^2] = \mathbf{E}[\lim_{t\to\infty}B_{t\wedge\tau}^2] = \mathbf{E}[B_\tau^2]$. Since $\tau\wedge t$ is a non-decreasing sequence and as $t\to\infty$, $t\wedge\tau\to\tau$ almost surely, as $\tau<\infty$, by the monotone convergence theorem, $\lim_{t\to\infty}\mathbf{E}[t\wedge\tau] = \mathbf{E}[\tau]$.

Example 5.22. (First passage time of Brownian Motion.) We can use the previous two examples to get some very interesting information on the first passage time:

$$\tau_a = \inf\{t\geq 0 : B_t \geq a\}$$

Let $\tau = \tau_a\wedge\tau_{-b}$ be as in the previous examples with $\tau_{-b} = \inf\{t\geq 0 : B_t\leq -b\}$. Note that $(\tau_{-b}, b\in\mathbf{R}_+)$ is a sequence of random variables that is increasing in b. A brownian motion path must cross through -1 before it hits -2 for the first time and in general $\tau_{-n}(\omega)\leq\tau_{-(n+1)}(\omega)$. Moreover, we have $\tau_{-b}\to\infty$ almost surely as $b\to\infty$. That's because, $\mathbb{P}\{\tau<\infty\} = 1$. Moreover, the event $\{B_\tau = a\}$ is the same as $\{\tau_a<\tau_{-b}\}$. Now, the events $\{\tau_a<\tau_{-b}\}$ are increasing in b, since if a path reaches a before $-b$, it will do so as well for a more negative value of $-b$. On one hand, this means by the continuity of probability measure lemma (1.8) that:

$$\begin{aligned}\lim_{b\to\infty}\mathbb{P}\{\tau_a<\tau_{-b}\} &= \mathbb{P}\{\lim_{b\to\infty}\tau_a<\tau_{-b}\}\\ &= \mathbb{P}\{\tau_a<\infty\}\end{aligned}$$

On the other hand, we have by example (5.20)

$$\begin{aligned}\lim_{b\to\infty}\mathbb{P}\{\tau_a<\tau_{-b}\} &= \lim_{b\to\infty}\mathbb{P}\{B_\tau = a\}\\ &= \lim_{b\to\infty}\frac{b}{b+a}\\ &= 1\end{aligned}$$

We just showed that:

$$\mathbb{P}\{\tau_a<\infty\} = 1 \tag{5.19}$$

In other words, every Brownian path will reach a, no matter how large a is!

How long will it take to reach a on average? Well, we know from example (5.21) that $\mathbf{E}[\tau_a\wedge\tau_{-b}] = ab$. On one hand this means,

$$\lim_{b\to\infty}\mathbf{E}[\tau_a\wedge\tau_{-b}] = \lim_{b\to\infty}ab = \infty$$

On the other hand, since the random variables τ_{-b} are increasing,

$$\lim_{b\to\infty} \mathbf{E}[\tau_a \wedge \tau_{-b}] = \mathbf{E}\left[\lim_{b\to\infty} \tau_a \wedge \tau_{-b}\right] = \mathbf{E}[\tau_a]$$

by the monotone convergence theorem (2.3). We just proved that:

$$\mathbf{E}[\tau_a] = \infty$$

In other words, any Brownian motion path will reach a, but the expected waiting time for this to occur is infinite, no matter, how small a is! What is happening here? No matter, how small a is, there is always paths that reach very large negative values before hitting a. These paths might be unlikely. However, the first passage time for these paths is so large that they affect the value of the expectation substantially. In other words, τ_a is a *heavy-tailed random variable*. We look at the distribution of τ_a in more detail in the next section.

Example 5.23. (When option stopping fails). Consider τ_a, the first passage time at $a > 0$. The random variable B_{τ_a} is well-defined since $\tau_a < \infty$. In fact, we have $B_{\tau_a} = a$ with probability one. Therefore, the following must hold:

$$\mathbf{E}[B_{\tau_a}] = a \neq B_0$$

Optional stopping theorem corollary (5.3) does not apply here, since the stopped process $(B_{t\wedge\tau_a} : t \geq 0)$ is not bounded. $B_{t\wedge\tau_a}$ can become infinitely negative before hitting a.

5.5 Reflection principle for Brownian motion.

Proposition 5.3. *(Bachelier's formula). Let $(B_t : t \leq T)$ be a standard brownian motion on $[0, T]$. Then, the CDF of the random variable $\sup_{0\leq t\leq T} B_t$ is:*

$$\mathbb{P}\left(\sup_{0\leq t\leq T} B_t \leq a\right) = \mathbb{P}\left(|B_T| \leq a\right)$$

In particular, its PDF is:

$$f_{\max}(a) = \frac{2}{\sqrt{2\pi T}} e^{-\frac{a^2}{2T}}$$

Remark. We can verify these results empirically. Note that the paths of the random variables $\max_{0\leq s\leq t} B_s$ and $|B_t|$ are very different as t varies for a given ω. One is increasing and the other is not. The equality holds in distribution for a fixed t. As a bonus corollary, we get the distribution of the first passage time at a.

Corollary 5.2. *Let $a \geq 0$ and $\tau_a = \inf\{t \geq 0 : B_t \geq a\}$. Then:*

$$\mathbb{P}\left(\tau_a \leq T\right) = \mathbb{P}\left(\max_{0\leq t\leq T} B_t \geq a\right) = \int_a^\infty \frac{2}{\sqrt{2\pi T}} e^{-\frac{x^2}{2T}}\, dx$$

In particular, the random variable τ_a has the PDF:

$$f_{\tau_a}(t) = \frac{a}{\sqrt{2\pi}} \frac{e^{-\frac{a^2}{2t}}}{t^{3/2}}, \quad t > 0$$

This implies that it is heavy-tailed with $\mathbf{E}[\tau_a] = \infty$.

Proof. The maximum on $[0,T]$ is larger than or equal to a if and only if $\tau_a \leq T$. Therefore, the events $\{\max_{0\leq t\leq T} B_t \geq a\}$ and $\{\tau_a \leq T\}$ are the same. So, the CDF $\mathbb{P}(\tau_a \leq t)$ of τ_a, by proposition (5.3) $\int_a^\infty f_{\max}(x)dx = \int_a^\infty \frac{2}{\sqrt{2\pi T}} e^{-\frac{x^2}{2T}} dx$.

$$\begin{aligned} f_{\tau_a}(t) &= -2\phi(a/\sqrt{t}) \cdot a \cdot \left(-\frac{1}{2t^{3/2}}\right) \\ &= \frac{a}{t^{3/2}} \phi\left(\frac{a}{\sqrt{t}}\right) \\ &= \frac{a}{t^{3/2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2t}} \end{aligned}$$

To estimate the expectation, it suffices to realize that for $t \geq 1$, $e^{-\frac{a^2}{2t}}$ is larger than $e^{-\frac{a^2}{2}}$. Therefore, we have:

$$\mathbf{E}[\tau_a] = \int_0^\infty t \frac{a}{\sqrt{2\pi}} \frac{e^{-a^2/2t}}{t^{3/2}} dt \geq \frac{ae^{-a^2/2}}{\sqrt{2\pi}} \int_1^\infty t^{-1/2} dt$$

This is an improper integral and it diverges like $\sqrt{t}$ and is infinite as claimed. ☐

To prove proposition (5.3), we will need an important property of Brownian motion called the *reflection principle*. To motivate it, recall the reflection symmetry of Brownian motion at time s in proposition (4.4). It turns out that this reflection property also holds if s is replaced by a stopping time.

Lemma 5.2. *(Reflection principle). Let $(B_t : t \geq 0)$ be a standard Brownian motion and let τ be a stopping time for its filtration. Then, the process $(\tilde{B}_t : t \geq 0)$ defined by the reflection at time τ:*

$$\tilde{B}_t = \begin{cases} B_t & \text{if } t \leq \tau \\ B_\tau - (B_t - B_\tau) & \text{if } t > \tau \end{cases}$$

is also a standard brownian motion.

Remark. We defer the proof of the reflection property of Brownian motion to a further section. It is intuitive and instructive to quickly picture this in the discrete-time setting. I adopt the approach as in Shreve-I.

We repeatedly toss a fair coin (p, the probability of H on each toss, and $q = 1 - p$, the probability of T on each toss, are both equal to $\frac{1}{2}$). We denote the successive outcomes of the tosses by $\omega_1\omega_2\omega_3 \ldots$. Let

$$X_j = \begin{cases} -1 & \text{if } \omega_j = H \\ +1 & \text{if } \omega_j = T \end{cases}$$

and define $M_0 = 0$, $M_n = \sum_{j=1}^n X_n$. The process $(M_n : n \in \mathbf{N})$ is a symmetric random walk.

Suppose we toss a coin an odd number $(2j-1)$ of times. Some of the paths will reach level 1 in the first $2j-1$ steps and other will not reach. In the case of 3 tosses, there are $2^3 = 8$ possible paths and 5 of these reach level 1 at some time $\tau_1 \leq 2j - 1$. From that moment on, we can create a reflected path, which steps up each time the original path steps down and steps down each time the original path steps up. If the original path ends above 1 at the final time $2j-1$, the reflected path ends below 1 and vice versa. If the original path ends at 1, the reflected path does also. In fact, the reflection at the first hitting time has the same distribution as the original random walk.

The key here is, out of the 5 paths that reach level 1 at some time, there are as many reflected paths that exceed 1 at time $(2j-1)$ as there are original paths that exceed 1 at time $(2j-1)$. So, to count the total number of paths that reach level 1 by time $(2j-1)$, we can count the paths that are at 1 at time $(2j-1)$ and then add on *twice* the number of paths that exceed 1 at time $(2j-1)$.

With this new tool, we can now prove proposition (5.3).

Proof. Consider $\mathbb{P}(\max_{t\leq T} B_t \geq a)$. By splitting this probability over the event of the endpoint, we have:

$$\mathbb{P}\left(\max_{t\leq T} B_t \geq a\right) = \mathbb{P}\left(\max_{t\leq T} B_t \geq a, B_T > a\right) + \mathbb{P}\left(\max_{t\leq T} B_t \geq a, B_T \leq a\right)$$

Note also, that $\mathbb{P}(B_T = a) = 0$. Hence, the first probability equals $\mathbb{P}(B_T \geq a)$. As for the second, consider the time τ_a. On the event considered, we have $\tau_a \leq T$ and using lemma (5.2) at that time, we get

$$\mathbb{P}\left(\max_{t\leq T} B_t \geq a, B_T \leq a\right) = \mathbb{P}\left(\max_{t\leq T} B_t \geq a, \tilde{B}_T \geq a\right)$$

Observe that the event $\{\max_{t\leq T} B_t \geq a\}$ is the same as $\{\max_{t\leq T} \tilde{B}_T \geq a\}$. (A rough picture might help here.) Thereforem the above probability is

$$\mathbb{P}\left(\max_{t\leq T} B_t \geq a, B_T \leq a\right) = \mathbb{P}\left(\max_{t\leq T} \tilde{B}_t \geq a, \tilde{B}_T \geq a\right) = \mathbb{P}\left(\max_{t\leq T} B_t \geq a, B_T \geq a\right)$$

where the last equality follows from the reflection principle ($\tilde{B}_t$ is also a standard brownian motion, and B_T and $\tilde{B}_T$ have the same distribution.) But, as above, the last probability is equal to $\mathbb{P}(B_T \geq a)$. We conclude that:

$$\mathbb{P}\left(\max_{t\leq T} B_t \geq a\right) = 2\mathbb{P}(B_T \geq a) = \frac{2}{\sqrt{2\pi T}}\int_a^\infty e^{-\frac{x^2}{2T}}\,dx = \mathbb{P}(|B_T| \geq a)$$

This implies in particular that $\mathbb{P}\left(\max_{t\leq T} B_t = a\right) = 0$. Thus, we also have $\mathbb{P}(\max_{t\leq T} B_t \leq a) = \mathbb{P}(|B_T| \leq a)$ as claimed. ☐

Example 5.24. (Simulating Martingales) Sample 10 paths of the following process with a step-size of 0.01:

(a) $B_t^2 - t, t \in [0,1]$

(b) Geometric Brownian motion : $S_t = \exp(B_t - t/2), t \in [0,1]$.

Let's write a simple `BrownianMotion` class, that we shall use to generate sample paths.

Listing 10: 10 paths of $B_t^2 - t$

```
import numpy as np
import matplotlib.pyplot as plt

import attrs
from attrs import define, field

@define
class BrownianMotion:
    _step_size = field(validator=attrs.validators.and_(attrs.validators.instance_of(float),
                                                       attrs.validators.ge(0.0)))
    # Time T
    _T = field(validator=attrs.validators.and_(attrs.validators.instance_of(float),
                                               attrs.validators.ge(0.0)))
    # number of paths
    _N = field(validator=attrs.validators.and_(attrs.validators.instance_of(int),
                                               attrs.validators.gt(0)))
```

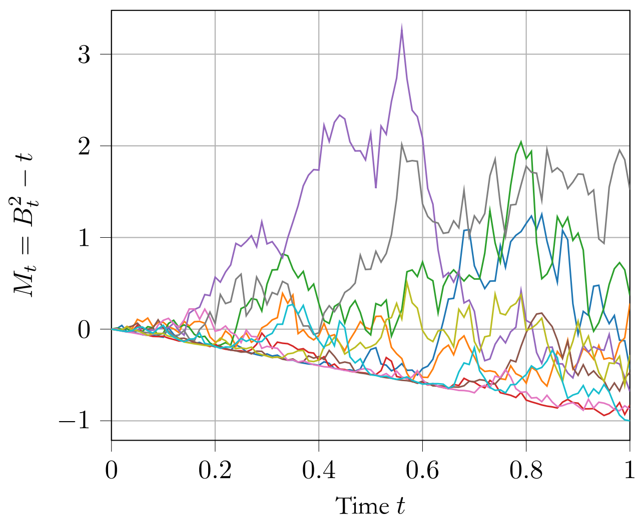
```python
    _num_steps = field(init=False)

    def __attrs_post_init__(self):
        self._num_steps = int(self._T/self._step_size)

    def covariance_matrix(self):
        C = np.zeros((self._num_steps,self._num_steps))

        for i in range(self._num_steps):
            for j in range(self._num_steps):
                s = (i+1) * self._step_size
                t = (j+1) * self._step_size
                C[i,j] = min(s,t)
        return C

    # Each column vector represents a sample path
    def generate_paths(self):
        C = self.covariance_matrix()
        A = np.linalg.cholesky(C)
        Z = np.random.standard_normal((self._num_steps, self._N))
        X = np.matmul(A,Z)
        X = np.concatenate((np.zeros((1,self._N)),X),axis=0)
        return X.transpose()
```

Now, the process $B_t^2 - t$ can be sampled as follows:

Listing 11: 10 paths of $B_t^2 - t$

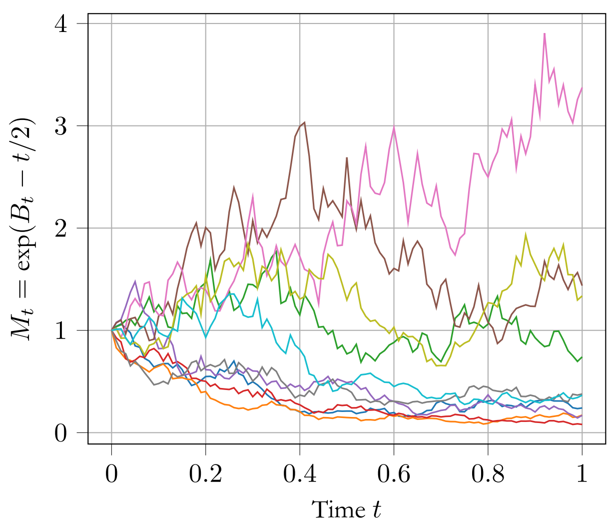
```python
def generateSquareOfBMCompensated(numOfPaths,stepSize,T):
    N = int(T/stepSize)

    X = []
    brownianMotion = BrownianMotion(stepSize,T)
    for n in range(numOfPaths):

        B_t = brownianMotion.samplePath()

        B_t_sq = np.square(B_t)

        t = np.linspace(start=0.0,stop=1.0,num=N+1)
        M_t = np.subtract(B_t_sq,t)
        X.append(M_t)

    return X
```

10 paths of the compensated squared BM on $[0, 1]$



The gBM process can be sampled similarly, with M_t `= np.exp(np.subtract(`B_t`,t/2))`.

10 paths of gBM on $[0, 1]$



Example 5.25. (Maximum of Brownian Motion.) Consider the maximum of Brownian motion on $[0, 1]$: $\max_{s \leq 1} B_s$.

(a) Draw the histogram of the random variable $\max_{s \leq 1} B_s$using $10, 0000$ sampled Brownian paths with a step size of 0.01.

(b) Compare this to the PDF of the random variable $|B_1|$.

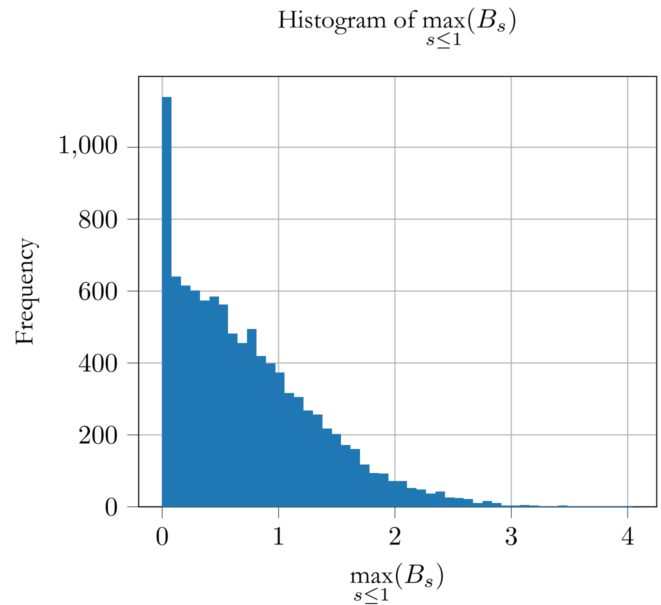
Solution.

I use the `itertools` python library to compute the running maximum of a brownian motion path.

Listing 12: The process $\sup_{s<1} B_s$

```
brownianMotion = BrownianMotion(stepSize=0.01,T=1)
data = []

for i in range(10000):
```

```
    B_t = brownianMotion.samplePath()
    max_B_t = list(itertools.accumulate(B_t,max))
    data.append(max_B_t[100])
```

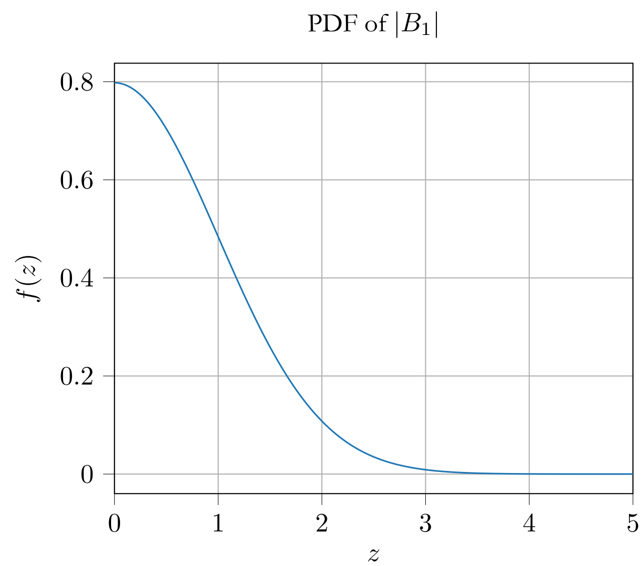


Analytically, we know that B_1 is a gaussian random variable with mean 0 and variance 1.

$$\begin{aligned}
\mathbb{P}(|B_1| \le z) &= \mathbb{P}(|Z| \le z) \\
&= \mathbb{P}(-z \le Z \le z) \\
&= \mathbb{P}(Z \le z) - \mathbb{P}(Z \le -z) \\
&= \mathbb{P}(Z \le z) - (1 - \mathbb{P}(Z \le z)) \\
F_{|B_1|}(z) &= 2\Phi(z) - 1
\end{aligned}$$

Differentiating on both sides, we get:

$$f_{|B_1|}(z) = 2\phi(z) = \frac{2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z \in [0, \infty)$$



Example 5.26. (First passage time.) Let $(B_t : t \geq 0)$ be a standard brownian motion. Consider the random variable:

$$\tau = \min\{t \geq 0 : B_t \geq 1\}$$

This is the first time that B_t reaches 1.

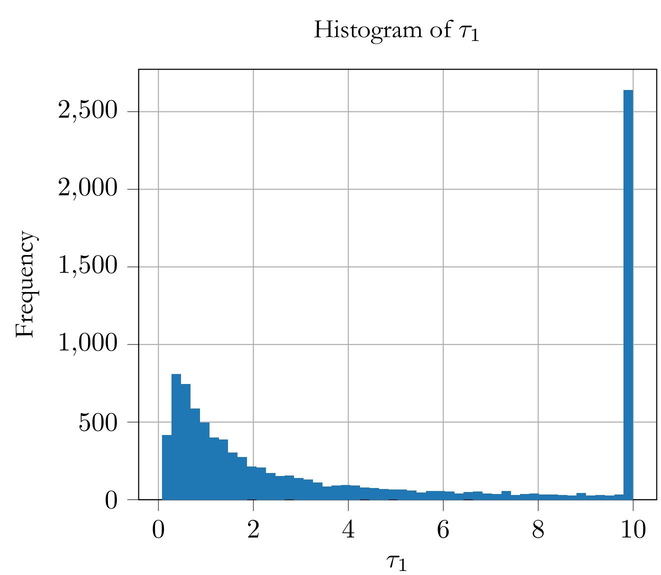
(a) Draw a histogram for the distribution of $\tau \wedge 10$ on the time-interval $[0, 10]$ using $10,000$ brownian motion paths on $[0, 10]$ with discretization 0.01.

The notation $\tau \wedge 10$ *means that if the path does not reach* 1 *on* $[0, 10]$, *then give the value* 10 *to the stopping time.*

(b) Estimate $\mathbf{E}[\tau \wedge 10]$.

(c) What proportion of paths never reach 1 in the time interval $[0, 10]$?

Solution.



To compute the expectation, we classify the hitting times of all paths into 50 bins. I simply did

```
frequency, bins = np.histogram(firstPassageTimes,bins=50,range=(0,10))
```

and then computed

```
expectation=np.dot(frequency,bins[1:])/10000.
```

This expectation estimate on my machine is $\mathbf{E}[\tau \wedge 10] = 4.34$ secs. There were approximately 2600 paths out of $10,000$ that did not reach 1.

Example 5.27. Gambler's ruin at the French Roulette. Consider the scenario in which you are gambling $\$1$ at the French roulette on the reds: You gain $\$1$ with probability $18/38$ and you lose a dollar with probability $20/38$. We estimate the probability of your fortune reaching $\$200$ before it reaches 0.

(a) Write a function that samples the simple random walk path from time 0 to time $5,000$ with a given starting point.

(b) Use the above to estimate the probability of reaching $\$200$ before $\$0$ on a sample of 100 paths if you start with $\$100$.

Example 5.28. Doob's maximal inequalities. We prove the following: Let $(M_k : k \geq 1)$ be positive submartingale for the filtration $(\mathcal{F}_k : k \in \mathbf{N})$. Then, for any $1 \leq p < \infty$ and $a > 0$

$$\mathbb{P}\left(\max_{k\leq n} M_k > a\right) \leq \frac{1}{a^p}\mathbf{E}[M_n^p]$$

(a) Use Jensen's inequality to show that if $(M_k : k \geq 1)$ is a positive submartingale, then so is $(M_k^p : k \geq 1)$ for $1 \leq p < \infty$. Conclude that it suffices to prove the statement for $p = 1$.

Solution.

The function $f(x) = x^p$ is convex. By conditional Jensen's inequality,

$$(\mathbf{E}[M_{k+1}|\mathcal{F}_k])^p \leq \mathbf{E}[M_k^p|\mathcal{F}_k]$$

Thus,

$$\mathbf{E}[M_{k+1}^p|\mathcal{F}_k] \geq (\mathbf{E}[M_{k+1}|\mathcal{F}_k])^p \geq M_k^p$$

where the last inequality follows from the fact that $(M_k : k \geq 1)$ is a positive submartingale, so $\mathbf{E}[M_{k+1}|\mathcal{F}_k] \geq M_k$. Consequently, $(M_k^p : k \geq 1)$ is also a positive submartingale.

(b) Consider the events

$$B_k = \bigcap_{j<k}\{\omega : M_j(\omega) \leq a\} \cap \{\omega : M_k(\omega) > a\}$$

Argue that the B_k's are disjoint and that $\bigcup_{k\leq n} B_k = \{\max_{k\leq n} M_k > a\} = B$.

Solution.

Clearly, B_k is the event that the first time to cross a is k. If B_k occurs, $B_{k+1}, B_{k+2}, \ldots$ fail to occur. Hence, all $B_k's$ are pairwise disjoint. The event $\bigcup_{k\leq n} B_k$ is the event that the random walk crosses a at any time $k \leq n$. Thus, the running maximum of the Brownian motion at time n exceeds a.

(c) Show that

$$\mathbf{E}[M_n] \geq \mathbf{E}[M_n \mathbf{1}_B] \geq a \sum_{k \leq n} \mathbb{P}(B_k) = a\mathbb{P}(B)$$

by decomposing B in B_k's and by using the properties of expectations, as well as the submartingale property.

Solution.

Clearly, $M_n \geq M_n \mathbf{1}_B \geq a\mathbf{1}_B$. And M_n is a positive random variable. By monotonicity of expectations, $\mathbf{E}[M_n] \geq \mathbf{E}[M_n \mathbf{1}_B] \geq a\mathbf{E}[\mathbf{1}_B] = a\mathbb{P}(B) = a\sum_{k\leq n} \mathbb{P}(B_k)$, where the last equality holds because the B_k's are disjoint.

(d) Argue that the inequality holds for continuous paths by discretizing time and using convergence theorems : If $(M_t : t \geq 0)$ is a positive submartingale with continuous paths for the filtration $(\mathcal{F}_t : t \geq 0)$, then for any $1 \leq p < \infty$ and $a > 0$:

$$\mathbb{P}\left(\max_{s \leq t} M_s > a\right) \leq \frac{1}{a^p}\mathbf{E}[M_t^p]$$

Solution.

Let $(M_t : t \geq 0)$ be a positive submartingale with continuous paths for the filtration $(\mathcal{F}_t : t \geq 0)$. Consider a sequence of partitions of the interval $[0, t]$ into 2^r subintervals :

$$D_r = \left\{\frac{kt}{2^r} : k = 0, 1, 2, \ldots, 2^n\right\}$$

And consider a sequence of discrete positive sub-martingales:

$$M^{(r)}_{kt/2^r} = M_{kt/2^r}, \quad k \in \mathbf{N}, 0 \leq k \leq 2^r$$

Next, we define for $r = 1, 2, 3, \ldots$

$$A_r = \left\{\sup_{s \in D_r} |M_s^{(r)}| > a\right\}$$

By using the maximal inequality in discrete time, gives us:

$$\mathbb{P}(A_r) = \mathbb{P}\left\{\sup_{s \in D_r} |M_s^{(r)}| > a\right\} \leq \frac{1}{a^p}\mathbf{E}\left[\left(M_s^{(r)}\right)^p\right] = \frac{1}{a^p}\mathbf{E}\left[M_t^p\right]$$

$$\begin{aligned}
\mathbb{P}\left(\max_{s \leq t} M_s > a\right) &= \mathbb{P}\left(\bigcup_{r=1}^{\infty} A_r\right) \\
&= \lim_{r \to \infty} \mathbb{P}\left(A_r\right) \\
&\{\text{Continuity of probability measure}\} \\
&\leq \lim_{r \to \infty} \frac{1}{a^p}\mathbf{E}\left[M_t^p\right]
\end{aligned}$$

6 Ito Calculus.

The Riemann-Stieltjes integral of g with respect to f is understood to be limit of the sums:

$$\int_0^t g(s) \cdot df(s) = \lim_{n\to\infty} \sum_{j=0}^{n-1} g(t_j)(f(t_{j+1}) - f(t_j))$$

The goal is to make sense of the above, when f is replaced by a Brownian motion $(B_t : t \geq 0)$.

$$\int_0^t g(s) \cdot dB_s = \lim_{n\to\infty} \sum_{j=0}^{n-1} g(t_j)(B_{t_{j+1}} - B_{t_j})$$

The major hurdle here is not the fact that the Brownian motion paths are random, but instead that these paths have *unbounded variation.* This means that the classical construction does not apply for a given path.

Note that the sum $\sum_{j=0}^{n-1} g(t_j)(B_{t_{j+1}} - B_{t_j})$ is a random variable. If the end-point $t_n = t$ is varied, it can be seen as a stochastic process. Since Brownian motion paths are continuous, this new stochastic process also has continuous paths. As we shall see, this stochastic process is in fact a continuous martingale. It turns out that these properties remain in the limit as $n \to \infty$.

What is the intepretation of the stochastic integral? If we think of $(B_t : t \geq 0)$ as modelling the price of a stock, then $\sum_{j=0}^{n-1} g(t_j)(B_{t_{j+1}} - B_{t_j})$ gives the value of a portfolio at time t that implements the following strategy: At t_j we buy $g(t_j)$ shares of the stock that we sell at time t_{j+1}. We do this for every $j \leq n-1$. The net gain or loss of this strategy is the sum over j of $g(t_j)(B_{t_{j+1}} - B_{t_j})$. Of course, in this implementation, the number of shares $g(t_j)$ put in play could be random and depend on the past information of the path upto time t_j.

6.1 Martingale Transform.

Let $(M_t, t \leq T)$ be a continuous square-integrable martingale on $[0,T]$ for the filtration $(\mathcal{F}_t : t \leq T)$ defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The idea of the martingale transform is to modify the amplitude of each increment in such a way as to produce a martingale when these new increments are summed up. The martingale transforms are to the Ito Integral, what Riemann sums are for the Riemann integral.

More precisely, let $(t_j, j \leq n)$ be a sequence of partitions of $[0,T]$ with $t_0 = 0$ and $t_n = T$. For example, we can take $t_j = \frac{jT}{n}$. Consider n fixed numbers $(Y_{t_0}, Y_{t_1}, \ldots, Y_{t_{n-1}})$. It is convenient to construct a function of time X_t from these:

$$X_t = Y_{t_j} \quad \text{if } t \in (t_j, t_{j+1}]$$

This can also be written as a sum of indicator functions:

$$X_t = \sum_{j=0}^{n-1} Y_{t_j} \mathbf{1}_{(t_j, t_{j+1}]}(t), \quad t \leq T \tag{6.1}$$

The integral of $(X_t : t \leq T)$ with respect to the martingale M on $[0,T]$ also called a *martingale transform,* is the sum of the increments of the martingale modulated by X; that is:

$$I_T = Y_0(M_1 - M_0) + \ldots + Y_{n-1}(M_T - M_{t_{n-1}}) = \sum_{j=0}^{n-1} Y_j(M_{t_{j+1}} - M_{t_j})$$

This is a random variable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$, since it is a linear combination of random variables in L^2. Note that, we recover M_T when X_{t_j} is 1 for all intervals. We may think of $(M_s : s \leq T)$ as the price of an asset, say a stock, on a time interval $[0, T]$. Then, the term:

$$Y_{t_j}(M_{t_{j+1}} - M_{t_j})$$

can be seen as the gain/loss in the time interval $(t_j, t_{j+1}]$ of buying Y_{t_j} units of the asset at time t_j at price M_{t_j} and selling these at time t_{j+1} at price $M_{t_{j+1}}$. Summing these terms over time gives the value of implementing the investment strategy X on the interval $[0, T]$. It is not hard to modify the definition to obtain a stochastic process on the whole interval $[0, T]$. For $t \leq T$, we simply sum up the increments up to t. This can be written down as:

$$I_t = Y_{t_0}(M_{t_1} - M_{t_0}) + \ldots + Y_{t_j}(M_t - M_{t_j}), \quad \text{if } t \in (t_j, t_{j+1}] \tag{6.2}$$

Example 6.1. (Integral of a simple process). Consider a standard Brownian motion $(B_t : t \in [0, 1])$ on the time interval $[0, 1]$. We know very well by now, that it is a martingale. We look at the simple integral constructed from it. We take the following integrand:

$$X_t = \begin{cases} 10 & \text{if } t \in (0, 1/3] \\ 5 & \text{if } t \in (1/3, 2/3] \\ 2 & \text{if } t \in (2/3, 1] \end{cases}$$

Then the integrals I_t as in equation (6.2) forms a process $(I_t : t \in [0, 1])$ of the form:

$$I_t = \begin{cases} 10B_t & \text{if } t \in (0, 1/3] \\ 10B_{1/3} + 5(B_t - B_{1/3}) & \text{if } t \in (1/3, 2/3] \\ 10B_{1/3} + 5(B_{2/3} - B_{1/3}) + 2(B_t - B_{2/3}) & \text{if } t \in (2/3, 1] \end{cases}$$

We make three important observations. First, the paths of the process $(I_t, t \in [0, 1])$ are continuous, because Brownian paths are. Second the process is a square-integrable martingale. It is easy to see that it is adapted and square-integrable, because I_t is a sum of square-integrable random variables. The martingale property is also not hard to verify. For example, we have for $t \in (2/3, 1]$:

$$\mathbf{E}[I_t | \mathcal{F}_{2/3}] = 10B_{1/3} + 5(B_{2/3} - B_{1/3}) + 2\mathbf{E}[B_t - B_{2/3} | \mathcal{F}_{2/3}] = I_{2/3}.$$

since $\mathbf{E}[B_t - B_{2/3} | \mathcal{F}_{2/3}] = 0$ by the martingale property of Brownian motion.

We can generalize the integrand or the investing strategy X by considering values X_{t_j} that depend on the process, hence are random, but predictable in a way. Namely, we can take X to be a random vector such that X_{t_j} is $\mathcal{F}_{t_j}$ measurable. In other words, X_{t_j} may be random, but it must depend only on the information available up to time t_j. Common sense dictates that the number of shares you buy today should not depend on the information in the future. With this in mind, for a given filtration, we define the space of simple (that is, discrete) adapted processes on $[0, T]$ as:

$$\mathcal{S}(T) = \left\{ (X_t : t \leq T) : X_t = \sum_{j=0}^{n-1} Y_{t_j} \mathbf{1}_{(t_j, t_{j+1}]}(t), Y_{t_j} \text{ is } \mathcal{F}_{t_j} \text{measurable}, \mathbf{E}[Y_{t_j}^2] < \infty \right\} \tag{6.3}$$

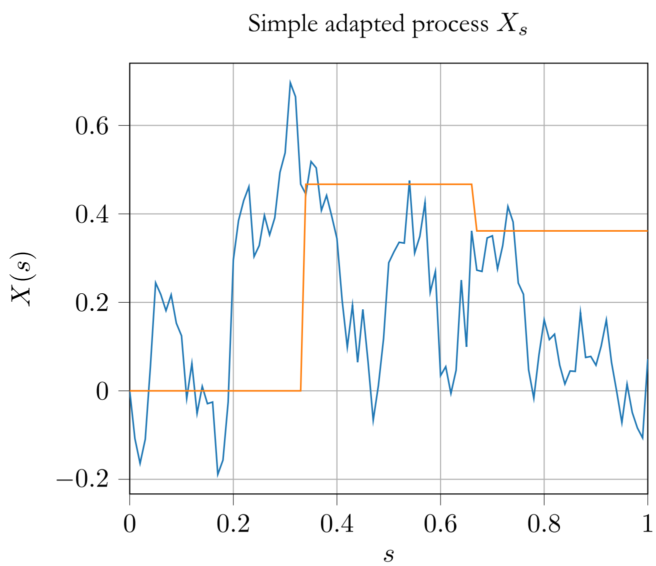
In other words, the processes in $\mathcal{S}(T)$ have paths that are piecewise constant on a finite number of intervals of $[0,T]$. The values $Y_{t_j}(\omega)$ on each time interval might vary depending on the paths ω. As random variables, the Y_{t_j}'s depend only on the information available upto time $t_j$and have a finite second moment : $\mathbf{E}[Y_{t_j}^2] < \infty$. Note that, $\mathcal{S}(T)$ is a linear space. If X, X' belong to $\mathcal{S}(T)$, then $aX + bX' \in \mathcal{S}(T)$ for all $a, b \in \mathbf{R}$. Indeed, if the paths of X, X' take a finite number of values, then so are the ones of $aX + bX'$.

Example 6.2. (An example of a simple adapted process). Let $(B_t : t \leq 1)$ be a standard Brownian motion. For the interval $[0,1]$, consider the investing strategy X in $\mathcal{S}(1)$ given by the position of the Brownian path at times $0, 1/3, 2/3$:

$$X_s = \begin{cases} 0 & \text{if } s \in [0, 1/3] \\ B_{1/3} & \text{if } s \in (1/3, 2/3] \\ B_{2/3} & \text{if } s \in (2/3, 1] \end{cases}$$

Clearly, X is simple and adapted to the Brownian filtration. For example, the value at $s = 3/4$ is $B_{2/3}$. In particular, it depends only on the information prior to the time $3/4$.

For a simple adapted process X, the integral I_t of X with respect to the martingale $(M_t : t \leq T)$ is the same as equation (6.2).



Definition 6.1. Let $(M_t : t \leq T)$ be a continuous square-integrable martingale for the filtration $(\mathcal{F}_t : t \leq T)$. Let $X \in \mathcal{S}(T)$ be a simple, adapted process $X = \sum_{j=0}^{n-1} Y_{t_j} \mathbf{1}_{(t_j, t_{j+1}]}$ on $[0,T]$. The martingale transform $I_t(X)$ is:

$$I_t(X) = \int_0^t X_s dM_s = \sum_{j=0}^{n-1} Y_{t_j}(M_{t_{j+1}} - M_{t_j})$$

It defines a process $(I_t : t \leq T)$.

Example 6.3. (Another integral of a simple process). Consider the simple process X of (6.2) defined on a Brownian motion. The integral of X as a process on $[0,1]$ is:

$$I_s(X) = \begin{cases} 0 & \text{if } s \in [0, 1/3] \\ B_{1/3}(B_s - B_{1/3}) & \text{if } s \in (1/3, 2/3] \\ B_{1/3}(B_{2/3} - B_{1/3}) + B_{2/3}(B_s - B_{2/3}) & \text{if } s \in (2/3, 1] \end{cases}$$

As in example (6.2), the paths of $I_s(X)$ are continuous for all $s \in [0,1]$, since the paths of B_s are continuous! This is also true at the integer times $s = 1/3, 2/3$, if we approached from the left or right. The process $(I_s : s \leq 1)$ is also a martingale for the Brownian filtration. The key here is that the value multiplying the increment on the interval $(t_j, t_{j+1}]$ is $\mathcal{F}_{t_j}-$measurable. For example, take $t > 2/3$ and $1/3 < s < 2/3$. The properties of conditional expectation in (5.1) and the fact that Brownian motion is a martingale give:

$$\begin{aligned}
\mathbf{E}[I_t|\mathcal{F}_s] &= \mathbf{E}[B_{1/3}(B_{2/3} - B_{1/3}) + B_{2/3}(B_t - B_{2/3})|\mathcal{F}_s] \\
&= \mathbf{E}[B_{1/3}(B_{2/3} - B_{1/3})|\mathcal{F}_s] + \mathbf{E}[B_{2/3}(B_t - B_{2/3})|\mathcal{F}_s] \\
&= B_{1/3}(B_s - B_{1/3}) + \mathbf{E}[\mathbf{E}[B_{2/3}(B_t - B_{2/3})|\mathcal{F}_{2/3}]|\mathcal{F}_s] \\
&= B_{1/3}(B_s - B_{1/3}) + \mathbf{E}[B_{2/3}\mathbf{E}[B_t - B_{2/3}|\mathcal{F}_{2/3}]|\mathcal{F}_s] \\
&= B_{1/3}(B_s - B_{1/3}) + \mathbf{E}[B_{2/3}(B_{2/3} - B_{2/3})|\mathcal{F}_s] \\
&= B_{1/3}(B_s - B_{1/3}) \\
&= I_s
\end{aligned}$$

Note that it was crucial to use the tower property in the third equality and that we took out what is known at $t = 2/3$ in the fourth equality.

Martingale transforms are always themselves martingales. In particular, it is not possible in this setup to design an investment strategy who value would be increasing on average.

Proposition 6.1. *Martingale transforms are martingales. Let $(M_t : t \leq T)$ be a continuous square-integrable martingale for the filtration $(\mathcal{F}_t : t \leq T)$ and let $X \in \mathcal{S}(T)$ be a simple process as in equation (6.3). Then, the martingale transform $(I_t : t \leq T)$ is a continuous martingale on $[0,T]$ for the same filtration.*

Proof. The fact that $I_t(X)$ is $\mathcal{F}_t-$measurable for $t \leq T$ is clear from the construction in equation (6.2). Indeed, the increments $M_{t_{j+1}} - M_{t_j}$ are $\mathcal{F}_t-$measurable for $t_{j+1} \leq t$ since the martingale is adapted. The integrand X is also adapted. Moreover, $I_t(X)$ is integrable since:

$$\begin{aligned}
\mathbf{E}[|I_t|] \leq \mathbf{E}[|I_T|] &= \mathbf{E}\left[\left|\sum_{j=0}^{n-1} Y_{t_j}(M_{t_{j+1}} - M_{t_j})\right|\right] \\
&\leq \mathbf{E}\left[\sum_{j=0}^{n-1} \left|Y_{t_j}(M_{t_{j+1}} - M_{t_j})\right|\right] = \sum_{j=0}^{n-1} \mathbf{E}[|Y_{t_j}||(M_{t_{j+1}} - M_{t_j})|] \\
&\leq \sum_{j=0}^{n-1} \left(\mathbf{E}[Y_{t_j}^2]\right)^{1/2} \left(\mathbf{E}[(M_{t_{j+1}} - M_{t_j})^2]\right)^{1/2} \\
&\quad \{\text{Cauchy-Schwarz}\}
\end{aligned}$$

Now, since both M_{t_j} and $M_{t_{j+1}}$ both belong to L^2, and L^2 is a linear space, their difference also belongs to L^2. Moreover, $\mathbf{E}[Y_{t_j}^2] < \infty$. Hence, the above sum is finite.

As for continuity, since $(M_t : t \leq T)$ is continuous, the only possible issue could be at the points t_j for some j. But in that case, we have $t > t_j$ but close and any outcome ω:

$$I_t(\omega) = \sum_{i=0}^{j-1} Y_{t_i}(M_{t_{i+1}}(\omega) - M_{t_i}(\omega)) + Y_j(M_t(\omega) - M_{t_j}(\omega))$$

as $t \to t_j^+$, $I_t \to I_{t_j}$ by continuity of $M_t(\omega)$. A similar argument holds for $t \to t_j^-$. If both the left- and right- limits exist and are equal to I_{t_j}, then I_t is continuous at t_j.

To prove the martingale property, consider $s < t$. We want to show that $\mathbf{E}[I_t|\mathcal{F}_s] = I_s$. Suppose that $t \in (t_j, t_{j+1}]$ for some $t_j < T$. By linearity of conditional expectations, we have:

$$\mathbf{E}[I_t|\mathcal{F}_s] = \sum_{i=0}^{j} \mathbf{E}[Y_{t_i}(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_s] \tag{6.4}$$

where it is understood that $t = t_{j+1}$ in the above to simplify notation. We can now handle each summand. There are three possibilities $s \geq t_{i+1}$, $s \in (t_i, t_{i+1})$ and $s < t_i$. It all depends on proposition (5.1). In the case $s \geq t_{i+1}$, we have:

$$\mathbf{E}[Y_{t_i}(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_s] = Y_{t_i}(M_{t_{i+1}} - M_{t_i})$$

since the whole summand is $\mathcal{F}_s-$measurable. In the case $s \in (t_i, t_{i+1})$, we have that Y_{t_i} is $\mathcal{F}_s-$measurable; therefore:

$$\mathbf{E}[Y_{t_i}(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_s] = Y_{t_i}\mathbf{E}[(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_s] = Y_{t_i}(M_s - M_{t_i})$$

by the martingale property. In the case, $s < t_i$, we use the tower property to get:

$$\begin{aligned}
\mathbf{E}[Y_{t_i}(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_s] &= \mathbf{E}[\mathbf{E}[Y_{t_i}(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_{t_i}]|\mathcal{F}_s] \\
&= \mathbf{E}[Y_{t_i}\mathbf{E}[(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_{t_i}]|\mathcal{F}_s] \\
&= \mathbf{E}[Y_{t_i}(M_{t_i} - M_{t_i})|\mathcal{F}_{t_i}]|\mathcal{F}_s] \\
&= 0
\end{aligned}$$

since $\mathbf{E}[(M_{t_{i+1}} - M_{t_i})|\mathcal{F}_{t_i}] = 0$ by the martingale property. Putting all the cases together, in (6.4) gives for $s \in (t_k, t_{k+1}]$, say:

$$\mathbf{E}[I_t|\mathcal{F}_s] = Y_{t_0}(M_{t_1} - M_{t_0}) + \ldots + Y_{t_{k-1}}(M_{t_k} - M_{t_{k-1}}) + Y_{t_k}(M_s - M_{t_k}) = I_s$$

□

Exercise 6.1. Let $(M_n : n \in \mathbf{N})$ be a martingale in discrete time for the filtration $(\mathcal{F}_n : n \geq 0)$. Let τ be a stopping time for the same filtration. Use the Martingale transform with the process:

$$X_n(\omega) = \begin{cases} +1 & \text{if } n < \tau(\omega) \\ 0 & \text{if } n \geq \tau(\omega) \end{cases}$$

to show that the stopped martingale $(M_{\tau \wedge n}, n \in \mathbf{N})$ is a martingale.

Proof. Let n be an arbitrary time. At any given time n, by definition of stopping times, we know if the event $\{\tau(\omega) \leq n\}$ has occurred. Thus, $X_n = \mathbf{1}_{\{\tau(\omega) \leq n\}}$ is $\mathcal{F}_n-$measurable. Also, $\mathbf{E}[X_n^2] \leq 1$. So, X_n is a simple adapted process.

Consider the martingale transform of the process $(X_n : n \in \mathbf{N})$ defined above:

$$I(n) = \sum_{i=1}^{n-1} X_i(M_{i+1} - M_i)$$

We have:

$$I_n = \begin{cases} M_n & \text{if } n < \tau(\omega) \\ M_\tau & \text{if } n \geq \tau(\omega) \end{cases}$$

That is, $I_n = M_{n\wedge\tau}$. By proposition (6.1), martingale transforms are martingales. So, a stopped martingale is also a martingale. □

6.2 The Ito Integral.

We now turn to martingale transforms where the underlying martingale is a standard Brownian motion $(B_t : t \geq 0)$. This gives our first definition of the Ito integral.

Definition 6.2. (Ito Integral on $\mathcal{S}(T)$). Let $(B_t : t \leq T)$ be a standard brownian motion on $[0, T]$ and let $X \in \mathcal{S}(T)$ be a simple process $X = \sum_{j=0}^{n-1} Y_{t_j} \mathbf{1}_{(t_j, t_{j+1}]}$ on $[0, T]$ adapted to the Brownian filtration. The Ito integral of X with respect to the Brownian motion is defined as the martingale transform:

$$\int_0^T X_s dB_s = \sum_{j=0}^{n-1} Y_{t_j}(B_{t_{j+1}} - B_{t_j})$$

and similarly for any $t \leq T$,

$$\int_0^t X_s dB_s = Y_{t_0}(B_{t_1} - B_{t_0}) + \ldots + Y_{t_j}(B_t - B_{t_j}) \quad \text{if } t \in (t_j, t_{j+1}]$$

Note again the similarities with Riemann sums. The interpretation of the Ito integral is as follows:

The value of implementing the strategy X on the underlying asset with price given by the Brownian motion.

The martingale transform with Brownian motion has more properties than with a generic martingale as given in definition (6.1). This is because Brownian motion increments are independent. We gather the properties of the Ito integral for $X \in \mathcal{S}(T)$ in an important proposition. The same exact result will hold for continuous strategies.

Proposition 6.2. *(Properties of the Ito Integral). Let $(B_t : t \leq T)$ be a standard Brownian motion on $[0, T]$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The Ito integral in the definition (6.2) has the following properties:*

- **Linearity.** If $X, X' \in \mathcal{S}(T)$ and $a, b \in \mathbf{R}$, then for all $t \leq T$,

$$\int_0^t (aX_s + bX'_s) dB_s = a \int_0^t X_s dB_s + b \int_0^t X'_s dB_s$$

- **Continuous martingale.** The process $(\int_0^t X_s dB_s, t \leq T)$ is a continuous martingale on $[0, T]$ for a Brownian filtration.

- **Ito Isometry.** The random variable $\int_0^t X_s dB_s$ is in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ with mean 0 and variance:

$$\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] = \int_0^t \mathbf{E}[X_s^2] ds = \mathbf{E}\left[\int_0^t X_s^2 ds\right], \quad t \leq T$$

It is very important for the understanding of the theory to keep in mind that $\int_0^t X_s dB_s$ is a random variable. We should walk away from the temptation to use the reflexes of classical calculus to manipulate it as if it were a Riemann Integral. The reason we use the integral sign to denote the random variable $\int_0^t X_s dB_s$ is because it shares the linearity property with the Riemann integral.

It turns out that Ito's isometry not only yields the mean and variance of the random variable $\int_0^t X_s dB_s$, but also the covariances of these random variables at different times, and the covariances for two integrals built with two different strategies on the same Brownian motion. What about the distribution of $\int_0^t X_s dB_s$? It turns out that the random variable $\int_0^t X_s dB_s$ is not Gaussian in general. However, if the process X is not random, then it will be.

Proof. The linearity property is clear from the definition of the martingale transform. The continuity and the martingale property follow from proposition (6.1).

We now prove Ito's isometry. We will use the properties of conditional expectation many times. To simplify notation, for fixed $t \in [0, T]$, we can suppose that the partition $(t_j, j \leq n)$ is a partition of $[0, t]$ with $t_n = t$. Since Y_{t_j} is $\mathcal{F}_{t_j}$-measurable, we have:

$$
\begin{aligned}
\mathbf{E}[Y_{t_j}(B_{t_{j+1}} - B_{t_j})] &= \mathbf{E}[\mathbf{E}[Y_{t_j}(B_{t_{j+1}} - B_{t_j})|\mathcal{F}_{t_j}]] \\
&= \mathbf{E}[Y_{t_j}\mathbf{E}[B_{t_{j+1}} - B_{t_j}|\mathcal{F}_{t_j}]] \\
&= 0
\end{aligned}
$$

since $\mathbf{E}[Y_{t_j}(B_{t_{j+1}} - B_{t_j})] = 0$ as Brownian motion is a martingale. Therefore, it follows that:

$$
\mathbf{E}\left[\int_0^t X_s dB_s\right] = \sum_{j=0}^{n-1} \mathbf{E}[Y_{t_j}(B_{t_{j+1}} - B_{t_j})] = 0
$$

As for the variance, we have by conditioning on $\mathcal{F}_{t_j}$, that for $t_i < t_j$:

$$
\begin{aligned}
\mathbf{E}[Y_{t_j}Y_{t_i}(B_{t_{j+1}} - B_{t_j})(B_{t_{i+1}} - B_{t_i})] &= \mathbf{E}[Y_{t_i}Y_{t_j}(B_{t_{i+1}} - B_{t_i})\mathbf{E}[(B_{t_{j+1}} - B_{t_j})|\mathcal{F}_{t_j}]] \\
&= 0
\end{aligned}
$$

since $\mathbf{E}[B_{t_{j+1}} - B_{t_j}|\mathcal{F}_{t_j}] = 0$ and since all factors but $B_{t_{j+1}} - B_{t_j}$ are $\mathcal{F}_{t_j}$-measurable. Thus, this yields:

$$
\begin{aligned}
\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] &= \sum_{i,j=0}^{n-1} \mathbf{E}[Y_{t_j}Y_{t_i}(B_{t_{j+1}} - B_{t_j})(B_{t_{i+1}} - B_{t_i})] \\
&= \sum_{j=0}^{n-1} \mathbf{E}[Y_{t_j}^2\mathbf{E}[(B_{t_{j+1}} - B_{t_j})^2|\mathcal{F}_{t_j}]]
\end{aligned}
$$

by the previous equation and the fact that Y_{t_j} is $\mathcal{F}_{t_j}$-measurable. Since the increment $B_{t_{j+1}} - B_{t_j}$ is independent of $\mathcal{F}_{t_j}$, we have:

$$
\mathbf{E}[(B_{t_{j+1}} - B_{t_j})^2|\mathcal{F}_{t_j}] = \mathbf{E}[(B_{t_{j+1}} - B_{t_j})^2] = t_{j+1} - t_j
$$

Therefore, we conclude that:

$$\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] = \sum_{j=0}^{n-1} \mathbf{E}[Y_{t_j}^2](t_{j+1} - t_j)$$

From the definition of X as a simple process in equation (6.1), we have $\int_0^t \mathbf{E}[X_s^2]ds = \sum_{j=0}^{n-1} \mathbf{E}[Y_{t_j}^2](t_{j+1} - t_j)$ since $X_s = Y_{t_j}$ on the whole interval $(t_j, t_{j+1}]$. $\square$

Example 6.4. We go back to the Ito integral in example (6.2). The mean of $I_t(X)$ is 0 by proposition (6.2) or by direct computation. It is not hard to compute the variance. For example at $t = 1$, it is:

$$\begin{aligned}\mathbf{E}[I_1^2(X)] &= \int_0^1 \mathbf{E}[X_u^2]du \\ &= \mathbf{E}[B_0^2] \cdot \frac{1}{3} + \mathbf{E}[B_{1/3}^2] \cdot \frac{1}{3} + \mathbf{E}[B_{2/3}^2]\frac{1}{3} = \frac{1}{9} + \frac{2}{9} = \frac{1}{3}\end{aligned}$$

Consider now another process Y on $[0, 1]$ defined on the same Brownian motion:

$$Y_t = B_0^2 \mathbf{1}_{[0,1/3]}(t) + B_{1/3}^2 \mathbf{1}_{(1/3,2/3]}(t) + B_{2/3}^2 \mathbf{1}_{(2/3,1]}(t)$$

Again the Ito integral $J_t = \int_0^t Y_s dB_s$ is well-defined as a process on $[0, 1]$:

$$J_t = \begin{cases} 0 & \text{if } t \in [0, 1/3] \\ B_{1/3}^2(B_t - B_{1/3}) & \text{if } t \in (1/3, 2/3] \\ B_{1/3}^2(B_{2/3} - B_{1/3}) + B_{2/3}^2(B_t - B_{2/3}) & \text{if } t \in (2/3, 1] \end{cases}$$

The covariance between the random variables I_1 and J_1 can be computed easily by using the independence of the increments and suitable conditioning. Indeed, we have:

$$\mathbf{E}[I_1(X)J_1(Y)] = \sum_{i,j=0}^{3} \mathbf{E}[B_{i/3}B_{j/3}^2(B_{(i+1)/3} - B_{i/3})(B_{(j+1)/3} - B_{j/3})]$$

If $j > i$, we can condition on $\mathcal{F}_{j/3}$ in the above summand to get:

$$\begin{aligned}&\mathbf{E}[B_{i/3}B_{j/3}^2(B_{(i+1)/3} - B_{i/3})(B_{(j+1)/3} - B_{j/3})|\mathcal{F}_{j/3}] \\ =&B_{i/3}B_{j/3}^2(B_{(i+1)/3} - B_{(j+1)/3})\mathbf{E}[(B_{(j+1)/3} - B_{j/3})|\mathcal{F}_{j/3}] = 0\end{aligned}$$

The same holds for $i > j$ by conditioning on $\mathcal{F}_{i/3}$. The only remaining terms are $i = j$:

$$\begin{aligned}\mathbf{E}[I_1 J_1] &= \sum_{i=0}^{3} \mathbf{E}[B_{i/3}^3(B_{(i+1)/3} - B_{i/3})^2] \\ &= \sum_{i=0}^{3} \mathbf{E}[B_{i/3}^3] \cdot \mathbf{E}[(B_{(i+1)/3} - B_{i/2})^2]\end{aligned}$$

by independence of increments. The first factor of each term is zero (due to the nature of odd moments of a Gaussian centered at 0). Therefore, the variables I_1 and J_1 are uncorrelated.

Remark. An *isometry* is a mapping between metric spaces(that is, with a distance) that actually preserves the distance between two points. (It literally means the *same measure* in Greek.) In case of Ito's isometry, the mapping is the one that sends the integrand X to the square-integrable random variable given by the integral:

$$I : \mathcal{S}(T) \to L^2(\Omega, \mathcal{F}, \mathbb{P})$$
$$X \mapsto \int_0^T X_s dB_s$$

The L^2-norm of $\int_0^T X_s dB_s$ is $\left\{\mathbf{E}\left[\left(\int_0^T X_s dB_s\right)^2\right]\right\}^{1/2}$. It turns out that the space $\mathcal{S}(T)$ is also a linear space with the norm $\|X\|_{\mathcal{S}} = \left(\int_0^T \mathbf{E}[X_s]^2 ds\right)^{1/2}$. Ito's isometry says that these two norms (and hence the lengths) are equal. In fact, this isometry extends in part to the L^2-space of functions on $\Omega \times [0, T]$, for which $\mathcal{S}(T)$ is a subspace. We will see that this isometry is central to the extension of the Ito integral in the limit as $n \to \infty$.

The next goal is to extend the Ito integral to processes X other than simple processes. The integral will be defined as a limit of the integrals of simple processes, much like the Riemann integral is a limit of the Riemann sums. But first, we need a good class of integrands.

Definition 6.3. For a given Brownian filtration $(\mathcal{F}_t : t \leq T)$, we consider the class of processes $\mathcal{L}_c^2(T)$ of processes $(X_t : t \leq T)$ such that the following hold:

(1) X_t is adapted. That is, X_t is $\mathcal{F}_t$-measurable.

(2) The norm of X_t:

$$\|X\|_{\mathcal{L}_c^2}^2 = \int_0^T \mathbf{E}[X_t^2]dt = \mathbf{E}\left[\int_0^T X_t^2 dt\right] < \infty$$

(3) (X_t) is almost surely continuous.

It is not hard to check that the processes $(B_t : t \leq T)$ and $(B_t^2 : t \leq T)$ are in$\mathcal{L}_c^2(T)$. In fact, if f is a continuous function and $\int_0^T \mathbf{E}[f(B_t)^2]dt < \infty$, then the process $(f(B_t) : t \leq T)$ is in $\mathcal{L}_c^2(T)$. Indeed, $f(B_t)$ is $\mathcal{F}_t$-measurable, since it is an explicit function of B_t. Moreover, the second condition is by assumption. The third simply holds because the composition of continuous functions is continuous. The main advantage of processes in $\mathcal{L}_c^2(T)$ is that they are easily approximated by simple adapted processes.

Lemma 6.1. (Approximation Lemma*).* *Let* $X \in \mathcal{L}_c^2(T)$*. Then, there exists a sequence* $(X^{(n)})$ *of simple step adapted processes in* $\mathcal{S}(T)$*, such that:*

$$\lim_{n\to\infty} \int_0^T \mathbf{E}[(X_t^{(n)} - X_t)^2]dt = 0$$

Proof. (1) For a given n, consider the partition $\{\frac{jT}{2^n}, \frac{(j+1)T}{2^n}\}$ and the simple step adapted process given by :

$$X_t^{(n)} = \sum_{j=0}^{n} X_{t_j} \mathbf{1}_{(t_j, t_{j+1}]}(t)$$

In other words, we give the constant value X_{t_j} on the whole interval $(t_j, t_{j+1}]$. By continuity of paths of X, it is clear that $X_t^{(n)}(\omega) \to X_t(\omega)$ at any $t \leq T$ and for any ω.

$(\star)$ **Justification**.

For any $s \in [0, T]$, let A_s be the set of all paths ω, such that $\lim_{t \to s} X(t, \omega) = X(s, \omega)$. Then, $\mathbb{P}(A_s) = 1$.

Pick an arbitrary $\epsilon > 0$ and fix a point $c \in [0, T]$. By definition of continuity, $(\exists \delta > 0)$ such that for all $|t - c| < \delta$ implies $|X_t - X_c| < \epsilon$. By the Archimedean property, there exists $N \in \mathbf{N}$, such that $\frac{1}{2^N} < \delta$.

We divide the interval $[0, T]$ into 2^n subintervals of size $\frac{T}{2^n}$.

We construct the simple process $X^{(n)}$, such that it takes the (random) but constant value $X^{(n)}_{t_j} = X_{\frac{jT}{2^n}}$ on the interval $\frac{jT}{2^n} < t \leq \frac{(j+1)T}{2^n}$.

$$X_t^{(n)} = \sum_{j=0}^{n-1} X_{\frac{jT}{2^n}} \cdot 1_{t \in \left[\frac{jT}{2^n}, \frac{(j+1)T}{2^n}\right)}$$

Now, there exists a sequence of dyadic intervals $I_N \subseteq I_{N+1} \subseteq \ldots$ always containing the point c.

From the almost sure continuity of $(X_t, t \leq T)$, it follows that, for all $n \geq N$, since $l(I_n) < \delta$, it follows that $|X_{\frac{jT}{2^n}} - X_c| < \epsilon$ with probability 1. Now, for all $n \geq N$, $X_c^{(n)} = X_{\frac{jT}{2^n}}$. Consequently, for all $n \geq N$, $\left|X_c^{(n)} - X_c\right| < \epsilon$ with probability 1.

Thus, $X_t^{(n)} \xrightarrow{a.s.} X_t$.

(2) Assume that $(X_t^{(n)} - X_t)$ is uniformly bounded. $(\exists M)$ $(\forall \omega)$ s.t. $|X_t^{(n)}(\omega) - X_t(\omega)| \leq M$.

$$\left\|X_t^{(n)} - X_t\right\|^2_{\mathcal{L}^2_c} = \int_0^T \mathbf{E}[(X_t^{(n)} - X_t)^2] dt$$

Passing to the limit on both sides, by the dominated convergence theorem:

$$\lim \left\|X_t^{(n)} - X_t\right\|_{\mathcal{L}^2_c} = \lim_{n \to \infty} \int_0^T \mathbf{E}[(X_t^{(n)} - X_t)^2] dt = 0$$

$\square$

Lemma 6.2. *(Convergence in L^2 implies convergence of the first and second moments). Let $(X_n : n \geq 0)$ be a sequence of random variables that converge to X in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.*

(a) Show that $\mathbf{E}[X_n^2]$ converges to $\mathbf{E}[X^2]$.

Hint: Write $X = (X - X_n) + X_n$. The Cauchy Schwarz inequality might be useful.

(b) Show that $\mathbf{E}[X_n]$ converges to $\mathbf{E}[X]$.

Hint: Write $|\mathbf{E}[X_n] - \mathbf{E}[X]|$ and use Jensen's inequality twice.

(a) We are given that $\lim_{n \to \infty} \mathbf{E}[|X_n - X|^2] \to 0$. Firstly, let $c(x) = |x|$. Let $p \in (0, 1)$. We have:

$$\begin{aligned} c(pa + (1-p)b) &= |pa + (1-p)b| \\ &\leq p|a| + (1-p)|b| \\ &= pc(a) + (1-p)c(b) \end{aligned}$$

Hence, $|x|$ is a convex function. Consequently, $0 \leq |\mathbf{E}[X^2 - X_n^2]| \leq \mathbf{E}[|X^2 - X_n^2]$. Therefore, we can write:

$$
\begin{aligned}
0 \le |\mathbf{E}[X^2 - X_n^2]| &\le \mathbf{E}[|X^2 - X_n^2|] \\
&= \mathbf{E}[|((X - X_n) + X_n)^2 - X_n^2|] \\
&= \mathbf{E}[|(X - X_n)^2 + 2(X - X_n)X_n + X_n^2 - X_n^2|] \\
&\le \mathbf{E}[|X - X_n|^2] + 2\mathbf{E}[|(X - X_n)(X_n)|] \\
&\le \mathbf{E}[|X - X_n|^2] + 2\left(\mathbf{E}[|(X - X_n)|^2]\right)^{1/2}\left(\mathbf{E}[|X_n|^2]\right)^{1/2} \\
&\{\text{Cauchy-Schwarz}\}
\end{aligned}
$$

Passing to the limit on both sides as $n \to \infty$, it follows that lim $|\mathbf{E}[X_n^2] - \mathbf{E}[X^2]| \to 0$. Consequently, $\mathbf{E}[X_n^2] \to \mathbf{E}[X^2]$.

(b) We have:

$$
\begin{aligned}
0 \le |\mathbf{E}[X_n] - \mathbf{E}[X]| &= |\mathbf{E}[X_n - X]| \\
&\le \mathbf{E}[|X_n - X|] \\
&\{\text{since } |x| \text{ is a convex function}\} \\
&\le \left(\mathbf{E}[|X_n - X|^2]\right)^{1/2}
\end{aligned}
$$

Passing to the limit on both sides, as $n \to \infty$, $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

Lemma 6.3. *We prove that the space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is complete; that is, if $(X_n : n \ge 1)$ is a Cauchy sequence in L^2, then there exists $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ such that $X_n \to X$ in L^2.*

(a) Argue from the definition of Cauchy sequence that we can find a subsequence $(X_{n_k} : k \ge 0)$ such that $\|X_m - X_{n_k}\| \le 2^{-k}$ for all $m > n_k$ where $\|\cdot\|$ is the L^2 norm.

Proof. We are given that $(X_n : n \in \mathbf{N})$ is a Cauchy sequence. For a Cauchy sequence, given any $\epsilon > 0$, we can find $N(\epsilon)$, such that for all $m > n \ge N(\epsilon)$, $||X_n - X_m|| < \epsilon$. Let $\epsilon_k = \frac{1}{2^k}$. There exists n_1 such that for all $m \ge n_1$, $||X_{n_1} - X_m|| < \epsilon$. There exists n_2 such that for all $m \ge n_2$, $||X_{n_2} - X_m|| < \frac{\epsilon}{2}$. In general, there exists n_k, such that for all $m \ge n_k$,

$$
\|X_{n_k} - X_m\|_{L^2} < \epsilon_k = \frac{1}{2^k}
$$

□

(b) Consider the candidate limit $\sum_{j=0}^{\infty}(X_{n_{j+1}} - X_{n_j})$ with $X_{n_0} = 0$. Show that this sum converges almost surely (so X is well-defined) by considering

$$
\sum_{j=0}^{k} \mathbf{E}\left[\left|X_{n_{j+1}} - X_{n_j}\right|\right]
$$

Proof. Firstly, by Jensen's inequality, we have $(\mathbf{E}[|X|])^2 \le \mathbf{E}[|X|^2]$. So, $\mathbf{E}[|X|] \le \mathbf{E}[|X|^2]^{\frac{1}{2}}$ or $\|X\|_{L^1} \le \|X\|_{L^2}$.

Let $A_k(\epsilon)$ be the event that there is an excursion at n_k, $\left\{\left|X_{n_{k+1}} - X_{n_k}\right| > \epsilon\right\}$. And let $B_k(\epsilon)$ be the event that atleast one of $A_k, A_{k+1}, \ldots$ occurs, $\bigcup_{m \ge k} A_m(\epsilon)$.

$$\begin{aligned}
\mathbb{P}(B_k) &= \mathbb{P}(\cup_{m\geq k} A_m) \\
&\leq \sum_{m\geq k} \mathbb{P}(A_m) \\
&\{\text{Union bound}\} \\
&\leq \frac{1}{\epsilon} \sum_{m\geq k} \mathbb{E}[|X_{n_{m+1}} - X_{n_m}|] \\
&\{\text{Chebyshev Inequality}\} \\
&\leq \frac{1}{\epsilon} \sum_{m\geq k} \left\| X_{n_{m+1}} - X_{n_m} \right\|_{L^2} \\
&\leq \frac{1}{\epsilon} \sum_{m\geq k} \frac{1}{2^m}
\end{aligned}$$

The series on the right is convergent.

$$\lim_{k\to\infty} \mathbb{P}(B_k) \leq \frac{1}{\epsilon} \lim_{k\to\infty} \sum_{m\geq k} \frac{1}{2^m}$$

The tail sum of a convergent series approaches zero. So, $\lim_{k\to\infty} \mathbb{P}(B_k) = 0$.

By the necessary and sufficient condition for almost sure convergence, the series $\sum \left(X_{n_{k+1}} - X_{n_k}\right)$ converges almost surely. $\square$

(c) Show that $\|X - X_{n_k}\|_{L_2} \to 0$ as $k \to \infty$. Conclude that $\|X\| < \infty$. (This shows the convergence in L^2 along the subsequence!)

Proof. We have shown that the sum $\sum_{k=0}^{\infty}(X_{n_{k+1}} - X_{n_k})$ converges and let the candidate limit be X. Therefore, $\lim_{N\to\infty} \sum_{j=0}^{N}(X_{n_{j+1}} - X_{n_j}) = X$ almost surely. Therefore:

$$X - X_{n_k} = \lim_{N\to\infty} \sum_{j=k}^{N}(X_{n_{j+1}} - X_{n_j}) = \liminf_{N\to\infty} \sum_{j=k}^{N}(X_{n_{j+1}} - X_{n_j})$$

since $X_{n_k} = \sum_{j=0}^{k-1}(X_{n_{j+1}} - X_{n_j})$ and $\liminf a_n = \lim a_n = \limsup a_n$ for a convergent sequence (a_n). We have:

$$\|X - X_{n_k}\|_2 = \left\| \liminf_{N\to\infty} \sum_{j=k}^{N}(X_{n_{j+1}} - X_{n_j}) \right\|_2$$

By Fatou's lemma,

$$
\begin{aligned}
\left\| \liminf_{N\to\infty} \sum_{j=k}^{N} (X_{n_{j+1}} - X_{n_j}) \right\|_2 &\leq \liminf_{N\to\infty} \left\| \sum_{j=k}^{N} (X_{n_{j+1}} - X_{n_j}) \right\|_2 \\
&\leq \liminf_{N\to\infty} \sum_{j=k}^{N} \left\| (X_{n_{j+1}} - X_{n_j}) \right\|_2 \\
&\{\text{Triangle Inequality}\} \\
&= \frac{1}{2^k} \left(1 + \frac{1}{2} + \frac{1}{2^2} + \dots \right) \\
&= \frac{1}{2^{k-1}}
\end{aligned}
$$

Thus, as $k \to \infty$, $\|X - X_{n_k}\|_2 \to 0$.

Similarly,

$$
X = \lim_{N\to\infty} \sum_{j=0}^{N} (X_{n_{j+1}} - X_{n_j}) = \liminf_{N\to\infty} \sum_{j=0}^{N} (X_{n_{j+1}} - X_{n_j})
$$

So,

$$
\begin{aligned}
\|X\|_2 &= \left\| \liminf_{N\to\infty} \sum_{j=0}^{N} (X_{n_{j+1}} - X_{n_j}) \right\|_2 \\
&\leq \liminf_{N\to\infty} \left\| \sum_{j=0}^{N} (X_{n_{j+1}} - X_{n_j}) \right\|_2 \\
&\{\text{Fatou's Lemma}\} \\
&\leq \liminf_{N\to\infty} \sum_{j=0}^{N} \left\| (X_{n_{j+1}} - X_{n_j}) \right\|_2 \\
&\{\text{Triangle Inequality}\} \\
&= 1 + \frac{1}{2} + \frac{1}{2^2} + \dots \\
&< \infty
\end{aligned}
$$

(d) Use again the Cauchy definition and the subsequence to show convergence of the whole sequence that is, $\|X - X_n\| \to 0$.

We have:

$$
\begin{aligned}
\|X - X_n\| &= \|X - X_{n_k} + X_{n_k} - X_n\| \\
&\leq \|X - X_{n_k}\| + \|X_{n_k} - X_n\| \\
&\{\text{Triangle inequality}\}
\end{aligned}
$$

Pick an arbitrary $\epsilon > 0$. There exists $K_1(\epsilon)$ such that, $\left\|X - X_{n_{K_1}}\right\| < \epsilon/2$. There exists $K_2(\epsilon)$ such that for all $n > n_{K_2}$, $\left\|X_{n_{K_2}} - X_n\right\| < \epsilon/2$. Pick $n_K = \max\{n_{K_1}, n_{K_2}\}$. Then, for all $n > n_K$, $\|X - X_n\| < \epsilon$. Consequently, $\|X - X_n\| \to 0$. $\square$

Theorem 6.1. *Let $(B_t : t \leq T)$ be a standard Brownian motion on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $(X_t : t \leq T)$ be a process in $\mathcal{L}_c^2(T)$. There exist random variables $\int_0^t X_s dB_s$, $t \leq T$ with the following properties:*

(1) Linearity: If $X, Y \in \mathcal{L}_c^2(T)$ and $a, b \in \mathbf{R}$, then

$$\int_0^t (aX_s + bY_s) dB_s = a \int_0^t X_s dB_s + b \int_0^t Y_s dB_s, \quad t \leq T$$

(2) Continuous Martingale: The process $(\int_0^t X_s dB_s, t \leq T)$ is a continuous martingale for the Brownian filtration.

(3) Ito's Isometry: The random variable $\int_0^t X_s dB_s$ is in $L^2(\Omega, \mathcal{F}, \mathbb{P})$ with mean 0 and variance

$$\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] = \int_0^t \mathbf{E}[X_s^2] ds = \mathbf{E}\left[\int_0^t X_s^2 ds\right], \quad t \leq T$$

In other words,

$$\left\| I^X(t) \right\|_{L^2} = \|X\|_{\mathcal{L}_c^2}$$

Proof. Consider the process $X = (X_t : t \leq T)$ in $\mathcal{L}_c^2(T)$. By the approximation lemma (6.1), we can approximate X by a sequence of simple adapted processes $(X_t^{(n)} : t \leq T)$. In particular, that the sequence is Cauchy for the metric

$$\left\| X^{(n)} - X^{(m)} \right\|_{\mathcal{L}_c^2} = \left(\int_0^t \mathbf{E}[(X_s^{(n)} - X_s^{(m)})^2] dt \right)^{1/2} \tag{6.5}$$

The key step is the following. We know that the integral $I^{X^{(n)}}(t) = \int_0^t X_s^{(n)} dB_s$ is well defined as a random variable in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Moreover, we know from Ito Isometry, that the L^2-distance of the processes in equation (6.5) is the same as the L^2 distance of the $I^{X^{(n)}}$'s. Since $(X^{(n)})$ is Cauchy in $\mathcal{L}_c^2$, it means that the sequence $(I^{X^{(n)}}, n \in \mathbf{N})$ is Cauchy in L^2. By Cauchy completeness property, $I^{X^{(n)}}$ converges in L^2 to a random variable that we denote by $I^X(t)$ or $\int_0^t X_s dB_s$. Furthermore, the limit $I^X(t)$ does not depend on the approximating sequence $(X^{(n)})$. We could have taken any other sequence to approximate X and Ito isometry guarantees, that the corresponding integrals will converge to the same random variable.

We now prove the properties.

(1) Linearity. It follows by using linearity property in proposition (6.2) for $X^{(n)}$ and $Y^{(n)}$, the two sequences of approximating processes of X and Y that:

$$\begin{aligned} I(aX^{(n)} + bY^{(n)}) &= aI(X^{(n)}) + bI(Y^{(n)}) \\ \lim_{n\to\infty} I(aX^{(n)} + bY^{(n)}) &= a \lim_{n\to\infty} I(X^{(n)}) + b \lim_{n\to\infty} I(Y^{(n)}) \\ I(aX + bY) &= aI(X) + bI(Y) \end{aligned}$$

(2) Isometry. We refer lemma (6.2). The variance property now follows from the following facts:

The convergence of a sequence (X_n) to X in L^2, implies the convergence of the first and second moments.

$\mathcal{L}_c^2$ is a subspace of L^2. Since, $X^{(n)} \to X$ in $\mathcal{L}_c^2$, it follows that $||X^{(n)} - X|| \to 0$. Thus,$\int_0^t \mathbf{E}\left[\left(X_s^{(n)}\right)^2\right] ds$ converges to $\int_0^t \mathbf{E}[X_s^2] ds$.

Now, $I_t^{X^{(n)}} \to I_t$ in L^2. So, $\mathbf{E}[I_t^{X^{(n)}}] \to \mathbf{E}[I_t^X]$ and $\mathbf{E}[\left(I_t^{X^{(n)}}\right)^2] \to \mathbf{E}[\left(I_t^X\right)^2]$. That is, $\mathbf{E}\left[\left(\int_0^t X_s^{(n)} dB_s\right)^2\right]$ converges to $\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right]$.

Since, Ito isometry for simple adapted processes implies

$$\int_0^t \mathbf{E}\left[\left(X_s^{(n)}\right)^2\right] ds = \mathbf{E}[\left(I_t^{X^{(n)}}\right)^2]$$

these sequences are equal and have the same limits. Consequently,

$$\int_0^t \mathbf{E}[X_s^2] ds = \mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right]$$

(3) Continuous Martingale. Write $I_t = \int_0^t X_s dB_s$. We must show that $\mathbf{E}[I_t|\mathcal{F}_s] = I_s$ for any $t > s$. To see this, we go back to the definition of conditional expectations. The random variable I_t must be $\mathcal{F}_t$ measurable by construction. Now, for a bounded random variable W that is $\mathcal{F}_s$-measurable, we need to show that:

$$\mathbf{E}[WI_t] = \mathbf{E}[W\mathbf{E}[I_t|\mathcal{F}_s]] = \mathbf{E}[WI_s]$$

This is clear for $I_t^{(n)}$, the approximating integrals, because $(I_t^{(n)}, t \le T)$ is a martingale. The above then follows from the fact that $WI_s^{(n)}$ converges to WI_s in L^2 (and thus the expectation converges) and the same way for t. The fact that the path $t \to I_t(\omega)$ is continuous with probability one is a bit more involved. It uses Doob's maximal inequality. $\square$

Example 6.5. (Sampling Ito Integrals) How can we sample paths of processes given by Ito integrals? A very simple method is to go back to the integral on simple processes. Consider the process $I_t = \int_0^t X_s dB_s, t \le T$ constructed from $X \in \mathcal{L}_c^2(T)$ and from a standard brownian motion $(B_t, t \ge 0)$. To simulate paths, we fix the endpoint, say T and a step-size $1/n$. Then, we can generate the process at every $t_j = \frac{jT}{n}$ by taking

$$I_{t_j} = \sum_{i=0}^{j-1} X_{t_i}(B_{t_{i+1}} - B_{t_i}), \quad j \le n$$

Here are two observations that makes this expression more palatable. First note that the increment $B_{t_{i+1}} - B_{t_i}$ is a Gaussian random variable of mean 0 and variance $\frac{T}{n}$ for every i. Second, we have $I_{t_j} - I_{t_{j-1}} = X_{t_{j-1}}(B_{t_j} - B_{t_{j-1}})$, so the values I_{t_j} can be computed recursively.

Once the conclusions of theorem (6.1) are accepted, we are free to explore the beauty and the power of Ito Calculus. As a first step, we observe that with Ito's isometry, we can compute not only variances, but also covariances between integrals. This is because an isometry also preserves the inner product in L^2 spaces.

Example 6.6. Increments of martingales are uncorrelated.

(a) Let $(M_t : t \ge 0)$ be a square integrable martingale for the filtration $(\mathcal{F}_t : t \ge 0)$. Use the properties of conditional expectation to show that for $t_1 \le t_2 \le t_3 \le t_4$, we have:

$$\mathbf{E}[(M_{t_2} - M_{t_1})(M_{t_4} - M_{t_3})] = 0$$

(b) Let $(B_t : t \geq 0)$ be a standard brownian motion, and let $(X_t : t \leq T)$ be a process in $\mathcal{L}_c^2(T)$. Use part(a) to show that the covariance between integrals at different times $t < t'$ is:

$$\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)\left(\int_0^{t'} X_s dB_s\right)\right] = \int_0^{t \wedge t'} \mathbf{E}[X_s^2] ds, \quad t, t \leq T$$

Solution.

(a) We have:

$$\begin{aligned}
\mathbf{E}\left[(M_{t_2} - M_{t_1})(M_{t_4} - M_{t_3})\right] &= \mathbf{E}\left[(M_{t_2} - M_{t_1})\mathbf{E}[(M_{t_4} - M_{t_3})|\mathcal{F}_{t_3}]\right] \\
&= \mathbf{E}\left[(M_{t_2} - M_{t_1})(M_{t_3} - M_{t_3})|\mathcal{F}_{t_3}\right] \\
&= 0
\end{aligned}$$

(b) Let $(X_t : t \leq T)$ be a process in $\mathcal{L}_c^2(T)$. We know that the Ito integral $I^X(t) = \int_0^t X_s dB_s$ is a continuous martingale. With $0 \leq t \leq t'$, consider the increments: $(I^X(t) - 0)$, $(I^X(t') - I^X(t))$. These increments are uncorrelated. Hence,

$$\begin{aligned}
\mathbf{E}\left[I^X(t)(I^X(t') - I^X(t))\right] &= 0 \\
\mathbf{E}[I^X(t)I^X(t')] &= \mathbf{E}[(I^X(t))^2] \\
\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)\left(\int_0^{t'} X_s dB_s\right)\right] &= \mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] \\
&= \int_0^{t \vee t'} \mathbf{E}[X_s]^2 ds
\end{aligned}$$

This closes the proof.

Corollary 6.1. *Let $(B_t : t \leq T)$ be a standard brownian motion, and let $X \in \mathcal{L}_c^2(T)$. We have:*

$$\mathbf{E}\left[I_t I_{t'}\right] = \mathbf{E}\left[\left(\int_0^t X_s dB_s\right)\left(\int_0^{t'} X_s dB_s\right)\right] = \int_0^{t \wedge t'} \mathbf{E}[X_s^2] ds, \quad t, t \leq T$$

for any $Y \in \mathcal{L}_c^2(T)$, and

$$\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)\left(\int_0^t Y_s dB_s\right)\right] = \int_0^t \mathbf{E}[X_s Y_s] ds, \quad t \leq T$$

Note that, when X is just a constant 1, we recover from the first equation the covariance of the Brownian motion.

Proof. We just proved assertion (1) in the example (6.6). As for the second, we have on one hand by Ito's isometry:

$$
\begin{aligned}
\mathbf{E}\left[\left(\int_0^t \{X_s+Y_s\}dB_s\right)^2\right] &= \int_0^t \mathbf{E}[(X_s+Y_s)^2]ds \\
&= \int_0^t \mathbf{E}[X_s^2]ds + \int_0^t \mathbf{E}[Y_s^2]ds + 2\int_0^t \mathbf{E}[X_sY_s]ds
\end{aligned}
$$

On the other hand, by linearity of Ito integral and of the expectation, we have:

$$
\begin{aligned}
\mathbf{E}\left[\left(\int_0^t \{X_s+Y_s\}dB_s\right)^2\right] &= \mathbf{E}\left[\left(\int_0^t X_sdB_s + \int Y_sdB_s\right)^2\right] \\
&= \mathbf{E}\left[\left(\int_0^t X_sdB_s\right)^2\right] + \mathbf{E}\left[\left(\int_0^t Y_sdB_s\right)^2\right] \\
&+ 2\mathbf{E}\left[\left(\int_0^t X_sdB_s\right)\left(\int_0^t Y_sdB_s\right)\right]
\end{aligned}
$$

By Ito's Isometry, $\|I_s(X)\|_{L^2} = \|X\|_{\mathcal{L}_c^2}$ and $\|I_s(Y)\|_{L^2} = \|Y\|_{\mathcal{L}_c^2}$. Hence, by equating the above two expressions, we conclude that:

$$
\mathbf{E}\left[\left(\int_0^t X_sdB_s\right)\left(\int_0^t Y_sdB_s\right)\right] = \int_0^t \mathbf{E}[X_sY_s]ds
$$

Thus, Ito isometry also preserves inner products. $\langle I_t(X), I_t(Y)\rangle_{L^2} = \langle X_t, Y_t\rangle_{\mathcal{L}_c^2}$. □

Example 6.7. Consider the processes $(B_t : t \le T)$ and $(B_t^2 : t \le T)$ for a given standard Brownian motion. Note that these two processes are in $\mathcal{L}_c^2(T)$ for any $T > 0$. By the existence theorem (6.1), the random variables

$$
I_t = \int_0^t B_sdB_s, \quad J_t = \int_0^t B_s^2dB_s
$$

exist and are in $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Their mean is 0 and they have variances

$$
\mathbf{E}[I_t^2] = \int_0^t \mathbf{E}[B_s^2]ds = \int_0^t sds = \frac{t^2}{2}
$$

and

$$
\mathbf{E}[J_t^2] = \int_0^t \mathbf{E}[B_s^4]ds = \int_0^t 3s^2ds = t^3
$$

The covariance by corollary (6.1) is :

$$
\mathbf{E}[I_tJ_t] = \int_0^t \mathbf{E}[B_s^3]ds = 0
$$

The variables are uncorrelated.

Example 6.8. (A path-dependent integral) Consider the process $X_t = \int_0^t B_s dB_s$ on $[0, T]$ as in example (6.7). Note that the process $(X_t : t \leq T)$ is itself in $\mathcal{L}_c^2(T)$. In particular, the integral $\int_0^t X_s dB_s$ is well-defined! (Note that the integrand X_t is $\mathcal{F}_t-$measurable but its value depends on the whole Brownian motion upto time t). The mean of the integral is 0 and its variance is obtained by applying Ito's isometry twice:

$$
\begin{aligned}
\mathbf{E}\left[\left(\int_0^t X_s dB_s\right)^2\right] &= \int_0^t \mathbf{E}[X_s^2] ds \\
&= \int_0^t \left(\int_0^s \mathbf{E}[B_s^2] ds\right) ds \\
&= \int_0^t \left(\int_0^s s ds\right) ds \\
&= \int_0^t (s^2/2) ds \\
&= \frac{t^3}{6}
\end{aligned}
$$

In general, the Ito integral is not Gaussian. However, if the integrand X is not random, the process is actually Gaussian. In this particular case, the integral is sometimes called the *Wiener Integral.*

Corollary 6.2. *(Wiener Integral.) Let* $(B_t : t \leq T)$ *be a standard Brownian motion and let* $f : [0, T] \to \mathbf{R}$ *be a function such that* $\int_0^T f^2(s) ds < \infty$. *Then, the process* $(I_t(f) : t \leq T) = (\int_0^t f(s) dB_s : t \leq T)$ *is Gaussian with mean* 0 *and covariance:*

$$
Cov\left(\int_0^t f(s) dB_s, \int_0^{t'} f(s) dB_s\right) = \int_0^{t \wedge t'} f(s)^2 ds
$$

Proof. We prove the case when f is continuous. In this case, we can use the proof of the approximation Lemma (6.1). Let $(t_j : j \leq 2^n)$ be a partition of 2^n intervals. The lemma shows that the sequence of functions:

$$
f^{(n)}(t) = \sum_{j=0}^{2^n-1} f(t_j) \mathbf{1}_{(t_j, t_{j+1}]}(t), \quad t \leq T
$$

approximates f. The Ito integral of $f^{(n)}$ is:

$$
I_t^{(n)} = \sum_{i=0}^{2^n-1} f(t_j)(B_{t_{j+1}} - B_{t_j}), \quad t \in (t_j, t_{j+1}]
$$

This is a Gaussian process for any n. This is because for any choice of times $s_1, \ldots, s_m$, the vector $(I_{s_1}^{(n)}, I_{s_2}^{(n)}, \ldots, I_{s_m}^{(n)})$ is Gaussian, since it reduces to linear combinations of Brownian motion increments at fixed times. Moreover, the random variable $\int_0^t f(s) dB_s$ is the L^2 limit of $I_t^{(n)}$ by existence theorem (6.1). It remains to show that an L^2-limit of a sequence of Gaussian vectors is Gaussian. This is sketched in the example below. The expression for covariances follows from the collary on covariance of Ito integrals (6.1). □

Exercise 6.2. L^2-limit of Gaussians is Gaussian. Let $(X_n : n \geq 0)$ be a sequence of Gaussian random variables that converge to X in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

(a)Show that X is also Gaussian.

Hint: Use the characteristic function of a Gaussian random variable. Use also the fact that there is a subsequence that converges almost surely.

(b) Find its mean and variance in terms of X.

Solution.

(a)The sequence (X_n) converges to X in L^2. Thus, $\mathbf{E}X_n \to \mathbf{E}X$ and $\mathbf{E}X_n^2 \to \mathbf{E}X^2$.

Also, mean square convergence implies convergence in probability, which in turn also implies that there exists a subsequence (X_{n_k}) that converges to X almost surely.

The characteristic function of X_{n_k} is:

$$M_{X_{n_k}}(s) = \exp\left[\frac{1}{2}\mathbf{E}[(X_{n_k} - \mathbf{E}X_{n_k})^2]s^2\right]$$

Moreover, as $X_{n_k} \to X$ in distribution, their characteristic functions converge. So,

$$M_X(s) = \exp\left[\frac{1}{2}\mathbf{E}[(X - \mathbf{E}X)^2]s^2\right]$$

This shows that X is a Gaussian random variable.

Example 6.9. (Ornstein-Uhlenbeck process as an Ito Integral). Consider the function $f(s) = e^s$. The Ornstein-Uhlenbeck process starting at X_0 can also be written as:

$$Y_t = e^{-t}\int_0^t e^s dB_s, \quad t \geq 0$$

To see this mathematically, not that $(Y_t : t \geq 0)$ is a Gaussian process by corollary (6.2). The mean is 0 and the covariance by corollary (6.1) is:

$$\begin{aligned}
\mathbf{E}[Y_tY_s] &= e^{-t} \cdot e^{-s}\int_0^s e^{2u}du = e^{-t-s}\cdot\left[\frac{e^{2u}}{2}\right]_0^s \\
&= e^{-t-s}\cdot\left(\frac{e^{2s}}{2} - \frac{1}{2}\right) \\
&= \frac{1}{2}(e^{-(t-s)} - e^{-(t+s)}), \quad s \leq t
\end{aligned}$$

In this case, the process is stationary in the sense that $(Y_t : t \geq 0)$ has the same distribution as $(Y_{t+a} : t \geq 0)$ for any $a \geq 0$.

Exercise 6.3. Another application of Doob's maximal inequality. Let $(B_t : t \in [0,1])$ be a Brownian motion defined on $(\Omega, \mathcal{F}, \mathbb{P})$. The Brownian bridge $(Z_t : t \in [0,1])$ is the stochastic process with the distribution defined in example (3.18). Another way to construct a Brownian bridge is as follows:

$$Z_t = (1-t)\int_0^t \frac{1}{1-s}dB_s, \quad t < 1$$

In this exercise, we prove that $\lim_{t\to 1} Z_t = 0$ almost surely as expected.

(a) Show that $\lim_{t\to 1} Z_t = 0$ in $L^2(\Omega, \mathcal{F}, \mathbb{P})$.

Solution.

By Ito Isometry,

$$\begin{aligned}
\mathbf{E}[Z_t^2] &= (1-t)^2 \int_0^t \frac{1}{(1-s)^2} ds \\
&= (1-t)^2 \left[\frac{1}{(1-s)} \right]_0^t \\
&= (1-t)^2 \left(\frac{1}{1-t} - 1 \right) \\
&= (1-t)^2 \left(\frac{t}{(1-t)} \right) \\
&= t(1-t)
\end{aligned}$$

Thus, $\lim_{t\to 1} \mathbf{E}[Z_t^2] = 0$. Hence, $\lim_{t\to 1} Z_t \xrightarrow{L^2} 0$.

(b) Using the Doob's maximal inequality of example (5.28) show that:

$$\mathbb{P}\left(\max_{t \in \left[1-\frac{1}{2^n}, 1-\frac{1}{2^{n+1}}\right]} |Z_t| > \delta \right) < \frac{1}{\delta^2} \frac{1}{2^{n-1}}$$

Solution.

Pick an arbitrary $\delta > 0$. Let $A_n(\delta)$ be the event that Z_t exceeds δ in the interval $\left[1-\frac{1}{2^n}, 1-\frac{1}{2^{n+1}}\right]$. By Doob's maximal inequality, the probability of this event is bounded by:

$$\begin{aligned}
\mathbb{P}\left(A_n(\delta)\right) &\leq \frac{1}{\delta^2} \mathbf{E}[Z^2_{1-\frac{1}{2^{n+1}}}] \\
&= \frac{1}{\delta^2} \left(1 - \frac{1}{2^{n+1}}\right) \frac{1}{2^{n+1}} \\
&\leq \frac{1}{\delta^2} \cdot \frac{1}{2^{n+1}} \\
&\leq \frac{1}{\delta^2} \cdot \frac{1}{2^{n-1}}
\end{aligned}$$

(c) Deduce that $\lim_{t\to 1} Z_t = 0$ almost surely using Borel-Cantelli Lemma.

Solution.

Consider the infinite series $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\delta))$. We have:

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}(A_n(\delta)) &\leq \frac{1}{\delta^2} \sum_{n=1}^{\infty} \frac{1}{2^{n-1}} \\
&= \frac{2}{\delta^2}
\end{aligned}$$

Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n(\delta)) < \infty$, by BCL1(Borel-Cantelli Lemma 1), the event $A_n(\delta)$ occurs finitely many times, almost surely. There exists $n_0 \in \mathbf{N}$, such that for all $n \geq n_0$, $\max_{t \in \left[1-\frac{1}{2^n}, 1-\frac{1}{2^{n+1}}\right]} |Z_t| \leq \delta$ with probability 1. But, $\lim_{n\to\infty} \max_{t \in \left[1-\frac{1}{2^n}, 1-\frac{1}{2^{n+1}}\right]} |Z_t| = Z_1$. Consequently, $(\forall \delta > 0), \mathbb{P}(Z_1 \leq \delta) = 1$.

Example 6.10. (Brownian bridge as an Ito Integral) We know that another way to construct a Brownian bridge process is as follows:

$$Z_t = (1-t)\int_0^t \frac{1}{1-s} dB_s$$

We know that $\lim_{t\to 1} Z_t = 0$ almost surely. The process Z is a Gaussian process by corollary (6.2). The mean is zero and the covariance is, by corollary (6.1) is given by:

$$\begin{aligned}
\mathbf{E}[Z_t Z_s] &= (1-s)(1-t)\int_0^s \frac{1}{(1-s)^2} dB_s \\
&= (1-s)(1-t)\left[\frac{1}{(1-s)}\right]_0^s \\
&= (1-s)(1-t)\left(\frac{1}{1-s} - 1\right) \\
&= (1-s)(1-t)\left(\frac{s}{1-s}\right) \\
&= s(1-t)
\end{aligned}$$

The above representations of the Ornstein-Uhlenbeck process and the brownian bridge implies that they are not martingales.

6.3 Ito's Formula.

The Ito integral was constructed in the last section in a rather abstract way. It is the limit of a sequence of random variables constructed from Brownian motion. It is good to remind ourselves that the classical Riemann integral is also very abstract! It is defined as the limit of the sequence of Riemann sums. It does not always have an explicit form. For example, the CDF of a Gaussian random variable

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^x e^{(-y^2/2)} dy$$

is a well-defined function of x, but the integral cannot be expressed in terms of the typical elementary functions of calculus. But, in some cases a Riemann integral can be written explicitly in terms of such functions. This is the content of the fundamental theorem of calculus. It is useful to recall the theorem, as Ito's formula is built upon it.

Let $f : [0,T] \to \mathbf{R}$ be a function for which the derivative f' exists and is a continuous function on $[0,T]$. We will say that such a function is in $\mathcal{C}^1([0,T])$. The fundamental theorem of calculus says that we can write:

$$f(t) - f(0) = \int_0^t f'(s) ds, \quad t \leq T \tag{6.6}$$

Note that, we often write this result in the differential form:

$$df(t) = f'(t)dt$$

The differential form has no rigorous meaning in itself. It is simply a compact and convenient notation that encodes FTC (6.6).

The stochastic equivalent of the fundamental theorem of calculus is the Ito's formula provided below. It related the Ito integral to an explicit function of Brownian motion. Note that the function f must be in $\mathcal{C}^2(\mathbf{R})$, that is, f' and f'' exist and are continuous on the whole space $\mathbf{R}$.

Theorem 6.2. *(Ito's Formula) Let $(B_t : t \leq T)$ be a standard Brownian motion. Consider $f \in \mathcal{C}^2(\mathbf{R})$. Then, with probability one, we have:*

$$f(B_t) - f(B_0) = \int_0^t f'(B_s)dB_s + \frac{1}{2}\int_0^t f''(B_s)ds, \quad t \leq T \tag{6.7}$$

We will see other variations in proposition and later sections. Before giving an idea of the proof, we make some important observations:

Remark. (1) Equation (6.7) is an equality of processes, which is much stronger than equality in distribution. In other words, if you take a path of the process on the left constructed on a given Brownian motion, then this path will be the same as the path of the on the right constructed on the same Brownian motion. This equality holds in the limit where the mesh of the partition of the interval $[0, T]$ goes to 0.

(2) Note the similarity with the classical formulation in (6.6), if we replace the Riemann integral by Ito's integral. We do have the additional integral of $f''(B_s)$. As we will see in the proof, this additional term comes from the quadratic term in the Taylor's approximation and from the quadratic variation of Brownian motion seen in theorem (4.5). As in the classical case, it is very convenient to summarize the conclusion of Ito's formula in differential form:

$$df(B_t) = f'(B_t)dB_t + \frac{1}{2}f''(B_t)dt \tag{6.8}$$

We stress that the differential form has no meaning by itself. It is a compact way to express the two integrals in Ito's formula and a powerful device for computations.

(3) An important consequence of Ito's formula is that it provides a systematic way to construct martingales as explicit functions of Brownian motion. To make sure that, $\int_0^t f'(B_s)dB_s, t \leq T$ defines a continuous square integrable martingale on $[0, T]$, we might need to check that $(f'(B_t), t \leq T) \in \mathcal{L}_c^2(T)$. In general, the Ito integral makes sense as a local martingale.

Corollary 6.3. *(Brownian Martingales). Let $(B_t : t \leq T)$ be a standard brownian motion. Consider $f \in \mathcal{C}^2(\mathbf{R})$ such that $\int_0^T \mathbf{E}[f'(B_s)^2]ds < \infty$. Then the process:*

$$\left(f(B_t) - \frac{1}{2}\int_0^t f''(B_s)ds, \quad t \leq T\right)$$

is a martingale for the Brownian motion.

Proof. This is straightforward from the Ito's formula:

$$f(B_t) - \frac{1}{2}\int_0^t f''(B_s)ds = f(B_0) + \int_0^t f'(B_s)dB_s$$

The first term is a constant and the second term is a continuous martingale by proposition 6.1. □

The integral we subtracted from $f(B_t)$ is called the *compensator*. A simple case is given by the function $f(x) = x^2$. For this function the corollary gives that the process $B_t^2 - t, t \geq 0$ is a martingale, as we already observed. The compensator was then simply t. In general, a compensator might be random.

(4) The compensator is the Riemann integral $\int_0^t f''(B_s)ds$. It might seem to be a strange object at first. The function $f''(B_s)$ is random (it depends on ω), so the integral is a random variable. There is no problem in integrating the random function $f''(B_s)$ since by assumption it is a continuous function of s, since f'' and $B_s(\omega)$ are continuous. In fact, the paths

of $\int_0^t f''(B_s)ds$ are much smoother than the ones of Brownian motion in general: the paths are differentiable everywhere (the derivative is $f''(B_t)$) and in particular the paths have bounded variation.

To sum it up, Ito's formula says that $f(B_t)$ can be expressed as a sum of two processes: one with bounded variation (the Riemann integral) and a (local) martingale with finite quadratic variation (the Ito integral). In the next section, we study Ito processes in more generality, which are processes that can be expressed as a sum of a Riemann integral and an Ito integral.

Example 6.11. Let

$$f(x) = x^3$$

In this case, Ito's formula yields:

$$\begin{aligned} B_t^3 &= \int_0^t 3B_s^2 dB_s + \frac{1}{2}\int_0^t 6B_s ds \\ &= 3\int_0^t B_s^2 dB_s + 3\int_0^t B_s ds \end{aligned}$$

We can look at a sample of a single path of each of these processes constructed from the same Brownian motion. Note that they are almost equal (the discrepancy is only due to discretization in the numerics)! From the above equation, we conclude that $B_t^3 - 3\int_0^t B_s ds$ is a martingale. See the figure below for a sample of its paths. The process $(\int_0^t B_s ds, \quad t \geq 0)$ is not complicated. It is a Gaussian process since the integral is the limit (almost sure and L^2) of the Riemann sums :

$$\sum_{j=0}^{n-1} B_{t_j}(t_{j+1} - t_j)$$

and each term of the sum is a Gaussian random variable. Clearly, the mean of $\int_0^t B_s ds$ is 0. The covariance of the process can be calculated directly by interchanging the integrals and the expectation:

$$\mathbf{E}\left[\left(\int_0^t B_s ds\right)\left(\int_0^{t'} B_u du\right)\right] = \int_0^t \int_0^{t'} \mathbf{E}[B_s B_u] ds du = \int_0^t \int_0^{t'} (s \wedge u) ds du$$

Here the domain of integration is $D = [0,t] \times [0,t']$. Assume that $t < t'$. We can divide the domain into two sub-domains $D_1 = \{(x,y) : 0 \leq x \leq t, 0 \leq y \leq x\}$ and $D_2 = \{(x,y) : 0 \leq x \leq t, x \leq y \leq t'\}$. Consequently, we can evaluate the above double integral as:

$$\begin{aligned} I &= \int_0^t \int_0^{t'} \min(x,y) dx dy \\ &= \int_0^t \int_0^x y dy dx + \int_0^t \int_x^{t'} x dy dx \\ &= \int_0^t \left[\frac{y^2}{2}\right]_0^x dx + \int_0^t x\,[y]_x^{t'}\, dx \\ &= \int_0^t \frac{x^2}{2} dx + \int_0^t x(t' - x) dx \\ &= \left[\frac{x^3}{6}\right]_0^t + t'\left[\frac{x^2}{2}\right]_0^t - \left[\frac{x^3}{3}\right]_0^t \\ &= \frac{t^3}{6} + \frac{t't^2}{2} - \frac{t^3}{3} \\ &= -\frac{t^3}{6} + \frac{t't^2}{2} \end{aligned}$$

In particular the variance at time t is $t^3/3$. The paths of this process are very smooth, as can be observed in the figure below. In fact, the paths are differentiable and the derivative at time t is B_t.

Example 6.12. Let

$$f(x) = \cos x$$

In this case, the Ito's formula gives:

$$\cos B_t - \cos B_0 = \int_0^t - \sin(B_s) dB_s + \frac{1}{2} \int_0^t (-\cos(B_s) ds$$

In particular, the process

$$M_t = \cos B_t + \frac{1}{2} \int_0^t \cos B_s ds = 1 - \int_0^t \sin B_s dB_s, \quad t \geq 0$$

is a continuous martingale starting at $M_0 = 1$. It is easy to check that the process $(\sin B_t, t \leq T)$ is in $\mathcal{L}_c^2(T)$ for any T. $\sin B_t$ is $\mathcal{F}_t-$measurable since it is a function of B_t. Moreover, Also, $\sin(x)$ is continuous, and the composition of continuous functions is continuous.

Where does Ito's formula come from? It is the same idea as for the proof of the Fundamental Theorem of Calculus(FTC). Let's start with the latter. Suppose $f \in \mathcal{C}^1(\mathbf{R})$; that is: f is differentiable with a continuous derivative. Then, f admits a Taylor approximation around s of the form:

$$f(t) - f(s) = f'(s)(t-s) + \mathcal{E}(s,t) \tag{6.9}$$

(This is in spirit of the mean value theorem) Here, $\mathcal{E}(s,t)$ is an error term that goes to 0 faster than $(t-s)$ as $s \to t$ (for example $(t-s)^2$). Now, for a partition $(t_j, j \leq n)$ of $[0,t]$, say $t_j = \frac{jt}{n}$, we can trivially write for any n:

$$f(t) - f(0) = \sum_{j=0}^{n} f(t_{j+1}) - f(t_j)$$

Now, we can use the equation (6.9) at $s = t_j$:

$$f(t_{j+1}) - f(t_j) = f'(t_j)(t_{j+1} - t_j) + \mathcal{E}(t_j, t_{j+1})$$

Therefore, we have by taking the limit of large n:

$$f(t) - f(0) = \lim_{n \to \infty} \sum_{j=0}^{n} f'(t_j)(t_{j+1} - t_j) + \sum_{j=0}^{n} \mathcal{E}(t_j, t_{j+1}) = \int_0^t f'(s) ds + 0$$

The idea for Ito's formula is similar to the above with two big differences : first we will consider a function of *space* and *time*. Second, we shall need a Taylor approximation to the second order around a point x: if $f \in \mathcal{C}^2(\mathbf{R})$ we have:

$$f(y) - f(x) = (y-x)f'(x) + \frac{1}{2}(y-x)^2 f''(x) + \mathcal{E}(x,y) \tag{6.10}$$

where $\mathcal{E}(x,y)$ is the error term that converges to 0 faster than $(x-y)^3$ as $y \to x$.

Proof. Recall that by assumption $f \in \mathcal{C}^2(\mathbf{R})$. We will prove the particular case, where f is 0 outside a bounded interval. This implies that both the derivatives are bounded, since by the preservation of the compact set theorem, continuous functions preserve compact sets. We first prove the formula for a fixed t. Then, we generalize to processes on $[0,T]$. Consider a partition $(t_j : j \leq n)$ of $[0,t]$. From the Taylor's series expansion above:

$$f(B_t) - f(B_0) = \sum_{j=0}^{n-1} f'(B_{t_j})(B_{t_{j+1}} - B_{t_j}) + \frac{1}{2}\sum_{j=0}^{n-1} f''(B_{t_j})(B_{t_{j+1}} - B_{t_j})^2 + \sum_{j=0}^{n} \mathcal{E}(B_{t_j}, B_{t_{j+1}}) \tag{6.11}$$

As $n \to \infty$, the first term converges (as a random variable in L^2) to the Ito integral. This is how we proved proposition (6.1) using simple processes. We claim that the second term converges to the Riemann integral. To see this, consider the corresponding Riemann sum :

$$\sum_{j=0}^{n-1} f''(B_{t_j})(t_{j+1} - t_j)$$

This term converges almost surely to the Riemann integral $\int_0^t f''(B_s)ds$ since f'' is continuous. It also converges in L^2 by theorem by the dominated convergence theorem, since $f''(\cdot)$ is bounded by assumption. Therefore, to show that the second term converges to the same limit, it suffices to show that the L^2-distance between the second term and the Riemann sum goes to 0. That is,

$$\lim_{n\to\infty} \mathbf{E}\left[\left(\sum_{j=0}^{n-1} f''(B_{t_j})\left\{\left(B_{t_{j+1}} - B_{t_j}\right)^2 - (t_{j+1} - t_j)\right\}\right)^2\right] \tag{6.12}$$

This is in the same spirit as the proof of the quadratic variation of the Brownian motion in theorem (4.5). To lighten the notation, define the variables :

$$X_j := (B_{t_{j+1}} - B_{t_j})^2 - (t_{j+1} - t_j), \quad j \leq n-1$$

We expand the square in (6.12) to get:

$$\sum_{j=0}^{n-1}\sum_{k=0}^{n-1} \mathbf{E}\left[f''(B_{t_j})f''(B_{t_k})X_j X_k\right]$$

For $j < k$, we co‘ndition on $\mathcal{F}_{t_k}$ to get :

$$\begin{aligned}\sum_{j=0}^{n-1}\sum_{k=0}^{n-1} \mathbf{E}\left[f''(B_{t_j})f''(B_{t_k})X_j X_k\right] &= 2\sum_{j<k} \mathbf{E}\left[\mathbf{E}\left[f''(B_{t_j})f''(B_{t_k})X_j X_k \middle| \mathcal{F}_{t_k}\right]\right] \\ &\quad + \sum_{j=0}^{n-1} \mathbf{E}\left[\left(f''(B_{t_j})\right)^2 X_j^2\right]\end{aligned}$$

The first term on the right hand can be expressed as :

$$2\sum_{j<k}^{n-1}\mathbf{E}\left[\mathbf{E}\left[f''(B_{t_j})f''(B_{t_k})X_jX_k\middle|\mathcal{F}_{t_k}\right]\right] = 2\sum_{j<k}^{n-1}\mathbf{E}\left[f''(B_{t_j})f''(B_{t_k})X_j\mathbf{E}\left[X_k\middle|\mathcal{F}_{t_k}\right]\right]$$
$$\{\text{Taking out what is known}\}$$

The random variable $\mathbf{E}\left[X_k\middle|\mathcal{F}_{t_k}\right]$ turns out to be:

$$\begin{aligned}\mathbf{E}\left[X_k\middle|\mathcal{F}_{t_k}\right] &= \mathbf{E}\left[(B_{t_{k+1}}-B_{t_k})^2\right]-(t_{k+1}-t_k)\\ &\{\because B_{t_{k+1}}-B_{t_k} \text{ is independent of } \mathcal{F}_{t_k}\}\\ &= (t_{k+1}-t_k)-(t_{k+1}-t_k)\\ &= 0\end{aligned}$$

So, the entire summand of the first term equals 0, and we are left with:

$$\begin{aligned}&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2X_j^2\right]\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\left\{\left(B_{t_{j+1}}-B_{t_j}\right)^2-(t_{j+1}-t_j)\right\}^2\right]\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\left\{\left(B_{t_{j+1}}-B_{t_j}\right)^4-2\left(B_{t_{j+1}}-B_{t_j}\right)^2(t_{j+1}-t_j)+(t_{j+1}-t_j)^2\right\}\right]\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\left\{\left(B_{t_{j+1}}-B_{t_j}\right)^4-2\left(B_{t_{j+1}}-B_{t_j}\right)^2(t_{j+1}-t_j)+(t_{j+1}-t_j)^2\right\}\middle|\mathcal{F}_{t_j}\right]\right]\\
&\{\text{Conditioning on } \mathcal{F}_{t_j}\}\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\mathbf{E}\left[\left\{\left(B_{t_{j+1}}-B_{t_j}\right)^4-2\left(B_{t_{j+1}}-B_{t_j}\right)^2(t_{j+1}-t_j)+(t_{j+1}-t_j)^2\right\}\middle|\mathcal{F}_{t_j}\right]\right]\\
&\{\text{ Taking out what is known }\}\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\mathbf{E}\left[\left\{\left(B_{t_{j+1}}-B_{t_j}\right)^4-2\left(B_{t_{j+1}}-B_{t_j}\right)^2(t_{j+1}-t_j)+(t_{j+1}-t_j)^2\right\}\right]\right]\\
&\{\text{ Independence }\}\\
=&\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\left(3(t_{j+1}-t_j)^2-2(t_{j+1}-t_j)^2+(t_{j+1}-t_j)^2\right)\right]\\
=&2\sum_{j=0}^{n-1}\mathbf{E}\left[\left(f''(B_{t_j})\right)^2\right](t_{j+1}-t_j)^2\end{aligned}$$

Since $f''(x)$ is bounded, the last result approaches 0, as the mesh size becomes finer and finer and $n\to\infty$. It remains to handle the error term in (6.11). This follows the same idea as for the second term and we omit it.

To extend the formula to the whole interval $[0,T]$, notice that the processes of both sides of the equation (6.7) have continuous paths. Since they are equal with probability one at any fixed time by the above argument, they must be equal for any countable set of times. It suffices to consider the processes on the rational times in $[0,T]$, which are dense in $[0,T]$. Since the paths are continuous and they are equal on these times, they must be equal at all times on $[0,T]$. $\square$

Recall from equation (6.8), that Ito's formula can be written in the differential form:

$$df(B_t) = f'(B_t)dB_t + \frac{1}{2}f''(B_t)dt$$

This notation has no meaning by itself. It is a compact way to write (6.7). This allows us to derive an easy and useful computational formula: if we blindly apply the classical differential to f to second order in the Taylor expansion, we formally obtain:

$$df(B_t) = f'(B_t)dB_t + \frac{1}{2}f''(B_t)(dB_t)^2 \tag{6.13}$$

Therefore, Ito's formula is equivalent to applying the rule $dt = dB_t \cdot dB_t$. In fact, it is counterproductive to learn Ito's formula by heart. It is much better to simply compute the differential upto the second order and apply the following *simple* rules of Ito calculus:

$\cdot$	dt	dB_t
dt	0	0
dB_t	0	dt

It is not hard to extend Ito's formula to a function $f(t, x)$ of both time and space:

$$\begin{aligned} f : [0, T] \times \mathbf{R} &\mapsto \mathbf{R} \\ (t, x) &\mapsto f(t, x) \end{aligned}$$

Such functions have partial derivatives that are themselves functions of time and space. We will use the following notation for the partial derivatives:

$$\partial_t f(t, x) = \frac{\partial f}{\partial t}(t, x), \quad \partial_x f(t, x) = \frac{\partial f}{\partial x}(t, x), \quad \partial_{xx}(t, x) = \frac{\partial^2 f}{\partial x^2} f(t, x)$$

The reason for this notation is to avoid confusion between the variable that is being *differentiated* and the value of time and space at which the derivative is *being evaluated*. It might appear strange at first, but it will avoid confusion down the road (especially when dealing with several space variables in a later section). To apply Ito's formula, we will need that the partial derivative with respect to time $\partial_t f$ exists and is continuous as a function on $[0, T] \times \mathbf{R}$ and that the first and second partial derivatives in space $\partial_x f$ and $\partial_{xx} f$ exist and are continuous. We say that such a function f is in $\mathcal{C}^{1,2}[0, T] \times \mathbf{R}$. Then, with probability 1, we have for every $t \in [0, T]$:

Proposition 6.3. *(Ito's formula) Let $(B_t : t \leq T)$ be a standard brownian motion on $[0, T]$. Consider a function f of time and space with $f \in \mathcal{C}^{1,2}$([0,T]$\times$$\mathbf{R}$). Then, with probability one, we have for every $t \in [0, T]$,*

$$f(t, B_t) - f(0, B_0) = \int_0^t \partial_x f(s, B_s)dB_s + \int_0^t \left\{\partial_t f(s, B_s) + \frac{1}{2}\partial_{xx} f(s, B_s)\right\} ds$$

or in differential form we have:

$$df(t, B_t) = \partial_x f(t, B_t)dB_t + \left(\partial_t f(t, B_t) + \frac{1}{2}\partial_{xx} f(t, B_t)\right) dt$$

Proof. The idea of the proof is similar as for a function of space only, as it depends on a Taylor's approximation and on the quadratic variation. Here, however, we need to apply Taylor's approximation to second order in space and to the first order in time. We then get something of the form:

$$\begin{aligned}
f(t, B_t) - f(0, B_0) = \sum_{j=0}^{n-1} &\partial_x f(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j}) + \partial_t f(t_j, B_{t_j})(t_{j+1} - t_j) \\
&+ \frac{1}{2}\partial_{xx} f(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})^2 \\
&+ \partial_t \partial_x f(t_j, B_{t_j})(B_{t_{j+1}} - B_{t_j})(t_{j+1} - t_j) + \mathcal{E}
\end{aligned}$$

The first two lines becomes the integrals in the Ito's formula. We see a new animal on the last line: the mixed derivative $\partial_t \partial_x f$. This term is related to the limit in the cross variation between B_t and t given by:

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} (B_{t_{j+1}} - B_{t_j})(t_{j+1} - t_j)$$

It can be shown that it goes to 0 in a suitable sense.

Let $(t_j : j \leq n)$ be a sequence of partitions on $[0, t]$. We have:

$$\begin{aligned}
\mathbf{E}\left[\left(\sum_{j=0}^{n-1} (t_{j+1} - t_j)(B_{t_{j+1}} - B_{t_j})\right)^2\right] &\leq \|\Delta_n\|^2 \, \mathbf{E}\left[\left(\sum_{j=0}^{n-1} (B_{t_{j+1}} - B_{t_j})\right)^2\right] \\
&= \|\Delta_n\|^2 \sum_{j=0}^{n-1} \mathbf{E}\left[(B_{t+1} - B_{t_j})^2\right] \\
&\quad + 2\,\|\Delta_n\|^2 \sum_{j<k} \mathbf{E}\left[(B_{t+1} - B_{t_j})(B_{t_{k+1}} - B_{t_k})\right]
\end{aligned}$$

Since $\mathbf{E}[(B_{t_{j+1}} - B_{t_j})(B_{t_{k+1}} - B_{t_k})] = 0$ and $\mathbf{E}[(B_{t_{j+1}} - B_{t_j})^2] = (t_{j+1} - t_j)$, we find that the above variance is $\|\Delta_n\|^2 \cdot t$. As $n \to \infty$, $\|\Delta_n\|^2 \to 0$. Consequently, the cross-variation approaches 0 in the mean square sense.

This justifies the rule $dt \cdot dB_t = 0$. We can also justify the rule $dt \cdot dt = 0$.

$$\begin{aligned}
\sum_{j=0}^{n-1} (t_{j+1} - t_j)^2 &\leq \|\Delta_n\| \sum_{j=0}^{n-1} (t_{j+1} - t_j) \\
&= \|\Delta_n\| \cdot t
\end{aligned}$$

As $n \to \infty$, $\|\Delta_n\| \to 0$, and we get the desired result.

Once these facts are known, the rest of the proof is done similarly to the one for the function of space only. We do notice though that the formula is easy to derive once we accept the rules of Ito calculus. By writing the differential to second order in space, and to first order in time and applying the rules of Ito calculus, we get:

$$df(t, B_t) = \partial_x f(t, B_t) dB_t + \left\{\partial_t f(t, B_t) + \frac{1}{2}\partial_{xx} f(t, B_t)\right\} dt$$

□

As in the one variable case, we get a corollary to construct Martingales:

Corollary 6.4. *(Brownian Martingales) Let $(B_t : t \leq T)$ be a standard Brownian motion. Consider $f \in \mathcal{C}^{1,2}([0,T] \times \mathbf{R})$ such that the process $(\partial_x f(t, B_t) : t \leq T) \in \mathcal{L}_c^2(T)$. Then, the process*

$$\left(f(t, B_t) - \int_0^t \left\{ \partial_s f(s, B_s) + \frac{1}{2} \partial_{xx} f(s, B_s) \right\} ds, t \leq T \right)$$

is a martingale for the Brownian filtration. In particular, if $f(t,x)$ satisfies the partial differential equation $\partial_t f = -\frac{1}{2}\partial_{xx} f$, then the process $(f(t, B_t), t \leq T)$ is itself a martingale.

We now catch a glimpse of the powerful connection between two fields of mathematics: *the study of martingales is closely connected to the study of differential equations.* We will see this connection in action in the gambler's ruin problem in the next section.

Example 6.13. Consider the function $f(t,x) = tx$. In this case, we have: $\partial_t f = x$, $\partial_x f = t$ and $\partial_{xx} f = 0$. Ito's formula yields:

$$d(tB_t) = tdB_t + xdt$$

Therefore, the process $M_t = tB_t - \int_0^t B_s ds$ is a martingale for the Brownian filtration. It is also a Gaussian process by corollary (6.2). The mean is 0 and the covariance by corollary (6.7) is:

$$\mathbf{E}[M_t M_{t'}] = \int_0^{t \wedge t'} s^2 ds = \frac{(t \wedge t')^3}{3}$$

Example 6.14. (Vasicek Interest Rate Model) Let $(B_t : t \leq T)$ be a standard Brownian motion. Vasicek assumed that the instantaneous spot rate under the real-world measure evolves as an Ornstein-Uhlenbeck process with constant coefficients. Thus:

$$dr_t = k(\theta - r_t)dt + \sigma dB_t \tag{6.14}$$

Rearranging the equation, multiplying both sides by the integrating factor and integrating from s to t, we have:

$$\begin{aligned}
dr_t &= k\theta dt - kr_t dt + \sigma dB_t \\
dr_t + kr_t dt &= k\theta dt + \sigma dB_t \\
e^{kt} dr_t + kr_t e^{kt} dt &= k\theta e^{kt} dt + \sigma e^{kt} dB_t \\
d(e^{kt} r_t) &= k\theta e^{kt} dt + \sigma e^{kt} dB_t \\
e^{kt} r_t - e^{ks} r_s &= \theta(e^{kt} - e^{ks}) + \sigma \int_s^t e^{kt} dB_t \\
e^{kt} r_t &= r_s e^{ks} + \theta(e^{kt} - e^{ks}) + \sigma \int_s^t e^{kt} dB_t \\
r_t &= r_s e^{-k(t-s)} + \theta(1 - e^{-k(t-s)}) + \sigma \int_s^t e^{-k(t-u)} dB_u
\end{aligned}$$

for all t. By corollary (6.2), $\int_s^t e^{-k(t-u)} dB_u$ is a Gaussian process with mean 0 and variance:

$$\begin{aligned}\int_s^t e^{-2k(t-u)}du &= e^{-2kt}\int_s^t e^{2ku}du \\ &= \frac{e^{-2kt}}{2k}[e^{2ku}]_s^t \\ &= \frac{e^{-2kt}}{2k}[e^{2kt}-e^{2ks}] \\ &= \frac{1}{2k}(1-e^{-2k(t-s)})\end{aligned}$$

Thus, the Vasicek process is Gaussian with mean:

$$\mathbf{E}[r_t] = r_s e^{-k(t-s)} + \theta(1-e^{-k(t-s)})$$

and variance:

$$Var[r_t] = \frac{\sigma^2(1-e^{-2k(t-s)})}{2k}$$

Thus, r_t can be negative with positive probability. The possibility of negative rates is indeed a major drawback of the Vasicek model. However, the analytical tractability that is implied by a Gaussian density is hardly achieved when assuming other distributions for the process r. If we let $t \to \infty$, we get $\mathbf{E}[r_t] = \theta$. So, the drift of process $(r_t : t \leq T)$ is positive, whenever $r_t < \theta$ and whilst it is negative, whenever $r_t > \theta$ and so it is pushed everytime, to be closer on average to the level θ. Hence, it is mean reverting.

The solution of the stochastic differential equation (6.14) can also be verified using Ito's lemma. Let:

$$r_t = r_0 e^{-kt} + \theta(1-e^{-kt}) + \sigma\int_0^t e^{-k(t-u)}dB_u$$

And consider the function:

$$f(t,x) = r_0 e^{-kt} + \theta(1-e^{-kt}) + \sigma e^{-kt}x$$

where $(X_t, t \leq T) = \int_0^t e^{ku}dB_u$

Then,

$$\begin{aligned}\partial_x f(t,X_t) &= \sigma e^{-kt} \\ \partial_t f(t,X_t) &= -r_0 k e^{-kt} + k\theta e^{-kt} - \sigma k X_t e^{-kt} = -kf(t,x) + k\theta \\ \partial_{xx} f(t,X_t) &= 0\end{aligned}$$

By Ito's Lemma:

$$\begin{aligned}df(t,X_t) &= \sigma e^{-kt}dX_t + (-kf(t,x)+k\theta)dt \\ df(t,X_t) &= k(\theta - f(t,x))dt + \sigma dB_t\end{aligned}$$

Example 6.15. (Cox-Ingersoll-Ross (CIR) Model). Let $(B_t : t \geq 0)$ be a Brownian motion. The Cox-Ingersoll-Ross model for the instantaneous spot interest rate process r_t is:

$$dr_t = k(\theta - r_t)dt + \sigma\sqrt{r_t}dB_t$$

They introduced a square-root term in the diffusion coefficient of the instantaneous short-rate dynamics proposed by Vasicek. The resulting model has been a benchmark for many years because of its analytical tractability and the fact, that contrary to Vasicek (1977) model, the instantaneous short rate is always positive. The condition $2k\theta > \sigma^2$ has to be imposed to ensure that the origin is inaccessible to the process. Unlike the Vasicek equation the CIR model does not have a closed-form solution.

Although, we cannot derive a closed-form solution , the expectation and variance of r_t can be be computed.

Consider the function $f(t,x) = e^{kt}x$, where $X_t = r_t$. We have:

$$\begin{aligned}
\partial_x f(t,x) &= e^{kt} \\
\partial_t f(t,x) &= ke^{kt}x \\
\partial_{xx} f(t,x) &= 0
\end{aligned}$$

By the Ito's-Lemma, we have:

$$\begin{aligned}
df(t,X_t) &= e^{kt}dX_t + ke^{kt}xdt \\
&= e^{kt}(k(\theta - r_t))dt + e^{kt}\sigma\sqrt{r_t}dB_t + ke^{kt}r_t dt \\
d(e^{kt}r_t) &= e^{kt}k\theta dt + e^{kt}\sigma\sqrt{r_t}dB_t
\end{aligned}$$

Integrating both sides of the equation, we have:

$$\begin{aligned}
e^{kt}r_t|_0^t &= k\theta \int_0^t e^{kt}dt + \sigma \int_0^t e^{kt}\sqrt{r_t}dB_t \\
e^{kt}r_t - r_0 &= \theta(e^{kt} - 1) + \sigma \int_0^t e^{kt}\sqrt{r_t}dB_t \\
r_t &= r_0 e^{-kt} + \theta(1 - e^{-kt}) + \sigma \int_0^t \sqrt{r_t}dB_t
\end{aligned}$$

We know that, $(I_t = \int_0^t \sqrt{r_t}dB_t, t \leq T)$ is an Ito integral with mean 0. So, the mean of the process $(r_t : t \leq T)$ is :

$$\mathbf{E}[r_t] = r_0 e^{-kt} + \theta(1 - e^{-kt})$$

The variance of the Ito integral is given by:

$$\begin{aligned}
Var[I_t] = \mathbf{E}[I_t^2] &= \int_0^t \mathbf{E}[\left(\sqrt{r}_t\right)^2]dt \\
&= \int_0^t \mathbf{E}[r_t]dt \\
&= \int_0^t (r_0 e^{-kt} + \theta(1-e^{-kt}))dt \\
&= (r_0-\theta)\int_0^t e^{-kt}dt + \theta\int_0^t dt \\
&= (r_0-\theta)\frac{(e^{-kt}-1)}{-k} + \theta t \\
&= \frac{(r_0-\theta)}{k}(1-e^{-kt}) + \theta t
\end{aligned}$$

Example 6.16. (Geometric Brownian Motion revisited). Consider an asset price process that satisfies

$$S_t = f(t, B_t) = S_0 e^{\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma B_t}$$

Thus, $f(t,x) = S_0 e^{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma x}$. $\partial_x f(t, B_t) = S_0 \sigma e^{(\mu-\sigma^2/2)t+\sigma B_t} = \sigma S_t$, $\partial_{xx} f(t,B_t) = S_0\sigma^2 e^{(\mu-\sigma^2/2)t+\sigma B_t} = \sigma^2 S_t$ and $\partial_t f(t,B_t) = \left(\mu - \frac{\sigma^2}{2}\right) S_t$. Therefore, by Ito's Lemma:

$$\begin{aligned}
dS_t &= \sigma S_t dB_t + \left\{\frac{1}{2}\sigma^2 S_t + \mu S_t - \frac{\sigma^2}{2} S_t\right\} dt \\
&= \mu S_t dt + \sigma S_t dB_t
\end{aligned}$$

In integral notation, the asset price $(S_t : t \leq T)$ is given by:

$$S_t = S_0 + \sigma \int_0^t f(t, B_t) dB_t + \int_0^t \mu f(t, B_t) dt$$

6.4 Gambler's ruin for Brownian Motion with a Drift.

We solved the Gambler's ruin problem for the standard Brownian motion in example (5.20). We now deal with the case where a drift is present. Consider the Brownian motion with a drift:

$$X_t = \sigma B_t + \mu t$$

where $(B_t, t \geq 0)$ is a standard Brownian motion.

7 Multivariate Ito Calculus.

7.1 Multidimensional Brownian motion.

Definition 7.1. (Brownian motion in $\mathbf{R}^d$). Take $d \in \mathbf{N}$. Let $B^{(1)}, \ldots, B^{(d)}$ be independent standard Brownian motions in $(\Omega, \mathcal{F}, \mathbb{P})$. The process $(B_t : t \geq 0)$ taking values in $\mathbf{R}^d$ defined by :

$$B_t = (B_t^{(1)}, \dots, B_t^{(d)}), \quad t \geq 0$$

is called a $d-$dimensional Brownian motion or a Brownian motion in $\mathbf{R}^d$.

The Brownian filtration $(\mathcal{F}_t : t \geq 0)$ is now composed of the information of all Brownian motions. In other words, it is given by the sigma-fields:

$$\mathcal{F}_t = \sigma(B_s^{(i)}, 1 \leq i \leq d, s \leq t)$$

For every outcome ω, the path of trajectory of a $d-$dimensional Brownian motion is a curve in space parametrized by the time t:

$$t \mapsto B_t(\omega) = (B_t^{(1)}(\omega), B_t^{(2)}(\omega), \dots, B_t^{(d)}(\omega))$$

Of course, this curve is continuous, since each coordinate is. The below numerical project gives an example of one path of a two-dimensional brownian motion. This is a very rugged and intertwined curve! We might wonder, what it does as $t \to \infty$. Does it wander around $(0, 0)$ ad infinitum or does it eventually escape to infinity? We will answer this question in a later section. For doing so, we shall need a version of Ito's formula for multi-dimensional Brownian motion. We finish this section by noticing that it is also easy to construct Brownian motions in higher dimensions for which the coordinates are correlated.

Example 7.1. (Example of Brownian motion with correlated coordinates) Let $(B_t : t \geq 0)$ be a two dimensional brownian motion. Let $-1 < \rho < 1$. We construct the two dimensional process as follows: $W_t = (W_t^{(1)}, W_t^{(2)})$ where:

$$\begin{aligned} W_t^{(1)} &= B_t^{(1)} \\ W_t^{(2)} &= \rho B_t^{(1)} + \sqrt{1-\rho^2} B_t^{(2)} \end{aligned}$$

$W_t^{(1)} = B_t^{(1)}$ is Gaussian with mean 0 and variance t. Since, $B_t^{(1)}$ and $B_t^{(2)}$ are independent gaussian random variables and the sum of IID Gaussians is Gaussian, $W_t^{(2)}$ is Gaussian with mean 0 and variance t. The covariance between $W_t^{(1)}$ and $W_t^{(2)}$ is:

$$\begin{aligned} \mathbf{E}[W_t^{(1)} W_t^{(2)}] &= \mathbf{E}[B_t^{(1)}(\rho B_t^{(1)} + \sqrt{1-\rho^2} B_t^{(2)})] \\ &= \mathbf{E}[\rho (B_t^{(1)})^2 + \sqrt{1-\rho^2} B_t^{(1)} B_t^{(2)}] \\ &= \rho t \end{aligned}$$

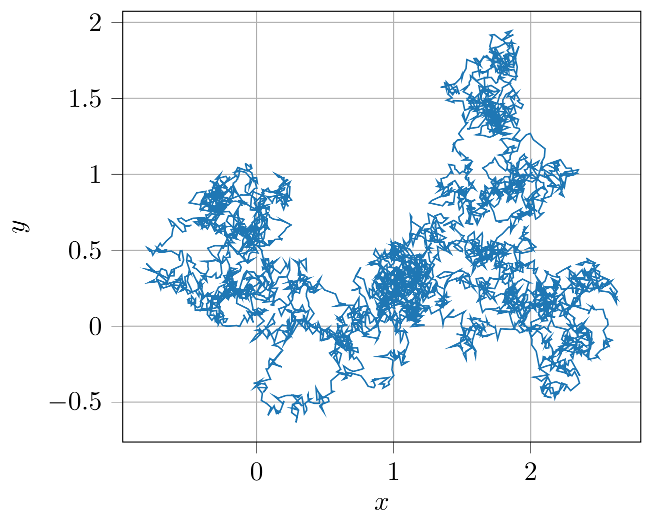
Hence, the coordinates at time t are not independent.

Exercise 7.1. 2D Brownian Motion. Consider a two-dimensional Brownian motion $(B_t^{(1)}, B_2^{(2)})$ starting at $(0, 0)$.

(a) Plot one path of this Brownian motion on the plane $\mathbf{R}^2$ on the plane in $\mathbf{R}^2$ on the time interval $[0, 5]$ using a discretization of 0.005 and 0.001.

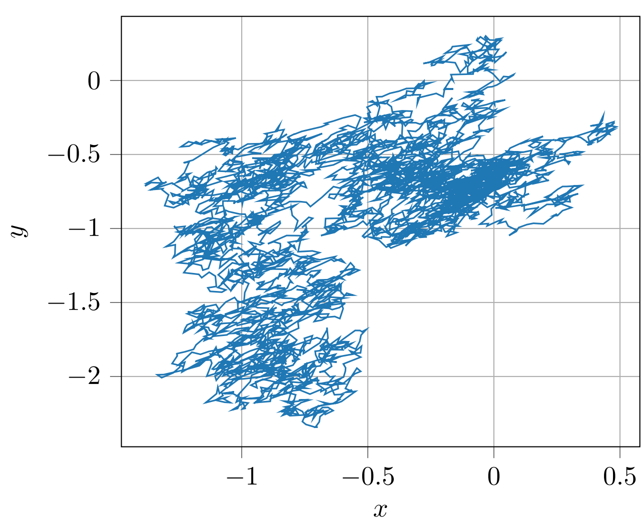
Solution.

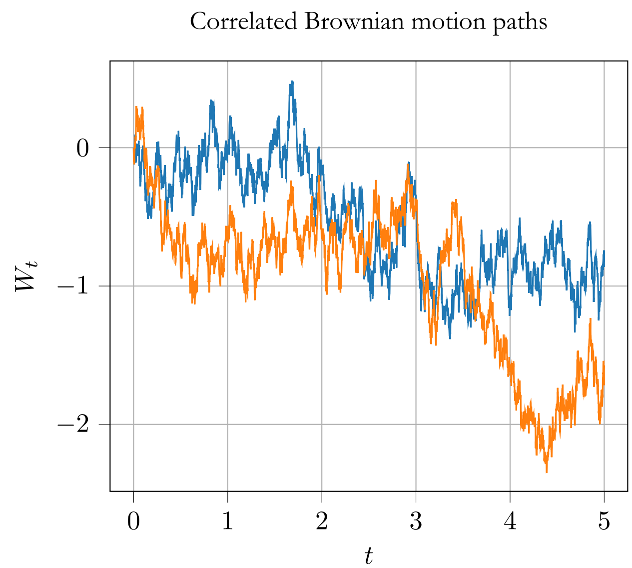
Two dimensional brownian motion



(b) Consider now the process $(W_t : t \geq 0)$ for $\rho = 1/2$ as in example (7.1). Plot one path of this process on the plane $\mathbf{R}^2$ on the time-interval $[0, 5]$ using a discretization of 0.001.

Correlated Brownian motion $W(t) = (W_t^{(1)}, W_t^{(2)})$





7.2 Ito's Formula.

Theorem 7.1. *Ito's Formula. Let* $(B_t : t \geq 0)$ *be a* $d-$*dimensional brownian motion. Consider* $f \in \mathcal{C}^2(\mathbf{R}^d)$. *Then, we have with probability one that for all* $t \geq 0$:

$$f(B_t) - f(B_0) = \sum_{i=1}^{d} \int_0^t \partial_{x_i} f(B_s) dB_s^{(i)} + \frac{1}{2} \int_0^t \sum_{i=1}^{d} \partial_{x_i}^2 f(B_s) ds \tag{7.1}$$

Remark. We stress that, as in the one-dimensional case, in theorem (6.7), Ito's formula is an equality of processes (and not an equality in distribution). Thus, the processes on both sides must agree for each path. Interestingly, the mixed partials $\partial_{x_i x_j} f(B_s)$, $i \neq j$ do not appear in the formula! We see from Ito's formula that the process $f(B_t)$ can be represented as a sum of $d+1$ processes: d Ito integrals and one Riemann integral (which is a process of finite variation). In vector notation, the formula takes the form:

$$f(B_t) - f(B_0) = \int_0^t \nabla f(B_s)^T dB_s + \frac{1}{2} \int_0^t \nabla^2 f(B_s) ds$$

where it is understood that the first term is the sum of the d Ito integrals in the equation. The symbol ∇^2 is the Laplacian of f: $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} f(B_s^{(1)}, \ldots, B_s^{(d)}) ds$

In differential form, Ito's formula becomes very neat:

$$df(B_t) = \sum_{i=1}^{d} \partial_{x_i} f(B_s) dB_t^{(i)} + \frac{1}{2} \sum_{i=1}^{d} \partial_{x_i}^2 f(B_s) dt = \nabla f(B_t)^T dB_s + \frac{1}{2} \nabla^2 f(B_t) dt$$

Example 7.2. Consider the functions (1) $f(x_1, x_2) = x_1^2 + x_2^2$ (2) $f(x_1, x_2) = e^{x_1} \cos x_2$ and the processes $(X_t : t \geq 0)$ and $(Y_t : t \geq 0)$. If we apply Ito's formula to the first process, we have:

$$\begin{aligned} X_t &= \int_0^t 2B_s^{(1)} dB_s^{(1)} + \int_0^t 2B_s^{(2)} dB_s^{(2)} + \frac{1}{2} \int_0^t (4dt) \\ &= \int_0^t 2B_s^{(1)} dB_s^{(1)} + \int_0^t 2B_s^{(2)} dB_s^{(2)} + 2t \end{aligned}$$

The second process gives:

$$
\begin{aligned}
Y_t &= \cos B_s^{(2)} \int_0^t e^{B_s^{(1)}} dB_s^{(1)} - e^{B_s^{(1)}} \int \sin B_s^{(2)} dB_s^{(2)} + \frac{1}{2}\int_0^t \left(e^{B_s^{(1)}} \cos B_s^{(2)} - e^{B_s^{(1)}} \cos B_s^{(2)} \right) dt \\
&= 1 + \cos B_s^{(2)} \int_0^t e^{B_s^{(1)}} dB_s^{(1)} - e^{B_s^{(1)}} \int \sin B_s^{(2)} dB_s^{(2)}
\end{aligned}
$$

Exercise 7.2. Cross-Variation of $B_t^{(1)}$ and $B_t^{(2)}$. Let $(t_j : j \leq n)$ be a sequence of partitions of $[0, t]$ such that $\max_j |t_{j+1} - t_j| \to 0$ as $n \to \infty$. Prove that:

$$
\lim_{n\to\infty} \sum_{j=0}^{n-1} (B_{t_{j+1}}^{(1)} - B_{t_j}^{(1)})(B_{t_{j+1}}^{(2)} - B_{t_j}^{(2)}) = 0 \quad \text{in } L^2
$$

This justifies the rule $dB_t^{(1)} \cdot dB_t^{(2)} = 0$.

Hint: Just compute the second moment of the sum.

Solution. We have:

$$
\begin{aligned}
&\mathbf{E}\left[\left(\sum_{j=0}^{n-1} (B_{t_{j+1}}^{(1)} - B_{t_j}^{(1)})(B_{t_{j+1}}^{(2)} - B_{t_j}^{(2)}) \right)^2 \right] \\
&= \sum_{j=0}^{n-1} \mathbf{E}[(B_{t_{j+1}}^{(1)} - B_{t_j}^{(1)})^2 (B_{t_{j+1}}^{(2)} - B_{t_j}^{(2)})^2] \\
&+ 2\sum_{j<k} \mathbf{E}\left[(B_{t_{j+1}}^{(1)} - B_{t_j}^{(1)})(B_{t_{k+1}}^{(1)} - B_{t_k}^{(1)})(B_{t_{j+1}}^{(2)} - B_{t_j}^{(2)})(B_{t_{k+1}}^{(2)} - B_{t_k}^{(2)}) \right]
\end{aligned}
$$

Both these expectations are zero, since the brownian motions are independent and non-overlapping increments are independent.

Consequently, $\sum_{j=0}^{n-1} (B_{t_{j+1}}^{(1)} - B_{t_j}^{(1)})(B_{t_{j+1}}^{(2)} - B_{t_j}^{(2)}) \to 0$ in the L^2 sense.

Proof. The proof of the formula follows the usual recipe: Taylor's theorem together with the quadratic variation and the cross-variation. In this case, we do get a cross-variation between the different Brownian motions. More precisely, consider a partition $(t_j : j \leq n)$ of $[0, t]$. Then we can write:

$$
f(B_t) - f(B_0) = \sum_{j=0}^{n-1} (f(B_{t_{j+1}}) - f(B_{t_j}))
$$

We can apply the Taylor's series expansion for each j to get:

$$
\begin{aligned}
f(B_t) - f(B_0) &= \sum_{j=0}^{n-1} \nabla f(B_{t_j})(B_{t_{j+1}} - B_{t_j}) \\
&\quad + \frac{1}{2} \sum_{j=0}^{n-1} (B_{t_{j+1}} - B_{t_j})^T H f(B_{t_j})(B_{t_{j+1}} - B_{t_j}) + \mathcal{E}
\end{aligned}
$$

where Hf is the Hessian matrix of f. We wrote the expansion using the vector notation to be economical. Let's keep in mind that each term is a sum over the derivatives. The first term will converge to d Ito integrals as in the one-dimensional case. Now, the summand in the second term is:

$$(B^{(1)}_{t_{j+1}} - B^{(1)}_{t_j}, \dots, B^{(d)}_{t_{j+1}} - B^{(d)}_{t_j}) \begin{bmatrix} \partial^2_{x_1} f(B_{t_j}) & \dots & \partial^2_{x_1 x_d} f(B_{t_j}) \\ \vdots & \ddots & \\ \partial^2_{x_d x_1} f(B_{t_j}) & & \partial^2_{x_d} f(B_{t_j}) \end{bmatrix} \begin{bmatrix} B^{(1)}_{t_{j+1}} - B^{(1)}_{t_j} \\ \vdots \\ B^{(d)}_{t_{j+1}} - B^{(d)}_{t_j} \end{bmatrix}$$

So, $(B^{(i)}_{t_{j+1}} - B^{(i)}_{t_j})$ is pre-multiplied with the term $\partial^2_{x_i x_k} f(B_{t_j})$ and it is post-multiplied $(B^{(k)}_{t_{j+1}} - B^{(k)}_{t_j})$. Consequently, the second term in the Taylor's series expansion can be re-written as:

$$\sum_{j=0}^{n-1} \left(\sum_{i=1}^{d} \partial^2_{x_i} f(B_{t_j})(B^{(i)}_{t_{j+1}} - B^{(i)}_{t_j})^2 + \sum_{1 \le i < k \le d} \partial^2_{x_i x_k} f(B_{t_j})(B^{(i)}_{t_{j+1}} - B^{(i)}_{t_j})(B^{(k)}_{t_{j+1}} - B^{(k)}_{t_j}) \right)$$

The second term on the right converges to 0 in the L^2 sense when $i \neq k$, from exercise (7.2). This explains why the mixed derivatives disappear in the multi-dimensional Ito's formula. As for the case $i = k$, it reduces to the quadratic variation as in the one-dimensional case. This is where the Riemann integral arises, after suitable conditioning on $\mathcal{F}_{t_j}$, the sigma-field generated by B_s, $s \le t_j$. □

As in the one-dimensional case,it is not necessary to learn Ito's formula by heart. It suffices to write the differential of the function f to second order. We can then apply the rules of multivariate Ito calculus:

$\cdot$	dt	$dB^{(1)}_t$	$dB^{(2)}_t$	$\dots$
dt	0	0	0	0
$dB^{(1)}_t$	0	dt	0	0
$dB^{(2)}_t$	0	0	dt	0
$\dots$	0	0	0	dt

Note that the rule $dB^{(i)}_t dB^{(j)}_t = 0$ for $i \neq j$ is being motivated by the cross-variation result (7.2).

How can we construct martingales using the Ito's formula? Recall that an Ito integral $(\int_0^t X_s dB_s, t \le T)$ is a martingale whenever the integrand is in $\mathcal{L}^2_c(T)$, the space of adapted processes with continuous paths and for which:

$$\int_0^T \mathbf{E}[X_s^2] ds < \infty$$

The only difference here is that the integrand is a function of many Brownian motions. However, the integrands involved in the Ito integrals of the multidimensional Ito's formula (7.1) are clearly adapted to the filtration $(\mathcal{F}_t : t \ge 0)$ of $(B_t : t \ge 0)$ as they are functions of the Brownian motion at the time. The arguments of Ito integral in (6.2) and (6.1) apply verbatim, if we take the definition of $\mathcal{L}^2_c(t)$ with the filtration $(\mathcal{F}_t : t \ge 0)$ of $(B_t : t \ge 0)$. With this in mind, we have the following corollary.

Corollary 7.1. *(Brownian Martingales) Let $(B_t : t \ge 0)$ be a Brownian motion in $\mathbf{R}^d$. Consider $f \in \mathcal{C}^2(\mathbf{R}^d)$ such that processes $(\partial_{x_i} f(B_t), t \le T) \in \mathcal{L}^2_c(T)$ for every $i \le d$. Then, the process :*

$$f(B_t) - \frac{1}{2}\int_0^t \nabla^2 f(B_s) ds, \quad t \le T$$

where $\nabla^2 = \sum_{i=1}^d \partial^2_{x_i}$ is the Laplacian, is a martingale for the Brownian filtration.

For example, consider the processes $X_t = (B_t^{(1)})^2 + (B_t^{(2)})^2$ and $Y_t = \exp(B_t^{(1)}) \cos(B_t^{(2)})$. Then, we have :

$$\frac{1}{2}\int_0^t \nabla^2 X_s ds = \frac{1}{2}\int_0^t 4 ds = 2t$$

and

$$\frac{1}{2}\int_0^t \nabla^2 Y_s ds = \frac{1}{2}\int_0^t 0 \cdot ds = 0$$

Thus, the processes $X_t - 2t$ and Y_t are martingales for the Brownian filtration. In one dimension, there are no interesting martingales constructed with functions of *space* only. Indeed, $(f(B_t) : t \geq 0)$ is a martingale if and only if $f''(x) = 0$ for all x. But, such functions are of the form $f(x) = ax + b$, $a, b \in \mathbf{R}$. In other words, in one dimension, Brownian martingales of the form $f(B_t)$ are simply $aB_t + b$. Not very surprising! The situation is very different in higher dimensions. Indeed, corollary (7.1) implies that $f(B_t)$ is a martingale whenever f is a harmonic function:

Definition 7.2. A function $f : \mathbf{R}^d \to \mathbf{R}$ is harmonic in $\mathbf{R}^d$ if and only if $\nabla^2 f(x) \equiv 0$ for all $x \in \mathbf{R}^d$. More generally, a function $f : \mathbf{R}^d \to \mathbf{R}$ is harmonic in the region $\mathcal{O} \subset \mathbf{R}^d$ if and only if $\nabla^2 f(x) \equiv 0$ for all $x \in \mathcal{O}$.

Note that the function $f(x) = e^{x_1} \cos x_2$ is harmonic in $\mathbf{R}^d$. This is why the process $Y_t = \exp(B_t^{(1)}) \cos(B_t^{(2)})$ is a martingale. The distinction to a subset of $\mathbf{R}^d$ in the above definition is important since it may happen that the function is harmonic only in a subset of the space; see for example equation. It is possible to define a Brownian martingale in such cases by considering the process until it exits the region. This will turn out to be important as we move ahead.

The multidimensional Ito's formula generalizes to functions of time and space as in proposition (6.3):

Definition 7.3. A function $f : [0, \infty) \times \mathbf{R}^d \to \mathbf{R}$ is in $\mathcal{C}^{1,2}([0, T] \times \mathbf{R}^d)$ if the partial derivative in *time* :

$$\frac{\partial}{\partial t} f(t, \mathbf{x})$$

exists and is continuous and the second order partial derivatives in space:

$$\frac{\partial^2}{\partial x_i^2} f(t, x_1, x_2, \ldots, x_i, \ldots, x_d), \quad 1 \leq i \leq d$$

exist and are continuous.

Theorem 7.2. *(Ito's formula) Let* $(B_t : t \leq T)$ *be a d-dimensional Brownian motion. Consider a function* $f \in \mathcal{C}^{1,2}([0, T] \times \mathbf{R}^d)$. *Then, we have with probability one for all* $t \leq T$*:*

$$f(t, B_t) - f(0, B_0) = \sum_{i=1}^d \int_0^t \partial_{x_i} f(s, B_s) dB_s^{(i)} + \int_0^t \left(\partial_t f(s, B_s) + \sum_{i=1}^d \partial_{x_i}^2 f(s, B_s) \right) ds$$

The martingale condition is then similar to the ones in corollary (7.1): if the processes $(\partial_{x_i} f(s, B_s), t \leq T) \in \mathcal{L}_c^2(T)$ for every $1 \leq i \leq d$, then the process

$$f(t, B_t) - \int_0^t \left\{ \partial_t f(s, B_s) + + \sum_{i=1}^d \partial_{x_i}^2 f(s, B_s) \right\} ds, \quad t \leq T$$

is a martingale for the Brownian filtration. In particular, if f satisfies the partial differential equation:

$$\frac{\partial f}{\partial t} + \frac{1}{2}\nabla^2 f = 0$$

then the process $(f(t, B_t) : t \leq T)$ itself is a martingale.

7.3 Recurrence and Transience of Brownian Motion.

In one dimension, we established in example

8 Ito Processes and Stochastic Differential Equations.

Let's start with the definition of Ito processes.

Definition 8.1. (Ito Process). Let $(B(t) : t \geq 0)$ be a standard brownian motion defined on $(\Omega, \mathcal{F}, \mathbb{P})$. An Ito process $(X(t) : t \geq 0)$ is of the form:

$$X(t) = X(0) + \int_0^t V(s)dB(s) + \int_0^t D(s)ds \tag{8.1}$$

where $(V(t), t \geq 0)$ and $(D(t), t \geq 0)$ are two adapted processes for which the integrals make sense in the sense of Ito and Riemann. We refer to $(V(t) : t \geq 0)$ as the *local volatility* and to $(D(t) : t \geq 0)$ as the *local drift*.

We will often denote an Ito process $(X(t) : t \geq 0)$ in differential form as:

$$dX(t) = D(t)dt + V(t)dB(t) \tag{8.2}$$

This form makes no rigorous sense; when we write it, we mean (8.1). Nevertheless, the differential equation has two great advantages:

(1) It gives some intuition on what drives the variation of $X(t)$. On one hand, there is a contribution of the Brownian increments which are modulated by the volatility $V(t)$. On the other hand, there is a smoother contribution coming from the time variation which is modulated by the drift $D(t)$.

(2) The differential notation has computational power. In particular, evaluating Ito's formula is reduced to computing differentials, as in classical calculus, but by doing it upto the second order.

An important class of Ito processes is given by processes for which the volatility and the drift are simply functions of the position of the process.

Definition 8.2. Let $(B(t) : t \geq 0)$ be a standard Brownian motion. An Ito process $(X(t) : t \geq 0)$ of the form

$$dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t), \quad X(0) = x \tag{8.3}$$

where μ and σ are functions from **R** to **R**, is called a time-homogenous diffusion. An Ito-process $(Y(t), t \geq 0)$ of the form:

$$dY(t) = \mu(t, X(t))dt + \sigma(t, X(t))dB(t) \quad Y(0) = y \tag{8.4}$$

where μ and σ are now functions $[0, \infty) \times \mathbf{R} \to \mathbf{R}$ is called a time-inhomogenous diffusion. The equations above are called *stochastic differential equations* (SDE) of the respective process $(X(t))$ and $(Y(t))$.

In other words, a diffusion $(X(t), t \geq 0)$ is called an Ito process whose local volatility $V(t)$ and local drift $D(t)$ at time t depend only on the position of the process at time t and possibly on the time t itself. It cannot depend on the path of the process before time t or on the explicit values of the driving Brownian motion at that time (which is not the process $X(t)$ itself). The class of diffusions, and of the Ito processes in general, constitutes a huge collection of stochastic processes for stochastic modelling.

Note that an SDE is a generalization of ordinary differential equations or ODEs. Indeed, if there were no randomness, that is, no Brownian motion, the SDE would be reduced to

$$dX(t) = \mu(X(t))dt$$

This can be written for $X(t) = f(t)$ as:

$$\frac{df}{dt} = \mu(f)$$

This is a first-order ordinary differential equation. It governs the deterministic evolution of the function $X(t) = f(t)$ in time. An SDE adds a random term to this evolution that is formally written as:

$$\frac{dX}{dt} = \mu(X(t)) + \sigma(X(t))\frac{dB(t)}{dt}$$

We know very well, that Brownian motion is not differentiable; hence the above is not well-defined. The ill-defined term $dB(t)/dt$ is sometimes called white noise. However, equation (8.3) is well-defined in the sense of the Ito process. These types of equations are well-suited to model phenomena with intrinsic randomness.

Here are some examples of diffusions:

Example 8.1. (Brownian Motion with a drift). If we take $X(t) = \sigma B(t) + \mu t$ for some $\sigma > 0$ and $\mu \in \mathbf{R}$, then we can write $X(t)$ as:

$$X(t) = \int_0^t \sigma dB(t) + \int_0^t \mu dt, \quad X(0) = 0$$

In the differential form this becomes

$$dX(t) = \mu dt + \sigma dB(t)$$

In this case, the local drift and the local volatility are constant.

Example 8.2. (Geometric Brownian Motion). We consider the process $S(t) = \exp((\mu - \sigma^2/2)t + \sigma B(t))$. To find the stochastic differential equation, we apply the Ito's Lemma to

$$f(t, x) = \exp((\mu - \sigma^2/2)t + \sigma x)$$

We have:

$$\begin{aligned} df(t,x) &= \left((\mu - \sigma^2/2) + \frac{1}{2}\sigma^2\right)\exp((\mu - \sigma^2/2)t + \sigma x)dt + \sigma \exp((\mu - \sigma^2/2)t + \sigma x)dB(t) \\ &= \mu S(t)dt + \sigma S(t)dB(t) \end{aligned}$$

Note that the local drift and the local volatility are now proportional to the position. So, the higher $S(t)$, the higher the volatility and drift.

Example 8.3. (Any smooth function of Brownian motion). Ito's formula gurarantees that any smooth function $f(t, B(t))$ of time and a Brownian motion is an Ito process with volatility $V(t) = \partial_t f(t, B(t))$ and drift $D(t) = \partial_x f(t, B(t)) + \frac{1}{2}\partial_{xx} f(t, B(t))$. We will see in further ahead, that, in general, any reasonable function of an Ito process remains an Ito process.

Example 8.4. (An Ito process that is not a diffusion) Consider the process

$$X(t) = \int_0^t B^2(s)dB(s)$$

This is an Ito process with local volatility $V(t) = B(t)^2$ and local drift $D(t) = 0$. However, it is not a diffusion, because the local volatility is not an explicit function of $X(t)$.

It turns out that the Brownian bridge is a time-inhomogenous diffusion and that the Ornstein-Uhlenbeck process is a time-homogenous diffusion. To understand these examples, we need to extend Ito's formula to Ito processes.

8.1 Ito's Formula.

The first step towards a general Ito's formula is the quadratic variation of an Ito process.

Proposition 8.1. *(Quadratic variation of an Ito process). Let* $(B(t), t \geq 0)$ *be a standard Brownian motion and* $(X(t) : t \geq 0)$ *be an Ito process of the form* $dX(t) = V(t)dB(t) + D(t)dt$*. Then, the quadratic variation of the process* $(X(t) : t \geq 0)$ *is:*

$$< X, X >_t = \lim_{n\to\infty} \sum_{j=0}^{n-1} (X(t_{j+1}) - X(t_j))^2 = \int_0^t V(s)^2 ds \tag{8.5}$$

for any partition $(t_j, j \leq n)$ *of* $[0, t]$*, where the limit is in probability.*

Remark. Note that the quadratic variation is increasing in t, but it is not deterministic in general! The quadratic variation is a smooth stochastic process. (It is differentiable) Observe that we recover the quadratic variation for the Brownian motion for $V(t) = 1$ as expected. We also notice that the formula follows easily from the rules of Ito Calculus, thereby showing the consistency of the theory. Indeed we have:

$$\begin{aligned} d < X, X >_t &= (dX(t))^2 = (V(t)dB(t) + D(t)dt)^2 \\ &= V(t)^2(dB(t))^2 + 2V(t)D(t)dB(t) \cdot dt + D^2(t)(dt)^2 \\ &= V(t)^2 dt \end{aligned}$$

Proof. The proof is involved, but it reviews some important concepts of stochastic calculus. We prove the case when the process V is in $\mathcal{L}_c^2(T)$ for some $T > 0$. We write $I(t) = \int_0^t V(s)dB(s)$ and $R(t) = \int_0^t D(s)ds$. We first show that only the Ito integral contributes to the quadratic variation and the Riemann integral does not contribute, so that:

$$< X, X >_t = < I, I >_t \tag{8.6}$$

We have that the increment square of $X(t)$ is:

$$(X(t_{j+1}) - X(t_j))^2 = (I(t_{j+1}) - I(t_j))^2 + 2(I(t_{j+1}) - I(t_j))(R(t_{j+1}) - R(t_j)) + (R(t_{j+1}) - R(t_j))^2$$

The Cauchy-Schwarz inequality implies :

$$\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))(R(t_{j+1})-R(t_j)) \leq \left(\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2\right)^{1/2}\left(\sum_{j=0}^{n-1}(R(t_{j+1})-R(t_j))^2\right)^{1/2}$$

Therefore, to prove equation (8.6), it suffices to show that $\sum_{j=0}^{n-1}(R(t_{j+1})-R(t_j))^2 \to 0$ almost surely. Since $D(s)$ is an almost surely continuous process, the stochastic process $R(t) = \int_0^t D(s)ds$ has continuous paths with probability 1. Therefore:

$$\sum_{j=0}^{n-1}(R(t_{j+1})-R(t_j))^2 = \max_{1\leq j\leq n}|R(t_{j+1})-R(t_j)|\sum_{j=0}^{n-1}(R(t_{j+1})-R(t_j))$$

Since, $R(t)$ is continuous on the compact set $[0,t]$, it is uniformly continuous a.s. So, as $|t_{j+1}-t_j| \to 0$, by uniform continuity it follows that $\max|R(t_{j+1})-R(t_j)| \to 0$ a.s.

It remains to prove that $< I, I >_t= \int_0^t V(s)^2 ds$. We first prove the case when $V \in \mathcal{S}(T)$ is a simple adapted process. Consider a partition $(t_j : j \leq n)$ of $[0,t]$. Without loss of generality, we can suppose that V is constant on each $[t_j, t_{j+1})$ by refining the partition. We then have:

$$\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2 = \sum_{j=0}^{n-1}V(t_j)^2(B(t_{j+1})-B(t_j))^2$$

Now, we have seen in the proof of Ito's formula (6.12) that $\mathbb{E}\left[\left\{\sum_{j=0}^{n-1}V(t_j)^2((B(t_{j+1})-B(t_j))^2-(t_{j+1}-t_j)\right\}^2\right] \to 0$, so $\sum_{j=0}^{n-1}V(t_j)^2(B(t_{j+1})-B(t_j))^2$ approaches $\sum_{j=0}^{n-1}V(t_j)^2(t_{j+1}-t_j)$ in the mean square sense. As the mesh size becomes finer, the L^2-limit is $\int_0^t V(t)^2 dt$.

The case $V \in \mathcal{L}_c^2(T)$ is proved by approximating V by a simple process in $\mathcal{S}(T)$. More precisely, we can find a simple process $V^{(\epsilon)}(t)$ that is ϵ-close to V in the sense:

$$||I^{(\epsilon)} - I|| = ||\int V^\epsilon dB(t) - \int V dB(t)|| = \int_0^t \mathbb{E}[(V^{(\epsilon)}(t)-V(t))^2]ds < \epsilon \tag{8.7}$$

To prove the claim, we need to show that for $t \leq T$,

$$\mathbb{E}\left[\left|\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2 - \int_0^t (V(s))^2 ds\right|\right] \to 0 \quad \text{as} \quad n\to\infty$$

L^1-convergence implies convergence in probability of the sequence $\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2$. We now introduce the $V^{(\epsilon)}(t)$ approximation inside the absolute value as well as its corresponding integral $I^{(\epsilon)}(t) = \int_0^t V^{(\epsilon)}(s)ds$. By the triangle inequality, we have:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2-\int_0^t (V(s))^2 ds\right|\right]\\
=&\mathbb{E}\left[\left|\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2-(I^{(\epsilon)}(t_{j+1})-I^{(\epsilon)}(t_j))^2+(I^{(\epsilon)}(t_{j+1})-I^{(\epsilon)}(t_j))^2-\int_0^t (V^{(\epsilon)}(s))^2 ds\right.\right.\\
&\left.\left.+\int_0^t (V^{(\epsilon)}(s))^2 ds-\int_0^t (V(s))^2 ds\right|\right]\\
\leq&\mathbb{E}\left[\left|\sum_{j=0}^{n-1}(I(t_{j+1})-I(t_j))^2-(I^{(\epsilon)}(t_{j+1})-I^{(\epsilon)}(t_j))^2\right|\right]+\mathbb{E}\left[\left|\sum_{j=0}^{n-1}(I^{(\epsilon)}(t_{j+1})-I^{(\epsilon)}(t_j))^2-\int_0^t (V^{(\epsilon)}(s))^2 ds\right|\right]\\
&+\mathbb{E}\left[\left|\int_0^t (V^{(\epsilon)}(s))^2 ds-\int_0^t (V(s))^2 ds\right|\right]
\end{aligned}
\tag{8.8}
$$

We show that the first and third terms converge uniformly and that the second term goes to 0 as $n \to \infty$.

The second term goes to 0 as $n \to \infty$ by the argument for simple processes. $< I^{(\epsilon)}, I^{(\epsilon)} >_t = \int_0^t V^{(\epsilon)}(s)^2 ds.$

For the third term, the linearity of the integral and the Cauchy Schwarz inequality (applied to $\mathbb{E}\int_0^t$) imply that it is:

$$
\mathbb{E}\left[\left|\int_0^t (V^{(\epsilon)}(s))^2 ds-\int_0^t (V(s))^2 ds\right|\right] \leq \mathbb{E}\left[\left|\int_0^t (V^{(\epsilon)}(s)-V(s))^2 ds\right|\right]^{1/2} \mathbb{E}\left[\left|\int_0^t (V^{(\epsilon)}(s)+V(s))^2 ds\right|\right]^{1/2}
$$

The first factor is smaller than the square root of ϵ by equation (8.7), whereas the second factor is bounded.

The first term in equation (8.8) is handled similarly. The linearity of the Ito integral and the Cauchy-Schwarz inequality applied to $\mathbb{E}\left[\sum_{j=0}^{n-1}\left(\int_{t_j}^{t_{j+1}}\cdot\right)\right]$ give that the first term is:

$$
\begin{aligned}
&\mathbb{E}\left[\left|\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(V(s))^2 dB(s)-\int_{t_j}^{t_{j+1}}(V^{\epsilon}(s))^2 dB(s)\right|\right]\\
=&\mathbb{E}\left[\left|\sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}(V(s)-V^{\epsilon}(s))dB(s)\int_{t_j}^{t_{j+1}}(V(s)+V^{\epsilon}(s))dB(s)\right|\right]\\
\leq&\mathbb{E}\left[\sum_{j=0}^{n-1}\left(\int_{t_j}^{t_{j+1}}(V(s)-V^{\epsilon}(s))dB(s)\right)^2\right]^{1/2}\mathbb{E}\left[\left|\sum_{j=0}^{n-1}\left(\int_{t_j}^{t_{j+1}}(V(s)+V^{\epsilon}(s))dB(s)\right)^2\right|\right]^{1/2}
\end{aligned}
$$

By Ito isometry, the first factor in the above expression can be simplified:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{j=0}^{n-1}\left(\int_{t_j}^{t_{j+1}}(V(s)-V^{\epsilon}(s))dB(s)\right)^2\right]^{1/2} &= \sum_{j=0}^{n-1}\mathbb{E}\left(\int_{t_j}^{t_{j+1}}(V(s)-V^{\epsilon}(s))dB(s)\right)^2\\
&= \sum_{j=0}^{n-1}\int_{t_j}^{t_{j+1}}\mathbb{E}[(V(s)-V^{\epsilon}(s))^2]ds
\end{aligned}
$$

By equation (8.7), this factor is smaller than ϵ. The second factor equals $\sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} \mathbb{E}[(V(s)+V^\epsilon(s))^2]ds$ by Ito-isometry and is uniformly bounded. This concludes the proof of the proposition. $\square$

Note that quadratic variation $< I, I >_t = \int (V(s))^2 ds$ is computed path-by-path and hence the result is random. On the other the variance of the Ito integral $Var(I(t)) = \mathbb{E}[I_t^2] = \int \mathbb{E}[V_s^2]ds$ is the mean value of all possible paths of the quadratic variation and hence is non-random. We are now ready to state Ito's formula for Ito processes. We write the result in differential form for conciseness.

Theorem 8.1. *(Ito's formula for Ito processes). Let $(B(t) : t \geq 0)$ be a standard brownian motion, and let $(X(t) : t \geq 0)$ be an Ito process of the form $dX(t) = V(t)dB(t) + D(t)dt$. Consider a function $f(t, x) \in \mathcal{C}^{1,2}([0, T] \times \mathbf{R})$. Then we have with probability one for all $t \leq T$:*

$$df(t, X(t)) = \partial_x f(t, X(t))dX(t) + \partial_t f(t, X(t))dt + \frac{1}{2}\partial_{xx} f(t, X(t))d < X, X >_t$$

This can also be written as:

$$df(t, X(t)) = \partial_x f(t, X(t))V(t)dB(t) + \left[\partial_x f(t, X(t))D(t) + \partial_t f(t, X(t)) + \frac{1}{2}(V(t))^2 \partial_{xx} f(t, X(t))\right] dt$$

The proof of the theorem (8.1) is again a Taylor approximation with the form of the quadratic variation of the process. We will omit it.

Example 8.5. (Ornstein-Uhlenbeck Process). Consider the Ornstein-Uhlenbeck process $(Y(t) : t \geq 0)$:

$$Y(t) = Y(0)e^{-t} + e^{-t} \int_0^t e^s dB(s)$$

Note that this process is an explicit function of t and of the Ito process $X(t) = Y(0) + \int_0^t e^s dB(s)$. Indeed, we have:

$$Y(t) = e^{-t} X(t)$$

Let $f(t, x) = e^{-t}x$. Then, $f_x(t, x) = e^{-t}$, $f_{xx}(t, x) = 0$ and $f_t(t, x) = -e^{-t}x$. So, by Ito's lemma,

$$\begin{aligned}
df(t, x) &= f_t(t, X(t))dt + f_x(t, X(t))dX(t) + \frac{1}{2} f_{xx}(t, X(t))d < X, X >_t \\
dY(t) &= -Y(t)dt + e^{-t}dX(t) \\
dY(t) &= -Y(t)dt + e^{-t}(e^t dB(t)) \\
dY(t) &= -Y(t)dt + dB(t) \qquad (8.9)
\end{aligned}$$

This is the SDE for the Ornstein Uhlenbeck process.

The SDE has a very nice interpretation: The drift is positive if $Y(t) < 0$ and negative if $Y(t) > 0$. Moreover, the drift is proportional to the position (exactly like a spring pulling the process back to the x-axis following the Hooke's law!). This is the mechanism that ensures that the process does not venture too far from 0 and is eventually stationary.

The SDE (8.9) is now easily generalized by adding two parameters for the volatility and the drift:

$$dY(t) = -kY(t)dt + \sigma dB(t), \quad k \in \mathbf{R}, \sigma > 0 \tag{8.10}$$

It is not hard to check that the solution to the SDE is:

$$Y(t) = Y(0)e^{-kt} + e^{-kt}\int_0^t e^{ks}\sigma dB(s) \tag{8.11}$$

Exercise 8.1. The Ornstein-Uhlenbeck process with parameters. Use the Ito's formula to show that the equation (8.11) is the solution to the Ornstein-Uhlenbeck SDE (8.10).

Solution.

Let $X(t) = Y(0) + \int_0^t e^{ks}\sigma dB(s)$, so $dX(t) = e^{kt}\sigma dB(t)$. Then, $Y(t) = e^{-kt}X(t)$. Let $f(t,x) = e^{-kt}x$. Then, by Ito's formula:

$$\begin{aligned}
df(t,x) &= -ke^{-kt}X(t)dt + e^{-kt}dX(t) \\
dY(t) &= -kY(t)dt + e^{-kt}e^{kt}\sigma dB(t) \\
dY(t) &= -kY(t)dt + \sigma dB(t)
\end{aligned}$$

The latest version of Ito's formula is another useful tool for producing martingales from a function of an Ito process. We start with two examples generalizing martingales for Brownian motion.

Example 8.6. (A generalization of $(B(t))^2 - t$). Let $(V(t) : t \leq T)$ be a process in $\mathcal{L}_c^2(T)$. Consider an Ito process $(X(t) : t \leq T)$ given by $dX(t) = V(t)dB(t)$. Note that $((X(t))^2 : t \leq T)$ is a submartingale by Jensen's inequality, since $\mathbb{E}[X^2(t)|\mathcal{F}_s] \geq (\mathbb{E}[X(t)|\mathcal{F}_s)^2 = X^2(s)$. We show that the compensated process

$$M(t) = X^2(t) - \int_0^t V^2(s)ds, \quad t \leq T$$

is a martingale for the Brownian filtration. (This is another instance of the Doob-Meyer decomposition). By the Ito's formula for $f(x) = x^2$, we have:

$$\begin{aligned}
df(x) &= f_x(X(t)dX(t) + \frac{1}{2}f_{xx}(X(t))d < X, X >_t \\
&= 2X(t)dX(t) + (V(t))^2 dt \\
df(X(t)) &= 2X(t)V(t)dB(t) + (V(t))^2 dt
\end{aligned}$$

In Integral form this implies:

$$\begin{aligned}
(X(t))^2 &= (X(0))^2 + 2\int_0^t X(s)V(s)dB(s) + \int_0^t (V(s))^2 ds \\
M(t) = (X(t))^2 - \int_0^t (V(s))^2 ds &= (X(0))^2 + 2\int_0^t X(s)V(s)dB(s)
\end{aligned}$$

We conclude that $(M(t) : t \leq T)$ is a martingale, provided $X(t)V(t) \in L_c^2(T)$.

There is another more direct way to prove that $(M(t) : t \leq T)$ is a martingale whenever $(V(t) : t \leq T) \in \mathcal{L}_c^2(T)$. This is by using increments: for $t' < t \leq T$,

$$\begin{aligned}
\mathbb{E}[X_{t'}^2|\mathcal{F}_t] &= \mathbb{E}[(X_t + (X_{t'} - X_t))^2|\mathcal{F}_t] \\
&= \mathbb{E}[X_t^2 + 2X_t(X_{t'} - X_t) + (X_{t'} - X_t)^2|\mathcal{F}_t] \\
&= X_t^2 + 2X_t\mathbb{E}[X_{t'} - X_t|\mathcal{F}_t] + \mathbb{E}[(X_{t'} - X_t)^2|\mathcal{F}_t]
\end{aligned}$$

Since $(X_t : t \geq 0)$ is a martingale, $\mathbb{E}[(X_{t'} - X_t)|M_t] = 0$, so the middle term equals zero and we are left with:

$$\mathbb{E}[X_{t'}^2|\mathcal{F}_t] = X_t^2 + \mathbb{E}[(X_{t'} - X_t)^2|\mathcal{F}_t]$$

By conditional Ito Isometry,

$$\mathbb{E}[(X_{t'} - X_t)^2|\mathcal{F}_t] = \int_0^{t'} V_s^2 ds - \int_0^t V_s^2 ds = \int_t^{t'} V_s^2 ds$$

Example 8.7. (A generalization of the geometric Brownian motion). Let $\sigma(t)$ be a continuous, deterministic function such that $|\sigma(t)| \leq 1, t \in [0, T]$. The process

$$M(t) = \exp\left(\int_0^t \sigma(s)dB(s) - \frac{1}{2}\int_0^t \sigma^2(s)ds\right), \quad t \leq T$$

is a martingale for the Brownian filtration. To see this, note that we can write $M(t) = f(t, X(t))$ where $f(t, x) = \exp(x - \frac{1}{2}\int \sigma^2(s)ds)$ and $X(t) = \int_0^t \sigma(s)dB(s)$, so $dX(t) = \sigma(t)dB(t)$. Ito's formula gives:

$$\begin{aligned}
df(t, x) &= f_t(t, X(t)dt + f_x(t, X(t))dX(t) + \frac{1}{2}f_{xx}(t, X(t))d < X, X >_t \\
dM(t) &= -\frac{1}{2}\sigma^2(t)M(t)dt + M(t)\sigma(t)dB(t) + \frac{1}{2}M(t)\sigma^2(t)dt \\
&= M(t)\sigma(t)dB(t) \\
M(t) &= M(0) + \int_0^t M(s)\sigma(s)dB(s)
\end{aligned}$$

Observe also that:

$$\mathbb{E}[M_t^2] = e^{-\int_0^t \sigma^2(s)ds}\mathbb{E}[e^{2\int_0^t \sigma(s)dB(s)}] = e^{\int_0^t \sigma^2(s)ds}$$

since $\int_0^t \sigma(s)dB(s)$ is a Gaussian random variable with mean 0 and variance $\int_0^t \sigma^2(s)ds$. So, $\mathbb{E}[e^{2\int_0^t \sigma(s)dB(s)}] = \exp[\frac{1}{2} \times 4 \times \int_0^t \sigma^2(s)ds] = \exp(2\int_0^t \sigma^2(s)ds)$.

We conclude from the equation that $(M(t), t \geq 0)$ is a martingale.

Example 8.8. (Martingales of Geometric Brownian Motion). Let

$$S(t) = S(0)\exp(\sigma B(t) - \sigma^2 t/2)$$

be a geometric brownian motion. We find a PDE satisfied by $f(t,x)$ for $f(t,S(t))$ to be a martingale. It suffices to apply Ito's formula of theorem (8.1). We get:

$$df(t,S(t)) = f_t(t,S(t))dt + f_x(t,S(t))dS(t) + \frac{1}{2}f_{xx}(t,S(t))dS(t)\cdot dS(t)$$

Now note from the earlier result that $dS(t) = S(t)\sigma dB(t)$. So, $dS(t)\cdot dS(t) = \frac{1}{2}\sigma^2(S(t))^2 dt$. So,

$$df(t,S(t)) = \left\{\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2(S(t))^2\frac{\partial^2 f}{\partial x^2}\right\}dt + \sigma S(t)\frac{\partial f}{\partial x}dB(t)$$

Finally, the PDE for $f(t,x)$ is obtained by setting the factor in front of dt to 0, because we want f to be a martingale process. It is important to keep in mind, that the PDE should always be written in terms of the time variable t and the space variable x. Therefore, the PDE of f as a function of time and space is:

$$\frac{1}{2}\sigma^2x^2\frac{\partial^2 f}{\partial x^2}(t,x) + \frac{\partial f}{\partial t}(t,x) = 0$$

No more randomness appears in the PDE!

Here is a specific case where we can apply the Ito's formula to construct martingales of Ito processes.

Example 8.9. Consider the process given by the SDE:

$$dX(t) = X(t)dB(t), \quad X(0) = 2$$

Let's find a PDE for which $f(t,X(t))$ is a martingale for the Brownian filtration. We have by Ito's formula that:

$$\begin{aligned} df(t,X(t)) &= f_t(t,X(t))dt + f_x(t,X(t))dX(t) + \frac{1}{2}f_{xx}(t,X(t))d<X,X>_t \\ &= \left(\frac{\partial f}{\partial t} + \frac{1}{2}(X(t))^2\frac{\partial^2 f}{\partial x^2}\right)dt + X(t)\frac{\partial f}{\partial x}dB(t) \end{aligned}$$

Setting the drift term to 0 gives the PDE:

$$\frac{\partial f}{\partial t} + \frac{1}{2}x^2\frac{\partial^2 f}{\partial x^2} = 0$$

It is then easy to check that $X(t)$ is a martingale and so is $t + \log(X(t))^2$, since the functions $f(t,x) = x$ and $f(t,x) = t + \log x^2$ satisfy the PDE. However, the process $tX(t)$ is not, as the function $f(t,x) = xt$ is not a solution of the PDE.

8.2 Multivariate Extension.

Ito's formula can be generalized to several Ito processes. Let's start by stating an example of a function of two Ito processes. Such a function $f(x_1,x_2)$ will be a function of two space variables. Not surprisingly, it needs to have two derivatives in each variable and they need to be a continuous function; we need $f \in \mathcal{C}^{2\times 2}(\mathbf{R}\times\mathbf{R})$.

Theorem 8.2. *(Ito's formula for many Ito processes) Let $(X(t) : t \geq 0)$ and $(Y(t) : t \geq 0)$ be two Ito processes of the form:*

$$\begin{aligned} dX(t) &= V(t)dB(t) + D(t)dt \\ dY(t) &= U(t)dB(t) + R(t)dt \end{aligned} \tag{8.12}$$

where $(B(t) : t \geq 0)$ is a standard Brownian motion. Then, for $f \in \mathcal{C}^{2\times 2}(\mathbf{R} \times \mathbf{R})$, we have:

$$\begin{aligned} df(X(t), Y(t)) = f_x(X(t), Y(t))dX(t) + f_y(X(t), Y(t))dY(t) + \frac{1}{2}f_{xx}(X(t), Y(t))d < X, X >_t \\ + f_{xy}(X(t), Y(t))d < X, Y >_t + \frac{1}{2}f_{yy}(X(t), Y(t))d < Y, Y >_t \end{aligned}$$

The idea of the proof is the same as in theorem 8.1 : Taylor's expansion and quadratic variation, together with the cross-variation of two processes.

$$\begin{aligned} dX(t) \cdot dY(t) &= (V(t)dB(t) + D(t)dt)(U(t)dB(t) + R(t)dt) \\ &= U(t)V(t)dt \end{aligned}$$

Example 8.10. (Product Rule) An important example of this formula is Ito's product rule. Let $X(t)$ and $Y(t)$ be as in equation (8.12). Then:

$$d(X(t)Y(t)) = Y(t)dX(t) + X(t)dY(t) + dX(t) \cdot dY(t)$$

Exercise 8.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(B_t : t \geq 0)$ be a standard brownian motion. Using integration by parts, show that

$$\int_0^t B(s)ds = \int_0^t (t-s)dB(s)$$

and prove that $\int_0^t B(s)ds \sim \mathcal{N}(0, t^3/3)$.

Is

$$X(t) = \begin{cases} 0 & t = 0 \\ \frac{\sqrt{3}}{t}\int_0^t B(s)ds & t > 0 \end{cases}$$

a standard Wiener process?

Solution.

Using integration by parts:

$$\int u\left(\frac{dv}{ds}\right)ds = uv - \int v\left(\frac{du}{ds}\right)ds$$

We set $u = B(s)$ and $dv/ds = 1$. Then:

$$
\begin{aligned}
\int_0^t B(s)ds &= sB(s)|_0^t - \int_0^t sdB(s) \\
&= tB(t) - \int_0^t sdB(s) \\
&= \int_{s=0}^{s=t} tdB(s) - \int_0^t sdB(s) \\
&= \int_0^t (t-s)dB(s)
\end{aligned}
$$

Thus, $\int_0^t B(s)ds$ is a Gaussian random variable with:

$$
\begin{aligned}
\mathbb{E}\left[\int_0^t B(s)ds\right] &= \mathbb{E}\left[\int_0^t (t-s)dB(s)\right] \\
&= 0
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\left(\int_0^t B(s)ds\right)^2\right] &= \int_0^t (t-s)^2 ds \\
&= \left.\frac{(t-s)^3}{-3}\right|_0^t \\
&= \frac{t^3}{3}
\end{aligned}
$$

Thus, using the properties of Ito Integral, $\int_0^t B(s)ds = \int_0^t (t-s)dB(s)$ is a martingale. Now the quadratic variation $< M, M >_t = 0$, and this can be a bit tricky. Remember, $\left\langle \int_0^t f(s, B_s)dB(s), \int_0^t f(s, B_s)dB(s) \right\rangle = \int_0^t f^2(s, B_s)ds$ if and only if f is a function of the time s and the position of the Brownian motion $B(s)$. Since, f is a function of t as well, this rule cannot be applied.

By first principles, we can show that, the quadratic variation is indeed 0:

$$
\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\sum_{j=0}^{n-1} (I(t_{j+1}) - I(t_j))^2\right] &= \lim_{n\to\infty} \mathbb{E}\left[\sum_{j=0}^{n-1} B_{t_j}^2 (t_{j+1} - t_j)^2\right] \\
&= \lim_{n\to\infty} \max_{1\leq j\leq n} |t_{j+1} - t_j| \cdot \mathbb{E}\left[\sum_{j=0}^{n-1} B_{t_j}^2 (t_{j+1} - t_j)\right]
\end{aligned}
$$

Since the paths of B_t are continuous, so are the paths B_t^2 on the compact interval $[0, t]$. So, $(B_s^2, s \in [0, t])$ is uniformly bounded. Thus, the expectation term is bounded. As $n \to \infty$, the mesh size approaches zero, and consequently the quadratic variation approaches zero.

Example 8.11. Let $X_t = \int_0^t B_s dB_s$ and $Y_t = \int_0^t B_s^2 dB_s$. Is $(X_t Y_t, t \geq 0)$ a martingale?

Solution.

By Ito's product rule, we have:

$$
\begin{aligned}
d(X_tY_t) &= X_t dY_t + Y_t dX_t + dX_t \cdot dY_t \\
&= X_t B_s^2 dB_s + Y_t B_s dB_s + (B_s dB_s) \cdot (B_s^2 dB_s) \\
&= X_t B_s^2 dB_s + Y_t B_s dB_s + B_s^3 dt \\
X_tY_t &= X_0Y_0 + \int_0^t X_t B_s^2 dB_s + \int_0^t Y_t B_s dB_s + \int_0^t B_s^3 dt
\end{aligned}
$$

The term in dt is not zero, so the product cannot be a martingale.

Example 8.12. (A generalization of Geometric Brownian Motion). Consider $(\int_0^t V_s dB_s, t \geq 0)$ an Ito process. Define the positive process:

$$
M_t = \exp\left(\int_0^t V_s dB_s - \frac{1}{2}\int_0^t V_s^2 ds\right), \quad t \geq 0 \tag{8.13}
$$

Solution.

Ito's formula applied to the processes $X_t = \int_0^t V_s dB_s$ and $Y_t = \frac{1}{2}\int_0^t V_s^2 ds$ with the function $f(x,y) = e^{x-y}$ yields:

$$
\begin{aligned}
df(x,y) &= f_x(X_t, Y_t) dX_t + f_y(X_t, Y_t) dY_t \\
&+ \frac{1}{2} f_{xx}(X_t, Y_t) dX_t \cdot dX_t + \frac{1}{2} f_{yy}(X_t, Y_t) dY_t \cdot dY_t \\
&+ f_{xy}(X_t, Y_t) dX_t \cdot dY_t
\end{aligned}
$$

Now, all first and second order derivatives are $\partial_x(e^{x-y}) = M_t$, $\partial_y(e^{x-y}) = -e^{x-y} = -M_t$. $dX_t = V_t dB_t$. $dY_t = \frac{1}{2}V_t^2 dt$. $dX_t \cdot dX_t = V_t^2 dt$, $dX_t \cdot dY_t = 0$, $dY_t \cdot dY_t = 0$. Consequently, we have:

$$
\begin{aligned}
dM_t &= M_t V_t dB_t - \frac{1}{2} M_t V_t^2 dt \\
&+ \frac{1}{2} M_t V_t^2 dt \\
&= M_t V_t dB_t
\end{aligned}
$$

Thus, $(M_t, t \geq 0)$ is a martingale.

Exercise 8.3. (Generalized Ito Integral). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $(B_t : t \geq 0)$ be a standard brownian motion. Given that f is a simple process, show that:

$$
\begin{aligned}
\int_0^t f(s, B_s) dB_s = B_t f(t, B_t) &- \int_0^t \left[B_s \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} + \frac{1}{2} B_s \frac{\partial^2 f}{\partial x^2}\right] ds \\
&- \int_0^t B_s \frac{\partial f}{\partial x} dB_s
\end{aligned}
$$

and

$$
\int_0^t f(s, B_s) ds = t f(t, B_t) - \int_0^t s \left[\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\right] ds - \int_0^t s \frac{\partial f}{\partial x} dB_s
$$

Solution.

I suppress (t, B_t) for simplicity. Applying the product rule to $B_t f$, we get:

$$\begin{aligned}
d(B_t f) &= f dB_t + B_t df + dB_t \cdot df \\
&= f dB_t + B_t \left(\frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (dB_t)^2 \right) \\
&+ dB_t \cdot \left(\frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (dB_t)^2 \right) \\
&= f dB_t + \left(B_t \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} + \frac{1}{2} B_t \right) dt + B_t \frac{\partial f}{\partial x} dB_t \\
B_t f &= \int_0^t f dB_s + \int_0^t \left(B_s \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} B_s \right) ds + \int_0^t B_s \frac{\partial f}{\partial x} dB_s \\
\int_0^t f dB_s &= B_t f - \int_0^t \left(B_s \frac{\partial f}{\partial t} + \frac{\partial f}{\partial x} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} B_s \right) ds - \int_0^t B_s \frac{\partial f}{\partial x} dB_s
\end{aligned}$$

Applying product rule to $tf(t, B_t)$, we get:

$$\begin{aligned}
d(tf) &= f dt + t df + dt \cdot df \\
&= f dt + t \left(\frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (dB_t)^2 \right) \\
&+ dt \left(\frac{\partial f}{\partial t} dt + \frac{\partial f}{\partial x} dB_t + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} (dB_t)^2 \right) \\
&= f dt + t \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \right) dt + t \frac{\partial f}{\partial x} dB_t \\
tf &= \int_0^t f ds + \int_0^t s \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \right) ds + \int_0^t s \frac{\partial f}{\partial x} dB_s \\
\int_0^t f ds &= tf - \int_0^t s \left(\frac{\partial f}{\partial t} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \right) ds - \int_0^t s \frac{\partial f}{\partial x} dB_s
\end{aligned}$$

The following example will be important when we discuss the Girsanov theorem.

8.3 Numerical Simulation of SDEs.

It is not too hard to implement iterative schemes to sample paths of a diffusion. Consider $(X_t : t \leq T)$ a solution to the SDE:

$$dX_t = \sigma(X_t) dB_t + \mu(X_t) dt, \quad X_0 = x$$

8.4 Existence and Uniqueness of solutions to SDEs.

Theorem 8.3. *(Existence and uniqueness of solutions to SDEs). Consider the SDE*

$$dX_t = \mu(X_t) dt + \sigma(X_t) dB_t, \quad X_0 = x, \quad t \in [0, T]$$

If the functions σ and μ grow not faster than Kx^2 for some $K > 0$ and are differentiable with bounded derivatives on **R**, *then there exists a unique solution $(X_t : t \in [0, T])$ to the SDE. In other words, there exists a continuous process $(X_t, t \leq T)$ adapted to the filtration of the brownian motion:*

$$X_t = x + \int_0^t \mu(X_s)ds + \int_0^t \sigma(X_s)dB_s, \quad t \leq T$$

Example 8.13. Consider the SDE:

$$dX_t = \sqrt{1 + X_t^2} dB_t + \sin X_t dt, \quad X_0 = 0$$

There exists a unique diffusion process $(X_t : t \geq 0)$ that is a solution of this SDE. To see this, we verify the conditions of theorem (8.3). We have:

$$\mu(x) = \sin x$$
$$\sigma(x) = \sqrt{1 + x^2}$$

Clearly, these functions satisfy the growth condition since μ is bounded and σ grows like $|x|$ for x large. As for the derivatives, we have $\sigma'(x) = \frac{1}{2\sqrt{1+x^2}}$ and $\mu'(x) = \cos x$. These two derivatives

8.5 Martingale Representation and Levy's characterization

We know, very well, by now that an Ito integral is a continuous martingale with respect to the Brownian fitration, whenever the integrand is in $\mathcal{L}_c^2(T)$. What can we say about the converse? In other words, if we have a martingale with respect to some Brownian filtration, can it be expressed as an Ito-integral for some integrand $(V(t) : t \in [0, T])$. Amazingly the answer to this question is yes.

Theorem 8.4. *(Martingale Representation Theorem). Let $(B(t) : t \in [0, T])$ be a Brownian motion with filtration $(\mathcal{F}_t : t \geq 0)$ on $(\Omega, \mathcal{F}, \mathbb{P})$. Consider a martingale $(M(t) : t \in [0, T])$ with respect to this filtration. Then, there exists an adapted process $(V_t : t \leq T)$ such that:*

$$M_t = M_0 + \int_0^t V_s dB_s, \quad t \leq T \tag{8.14}$$

One striking fact of this result is that $(M_t : t \leq T)$ ought to be continuous. In other words, we cannot construct a process with a jump that is a martingale adapted to a Brownian motion!

Instead of proving the theorem, we will see how the result is not too surprising with stronger assumptions. Instead of supposing that M_t is $\mathcal{F}_t$-measurable, take that M_t is $\sigma(B_t)$-measurable. In other words, $M_t = h(B_t)$ for some function h. In the case that h is smooth, then it is clear by the Ito's formula that the representation in equation (8.14) holds and $V_s = h'(B_t)$.

The relevance to hedging of this is that the only source of uncertainty in the model is the Brownian motion appearing in theorem (8.4), and hence there is only one source of uncertainty to be removed by hedging. This assumption implies that the martingale cannot have jumps because Ito integrals are continuous. If we want to have a martingale with jumps, we will need to build a model that includes sources of uncertainty different from or in addition to Brownian motion.

9 Change of Probability.

9.1 Change of Probability for a Random Variable.

Consider a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}[X] = 0$. We would like to change the mean of X so that $\mu \neq 0$. Of course, it is easy to change the mean of a random variable: If X has mean 0, then the random variable $X + \mu$ has mean μ. However, it might be that the variable $X + \mu$ does not share the same possible values as X. For example, take X to be a uniform random variable on $[-1, 1]$. While $X + 1$ has mean 1, the density of $X + 1$ would be non-zero on $[0, 2]$ instead of $[-1, 1]$.

Our goal is to find a good way to change the underlying probability $\mathbb{P}$, and thus the distribution of X, so that the set of outcomes is unchanged. If X is a discrete random variable, say with $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = 1/2$, we can change the probability in order to change the mean easily. It suffices to take $\tilde{\mathbb{P}}$ so that $\tilde{\mathbb{P}}(X = 1) = p$ and $\mathbb{P}(X = -1) = 1 - p$ for some appropriate $0 \leq p \leq 1$.

If X is a continuous random variable, with a PDF f_X, the probabilities can be changed by modifying the PDF. Consider the a new PDF:

$$\tilde{f}_X(x) = f_X(x)g(x)$$

for some function $g(x) > 0$ such that $\int f(x)g(x)dx = 1$. Clearly, $f_X(x)g(x)$ is also a PDF and $f_X(x) > 0$ if and only if $f_X(x)g(x) > 0$, so that the possible values of X are unchanged. A convenient (and important!) choice of function g is:

$$g(x) = \frac{e^{ax}}{\int_{\mathbf{R}} e^{ax} f_X(x)dx} = \frac{e^{ax}}{\mathbb{E}[e^{aX}]}, \quad a \in \mathbf{R} \tag{9.1}$$

assuming X has a well-defined MGF. Here a is a parameter that can be tuned to fit to a specific mean. The normalization factor in the denominator is the MGF of X. It ensures that $f_X(x)g(x)$ is a PDF. Note that if $a > 0$, the function g gives a bigger weight to large values of X. We say that g is biased towards the large values.

Example 9.1. (Biasing a uniform random variable). Let X be a uniform random variable on $[0, 1]$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Clearly, $\mathbb{E}[X] = 1/2$. How can we change the PDF of X so that the possible values are still $[0, 1]$, but the mean is $1/4$. We have that the PDF is $f_X(x) = 1$ if $x \in [0, 1]$ and 0 elsewhere. Therefore, the mean with the new PDF with parameter a as in the equation (9.1) is:

$$\begin{aligned}
\tilde{\mathbb{E}}[X] &= \int_0^1 x\tilde{f}(x)dx \\
&= \int_0^1 \frac{xe^{ax}}{\mathbb{E}[e^{aX}]}dx \\
&= \frac{a}{e^a - 1}\int_0^1 xe^{ax}dx \\
&= \frac{a}{e^a - 1}\left(\left[x\frac{e^{ax}}{a}\right]_0^1 - \frac{1}{a}\int_0^1 e^{ax}dx\right) \\
&= \frac{a}{e^a - 1}\left(\frac{e^a}{a} - \frac{1}{a}\frac{e^a - 1}{a}\right) \\
&= \frac{e^a}{e^a - 1} - \frac{1}{a}
\end{aligned}$$

For $\mathbb{E}[\tilde{X}]$to be equal to $1/4$, we get numerically $a \approx -3.6$. Note that the possible values of X remain the same under the new probability. However, the new distribution is no longer uniform! It has bias towards values closer to zero, as it should.

Example 9.2. (Biasing a Gaussian random variable). Let X be a Gaussian random variable with mean μ and variance σ^2. How can we change the PDF of X to have mean 0? Going back to (9.1), the mean μ under the new PDF with parameter a is:

$$\begin{aligned}
\tilde{\mathbb{E}}[X] &= \int_{-\infty}^{\infty} x\tilde{f}(x)dx \\
&= \int_{-\infty}^{\infty} xg(x)f(x)dx \\
&= \int_{-\infty}^{\infty} x \cdot \frac{e^{ax}}{\mathbb{E}[e^{aX}]} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \\
&= \frac{1}{e^{\mu a+\frac{1}{2}a^2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \cdot \exp\left[-\frac{1}{2}\left(\frac{x^2-2\mu x+\mu^2-2a\sigma^2 x}{\sigma^2}\right)\right] dx \\
&= \frac{1}{e^{\mu a+\frac{1}{2}a^2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \cdot \exp\left[-\frac{1}{2}\left(\frac{x^2-2(\mu+a\sigma^2)x+(\mu+a\sigma^2)^2-2\mu a\sigma^2-a^2\sigma^4}{\sigma^2}\right)\right] dx \\
&= \frac{e^{\mu a+a^2\sigma^2/2}}{e^{\mu a+\frac{1}{2}a^2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left[-\frac{1}{2}\left(\frac{x-(\mu+a\sigma^2)}{\sigma}\right)^2\right] dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left[-\frac{1}{2}\left(\frac{x-(\mu+a\sigma^2)}{\sigma}\right)^2\right] dx
\end{aligned}$$

For the specific choice of the parameter $a = \mu/\sigma^2$, we recover the PDF of a Gaussian random variable with mean 0. But, we can deduce more. The new PDF is also Gaussian. This was not the case for uniform random variables. In fact, the new PDF is exactly the same as the one of $X - \mu$. For if, $a = \mu/\sigma^2$, we have:

$$\tilde{\mathbb{E}}[X] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left[-\frac{x^2}{2\sigma^2}\right] dx$$

and observe that if $Y = X - \mu$, then:

$$\begin{aligned}
F_Y(x) &= \mathbb{P}(X-\mu < x) \\
&= \mathbb{P}(X \leq x+\mu) \\
&= F_X(x+\mu) \\
\frac{d}{dx}(F_Y(x)) &= \frac{d}{dx}(F_X(x+\mu)) \\
f_Y(x) &= f_X(x+\mu) \cdot \frac{d}{dx}(x+\mu) \\
f_Y(x) &= f_X(x+\mu) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x+\mu-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{x^2}{2\sigma^2}\right]
\end{aligned}$$

In other words:

For Gaussians, changing the mean by recentering is equivalent to changing the probability as in (9.1).

This is a very special property of the Gaussian distribution. The exponential and Poisson distributions have a similar property.

Example (9.2) is very important and we will state it as a theorem. Before doing so, we notice that the change of PDF (9.1) can be expressed more generally by changing the underlying probability measure(length, area, weights) $\mathbb{P}$ on the sample space Ω on which the random variables are defined. More precisely, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let X be a random variable defined on Ω. We define a new probability $\tilde{\mathbb{P}}$ on Ω as follows:

If $\mathcal{E}$ is an event in $\mathcal{F}$, then:

$$\begin{aligned}
\tilde{\mathbb{P}}(\mathcal{E}) = \tilde{\mathbb{E}}[1_{\mathcal{E}}] &= \int_{\mathbf{R}} 1_{\mathcal{E}} \cdot \tilde{f}(x)dx \\
&= \int_{\mathbf{R}} 1_{\mathcal{E}} \cdot g(x) f_X(x)dx \\
&= \int_{\mathbf{R}} 1_{\mathcal{E}} \cdot \frac{e^{ax}}{\mathbb{E}[e^{aX}]} f_X(x)dx \\
&= \mathbb{E}\left[1_{\mathcal{E}} \frac{e^{aX}}{\mathbb{E}[e^{aX}]}\right] \qquad (9.2)
\end{aligned}$$

Intuitively, we are changing the probability of each outcome $\omega \in \mathcal{E}$, by the factor

$$\frac{e^{aX(\omega)}}{\mathbb{E}[e^{aX}]} \qquad (9.3)$$

In other words, if $a > 0$, the outcomes ω for which X has large values are favored. Note that equation (9.1) for the PDF is recovered, since for any function h of X, we have:

$$\begin{aligned}
\tilde{\mathbb{E}}[h(X)] &= \mathbb{E}\left[\frac{e^{aX}}{\mathbb{E}[e^{aX}]} h(X)\right] \\
&= \int_{\mathbf{R}} h(x) \frac{e^{ax}}{\mathbb{E}[e^{aX}]} f_X(x)dx
\end{aligned}$$

In this setting, the above example becomes the preliminary version of the Cameron-Martin-Girsanov theorem:

Theorem 9.1. *Let X be a Gaussian random variable with mean μ and variance σ^2 defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then, under the probability $\tilde{\mathbb{P}}$ given by:*

$$\tilde{\mathbb{P}}(\mathcal{E}) = \mathbb{E}\left[1_{\mathcal{E}} e^{-\frac{\mu}{\sigma^2}X + \frac{1}{2}\frac{\mu^2}{\sigma^2}}\right], \quad \mathcal{E} \in \mathcal{F} \qquad (9.4)$$

the random variable X is Gaussian with mean 0 and variance σ^2.

Moreover, since X can be written as $X = Y + \mu$ where Y is Gaussian with mean 0 and variance σ^2 under $\mathbb{P}$, we have that $\tilde{\mathbb{P}}$ can be written as:

$$\tilde{\mathbb{P}}(\mathcal{E}) = \mathbb{E}\left[1_{\mathcal{E}} e^{-\frac{\mu}{\sigma^2}Y - \frac{1}{2}\frac{\mu^2}{\sigma^2}}\right], \quad \mathcal{E} \in \mathcal{F} \qquad (9.5)$$

It is good to pause for a second and look at the signs in the exponential of equations (9.4) and (9.5). The signs in the exponential might be very confusing and is the source of many mistakes in the Cameron-Martin-Girsanov theorem. A good trick is to say that, if we want to remove μ, then the sign in front of X or Y must be negative. Then, we add the exponential factor needed for $\tilde{\mathbb{P}}$ to be a probability. This is given by the MGF of X or Y depending on how we want to express it.

The probabilities $\mathbb{P}$ and $\tilde{\mathbb{P}}$, as defined in the equation (9.4) are obviously not equal since they differ by a factor in (9.3). However, they share some similarities. Most notably, if $\mathcal{E}$ is an event of positive $\mathbb{P}$-probability, $\mathbb{P}(\mathcal{E}) > 0$, then we must have $\tilde{\mathbb{P}}(\mathcal{E}) > 0$, since the factor in (9.3) is always strictly positive. The converse is also true: if $\mathcal{E}$ is an event of positive $\tilde{\mathbb{P}}$-probability, $\tilde{\mathbb{P}}(\mathcal{E}) > 0$, then we must have that $\mathbb{P}(\mathcal{E}) > 0$. This is because the factor in (9.3) can be inverted, being strictly positive. More precisely, we have:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{E}[1_{\mathcal{E}}] \\
&= \mathbb{E}\left[1_{\mathcal{E}} \frac{e^{aX(\omega)}}{\mathbb{E}[e^{aX}]} \left(\frac{e^{aX(\omega)}}{\mathbb{E}[e^{aX}]}\right)^{-1}\right] \\
&= \tilde{\mathbb{E}}\left[1_{\mathcal{E}} \left(\frac{e^{aX(\omega)}}{\mathbb{E}[e^{aX}]}\right)^{-1}\right]
\end{aligned}$$

The factor $\left(\frac{e^{aX(\omega)}}{\mathbb{E}[e^{aX}]}\right)^{-1}$ is also strictly positive, proving the claim. To sum it all up, the probabilities $\mathbb{P}$ and $\tilde{\mathbb{P}}$ essentially share the same possible outcomes. Such probability measures are said to be equivalent measures.

Definition 9.1. Consider the two probabilities $\mathbb{P}$ and $\tilde{\mathbb{P}}$ on $(\Omega, \mathcal{F})$. They are said to be equivalent, if for any event $\mathcal{E} \in \mathcal{F}$, we have $\mathbb{P}(\mathcal{E}) > 0$ if and only if $\mathbb{P}(\mathcal{E}) > 0$. Thus, $\mathbb{P}$ and $\tilde{\mathbb{P}}$ agree on the null sets. If $A \in \mathcal{F}$ is such that $\mathbb{P}(A) = 0$, then $\tilde{\mathbb{P}}(A) = 0$ and vice-versa.

Keep in mind that two probabilities that are equivalent might still be very far from being equal!

9.2 The Cameron-Martin Theorem.

Theorem 9.2. *(Cameron-Martin Theorem for constant drift). Let $(\tilde{B(t)}, t \in [0,T])$ be a $\mathbb{P}-$Brownian motion with constant drift θ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Consider the probability $\tilde{\mathbb{P}}$on Ω given by:*

$$\tilde{\mathbb{P}}(\mathcal{E}) = \mathbb{E}\left[e^{-\theta \tilde{B}(T) + \frac{\theta^2}{2}T} 1_{\mathcal{E}}\right], \quad \mathcal{E} \in \mathcal{F} \tag{9.6}$$

Then, the process $(\tilde{B}(t), t \in [0,T])$ under $\tilde{\mathbb{P}}$is distributed like a standard brownian motion. Moreover, since we can write $\tilde{B}_t = \theta t + B_t$ for some standard brownian motion $(B_t, t \in [0,T])$ on $(\Omega, \mathcal{F}, \mathbb{P})$, the probability $\tilde{\mathbb{P}}$ can also be written as:

$$\tilde{\mathbb{P}}(\mathcal{E}) = \mathbb{E}\left[e^{-\theta B(T) - \frac{\theta^2}{2}T} 1_{\mathcal{E}}\right] \tag{9.7}$$

It is a good idea to pause again and look at the signs in the exponential in equations (9.6) and (9.7). They behave the same way as in theorem (9.1). There is a minus sign in front of B_T to remove the drift. Before proving the theorem, we make some important remarks.

(1) **The end-point**. Note that only the endpoint $\tilde{B}(T)$ of the Brownian motion is involved in the change of probability. In particular, T cannot be $+\infty$. The Cameron-Martin theorem can only be applied on a finite interval.

(2) **A martingale.** The factor $M_T = e^{-\theta B(T) - \frac{\theta^2}{2}T} = e^{-\theta \tilde{B}(T) + \frac{1}{2}\theta^2 T}$ involved in the change of probability is the end-point of a $\mathbb{P}-$martingale, that is, it is a martingale under the original probability $\mathbb{P}$. To see this:

$$\begin{aligned}
\mathbb{E}[M_T|\mathcal{F}_t] &= \mathbb{E}\left[e^{-\theta B(T)-\frac{1}{2}\theta^2 T}|\mathcal{F}_t\right] \\
&= e^{-\theta B(t)}\mathbb{E}\left[e^{-\theta(B(T)-B(t))}|\mathcal{F}_t\right]e^{-\frac{\theta^2}{2}T} \\
&\{\text{Using } B(T)-B(t) \perp \mathcal{F}_t\} \\
&= e^{-\theta B(t)}\mathbb{E}\left[e^{-\theta(B(T)-B(t))}\right]e^{-\frac{\theta^2}{2}T} \\
&= e^{-\theta B(t)}e^{\frac{\theta^2}{2}(T-t)}e^{-\frac{\theta^2}{2}T} \\
&= e^{-\theta B(t)-\frac{\theta^2}{2}t}
\end{aligned}$$

In fact, since $B(t)$ is a $\mathbb{P}$-standard Brownian motion, $M(t) = e^{-\theta B(t)-\frac{\theta^2}{2}t}$ is a geometric brownian motion.

Interestingly, the drift of $\tilde{B}(t)$ becomes the volatility factor in M_T! $\mathbb{E}[M_T^2] = \mathbb{E}[e^{-2\theta B(T)-\theta^2 T}] = e^{-\theta^2 T}\cdot\mathbb{E}[e^{-2\theta B(T)}] = e^{-\theta^2 T}\cdot e^{2\theta^2 T} = e^{\theta^2 T}$.

The fact that $M(t)$ is a martingale is very helpful in calculations. Indeed, suppose we want to compute the expectation of a function $F(\tilde{B}(s))$ of a Brownian motion with drift at time $s < T$. Then, we have by theorem (9.2):

$$\begin{aligned}
\mathbb{E}[F(\tilde{B}(s))] &= \mathbb{E}[M_T M_T^{-1} F(\tilde{B}(s))] \\
&= \tilde{\mathbb{E}}[M_T^{-1}F(\tilde{B}(s))] \\
&= \tilde{\mathbb{E}}[e^{\theta\tilde{B}(T)-\frac{1}{2}\theta^2 T}F(\tilde{B}(s))]
\end{aligned}$$

Now, we know that under $\tilde{\mathbb{P}}$probability, $(\tilde{B}(t), t\in[0,T])$ is a standard brownian motion, or $\tilde{\mathbb{P}}$-standard brownian motion for short. Therefore, the process $e^{\theta\tilde{B}(t)-\frac{1}{2}\theta^2 t}$ is a martingale under the new probability measure $\tilde{\mathbb{P}}$, or a $\tilde{\mathbb{P}}$-martingale for short. By conditioning over $\mathcal{F}_s$ and applying the martingale property, we get:

$$\begin{aligned}
\mathbb{E}\left[F(\tilde{B}_s)\right] &= \tilde{\mathbb{E}}[e^{\theta\tilde{B}(T)-\frac{1}{2}\theta^2 T}F(\tilde{B}(s))] \\
&= \tilde{\mathbb{E}}[\tilde{\mathbb{E}}[e^{\theta\tilde{B}(T)-\frac{1}{2}\theta^2 T}F(\tilde{B}(s))|\mathcal{F}_s]] \\
&= \tilde{\mathbb{E}}[e^{\theta\tilde{B}(s)-\frac{1}{2}\theta^2 s}F(\tilde{B}(s))] \\
&= \mathbb{E}[e^{\theta B(s)-\frac{1}{2}\theta^2 s}F(B(s))]
\end{aligned}$$

The last equality may seem wrong as removed all the tildes. It is not! It holds because $(\tilde{B}(t))$ under $\tilde{\mathbb{P}}$ has the same distribution as $(B(t))$ under $\mathbb{P}$: a standard brownian motion. Of course, it would be possible to directly evaluate $\mathbb{E}[F(\tilde{B}(s))]$ here as we know the distribution of a Brownian motion with drift. However, when the function will involve more than one point (such as the maximum of the path), the Cameron-Martin theorem is a powerful tool to evaluate expectations.

(3) **The paths with or without the drift are the same.** Let $(B(t), t \leq T)$ be a standard Brownian motion defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Heuristically, it is fruitful to think of the sample space of Ω as the different continuous paths of Brownian motion. Since, the change of probability from $\mathbb{P}$ to $\tilde{\mathbb{P}}$simply changes the relative weights of the paths (and this change of weight is never zero, similarly to equation (9.3) for a single random variable), the theorem suggests that the paths of a standard Brownian motion and those of a Brownian motion with a constant drift θ (with volatility 1) are essentially the same.

The form of the factor $M_T = e^{-\theta\tilde{B}_T+\theta^2 T}$ can be easily understood at the heuristic level. For each outcome ω, it is proportional to $e^{-\theta\tilde{B}_T(\omega)}$ (The term $e^{(\theta^2/2)T}$ is simply to ensure that $\mathbb{P}(\Omega) = 1$) Therefore, the factor M_T penalizes the paths for which $\tilde{B}_T(\omega)$ is large and positive (if $\theta > 0$). In particular, it is conceivable that the Brownian motion with positive drift is reduced to standard Brownian motion under the new probability.

(4) **Changing the volatility.** What about the volatility? Is it possible to change the probability $\mathbb{P}$ to $\tilde{\mathbb{P}}$ in such a way that the Brownian motion under $\mathbb{P}$ has volatility $\sigma \neq 1$ under $\tilde{\mathbb{P}}$? The answer is no! The paths of the Brownian motions with different volatilities are inherently different. Indeed, it suffices to compute the quadratic variation. If $(B_t : t \in [0,T])$ has volatility 1 and $(\tilde{B}_t, t \in [0,T])$ has volatility 2. then the following convergence holds for ω in a set of probability one (for a partition fine enough, say $t_{j+1} - t_j = 2^{-n}$. Then $B_t = \int 1 \cdot dB_t$ and $\tilde{B}_t = \int 2 \cdot dB_t$

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} (B_{t_{j+1}}(\omega) - B_{t_j}(\omega))^2 = \int_0^T 1^2 \cdot ds = T$$

$$\lim_{n\to\infty} \sum_{j=0}^{n-1} (\tilde{B}_{t_{j+1}}(\omega) - \tilde{B}_{t_j}(\omega))^2 = \int_0^T 2^2 \cdot ds = 4T$$

In other words, the distribution of the standard brownian motion on $[0,T]$ is supported on paths whose quadratic variation is T, whereas the distribution of $(\tilde{B}_t, t \geq 0)$ is supported on paths where the quadratic variation is $4T$. These paths are very different. We conclude that the distributions of the two processes are not equivalent. Hence, a change of probability from $\mathbb{P}$ to $\tilde{\mathbb{P}}$ is not possible. In fact, we say that they are mutually singular, meaning the set of paths on which they are supported are disjoint.

Proof.

Let $(\tilde{B}_t : t \in [0,T])$ be a Brownian motion with constant drift θ defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Thus, $\tilde{B}_t = \theta t + B_t$.

Claim. $\tilde{B}_t$ is a $\tilde{\mathbb{P}}$-martingale.

Let

$$M_t = f(t, B_t) = \exp(-\theta B_t - (\theta^2/2)t)$$

So:

$$\begin{aligned} dM_t &= -\frac{\theta^2}{2} M_t dt - \theta M_t dB_t + \frac{1}{2}\theta^2 M(t) dt \\ &= -\theta M_t dB_t \end{aligned}$$

Consider the product $(M_t \tilde{B}_t)$. We have:

$$\begin{aligned} d(M_t \tilde{B}_t) &= \tilde{B}_t dM_t + M_t d\tilde{B}_t + dM_t \cdot d\tilde{B}_t \\ &= -\tilde{B}_t \theta M_t dB_t + M_t(\theta dt + dB_t) - \theta M_t dB_t(\theta dt + dB_t) \\ &= -\tilde{B}_t \theta M_t dB_t + \theta M_t dt + M_t dB_t - \theta M_t dt \\ &= (-\tilde{B}_t \theta + 1) M_t dB_t \end{aligned}$$

Thus, by the properties of Ito integral,$M_t \tilde{B}_t$ is a martingale under $\mathbb{P}$. By the abstract Bayes formula (9.4):

$$\begin{aligned} \tilde{\mathbb{E}}[\tilde{B}_t | \mathcal{F}_s] &= \frac{1}{M_s} \mathbb{E}[M_t \tilde{B}_t | \mathcal{F}_s] \\ &= \frac{1}{M_s} \cdot M_s \tilde{B}_s \\ &= \tilde{B}_s \end{aligned}$$

Thus, $\tilde{B}_t$ is a $\tilde{\mathbb{P}}$-martingale.

Claim. Our claim is that under the $\tilde{\mathbb{P}}$ measure, $\tilde{B}_t \sim \mathcal{N}^{\tilde{\mathbb{P}}}(0,t)$ and to do this we rely on the the moment-generating function.

By definition, for a constant Ψ:

$$\begin{aligned}
M_{\tilde{B}_t}(\Psi) &= \tilde{\mathbb{E}}\left[\exp\left(\Psi\tilde{B}_t\right)\right] \\
&= \mathbb{E}\left[M_T \exp\left(\Psi\tilde{B}_t\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\theta\tilde{B}_T + \frac{\theta^2}{2}T + \Psi\tilde{B}_t\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\theta(\theta T + B_T) + \frac{\theta^2}{2}T + \Psi(\theta t + B_t)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\theta B_T - \frac{\theta^2}{2}T + \Psi\theta t + \Psi B_t)\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\theta(B_T - B_t) - \frac{\theta^2}{2}T + \Psi\theta t + (\Psi-\theta)B_t)\right)\right] \\
&= \exp\left(-\frac{\theta^2}{2}T + \Psi\theta t\right)\mathbb{E}\left(-\theta(B_T - B_t)\right)\mathbb{E}\left((\Psi-\theta)B_t\right) \\
&= \exp\left(-\frac{\theta^2}{2}T + \Psi\theta t\right)\exp\left[\frac{1}{2}\theta^2(T-t)\right]\exp\left[\frac{1}{2}(\Psi-\theta)^2 t\right] \\
&= \exp\left[-\frac{1}{2}\left(\theta^2 - 2\Psi\theta - (\Psi-\theta)^2\right)t\right] \\
&= \exp\left[-\frac{1}{2}\left(\theta^2 - 2\Psi\theta - (\Psi^2 - 2\Psi\theta + \theta^2\right)t\right] \\
&= \exp(-\Psi^2 t)
\end{aligned}$$

Thus, $\tilde{B}_t \sim \mathcal{N}^{\tilde{\mathbb{P}}}(0,t)$.

Claim. Finally, to show that $\tilde{B}_t$ is indeed a $\tilde{\mathbb{P}}-$standard brownian motion, we have the following:

(a) $\tilde{B}_0 = \theta(0) + B_0 = 0$ and $\tilde{B}_t$ has almost surely continuous paths.

(b) We would like to prove that, for $s < t$, $\tilde{B}_t - \tilde{B}_s \sim \mathcal{N}^{\tilde{\mathbb{P}}}(0, t-s)$. We have:

$$\begin{aligned}
\mathbb{E}[\tilde{B}_t - \tilde{B}_s] &= \tilde{\mathbb{E}}[\tilde{B}_t] - \tilde{\mathbb{E}}[B_s] \\
&= 0
\end{aligned}$$

And,

$$\begin{aligned}
\tilde{\mathbb{E}}[(\tilde{B}_t - \tilde{B}_s)^2] &= \tilde{\mathbb{E}}[\tilde{B}_t^2 - 2\tilde{B}_t\tilde{B}_s + \tilde{B}_s^2] \\
&= \tilde{\mathbb{E}}[B_t^2] - 2\tilde{\mathbb{E}}[\tilde{B}_t\tilde{B}_s] + \tilde{\mathbb{E}}[\tilde{B}_s^2] \\
&= t + s - 2\tilde{\mathbb{E}}[\tilde{B}_t\tilde{B}_s]
\end{aligned}$$

(c) The non-overlapping increments of a $\tilde{\mathbb{P}}$-martingale are independent. To see this, suppose $t_1 \leq t_2 \leq t_3$:

$$\begin{aligned}\tilde{\mathbb{E}}[(B_{t_3}-B_{t_2})(B_{t_2}-B_{t_1})] &= \tilde{\mathbb{E}}[\tilde{\mathbb{E}}[(B_{t_3}-B_{t_2})(B_{t_2}-B_{t_1})|\mathcal{F}_{t_2}]] \\ &= \tilde{\mathbb{E}}[(B_{t_2}-B_{t_1})\tilde{\mathbb{E}}[(B_{t_3}-B_{t_2})|\mathcal{F}_{t_2}]] \\ &= \tilde{\mathbb{E}}[(B_{t_2}-B_{t_1})(B_{t_2}-B_{t_2})]] = 0\end{aligned}$$

Also, the covariance

$$\begin{aligned}\tilde{\mathbb{E}}[\tilde{B}_t\tilde{B}_s] &= \tilde{\mathbb{E}}[(\tilde{B}_t-\tilde{B}_s)\tilde{B}_s] + \tilde{\mathbb{E}}[\tilde{B}_s^2] \\ &= 0 + s\end{aligned}$$

So, $\mathbb{E}[(\tilde{B}_t - \tilde{B}_s)^2] = t + s - 2s = t - s$.

Consequently, $\tilde{B}_t$ is a $\tilde{\mathbb{P}}$-standard brownian motion.

Example 9.3. (Bachelier's formula for Brownian motion with a drift) One of the most interesting formulas we have seen so far is Bachelier's formula for the maximum of the Brownian motion in proposition

9.3 Radon-Nikodym Theorem.

Theorem 9.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathbb{Q}$ be another probability measure. Under the assumption that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, that is, $\mathbb{Q}(A) = 0 \iff \mathbb{P}(A) = 0$, there exists a non-negative random variable Z such that:*

$$Z := \frac{d\mathbb{Q}}{d\mathbb{P}}$$

and we call Z the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$.

At a heuristic level, as long as $\mathbb{P}$ and $\mathbb{Q}$ agree on the possible events (and null sets), we can define a likelihood ratio of the little probability elements $d\mathbb{Q}(\omega)$ and $d\mathbb{P}(\omega)$. This is an almost surely non-negative random variable with expectation 1.

It follows that:

$$\begin{aligned}\mathbb{Q}(\mathcal{E}) &= \mathbb{E}^{\mathbb{Q}}[1_{\mathcal{E}}] \\ &= \int_\Omega 1_{\{\omega\in\mathcal{E}\}} d\mathbb{Q}(\omega) \\ &= \int_\Omega 1_{\{\omega\in\mathcal{E}\}} Z d\mathbb{P}(\omega) \\ &= \mathbb{E}^{\mathbb{P}}[Z 1_{\mathcal{E}}]\end{aligned}$$

Since $\mathbb{Q}$ is a probability measure, $\mathbb{Q}(\Omega) = \mathbb{E}^{\mathbb{P}}[Z \cdot 1_\Omega] = \mathbb{E}^{\mathbb{P}}[Z] = 1$.

and

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}[X] &= \mathbb{E}^{\mathbb{P}}[ZX] \\ \mathbb{E}^{\mathbb{P}}[X] &= \mathbb{E}^{\mathbb{Q}}[\frac{1}{Z}X]\end{aligned}$$

Definition 9.2. *(Density Process).* We can define the Radon-Nikodym derivative(likelihood ratio) process:

$$Z(t) = \left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_t = \mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_t]$$

Then, $Z(t)$ is a $\mathbb{P}-$martingale with $Z(0) = \mathbb{E}^{\mathbb{P}}[Z(t)] = 1$. To see this:

$$\begin{aligned}
\mathbb{E}^{\mathbb{P}}[Z(t)|\mathcal{F}_s] &= \mathbb{E}^{\mathbb{P}}[\mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_t]|\mathcal{F}_s] \\
&= \mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_s] \\
&\{\text{Tower law}\} \\
&= Z(s)
\end{aligned}$$

Theorem 9.4. *(Abstract Bayes' Formula) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathbb{Q}$ be any other probability measure on it and suppose that $\mathbb{Q} << \mathbb{P}$. By the Radon-Nikodym theorem, $\exists Z = d\mathbb{Q}/d\mathbb{P}$, $Z \geq 0$ a.s. with $\mathbb{E}^{\mathbb{P}}[Z] = 1$. Then, we have:*

$$\mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t] = \frac{\mathbb{E}^{\mathbb{P}}[ZX|\mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_t]}$$

Proof. We use the definition of conditional expectations. Our claim is that for all $A \in \mathcal{F}_t$,

$$\int_A \mathbb{E}^{\mathbb{P}}[ZX|\mathcal{F}_t]d\mathbb{P} = \int_A \mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_t]\mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t]d\mathbb{P}$$

For the left side:

$$\begin{aligned}
\int_A \mathbb{E}^{\mathbb{P}}[ZX|\mathcal{F}_t]d\mathbb{P} &= \int_A ZX d\mathbb{P} \\
&\{\text{Definition of conditional expectations}\} \\
&= \int_A X d\mathbb{Q} \\
&\{\text{Radon-Nikodym Derivative}\}
\end{aligned}$$

For the right side:

$$\begin{aligned}
\int_A \mathbb{E}^{\mathbb{P}}[Z|\mathcal{F}_t]\mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t]d\mathbb{P} &= \int_A \mathbb{E}^{\mathbb{P}}[Z\mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t]|\mathcal{F}_t]d\mathbb{P} \\
&\{\text{Since } \mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t] \text{ is } \mathcal{F}_t - \text{measurable}\} \\
&= \int_A Z\mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t]d\mathbb{P} \\
&\{\text{Definition of conditional expectations}\} \\
&= \int_A \mathbb{E}^{\mathbb{Q}}[X|\mathcal{F}_t]d\mathbb{Q} \\
&\{\text{Radon-Nikodym Derivative}\} \\
&= \int_A X d\mathbb{Q}
\end{aligned}$$

□

9.4 Change of Measure for processes.

Theorem 9.5. *(Change of Measure). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $M(t)$ be any density process. Let $X(t)$ be any $\mathcal{F}_t$-measurable random variable. Then:*

$$\mathbb{E}^{\mathbb{Q}}[X(T)|\mathcal{F}_t] = \mathbb{E}^{\mathbb{P}}\left[\frac{M(T)}{M(t)}X(T)|\mathcal{F}_t\right]$$

Proof. Recall that $M = d\mathbb{Q}/d\mathbb{P}$. By the conditional Bayes' formula:

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}[X(T)|\mathcal{F}_t] &= \frac{\mathbb{E}^{\mathbb{P}}[MX(T)|\mathcal{F}_t]}{\mathbb{E}^{\mathbb{P}}[M|\mathcal{F}_t]} \\
&= \frac{\mathbb{E}^{\mathbb{P}}[\mathbb{E}[MX(T)|\mathcal{F}_T]|\mathcal{F}_t]}{M(t)} \\
&\{\text{Tower law; definition of density process} M(t)\} \\
&= \frac{\mathbb{E}^{\mathbb{P}}[X(T)\mathbb{E}[M|\mathcal{F}_T]|\mathcal{F}_t]}{M(t)} \\
&\{\text{Taking out what is known}\} \\
&= \frac{\mathbb{E}^{\mathbb{P}}[X(T)M(T)|\mathcal{F}_t]}{M(t)} \\
&= \frac{\mathbb{E}^{\mathbb{P}}[X(T)M(T)|\mathcal{F}_t]}{M(t)} \\
&= \mathbb{E}^{\mathbb{P}}\left[\frac{M(T)}{M(t)}X(T)|\mathcal{F}_t\right]
\end{aligned}$$

□

A density process may be used to artificially construct a new measure. Let $M(t)$ be any $\mathbb{P}$-martingale with $M(0) = 1$. We choose a final horizon time T and define the Radon-Nikodym derivative as $Z = M(T)$. The corresponding measure:

$$\mathbb{Q}(A) = \mathbb{E}^{\mathbb{Q}}[1_A] = \mathbb{E}^{\mathbb{P}}[M(T)1_A]$$

9.5 Black-Scholes Merton Option Pricing Formulae.

Let $(W(t), t \in [0,T])$ be a Brownian motion on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The Black-Scholes model consists of two assets (i) a stock and (ii) a risk-free bank account with dynamics as follows:

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t)$$

Let $f(x) = \ln x$. Then, by Ito's lemma:

$$
\begin{aligned}
df(x) &= f_x(x)dx + \frac{1}{2}f_{xx}(x)dx \cdot dx \\
df(S(t)) &= \frac{1}{S(t)}dS(t) - \frac{1}{2}\frac{1}{S(t)^2}dS(t)dS(t) \\
&= \mu dt + \sigma dW(t) - \frac{1}{2}\frac{1}{S(t)^2}\sigma^2 S^2(t)dt \\
d(\ln S(t)) &= \left(\mu - \frac{1}{2}\sigma^2\right)dt + \sigma dW(t) \\
\ln\left(\frac{S(t)}{S(0)}\right) &= \int_0^t \left(\mu - \frac{1}{2}\sigma^2\right)dt + \int_0^t \sigma dW(t) \\
S(t) &= S(0)\exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)r + \sigma W(t)\right]
\end{aligned}
$$

The dynamics of the locally risk-free bank account are:

$$
dB(t) = rB(t)dt
$$

The dynamics of the discounted stock price process are:

$$
\begin{aligned}
d(e^{-rt}S(t)) &= d(e^{-rt})S(t) + e^{-rt}dS(t) + d(e^{-rt})dS(t) \\
&= -re^{-rt}dtS(t) + e^{-rt}(\mu S(t)dt + \sigma S(t)dW(t)) \\
&= e^{-rt}S(t)(\mu - r)dt + e^{-rt}\sigma S(t)dW(t) \\
&= e^{-rt}\sigma S(t)\left\{\frac{(\mu - r)}{\sigma}dt + dW(t)\right\} \\
&= e^{-rt}\sigma S(t)(\theta dt + dW(t))
\end{aligned}
$$

Define

$$
\left.\frac{d\mathbb{Q}}{d\mathbb{P}}\right|_t = e^{-\theta W(t) - \frac{\theta^2}{2}t}
$$

and

$$
\begin{aligned}
W^{\mathbb{Q}}(t) &= W(t) + \theta t \\
dW^{\mathbb{Q}}(t) &= dW(t) + \theta dt
\end{aligned}
$$

Then, by the Girsanov theorem, $W^{\mathbb{Q}}(t)$ is a $\mathbb{Q}$-standard brownian motion. $W^{\mathbb{Q}}(t) \sim \mathcal{N}^{\mathbb{Q}}(0, t)$. Under $\mathbb{Q}$, the dynamics of the discounted stock price is:

$$
d(e^{-rt}S(t)) = e^{-rt}\sigma S(t)dW^{\mathbb{Q}}(t)
$$

The measure $\mathbb{Q}$ is said to be *risk-neutral* because it is equivalent to $\mathbb{P}$ ($\mathbb{Q}(A) = 0 \Longleftrightarrow \mathbb{P}(A) = 0$; they agree on null sets) and in addition it renders the discounted stock price into a martingale. Indeed:

$$e^{-rt}S(t) = S(0) + \int_0^t e^{-rt}\sigma S(u)dW^{\mathbb{Q}}(u)$$

and the process $\int_0^t e^{-rt}\sigma S(u)dW^{\mathbb{Q}}(u)$ is an Ito-integral and therefore a $\mathbb{Q}$-martingale.

The undiscounted stock price process $(S(t), t \in [0,T)$ is described by the $\mathbb{Q}$-dynamics:

$$\begin{aligned}
dS(t) &= \mu S(t)dt + \sigma S(t)(dW^{\mathbb{Q}}(t) - \theta dt) \\
&= \mu S(t)dt - \sigma S(t) \cdot \frac{\mu - r}{\sigma}dt + \sigma S(t)dW^{\mathbb{Q}}(t) \\
&= \mu S(t)dt - \mu S(t)dt + rS(t)dt + \sigma S(t)dW^{\mathbb{Q}}(t) \\
&= rS(t)dt + \sigma S(t)dW^{\mathbb{Q}}(t)
\end{aligned}$$

and

$$S(t) = S(0)\exp\left[(r - \sigma^2/2)t + \sigma W^{\mathbb{Q}}(t)\right]$$

By the risk-neutral pricing formula, the price of a derivative security with payoff $V(T)$ is:

$$V(t) = \mathbb{E}^{\mathbb{Q}}[e^{-\int_t^T r(u)du}V(T)|\mathcal{F}_t]$$

We have $r(u) \equiv r$. And further, for a european call option: $V(T) = ((S(T) - K) \cdot 1_{\{S_T > K\}}$. Thus:

$$\begin{aligned}
V(t) &= e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[(S(T) - K) \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t] \\
&= e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[S(T) \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t] - e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[K \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t] \\
&= e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[S(T) \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t] - Ke^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[1_{\{S(T)>K\}}|\mathcal{F}_t]
\end{aligned}$$

The second expectation is easily solved. We have:

$$\mathbb{E}^{\mathbb{Q}}[1_{\{S(T)>K}|\mathcal{F}_t] = \mathbb{Q}\{S(T) > K|\mathcal{F}_t]$$

Using the fact, that

$$\begin{aligned}
\mathbb{Q}\{S(T) > K|\mathcal{F}_t\} &= \mathbb{Q}\{\log S(T) > \log K|\mathcal{F}_t\} \\
&= \mathbb{Q}\{\log S(t) + \left(r - \frac{\sigma^2}{2}\right)(T-t) + \sigma(W^{\mathbb{Q}}(T) - W^{\mathbb{Q}}(t)) > \log K\}
\end{aligned}$$

Let Z be a standard normal random variable following $\mathcal{N}^{\mathbb{Q}}(0,1)$. Then, $W^{\mathbb{Q}}(T) - W^{\mathbb{Q}}(t) = \sqrt{T-t}Z$:

$$
\begin{aligned}
\mathbb{Q}\{S(T) > K|\mathcal{F}_t\} &= \mathbb{Q}\{\log S(t) + \left(r - \frac{\sigma^2}{2}\right)(T-t) + \sigma(W^{\mathbb{Q}}(T) - W^{\mathbb{Q}}(t)) > \log K\} \\
&= \mathbb{Q}\{\log S(t) + \left(r - \frac{\sigma^2}{2}\right)(T-t) + \sigma\sqrt{T-t}Z > \log K\} \\
&= \mathbb{Q}\{Z > \frac{\log K - \log S(t) - \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\} \\
&= \mathbb{Q}\{Z \leq \frac{\log \frac{S(t)}{K} + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\} \\
&= \Phi(d_-(\tau, S(t))), \quad \tau = T - t
\end{aligned}
$$

The first expectation is typically solved using a change of numeraire.

Let $\tilde{\mathbb{Q}}$ be another probability measure related to $\mathbb{Q}$ defined by the Radon-Nikodym derivative:

$$
M = \frac{d\tilde{\mathbb{Q}}}{d\mathbb{Q}} := \frac{S(T)e^{-rT}}{S(0)} = \exp\left[-\frac{\sigma^2}{2}T + \sigma W^{\mathbb{Q}}(T)\right]
$$

and correspondingly, let us define the Radon-Nikodym derivative process $(M(t), t \in [0, T])$ as:

$$
M(t) = \left.\frac{d\tilde{\mathbb{Q}}}{d\mathbb{Q}}\right|_t = \mathbb{E}^{\mathbb{Q}}[M|\mathcal{F}_t]
$$

Clearly, $M(T)$ is a non-negative random variable.

We note that:

$$
\begin{aligned}
\mathbb{E}^{\mathbb{Q}}[M(T)] &= \frac{\mathbb{E}^{\mathbb{Q}}[S(T)e^{-rT}]}{S(0)} \\
&= \frac{S(0)}{S(0)} = 1 \\
&\quad \{\text{Discount stock price is a martingale under } \mathbb{Q}\}
\end{aligned}
$$

Further, $M(t)$ is an exponential $\mathbb{Q}$-martingale. So:

$$
M(t) = \frac{S(t)e^{-rt}}{S(0)}
$$

By the change-of-measure theorem, the first expectation can be expressed as follows.

$$
\begin{aligned}
\mathbb{E}^{\tilde{\mathbb{Q}}}[1_{\{S(T)>K\}}|\mathcal{F}_t] &= \mathbb{E}^{\mathbb{Q}}\left[\frac{S(T)e^{-rT}}{S(t)e^{-rt}} \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t\right] \\
S(t)\mathbb{E}^{\tilde{\mathbb{Q}}}[1_{\{S(T)>K\}}|\mathcal{F}_t] &= \mathbb{E}^{\mathbb{Q}}\left[S(T)e^{-r(T-t)} \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t\right] \\
\implies \mathbb{E}^{\mathbb{Q}}\left[S(T)e^{-r(T-t)} \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t\right] &= S(t)\tilde{\mathbb{Q}}\{S(T) > K|\mathcal{F}_t\}
\end{aligned}
$$

So, the value of a European call option can be written as:

$$V(t) = \mathbb{E}^{\tilde{\mathbb{Q}}}[S(t) \cdot 1_{\{S(T)>K\}}|\mathcal{F}_t] - e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}[K \cdot 1_{\{S(T)>K\}}]$$

or equivalently:

$$V(t) = S(t)\tilde{\mathbb{Q}}\{S(T) > K|\mathcal{F}_t\} - Ke^{-r(T-t)}\mathbb{Q}\{S(T) > K|\mathcal{F}_t\}$$

Finally:

$$\begin{aligned}
\tilde{\mathbb{Q}}\{S(T) > K|\mathcal{F}_t\} &= \mathbb{E}^{\mathbb{Q}}\left[\frac{M(T)}{M(t)}1_{\{S(T)>K\}}|\mathcal{F}_t\right] \\
&= \mathbb{E}^{\mathbb{Q}}\left[e^{-\frac{\sigma^2}{2}(T-t)+\sigma(W^{\mathbb{Q}}(T)-W^{\mathbb{Q}}(t))}1_{\{S(T)>K\}}|\mathcal{F}_t\right]
\end{aligned}$$

Define $Y := \mathcal{N}^{\mathbb{Q}}(0,1)$. Then, $W^{\mathbb{Q}}(T) - W^{\mathbb{Q}}(t) = \sqrt{(T-t)}Y$. Now, it is easy to see that the event $\{S_T > K\}$ is the same as $\{Y < d_-(\tau, S(t))\}$. Thus:

$$\begin{aligned}
\tilde{\mathbb{Q}}\{S(T) > K|\mathcal{F}_t\} &= \int_{-\infty}^{d_-(\tau,S(t))} \exp\left(-\frac{\sigma^2}{2}\tau + \sigma\sqrt{\tau}y\right) f_Y^{\mathbb{Q}}(y)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_-(\tau,S(t))} \exp\left(-\frac{\sigma^2}{2}\tau + \sigma\sqrt{\tau}y\right) \exp\left(-\frac{y^2}{2}\right) dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_-(\tau,S(t))} \exp\left[-\frac{1}{2}\left(\sigma^2\tau + 2\sigma\sqrt{\tau}y + y^2\right)\right] dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_-(\tau,S(t))} \exp\left[-\frac{1}{2}\left(y + \sigma\sqrt{\tau}\right)^2\right] dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_-(\tau,S(t))+\sigma\sqrt{\tau}} \exp\left[-\frac{1}{2}z^2\right] dz
\end{aligned}$$

Let $d_+(\tau, S(t)) = d_-(\tau, S(t)) + \sigma\sqrt{\tau}$. Then:

$$\tilde{\mathbb{Q}}\{S(T) > K|\mathcal{F}_t\} = \Phi(d_+(\tau, S(t))$$