Martingale theory lecture notes

Márton Balázs*
Some proofs have been transferred over from *Probability 3 revision notes* by Aaron Smith.

August 11, 2025

These notes summarise the lectures and exercise classes *Martingale Theory with Applications* since Autumn 2021 in the University of Bristol.

Contents

---

*School of Mathematics, University of Bristol

1 A quick summary of some parts of measure theory

...from the probabilist point of view. This section mostly follows Shiryaev [1]. The aim is to build up a mathematical model of random experiments and measurements (random variables, that is) thereof. Probabilities of these random outcomes are also to be constructed.

The *power set* $\mathcal{P}(\Omega)$ of a set Ω is the set of all subsets of Ω, and $A^{\mathrm{c}} := \Omega - A$ denotes the complement of a set $A \subseteq \Omega$.

Definition 1.1. A family $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ of sets is called an *algebra*, if

- $\Omega \in \mathcal{A}$,

- for every A, $B \in \mathcal{A}$, $A \cup B \in \mathcal{A}$,

- for every $A \in \mathcal{A}$, $A^{\mathrm{c}} \in \mathcal{A}$.

This simple construction allows us to build a prototype of probabilities as follows.

Definition 1.2. Let $\mathcal{A}$ be an algebra. A set function $\mu : \mathcal{A} \to [0, \infty]$ is a *finitely additive measure* on $\mathcal{A}$, if for all disjoint A, $B \in \mathcal{A}$,
$$\mu(A \cup B) = \mu(A) + \mu(B).$$

As it turns out, the objects we defined so far are too general for our purposes, hence a refinement comes next.

Definition 1.3. A family $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ of sets is called a *σ-algebra*, if

- $\Omega \in \mathcal{F}$,

- for every countably many sets A_1, A_2, $\cdots \in \mathcal{F}$, $\bigcup_n A_n \in \mathcal{F}$,

- for every $A \in \mathcal{F}$, $A^{\mathrm{c}} \in \mathcal{F}$.

In this case the pair $(\Omega, \mathcal{F})$ is called a *measurable space*. Any set $A \in \mathcal{F}$ is said to be *$\mathcal{F}$-measurable*, or just *measurable*.

The novelty here is the requirement of $\mathcal{F}$ being closed for countably infinite unions, as opposed to finite unions only for an algebra.

Example 1.4. Here are examples of σ-algebras (check!).

- $\mathcal{F}_* = \{\emptyset, \Omega\}$ is called the *trivial σ-algebra*.

- If Ω is countable, very often $\mathcal{F}^* = \mathcal{P}(\Omega)$ is considered, which is a σ-algebra in this case.

- For a set $A \subset \Omega$, $\mathcal{F}_A = \{\emptyset, A, A^{\mathrm{c}}, \Omega\}$ is the *σ-algebra generated by A*.

Definition 1.5. A *measure* μ on an algebra $\mathcal{A}$ is a set function $\mu \colon \mathcal{A} \to [0, \infty]$ such that for any mutually disjoint sets A_1, A_2, $\cdots \in \mathcal{A}$ with $\bigcup_n A_n \in \mathcal{A}$,

$$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n) \tag{1.1}$$

holds. If $\mu(\Omega) = 1$, then we call μ a *probability measure*, and often use $\mathbb{P}$ instead. In this case the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*.

Notice that a σ-algebra is automatically an algebra, thus this definition applies for measures on σ-algebras. Property (1.1) is referred to as *σ-additivity*.

When modeling a random experiment, the set Ω is called the *sample space* of the experiment, it contains all *elementary outcomes*. Its measurable subsets $A \in \mathcal{F}$ are called *events*. These are exactly those sets of outcomes which have a probability. The empty set $\emptyset$ is always an event, called the *null event*. The above definitions imply that $\mathbb{P}(\emptyset) = 0$.

Here is a rather useful characterisation of probability measures.

Theorem 1.6. *Let $\mathbb{P}$ be a finitely additive measure on an algebra $\mathcal{A}$, and assume $\mathbb{P}(\Omega) = 1$. Then the following are equivalent.*

(a) $\mathbb{P}$ is a probability measure.

(b) If $(A_n)_{n\geq 1}$ is an increasing sequence of sets in $\mathcal{A}$ (that is, $A_n \subseteq A_{n+1}$ for $n \geq 1$) and $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$, then the limit below exists, and

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big).$$

(c) If $(A_n)_{n\geq 1}$ is a decreasing sequence of sets in $\mathcal{A}$ (that is, $A_n \supseteq A_{n+1}$ for $n \geq 1$) and $\bigcap_{n=1}^{\infty} A_n \in \mathcal{A}$, then the limit below exists, and

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

(d) If $(A_n)_{n\geq 1}$ is a decreasing sequence of sets in $\mathcal{A}$ and $\bigcap_{n=1}^{\infty} A_n = \emptyset$, then the limit below exists, and

$$\lim_{n\to\infty} \mathbb{P}(A_n) = 0.$$

Notice that the union of increasing sets, and the intersection of decreasing sets, are sometimes called the *limit* of the sets.

Proof. We first show that (a) implies (b). Let $B_1 = A_1$, $B_n = A_n \backslash A_{n-1}$ for all $n \geq 2$. Then $\{B_n\}_{n=1}^{\infty}$ satisfies $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} A_n$ and $B_i \cap B_j = \emptyset$ for all $i \neq j$. So $\{B_n\}_{n=1}^{\infty}$ forms a disjoint partition of $\bigcup_{n=1}^{\infty} A_n$. By the σ-additivity assumption in (a) we have that

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbb{P}(\bigcup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mathbb{P}(B_n) = \lim_{n\to\infty}\sum_{k=1}^{n} \mathbb{P}(B_k) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{k=1}^{n} B_k) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Now we show (b) implies (c). Since A_n is a decreasing sequence, $\Omega \backslash A_n$ is an increasing sequence. Moreover,

$$\mathbb{P}(\bigcap_{n=1}^{\infty} A_n) = 1 - \mathbb{P}(\bigcup_{n=1}^{\infty} \Omega\backslash A_n) = 1 - \lim_{n\to\infty} \mathbb{P}(\Omega\backslash A_n) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(c) implies (d) trivially, it is simply a special case. It remains to show that (d) implies (a), i.e., that if A_n is a decreasing sequence with $\bigcap_{n=1}^{\infty} A_n = \lim_{n\to\infty} A_n = \emptyset$ then $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$ implies $\mathbb{P}$ has the σ-additivity property.

Take any disjoint family of sets $\{A_k\}_{k=1}^{\infty}$. Then by finite additivity,

$$\sum_{k=1}^{\infty} \mathbb{P}(A_k) = \lim_{n\to\infty}\sum_{k=1}^{n} \mathbb{P}(A_k) = \lim_{n\to\infty} \mathbb{P}(\bigcup_{k=1}^{n} A_k) = \lim_{n\to\infty}[\mathbb{P}(\bigcup_{k=1}^{\infty} A_k) - \mathbb{P}(\bigcup_{k=n+1}^{\infty} A_k)].$$

Now $\bigcup_{k=n+1}^{\infty} A_k$ is a decreasing sequence in n where $\bigcap_{n=1}^{\infty} \bigcup_{k=n+1}^{\infty} A_k = \emptyset$. This is because if $\omega \in \bigcup_{k=n+1}^{\infty} A_k$ then $\omega \in A_N$ for unique N (by disjointness of the family). Hence ω is not in the intersection of all tail unions. By (c), $\mathbb{P}(\bigcup_{k=n+1}^{\infty} A_k) \to 0$. Moreover, σ-additivity holds. $\qquad\square$

Remark 1.7. *If A_n is an increasing sequence then $\mathbb{P}(A_n)$ is an increasing sequence. Indeed this is because $A_n \subseteq A_{n+1}$ implies that $A_{n+1} = A_{n+1}\backslash A_n \cup A_n$. By the σ-additivity of $\mathbb{P}$ we have that $\mathbb{P}(A_{n+1}) = \mathbb{P}(A_{n+1}\backslash A_n) + \mathbb{P}(A_n) \geq \mathbb{P}(A_n)$. Moreover $\mathbb{P}(A_n) \nearrow \mathbb{P}(\lim_{n\to\infty} A_n)$.*

This applies similarly to decreasing sequences. If A_n is decreasing then $\mathbb{P}(A_n)$ is a decreasing sequence and $\mathbb{P}(A_n) \searrow \mathbb{P}(\lim_{n\to\infty} A_n)$.

We now proceed with a short summary on how some commonly used σ-algebras are constructed. Here is the main tool for this:

Lemma 1.8. *Let $\mathcal{E} \subseteq \mathcal{P}(\Omega)$. Then*

- *there is a smallest algebra $\alpha(\mathcal{E})$ that contains all sets from $\mathcal{E}$;*

- *there is a smallest σ-algebra $\sigma(\mathcal{E})$ that contains all sets from $\mathcal{E}$.*

Proof. The intersection of algebras is an algebra, and the intersection of σ-algebras is a σ-algebra. To find the ones in the lemma, take the intersection of all algebras, respectively σ-algebras, that contain all sets in $\mathcal{E}$. $\qquad\square$

The above $\alpha(\mathcal{E})$ and $\sigma(\mathcal{E})$ are said to be *generated by* $\mathcal{E}$.

Let us now consider

$$
\mathcal{A} := \Big\{ \bigcup_{i=1}^{n} (a_i, b_i] : n < \infty, \text{ and } a_1 < b_1 \le a_2 < b_2 \le \cdots \le a_n < b_n \text{ in } \mathbb{R} \cup \{-\infty\};
$$

(1.2)

$$
\bigcup_{i=1}^{n} (a_i, b_i] \cup (c, \infty) : n < \infty, \text{ and } a_1 < b_1 \le a_2 < b_2 \le \cdots \le a_n < b_n \le c \text{ in } \mathbb{R} \cup \{-\infty\} \Big\}.
$$

This is an algebra (check!), but not a σ-algebra: each of $\left(0, 1 - \frac{1}{n}\right]$ is in $\mathcal{A}$, but the union of these sets for all n is $(0, 1)$, which is not in $\mathcal{A}$. However, this algebra can be used to generate the following σ-algebra:

Definition 1.9. The *Borel σ-algebra on* $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by (1.2). Sets in $\mathcal{B}(\mathbb{R})$ are said to be *Borel sets*.

This σ-algebra contains all subsets of $\mathbb{R}$ that are "of practical interest". I.e., it is not easy to come up with a non-Borel set in $\mathbb{R}$. Those interested can look up the Vitali set for an example.

In a similar way, n-dimensional rectangles:

$$
\big\{ (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_n, b_n] : a_1 < b_1, \ a_2 < b_2, \ldots, a_n < b_n \text{ in } \mathbb{R} \big\},
$$

rather than one-dimensional intervals, can be used to generate $\mathcal{B}(\mathbb{R}^n)$, the *Borel σ-algebra on* $\mathbb{R}^n$. This will contain "all n-dimensional sets of practical interest".

One can then proceed to $\mathbb{R}^\infty$, the set of real-valued sequences, by considering the σ-algebra $\mathcal{B}(\mathbb{R}^\infty)$ generated by rectangles of arbitrary finite dimension. Again, "practical sets", such as

$$
\big\{ (x_n) : \lim_{n \to \infty} x_n \text{ exists and is finite} \big\}; \ \big\{ (x_n) : \sup_n x_n > 5 \big\}; \ \big\{ (x_n) : \liminf_{n \to \infty} x_n > 5 \big\}
$$

all belong to $\mathcal{B}(\mathbb{R}^\infty)$.

One can even define $\mathcal{B}(\mathbb{R}^T)$ with an uncountable set T, for example σ-algebras on function spaces. This usually requires some restrictions on the family of functions considered.

The next theorem, which we cover without proof, allows to construct measures on generated σ-algebras.

Theorem 1.10 (Carathéodory). *Let $\mathcal{A}$ be an algebra on Ω. If μ_0 is a σ-additive measure on $(\Omega, \mathcal{A})$, then there exists a unique extension of it to $\big(\Omega, \sigma(\mathcal{A})\big)$ (the generated σ-algebra).*

Definition 1.11. Let $F : \mathbb{R} \to [0, 1]$ be a cumulative distribution function, and define the σ-additive measure $\mathbb{P}$ on (1.2) by $\mathbb{P}(a, b] := F(b) - F(a)$. This extends to the *Lebesgue-Stieltjes measure* on $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$.

When F is the Uniform$(0, 1)$ distribution, we obtain the *Lebesgue measure* on $\mathcal{B}([0, 1])$ this way.

We briefly mention that $\mathbb{P}$ can be extended to $\mathbb{R}^n$ in a natural way, then Kolmogorov's extension theorem can be used, under certain circumstances, to extend further to $\mathbb{R}^\infty$ or even $\mathbb{R}^T$.

The next task is to construct random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.12. Let $(\Omega, \mathcal{F})$ be a measurable space. A function $X : \Omega \to \mathbb{R}$ is called *measurable*, if for any $B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$. A (real-valued) *random variable* on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a measurable function from $X : \Omega \to \mathbb{R}$.

It should now be clear how probabilities associated with random variables work. The above definition exactly says that a random variable taking value in a Borel set $B \in \mathcal{B}(\mathbb{R})$ is an event in our probability space (i.e., belongs to $\mathcal{F}$). As an example, let us consider the distribution function F of a random variable. When first encountered, it is usually defined as $F(x) = \mathbb{P}\{X \le x\}$. With the above construction, it should rather be written as

$$
F(x) = \mathbb{P}\{\omega \in \Omega : X(\omega) \le x\} = \mathbb{P}\big\{X^{-1}(-\infty, x]\big\}.
$$

which of course has the same meaning, we just understand properly now what is behind the notation. The set $(-\infty, x] \in \mathcal{B}(\mathbb{R})$ is a Borel set, X is a measurable function which implies that $X^{-1}(-\infty, x] \in \mathcal{F}$, hence this is an event and it makes sense to talk about its probability in the sample space Ω.

We remark that limits, sums, differences, products, ratios (where defined), Borel-measurable functions of random variables are each random variables again, in other words these operations do not ruin measurability.

Next we briefly summarise without proofs how to construct expectations of random variables. (In measure theory, this would be called integrals of measurable functions.) We sometimes omit mentioning the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 1.13. A random variable X is called *simple*, if there exist $n > 0$, $x_1, x_2, \ldots, x_n \in \mathbb{R}$, and $A_1, A_2, \ldots, A_n \in \mathcal{F}$ with which

$$X(\omega) = \sum_{k=1}^{n} x_k \cdot \mathbf{1}_{A_k}(\omega).$$

Here $\mathbf{1}_A$ stands for the indicator function:

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

The next theorem we borrow from measure theory without proof.

Theorem 1.14.

(a) *For any random variable X there exists a sequence $X_1, X_2, \ldots$ of simple random variables such that $|X_n| \leq |X|$ for all n, and $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for all $\omega \in \Omega$.*

(b) *Moreover, if $X(\omega) \geq 0$ for every $\omega \in \Omega$, then X_n can be chosen to be non-decreasing in n for every fixed $\omega \in \Omega$ (denoted $X_n(\omega) \nearrow X(\omega)$ as $n \to \infty$ for all $\omega \in \Omega$).*

Definition 1.15 (Expectations)**.**

(a) If X is simple with $X = \sum_{k=1}^{n} x_k \cdot \mathbf{1}_{A_k}$, then $\mathbb{E} X := \sum_{k=1}^{n} x_k \cdot \mathbb{P}(A_k)$.

(b) If $X \geq 0$ is a random variable, then $\mathbb{E} X := \lim_{n \to \infty} \mathbb{E} X_n$, where $X_n \nearrow X$ are simple random variables. (Such sequence exists by the above, and it is a theorem that this limit does not depend on the choice of the sequence.) Notice that $\mathbb{E} X = \infty$ is possible.

(c) If X is a random variable, $\mathbb{E} X := \mathbb{E} X^+ - \mathbb{E} X^-$, unless both expectations on the right-hand side are infinite, in which case $\mathbb{E} X$ is not defined.

Here the positive and negative parts are used:

(1.3) $$x^+ = x \cdot \mathbf{1}\{x > 0\}, \qquad x^- = -x \cdot \mathbf{1}\{x < 0\}, \qquad x = x^+ - x^- \text{ for any } x \in \mathbb{R}.$$

Notice that options for $\mathbb{E} X$ are "not defined", $= \infty$, $= -\infty$, or $\in \mathbb{R}$.

2 Conditional expectation and a toy example

We start this section without proof with Kolmogorov's Theorem on conditional expectations from Williams [2]. Notation: $\mathbb{E}(\cdot\,;G) := \mathbb{E}(\cdot\mathbf{1}_G)$.

Theorem 2.1. *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}|X| < \infty$. Let $\mathcal{G}$ be a sub σ-algebra. Then there exists a random variable V such that*

(a) *V is $\mathcal{G}$-measurable,*

(b) *$\mathbb{E}|V| < \infty$,*

(c) *$\mathbb{E}(V\,;G) = \mathbb{E}(X\,;G)$ for any $G \in \mathcal{G}$.*

Such V is called a version of the conditional expectation $\mathbb{E}(X\,|\,\mathcal{G})$. *Indeed, two random variables V and V' with the above properties agree almost everywhere: $\mathbb{P}(V = V') = 1$.*

Our toy example will be the following. Let $\Omega = \{1, 2, \ldots, 12\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, and $\mathbb{P}$ be the uniform measure on the finite set Ω. Elementary outcomes in Ω will be denoted by ω. Define the random variables

$$Y := \left\lceil \frac{\omega}{4} \right\rceil = \begin{cases} 1, & \text{if } \omega = 1, 2, 3, 4, \\ 2, & \text{if } \omega = 5, 6, 7, 8, \\ 3, & \text{if } \omega = 9, 10, 11, 12, \end{cases} \qquad X := \left\lceil \frac{\omega}{2} \right\rceil = \begin{cases} 1, & \text{if } \omega = 1, 2, \\ 2, & \text{if } \omega = 3, 4, \\ 3, & \text{if } \omega = 5, 6, \\ 4, & \text{if } \omega = 7, 8, \\ 5, & \text{if } \omega = 9, 10, \\ 6, & \text{if } \omega = 11, 12. \end{cases}$$

The σ-algebra generated by Y is

$$\begin{aligned}
\mathcal{G} := \sigma(Y) &:= \sigma\big(Y^{-1}\big(\mathcal{B}(\mathbb{R})\big)\big) = \sigma\big(\{1, 2, 3, 4\}, \{5, 6, 7, 8\}, \{9, 10, 11, 12\}\big) \\
&= \big\{\emptyset, \{1, 2, 3, 4\}, \{5, 6, 7, 8\}, \{9, 10, 11, 12\}, \\
&\qquad\qquad \{1, 2, 3, 4, 5, 6, 7, 8\}, \{1, 2, 3, 4, 9, 10, 11, 12\}, \{5, 6, 7, 8, 9, 10, 11, 12\}, \Omega\big\}.
\end{aligned}$$

Similarly, the σ-algebra generated by X is

$$\mathcal{H} := \sigma(X) := \sigma\big(X^{-1}\big(\mathcal{B}(\mathbb{R})\big)\big) = \sigma\big(\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}, \{9, 10\}, \{11, 12\}\big).$$

We see that $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$. The σ-algebra $\mathcal{G}$ is *coarser* (contains less information), while $\mathcal{H}$ is *finer* (more information). We also see that

- Y is $\mathcal{G}$-measurable (by definition).

- Y is $\mathcal{H}$-measurable (due to $\mathcal{G} \subset \mathcal{H}$).

- X is $\mathcal{H}$-measurable (by definition).

- X is *not* $\mathcal{G}$-measurable (e.g., $X^{-1}\{1\} = \{1, 2\} \notin \mathcal{G}$).

Next, we find the conditional expectation $\mathbb{E}(X \,|\, \mathcal{G})$ based on the definition above. As $\mathcal{G} = \sigma(Y)$, an equivalent notation for this is $\mathbb{E}(X \,|\, \mathcal{G}) = \mathbb{E}(X \,|\, Y)$. Due to $|\Omega| = 12 < \infty$, finite mean of $V = \mathbb{E}(X \,|\, \mathcal{G})$ is not an issue. We look for a $\mathcal{G}$-measurable random variable V with $\mathbb{E}(V \,;\, G) = \mathbb{E}(X \,;\, G)$ for any $G \in \mathcal{G}$. An efficient choice for G is $\{1, 2, 3, 4\}$. As V is $\mathcal{G}$-measurable, and $\mathcal{G}$ has no set that distinguishes between these four outcomes, we find that $V(\omega)$ is the same for $\omega = 1, 2, 3, 4$. The above expectations turn into

$$V(1)\mathbb{P}\{1\} + V(2)\mathbb{P}\{2\} + V(3)\mathbb{P}\{3\} + V(4)\mathbb{P}\{4\} = X(1)\mathbb{P}\{1\} + X(2)\mathbb{P}\{2\} + X(3)\mathbb{P}\{3\} + X(4)\mathbb{P}\{4\}$$
$$V(1)\mathbb{P}\{1\} + V(1)\mathbb{P}\{2\} + V(1)\mathbb{P}\{3\} + V(1)\mathbb{P}\{4\} = X(1)\mathbb{P}\{1\} + X(2)\mathbb{P}\{2\} + X(3)\mathbb{P}\{3\} + X(4)\mathbb{P}\{4\}$$
$$V(1) = V(2) = V(3) = V(4) = \frac{1 \cdot \frac{1}{12} + 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 1.5.$$

Similarly, with the respective choices $G = \{5, 6, 7, 8\}$ and $G = \{9, 10, 11, 12\}$,

$$V(5) = V(6) = V(7) = V(8) = \frac{3 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12} + 4 \cdot \frac{1}{12} + 4 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 3.5,$$

$$V(9) = V(10) = V(11) = V(12) = \frac{5 \cdot \frac{1}{12} + 5 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12} + 6 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 5.5.$$

Hence the conditional expectation is the random variable

$$\mathbb{E}(X \,|\, \mathcal{G})(\omega) = V(\omega) = \begin{cases} 1.5, & \text{if } \omega = 1, 2, 3, 4, \\ 3.5, & \text{if } \omega = 5, 6, 7, 8, \\ 5.5, & \text{if } \omega = 9, 10, 11, 12, \end{cases}$$

being just the average of X over the smallest nontrivial respective units in $\mathcal{G}$.

In a similar way one can check

$$\mathbb{E}(Y \,|\, \mathcal{G})(\omega) = \begin{cases} 1, & \text{if } \omega = 1, 2, 3, 4, \\ 2, & \text{if } \omega = 5, 6, 7, 8, \\ 3, & \text{if } \omega = 9, 10, 11, 12 \end{cases} = Y(\omega),$$

and indeed it is always the case that $\mathbb{E}(Y \,|\, Y) = Y$ almost everywhere (a.e.).

Further examples are $\mathbb{E}\big(X \,|\, \{\emptyset, \Omega\}\big)$, where the random variable V we are looking for is measurable w.r.t. the trivial σ-algebra $\{\emptyset, \Omega\}$, in other words is a constant. Picking $G = \emptyset$ gives $\mathbb{E}(V \,;\, \emptyset) = 0 = \mathbb{E}(X \,;\, \emptyset)$, which is not very informative. The choice $G = \Omega$ on the other hand fixes the value of the constant V:

$$\begin{aligned} \mathbb{E}(V \,;\, \Omega) &= \mathbb{E}(X \,;\, \Omega) \\ \mathbb{E}\, V &= \mathbb{E}\, X \\ V &= \mathbb{E}\, X, \end{aligned}$$

that is, $\mathbb{E}\big(X \,|\, \{\emptyset, \Omega\}\big) = \mathbb{E}\, X$. This is again true a.e. in general, conditioning on the trivial σ-algebra always produces a full expectation.

If, on the other hand, one conditions on the full σ-algebra $\mathcal{F}$ that has all information that can be available in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then every event $G \in \mathcal{F}$ can be substituted, and the very detailed ones completely fix the conditional expectation. In our example we can e.g., take $\{7\}$ to obtain

$$\mathbb{E}(V ; \{7\}) = \mathbb{E}(X ; \{7\}),$$
$$V(7) \cdot \mathbb{P}\{7\} = X(7) \cdot \mathbb{P}\{7\},$$
$$V(7) = X(7) = 4.$$

Similarly, for any $\omega \in \Omega$ one has $V(\omega) = X(\omega)$, which leads us to $\mathbb{E}(X \,|\, \mathcal{F}) = V = X$. This is again a.e. true for general probability spaces: conditioning on the full information does not do any averaging and gives back the random variable instead.

Our final example is

$$\begin{aligned}
\mathcal{I} := \sigma\big(&\{1, 5, 9\}, \{3, 7, 11\}\big) \\
= \{&\emptyset, \{1, 5, 9\}, \{3, 7, 11\}, \{2, 4, 6, 8, 10, 12\}, \{1, 3, 5, 7, 9, 11\}, \\
&\{1, 2, 4, 5, 6, 8, 9, 10, 12\}, \{2, 3, 4, 6, 7, 8, 10, 11, 12\}, \Omega\}.
\end{aligned}$$

We compute $V = \mathbb{E}(Y \,|\, \mathcal{I})$ as before. This is $\mathcal{I}$-measurable, hence constant on $\{1, 5, 9\}$, as well as on $\{3, 7, 11\}$ and on $\{2, 4, 6, 8, 10, 12\}$. Substituting these as G (the rest in $\mathcal{I}$ will not provide additional help) in $\mathbb{E}(V ; G) = \mathbb{E}(Y ; G)$ results in

$$V(1) = V(5) = V(9) = \frac{Y(1)\mathbb{P}\{1\} + Y(5)\mathbb{P}\{5\} + Y(9)\mathbb{P}\{9\}}{\mathbb{P}\{1\} + \mathbb{P}\{5\} + \mathbb{P}\{9\}} = \frac{1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 2,$$

$$V(3) = V(7) = V(11) = \frac{Y(3)\mathbb{P}\{3\} + Y(7)\mathbb{P}\{7\} + Y(11)\mathbb{P}\{11\}}{\mathbb{P}\{3\} + \mathbb{P}\{7\} + \mathbb{P}\{11\}} = \frac{1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 2,$$

$$\begin{aligned}
V(2) = V(4) = V(6) \\
= V(8) = V(10) = V(12) &= \frac{Y(2)\mathbb{P}\{2\}+Y(4)\mathbb{P}\{4\}+Y(6)\mathbb{P}\{6\}+Y(8)\mathbb{P}\{8\}+Y(10)\mathbb{P}\{10\}+Y(12)\mathbb{P}\{12\}}{\mathbb{P}\{2\}+\mathbb{P}\{4\}+\mathbb{P}\{6\}+\mathbb{P}\{8\}+\mathbb{P}\{10\}+\mathbb{P}\{12\}} \\
&= \frac{1 \cdot \frac{1}{12} + 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12} + 3 \cdot \frac{1}{12}}{\frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12} + \frac{1}{12}} = 2.
\end{aligned}$$

We find that $\mathbb{E}(Y \,|\, \mathcal{I})$ is actually a constant, and in fact $= \mathbb{E}\, Y$.

We can repeat this calculation with any function $f : \mathbb{R} \to \mathbb{R}$ (in general this is chosen to be bounded and measurable) to find $\mathbb{E}\big(f(Y) \,|\, \mathcal{I}\big) = \mathbb{E}\big(f(Y)\big)$, a constant. This is when we say that the random variable Y is independent of the σ-algebra $\mathcal{I}$. Knowing which of the events $\{1, 5, 9\}$ and $\{3, 7, 11\}$ did or did not happen will not tell us any information about Y.

If $\mathcal{I}$ happens to be generated by yet another random variable Z, $\mathcal{I} = \sigma(Z)$, then the above is equivalent to variables Y and Z being independent.

An important property and tool with conditional expectations is the following:

Theorem 2.2 (Tower rule). *Let Z be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}|Z| < \infty$. Let $\mathcal{G} \subseteq \mathcal{H}$ be sub σ-algebras ($\mathcal{G}$ is* coarser *and $\mathcal{H}$ is finer). Then*

$$\mathbb{E}\big(\mathbb{E}(Z \,|\, \mathcal{G}) \,|\, \mathcal{H}\big) = \mathbb{E}\big(\mathbb{E}(Z \,|\, \mathcal{H}) \,|\, \mathcal{G}\big) = \mathbb{E}(Z \,|\, \mathcal{G}).$$

The proof follows from the definition of conditional expectations after a bit of manipulations, we leave this to the reader. Some special cases of interest:

- If $\mathcal{H} = \mathcal{F}$, the full σ-algebra in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then $\mathbb{E}(Z \,|\, \mathcal{H}) = \mathbb{E}(Z \,|\, \mathcal{F}) = Z$ for any random variable. The above then reads $\mathbb{E}(Z \,|\, \mathcal{G})$ for all three terms.

- If $\mathcal{G} = \{\emptyset, \Omega\}$, the trivial σ-algerba, then $\mathbb{E}(\cdot \,|\, \mathcal{G}) = \mathbb{E}(\cdot)$. The Tower rule then becomes

$$\mathbb{E}\big((\mathbb{E}\, Z) \,|\, \mathcal{H}\big) = \mathbb{E}\big(\mathbb{E}(Z \,|\, \mathcal{H})\big) = \mathbb{E}\, Z.$$

The first of these terms is uninteresting, but the second equality is very useful and might be familiar from earlier studies, especially when $\mathcal{H} = \sigma(V)$, the σ-algebra generated by another random variable V. In this case it reads $\mathbb{E}\big(\mathbb{E}(Z|V)\big) = \mathbb{E}\, Z$.

3 Probability toolbox

The following statements are widely used across probability, and will be built on in this unit. We always assume the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in the background.

We start with an important fact from calculus.

Lemma 3.1. *Let $a_k \in (0, 1)$ with $\lim_{k\to\infty} a_k = 0$. Then*

$$\prod_{k=1}^{\infty}(1 - a_k) = 0 \quad \Leftrightarrow \quad \sum_k a_k = \infty.$$

Proof. Convexity of the exponential function implies $1 - x \leq e^{-x}$ for any $x \in \mathbb{R}$. As terms in the product are non-negative,

$$0 \leq \prod_{k=1}^{\infty}(1 - a_k) \leq \prod_{k=1}^{\infty} e^{-a_k} = e^{-\sum_{k=1}^{\infty} a_k}.$$

This proves $\Leftarrow$.

The function e^{-2x} is smooth with value 1 and derivative -2 at $x = 0$. Hence for all small enough $x > 0$, $1 - x \geq e^{-2x}$. There is an index K that makes a_k small enough for this purpose for any $k \geq K$. Therefore

$$\prod_{k=K}^{\infty}(1 - a_k) \geq \prod_{k=K}^{\infty} e^{-2a_k} = e^{-2\sum_{k=K}^{\infty} a_k}.$$

If $\prod_{k=1}^{\infty}(1 - a_k) = 0$, then the left hand-side above is also zero, which proves $\Rightarrow$. $\qquad\square$

Definition 3.2. Let $A_1, A_2, \ldots$ be events. Then

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

By decoding the union and the intersection it becomes clear that this event describes that infinitely many of the A_n's occur, in other words *A_n's occur infinitely often (i.o.)*.

Theorem 3.3 (Borel-Cantelli lemmas)**.**

1. *If $A_1, A_2, \ldots$ are any events with $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(\limsup_n A_n) = 0$.*

2. *If $A_1, A_2, \ldots$ are **independent** events with $\sum_n \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(\limsup_n A_n) = 1$.*

Proof.

1. Notice that $\bigcup_{k=n}^{\infty} A_k$ is decreasing in n. Thus, by continuity of probability (Theorem 1.6) and Boole's inequality,

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\bigcup_{k=n}^{\infty} A_k\Big) \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0.$$

2. Notice that $\bigcap_{k=n}^{\infty} A_k^c$ is increasing in n. Thus,

$$\mathbb{P}\Big(\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\Big) = 1 - \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c\Big) = 1 - \lim_{n\to\infty} \mathbb{P}\Big(\bigcap_{k=n}^{\infty} A_k^c\Big) = 1 - \lim_{n\to\infty} \prod_{k=n}^{\infty}\big(1 - \mathbb{P}(A_k)\big).$$

   The product is 0 for any n due to Lemma 3.1, which completes the proof.

$\qquad\square$

We now turn to interchangeability of limits and expectations. The below are standard parts of measure theory, where they are treated for more general integrals than just expectations and sums as here.

Theorem 3.4 (Monotone convergence)**.** *Let $Y, X, X_1, X_2, X_3, \ldots$ be random variables.*

(a) *If $X_n \geq Y$ for each n, $\mathbb{E}\,Y > -\infty$, and $X_n \nearrow X$ for every $\omega \in \Omega$, then $\mathbb{E}\,X_n \nearrow \mathbb{E}\,X$.*

(b) *If $X_n \leq Y$ for each n, $\mathbb{E}\,Y < \infty$, and $X_n \searrow X$ for every $\omega \in \Omega$, then $\mathbb{E}\,X_n \searrow \mathbb{E}\,X$.*

Proof. Proof of (a) only; (b) follows similarly.

Suppose $Y \equiv 0$, i.e., $\forall \omega \in \Omega$, $Y(\omega) = 0$. Then as $X_n \geq 0$ (seen in the measure theory section) for each X_k there exists a sequence of simple (that is, constant on finitely many measurable sets) random variables $X_k^{(n)}$ such that $X_k^{(n)} \nearrow X_k$ as $n \to \infty$.

$$
\begin{array}{ccccccccc}
X_1 & \leq & X_2 & \leq & X_3 & \leq & \cdots & \leq & X \\
\vee| & & \vee| & & \vee| & & & & \\
\vdots & & \vdots & & \vdots & & & & \\
\vee| & & \vee| & & \vee| & & & & \\
X_1^{(3)} & & X_2^{(3)} & & X_3^{(3)} & & & & \\
\vee| & & \vee| & & \vee| & & & & \\
X_1^{(2)} & & X_2^{(2)} & & X_3^{(2)} & & & & \\
\vee| & & \vee| & & \vee| & & & & \\
X_1^{(1)} & & X_2^{(1)} & & X_3^{(1)} & & & &
\end{array}
$$

Define $Z^{(n)} := \max_{1 \leq j \leq n} X_j^{(n)}$. That is, $Z^{(n)}$ is the maximum value of the first n terms in the nth row from the bottom in the table above.

Properties of $Z^{(n)}$:

- For all $1 \leq k \leq n$ we have $X_k^{(n)} \leq Z^{(n)} \leq X_n$. The first inequality follows immediately from the definition of $Z^{(n)}$, it is simply the maximum of such values, and is hence an upper bound. The second inequality follows from chasing the column up within which the maximum lies and then across to the value X_n. Formally, for some $1 \leq k \leq n$, $Z^{(n)} = X_k^{(n)} \leq X_k \leq X_n$.

- $Z^{(n-1)} \leq Z^{(n)}$. Why? $Z^{(n-1)} = \max_{1 \leq j \leq n-1} X_j^{(n-1)} \leq \max_{1 \leq j \leq n-1} X_j^{(n)} \leq Z^{(n)}$. The inequality follows from the fact that for all $j \in \mathbb{N}$ we have $X_j^{(n-1)} \leq X_j^{(n)}$, and then we maximise over a larger domain.

Define $Z := \lim_{n \to \infty} Z^{(n)}$, which exists because $Z^{(n)}$ is an increasing sequence (the limit may possibly be infinite).

Since for all $1 \leq k \leq n$ we have $X_k^{(n)} \leq Z^{(n)} \leq X_n$, taking $n \to \infty$ we see that

$$
\lim_{n \to \infty} X_k^{(n)} \leq \lim_{n \to \infty} Z^{(n)} \leq \lim_{n \to \infty} X_n \implies X_k \leq Z \leq X \underset{k \to \infty}{\implies} Z = X.
$$

Note that, since the $Z^{(n)}$s are simple (indeed they are a maximum of simple random variables), by the definition of expectation of a limit simple random variables,

$$
\mathbb{E}\, X = \mathbb{E}\, Z = \mathbb{E} \lim_{n \to \infty} Z^{(n)} = \lim_{n \to \infty} \mathbb{E}\, Z^{(n)} \leq \lim_{n \to \infty} \mathbb{E}\, X_n.
$$

Thus it remains to show that $\mathbb{E}\, X \geq \lim_{n \to \infty} \mathbb{E}\, X_n$. Since $X_n \nearrow X$ we have that $X_n \leq X$ for all n, which implies that $\mathbb{E}\, X_n \leq \mathbb{E}\, X$. Hence,

$$
\lim_{n \to \infty} \mathbb{E}\, X_n \leq \mathbb{E}\, X.
$$

In the case where $Y \not\equiv 0$ then we repeat the above analysis with $X_n - Y$ which is a non-negative random variable. $\qquad\square$

For the next statement, notice that every sequence has a liminf.

Theorem 3.5 (Fatou's lemma)**.** *Let Y, X_1, X_2, X_3, ... be random variables with $X_n \geq Y$ for each n, $\mathbb{E}\, Y > -\infty$. Then $\liminf_n \mathbb{E}\, X_n \geq \mathbb{E} \liminf_n X_n$.*

It is sometimes convenient to pick $Y \equiv 0$ in the above theorems.

Proof. Define $Z_n := \inf_{m \geq n} X_m$. Then Z_n is an increasing sequence; indeed $\inf_{m \geq n} X_m \leq \inf_{m \geq n+1} X_m$ since the infinum is over a larger domain. Furthermore, $Z_n \nearrow Z := \liminf_{n \to \infty} X_n$; this follows from the fact that Z_n is increasing and by definition

$$
\liminf_{n \to \infty} X_n = \lim_{n \to \infty} \left(\inf_{m \geq n} X_m \right) = \lim_{n \to \infty} Z_n.
$$

Now $Z_n = \inf_{m \geq n} X_m \geq Y$ as $X_m \geq Y$ for all $m \in \mathbb{N}$. Thus we are in good shape to apply the monotone convergence theorem:

$$
\lim_{n \to \infty} \mathbb{E}\, Z_n = \mathbb{E}\, Z = \mathbb{E} \liminf_{n \to \infty} X_n
$$

But on the left-hand-side, as the limit exists it is equal to the lim inf. Now $Z_n = \inf_{m \geq n} X_m \leq X_n$ and thus

$$
\mathbb{E} \liminf_{n \to \infty} X_n = \lim_{n \to \infty} \mathbb{E}\, Z_n = \liminf_{n \to \infty} \mathbb{E}\, Z_n \leq \liminf_{n \to \infty} \mathbb{E}\, X_n. \qquad\square
$$

Theorem 3.6 (Dominated convergence)**.** *Let Y, X, X_1, X_2, X_3, ... be random variables, and assume $|X_n| \leq Y$ for each n, $\mathbb{E} Y < \infty$, and $X_n \to X$ almost surely (a.s., that is, $\mathbb{P}\{X_n \to X\} = 1$). Then $\mathbb{E}|X| < \infty$, $\mathbb{E} X_n \to \mathbb{E} X$, and $\mathbb{E}|X - X_n| \to 0$.*

Proof. To prove finiteness of the expectation note that $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$ implies that $|X_n| \xrightarrow{\text{a.s.}} |X|$ as $n \to \infty$ (mod is a continuous function). Furthermore, since $|X_n| \leq Y$, we have that $|X| = \lim_{n\to\infty} |X_n| \leq Y$ almost surely.

To prove convergence of the expectation we construct the following chain of inequalities

$$\mathbb{E} X = \mathbb{E} \lim_{n\to\infty} X_n = \mathbb{E} \liminf_{n\to\infty} X_n \underset{*}{\leq} \liminf_{n\to\infty} \mathbb{E} X_n \leq \limsup_{n\to\infty} \mathbb{E} X_n \underset{**}{\leq} \mathbb{E} \limsup_{n\to\infty} X_n = \mathbb{E} \lim_{n\to\infty} X_n = \mathbb{E} X,$$

where $*$ follows from Fatou's lemma and $**$ follows from Fatou's lemma on $-X_n$; indeed we have the relation $\liminf_{n\to\infty}(-X_n) = -\limsup_{n\to\infty} X_n$. Thus equality holds throughout the chain and we conclude that

$$\underbrace{\liminf_{n\to\infty} \mathbb{E} X_n = \limsup_{n\to\infty} \mathbb{E} X_n}_{\implies =\lim_{n\to\infty} \mathbb{E} X_n} = \mathbb{E} X.$$

Since the $\liminf$ and $\limsup$ agree, $\lim_{n\to\infty} \mathbb{E} X_n$ exists and is equal to $\mathbb{E} X$. Thus

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

where $\star$ follows from the fact that $\mathbb{E} \lim_{n\to\infty} X_n$ is also an element in the chain.

Finally, to prove $\mathbb{E}|X - X_n| \to 0$ we note that $|X_n - X| \leq |X_n| + |X| \leq 2Y$. Then we repeat the analysis above with $|X_n - X|$ and bounding random variable $Y \equiv 2Y$. $\qquad \square$

Example 3.7. Let

$$X_n = \begin{cases} n^2 - 1, & \text{with probability } \dfrac{1}{n^2}, \\ -1, & \text{with probability } 1 - \dfrac{1}{n^2}, \end{cases}$$

be independent. One easily checks $\mathbb{E} X_n = 0 \; \forall n$, hence $\lim_{n\to\infty} \mathbb{E} X_n = 0$. However, the probabilities in the first line are summable, hence Borel-Cantelli implies that a.s. $X_n \neq -1$ only happens for finitely many n. It follows that $X_n \to -1$ a.s., the limit does not swap with the expectation. Conditions of both Monotone and Dominated convergence fail.

Two important corollaries concern swapping sum and expectation. There is no issue with finite sums, but infinite sums require some thought. These will be important later on, hence the proof is provided.

Theorem 3.8 (Tonelli)**.** *Let $X_n \geq 0$ be random variables. Then $\mathbb{E} \sum_{k=1}^{\infty} X_k = \sum_{k=1}^{\infty} \mathbb{E} X_k$.*

Proof. First notice that $\sum_{k=1}^{n} X_k \geq 0$ and non-decreasing in n, hence the expectations and infinite sums are well-defined. The statement follows from Monotone convergence on the sequence $\sum_{k=1}^{n} X_k$ which converges monotonically to the infinite sum:

$$\mathbb{E} \sum_{k=1}^{\infty} X_k = \mathbb{E} \lim_{n\to\infty} \sum_{k=1}^{n} X_k = \lim_{n\to\infty} \mathbb{E} \sum_{k=1}^{n} X_k = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{E} X_k = \sum_{k=1}^{\infty} \mathbb{E} X_k.$$

$\qquad \square$

Theorem 3.9 (Fubini)**.** *Let X_n be random variables with $\mathbb{E} \sum_{k=1}^{\infty} |X_k| < \infty$. Then $\mathbb{E} \sum_{k=1}^{\infty} X_k = \sum_{k=1}^{\infty} \mathbb{E} X_k$.*

Proof. Recall (1.3), and notice $|x| = x^+ + x^-$ for any real x. By positivity,

$$(3.1) \qquad \infty > \mathbb{E} \sum_{k=1}^{\infty} |X_k| = \mathbb{E} \sum_{k=1}^{\infty} (X_k^+ + X_k^-) = \mathbb{E}\left(\sum_{k=1}^{\infty} X_k^+ + \sum_{k=1}^{\infty} X_k^- \right) = \mathbb{E} \sum_{k=1}^{\infty} X_k^+ + \mathbb{E} \sum_{k=1}^{\infty} X_k^-,$$

which also implies that both sums on the right are a.s. finite. Therefore, a.s.,

$$\sum_{k=1}^{\infty} X_k = \lim_{n\to\infty} \sum_{k=1}^{n} X_k = \lim_{n\to\infty} \sum_{k=1}^{n} (X_k^+ - X_k^-)$$

$$= \lim_{n\to\infty} \left(\sum_{k=1}^{n} X_k^+ - \sum_{k=1}^{n} X_k^- \right) = \lim_{n\to\infty} \sum_{k=1}^{n} X_k^+ - \lim_{n\to\infty} \sum_{k=1}^{n} X_k^- = \sum_{k=1}^{\infty} X_k^+ - \sum_{k=1}^{\infty} X_k^-.$$

By (3.1), we can apply $\mathbb{E}$ separately on this difference. Both sums on the right-hand side are of non-negative terms, hence Tonelli's theorem applies separately:

$$\mathbb{E}\sum_{k=1}^{\infty} X_k = \mathbb{E}\sum_{k=1}^{\infty} X_k^+ - \mathbb{E}\sum_{k=1}^{\infty} X_k^- = \sum_{k=1}^{\infty}\mathbb{E}\,X_k^+ - \sum_{k=1}^{\infty}\mathbb{E}\,X_k^-$$

$$= \lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}\,X_k^+ - \lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}\,X_k^- = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\mathbb{E}\,X_k^+ - \sum_{k=1}^{n}\mathbb{E}\,X_k^-\right) = \lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}\,X_k = \sum_{k=1}^{\infty}\mathbb{E}\,X_k.$$

When joining the two limits we used that, by (3.1) and the same application of Tonelli's theorem, each of $\sum_{k=1}^{n}\mathbb{E}\,X_k^+$ and $\sum_{k=1}^{n}\mathbb{E}\,X_k^-$ has a finite limit. $\qquad\square$

Here is a simple, but very useful theorem, the proof of which is again omitted.

Theorem 3.10 (Jensen's inequality). *Let X be a random variable with $\mathbb{E}\,|X| < \infty$, and g a convex $\mathbb{R} \to \mathbb{R}$ function. Then $g(\mathbb{E}\,X) \leq \mathbb{E}\,g(X)$.*

Proof. Since g is convex, for all $x_0 \in \mathbb{R}$ there exists λ such that $g(x) \geq g(x_0) + \lambda(x - x_0)$. (E.g., the tangent to the curve at all points is a lower bound for the curve if it happens to be differentiable, but this is not needed.) Hence $g(X) \geq g(x_0) + \lambda(X - x_0)$; in particular, for $x_0 = \mathbb{E}\,X \in \mathbb{R}$,

$$g(X) \geq g(\mathbb{E}\,X) + \lambda(X - \mathbb{E}\,X).$$

Note that λ is a constant which depends on the function g and the value of $\mathbb{E}\,X$ only (the slope of the bounding line is dependent only on the position x_0 on the curve) and thus it is not random. Therefore $\mathbb{E}\big(g(X)^-\big) < \infty$, hence $\mathbb{E}\,g(X)$ exists, and

$$\mathbb{E}\,g(X) \geq \mathbb{E}\,g(\mathbb{E}\,X) + \mathbb{E}[\lambda(X - \mathbb{E}\,X)] = g(\mathbb{E}\,X) + \lambda[\mathbb{E}\,X - \mathbb{E}\,X] = g(\mathbb{E}\,X). \qquad\square$$

Next we turn to expectations of powers of random variables.

Definition 3.11. Given the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a $p > 0$ real, we denote by $L^p(\Omega, \mathcal{F}, \mathbb{P})$ the set of random variables with finite p^{th} absolute moment. We also introduce the notation $||X||_p := \big(\mathbb{E}\,|X|^p\big)^{1/p}$, with the convention that here the pth power is inside the expectation, while the $1/p$ power is outside. Hence $L^p(\Omega, \mathcal{F}, \mathbb{P})$ is exactly the set of those random variables with finite $||X||_p$.

As we will see, often cases where $p \geq 1$ are relevant.

Next we explore useful properties of $||\cdot||_p$.

Theorem 3.12 (Ljapunov's inequality). *For any real $0 < p < q$ and any random variable, $||X||_p \leq ||X||_q$.*

Proof.

$$(\mathbb{E}\,|X|^s)^{\frac{1}{s}} = [(\mathbb{E}\,|X|^s)^{\frac{t}{s}}]^{\frac{1}{t}} \underset{\star\star}{\leq} [\mathbb{E}(|X|^s)^{\frac{t}{s}}]^{\frac{1}{t}} = (\mathbb{E}\,|X|^t)^{\frac{1}{t}},$$

where $\star\star$ follows from Jensen's inequality on the function $(\cdot)^{\frac{t}{s}} : \mathbb{R} \to \mathbb{R}$; which is convex since $t > s$ and therefore $\frac{t}{s} > 1$. $\qquad\square$

Theorem 3.13 (Hölder's inequality). *Let $p, q > 1$ that satisfy $\frac{1}{p} + \frac{1}{q} = 1$. If $||X||_p < \infty$ and $||Y||_q < \infty$, then $\mathbb{E}\,|XY| \leq ||X||_p \cdot ||Y||_q$.*

The case $p = q = 2$ should be familiar under the name Cauchy-Schwarz inequality.

Proof. Since $\log : \mathbb{R} \to \mathbb{R}$ is a concave function and $\frac{x}{p} + \frac{y}{q}$ is a convex combination of x and y we have that

$$\log\left(\frac{x}{p} + \frac{y}{q}\right) \geq \frac{\log(x)}{p} + \frac{\log(y)}{q} \implies \frac{x}{p} + \frac{y}{q} \geq x^{\frac{1}{p}} \cdot y^{\frac{1}{q}},$$

since the exponential function is increasing. Let $x = \frac{|X|^p}{\mathbb{E}\,|X|^p}$ and $y = \frac{|Y|^q}{\mathbb{E}\,|Y|^q}$. Then

$$\frac{1}{p}\frac{|X|^p}{\mathbb{E}\,|X|^p} + \frac{1}{q}\frac{|Y|^q}{\mathbb{E}\,|Y|^q} \geq \frac{|X|}{(\mathbb{E}\,|X|^p)^{\frac{1}{p}}} \cdot \frac{|Y|}{(\mathbb{E}\,|Y|^q)^{\frac{1}{q}}}.$$

Taking expectations of both sides,

$$1 \geq \frac{\mathbb{E}\,|XY|}{(\mathbb{E}\,|X|^p)^{\frac{1}{p}}(\mathbb{E}\,|Y|^q)^{\frac{1}{q}}}. \qquad\square$$

Notice that by $|X|^p \geq 0$, $||X_p|| = 0$ implies that $X = 0$ a.s. Also, $||\lambda X||_p = |\lambda| \cdot ||X||_p$ for any $\lambda \in \mathbb{R}$ is easily checked from the definition. This, together with the triangle inequality below, justifies the name *p-norm* for $|| \cdot ||_p$ when $p \geq 1$.

Theorem 3.14 (Minkowski's inequality). *Let $p \geq 1$, and $||X||_p < \infty$, $||Y||_p < \infty$. Then $||X + Y||_p \leq ||X||_p + ||Y||_p$.*

Proof. If either (or both) $||X||_p = \infty$ or $||Y||_p = \infty$ then the inequality holds trivially. Hence suppose that $||X||_p < \infty$ and $||Y||_p < \infty$.

For the case $p = 1$ this is trivial and follows immediately from the triangle inequality of the mod.

Now consider the case $p > 1$. Define $F(x) = (a + x)^p - 2^{p-1}(a^p + x^p)$, $x > 0$; where $a > 0$ is some constant. This has the derivative

$$F'(x) = p(a + x)^{p-1} - 2^{p-1}px^{p-1},$$

and so F is stationary at $x = a$. Furthermore,

$$F'(x) > 0 \iff p(a + x)^{p-1} - 2^{p-1}px^{p-1} > 0 \iff \left(\frac{a+x}{2x}\right)^{p-1} > 1 \iff x < a.$$

Similarly, $F'(x) = 0$ if and only if $x = a$ and $F'(x) < 0$ if and only if $x > a$. Thus F is an increasing function for $x < a$; reaching a global maximum at $x = a$ and then decreasing for $x > a$. Therefore, $F(x) \leq F(a) = 0$ for all $x \in \mathbb{R}$. We therefore have the inequality

$$(a + x)^p \leq 2^{p-1}(a^p + x^p)$$

for all $a > 0$, $x > 0$, $p > 1$. Applying this:

$$|X + Y|^p \leq (|X| + |Y|)^p \leq 2^{p-1}(|X|^p + |Y|^p).$$

Taking expectations,

$$\mathbb{E}\,|X + Y|^p \leq 2^{p-1}(\mathbb{E}\,|X|^p + \mathbb{E}\,|Y|^p) < \infty$$

since we assumed that both $||X||_p < \infty$ and $||Y||_p < \infty$. This verifies that $||X+Y||_p < \infty$, that is, $X + Y \in L^P(\Omega)$. Now we prove the Minkowski inequality.

$$\mathbb{E}\,|X + Y|^p = \mathbb{E}(|X + Y||X + Y|^{p-1}) \underset{*}{\leq} \mathbb{E}\left[\left(|X| + |Y|\right)|X + Y|^{p-1}\right] = \mathbb{E}(|X||X + Y|^{p-1}) + \mathbb{E}(|Y||X + Y|^{p-1}).$$

Above, $*$ follows from the triangle inequality on $\mathbb{R}$. Let q be such that $\frac{1}{p} + \frac{1}{q} = 1$; this implies $q = p/(p - 1)$. By Hölder's inequality

$$\mathbb{E}(|X||X + Y|^{p-1}) \leq (\mathbb{E}\,|X|^p)^{\frac{1}{p}}(\mathbb{E}\,|X + Y|^{(p-1)q})^{\frac{1}{q}} = (\mathbb{E}\,|X|^p)^{\frac{1}{p}}\underbrace{(\mathbb{E}\,|X + Y|^p)}_{<\infty \text{ by } *}^{\frac{1}{q}} = ||X||_p||X + Y||_p^{\frac{p}{q}},$$

$$\mathbb{E}(|Y||X + Y|^{p-1}) \leq (\mathbb{E}\,|Y|^p)^{\frac{1}{p}}(\mathbb{E}\,|X + Y|^{(p-1)q})^{\frac{1}{q}} = (\mathbb{E}\,|Y|^p)^{\frac{1}{p}}(\mathbb{E}\,|X + Y|^p)^{\frac{1}{q}} = ||Y||_p||X + Y||_p^{\frac{p}{q}}.$$

Plugging this into the above yields that,

$$\underbrace{\mathbb{E}\,|X + Y|^p}_{=||X+Y||_p^p} \leq \left(||X||_p + ||Y||_p\right)||X + Y||_p^{\frac{p}{q}}.$$

Dividing through by $||X + Y||_p^{\frac{p}{q}} (\geq 0)$ and noting that $p - p/q = 1$ by the definition of q (multiply by p), this is precisely the Minkowski inequality. $\qquad\square$

4 Modes of convergence

There are several ways to state that a sequence of random variables converges to a limit. We define the most commonly used modes and state some of their connections.

Definition 4.1.

- Random variables X_n *converge weakly* to X, denoted $X_n \xrightarrow{\text{w}} X$, if for every bounded and continuous $f : \mathbb{R} \to \mathbb{R}$ function, $\mathbb{E}\,f(X_n) \to \mathbb{E}\,f(X)$. This is equivalent to convergence of the distribution functions: $F_{X_n}(x) \to F_X(x)$ at every x where the limit distribution function F_X is continuous. (Those interested can look up Portmanteau's theorem.) Other commonly used notation for this is $X_n \Rightarrow X$.

- Random variables X_n *converge in probability* to X, denoted $X_n \xrightarrow{\mathbb{P}} X$, if for any $\varepsilon > 0$, $\mathbb{P}\{|X - X_n| \geq \varepsilon\} \to 0$.

- Random variables X_n *converge strongly*, or *almost surely* to X, if $\mathbb{P}\{X_n \to X\} = 1$.

- Fix a $p > 0$. Random variables X_n *converge in L^p*, denoted $X_n \xrightarrow{L^p} X$ if $||X - X_n||_p \to 0$.

Notice that for weak convergence one does not even need the random variables to be defined on the same probability space. This mode only features the distributions, not the actual values of the random variables. The other three modes compare values, hence require the random variables to be defined on a common probability space.

Theorem 4.2.

(a) $X_n \xrightarrow{a.s.} X$ if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k \geq n} |X_k - X| \geq \varepsilon) \to 0$ as $n \to \infty$.

(b) X_n is Cauchy almost surely if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k, \ell \geq n} |X_k - X_\ell| \geq \varepsilon) \to 0$ as $n \to \infty$. This is also equivalent to for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k \geq 0} |X_{n+k} - X_n| \geq \varepsilon) \to 0$.

Remark 4.3. *Note that (a) is like a 'boosted' version of convergence in probability, where we require that all points onwards from n are within ε of X.*

Proof. (a) Define the events $A_k^m := \{\omega \in \Omega : |X_k - X| \geq \frac{1}{m}\}$ and $A^m := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^m$, which is the event that $|X_k - X| \geq \frac{1}{m}$ for infinitely many k.

Note that $X_n \not\to X$ if for some $m \in \mathbb{N}$, A^m occurs; that is, for some $m > 0$, $|X_k - X| \geq \frac{1}{m}$ for infinitely many k. The event that this happens for at least one m is $\bigcup_{m=1}^{\infty} A^m$. Thus $X_n \xrightarrow{a.s.} X$ if and only if $\mathbb{P}(\bigcup_{m=1}^{\infty} A^m) = 0$. Since $\mathbb{P}(A^m) \leq \mathbb{P}(\bigcup_{m=1}^{\infty} A^m) \leq \sum_{m=1}^{\infty} \mathbb{P}(A^m)$, $\mathbb{P}(\bigcup_{m=1}^{\infty} A^m) = 0$ if and only if $\mathbb{P}(A^m) = 0$ for all $m \in \mathbb{N}$. $\mathbb{P}(A^m) = 0$ if and only if

$$0 = \mathbb{P}\left(\underbrace{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^m}_{\star} \right) = \lim_{n \to \infty} \mathbb{P}\left(\underbrace{\bigcup_{k=n}^{\infty} A_k^m}_{\star\star} \right) = \lim_{n \to \infty} \mathbb{P}(\sup_{k \geq n} |X_k - X| \geq 1/m).$$

$\star$ defines a decreasing sequence. $\star\star$ is the event that for some $k \geq n$ we have $|X_k - X| \geq \frac{1}{m}$, which occurs if and only if the supremum, $\sup_{k \geq n} |X_k - X| \geq 1/m$.

As the above m is arbitrarily large, the proof is complete.

To prove (b), we repeat exactly the same analysis with the event $B_{k,\ell}^m = \{\omega \in \Omega : |X_k - X_\ell| \geq \frac{1}{m}\}$. $\qquad \square$

Theorem 4.4.

(a) $X_n \to X$ a.s. implies $X_n \xrightarrow{\mathbb{P}} X$.

(b) Fix any $p > 0$. Then $X_n \xrightarrow{L^p} X$ implies $X_n \xrightarrow{\mathbb{P}} X$.

(c) $X_n \xrightarrow{\mathbb{P}} X$ implies $X_n \xrightarrow{w} X$.

Proof. To prove (a) we use the previous theorem, that is, that $X_n \xrightarrow{a.s.} X$ if and only if for all $\varepsilon > 0$, $\mathbb{P}(\sup_{k \geq n} |X_k - X| \geq \varepsilon) \to 0$ as $n \to \infty$. The event $|X_n - X| \geq \varepsilon$ implies that $\sup_{k \geq n} |X_k - X| \geq \varepsilon$, and thus

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \mathbb{P}(\sup_{k \geq n} |X_k - X| \geq \varepsilon) \to 0,$$

which implies $X_n \xrightarrow{\mathbb{P}} X$.

To prove (b) we use Markov's inequality, which says that for a non-negative random variable Z, $\mathbb{P}(Z \geq c) \leq \mathbb{E} Z/c$. Then for all $p > 0$,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}(|X_n - X|^p \geq \varepsilon^p) \leq \frac{\mathbb{E} |X_n - X|^p}{\varepsilon^p} = \frac{||X_n - X||_p^p}{\varepsilon^p} \to 0$$

as $n \to \infty$ since $X_n \xrightarrow{L^p}$ if and only if $||X_n - X||_p \to 0$.

Now we prove (c) which says that convergence in probability implies weak convergence (in distribution). This is the most difficult to prove. Fix f bounded and continuous such that $|f(x)| \leq c$ ($\spadesuit$) for some $c > 0$. We want to show that $\mathbb{E} f(X_n) \to \mathbb{E} f(X)$, or equivalently that $|\mathbb{E} f(X_n) - \mathbb{E} f(X)| \to 0$. Fix $\varepsilon > 0$.

Assuming X has a proper distribution, i.e., that X is finite almost surely, there exists $N > 0$ such that

$$(\star) \qquad\qquad\qquad \mathbb{P}(|X| > N) \leq \frac{\varepsilon}{4c}.$$

13

$[-2N, 2N]$ is compact, and thus f is uniformly continuous on $[-2N, 2N]$. Hence there exists $\delta > 0$ such that for all $x, y \in [-2N, 2N]$ with $|x - y| \leq \delta$ we have

$$(\dagger) \qquad\qquad\qquad |f(x) - f(y)| \leq \frac{\varepsilon}{2}.$$

Define $\mathbb{E}(Z; A) = \mathbb{E}(Z\mathbf{1}A)$. Then for a disjoint partition of the sample space $A_1, A_2, \ldots, A_n$, we may write $\mathbb{E}\, Z = \mathbb{E}(Z; A_1) + \mathbb{E}(Z; A_2) + \cdots + \mathbb{E}(Z; A_n)$. This is because $\mathbb{E}(Z; A_1) + \mathbb{E}(Z; A_2) + \cdots + \mathbb{E}(Z; A_n) = \mathbb{E}[Z(\mathbf{1}A_1 + \cdots + \mathbf{1}A_n)]$ by linearity of expectation and because $(\mathbf{1}A_1 + \cdots + \mathbf{1}A_n)(\omega) = 1$ for all $\omega \in \Omega$ by the disjointness of the partition.

$$
\begin{aligned}
|\mathbb{E}\, f(X_n) - \mathbb{E}\, f(X)| &= |\mathbb{E}[f(X_n) - f(X)]| \\
&\leq \mathbb{E}\,|f(X_n) - f(X)| \qquad\qquad\text{Jensen's inequality on } |\cdot| \\
&= \mathbb{E}\Big[\underbrace{|f(X_n) - f(X)|}_{\leq \frac{\varepsilon}{2} \;\Longleftarrow\; \dagger}; \underbrace{|X_n - X| \leq \delta; |X| \leq N}_{\Longrightarrow\; X_n, X \in [-2N, 2N], \mathbf{1} \leq 1}\Big] \\
&\quad + \mathbb{E}\Big[\underbrace{|f(X_n) - f(X)|}_{\spadesuit \;\Longrightarrow\; \leq 2c}; \underbrace{|X_n - X| \leq \delta}_{\mathbf{1} \leq 1}; |X| > N\Big] \\
&\quad + \mathbb{E}\Big[\underbrace{|f(X_n) - f(X)|}_{\spadesuit \;\Longrightarrow\; \leq 2c}; |X_n - X| > \delta\Big] \\
&\leq \frac{\varepsilon}{2} + 2c\underbrace{\mathbb{P}(|X| > N)}_{\star \;\Longrightarrow\; \leq \frac{\varepsilon}{4c}} + 2c\mathbb{P}(|X_n - X| > \delta) \\
&\leq \varepsilon + 2c\mathbb{P}(|X_n - X| > \delta).
\end{aligned}
$$

Since $X_n \xrightarrow{\mathbb{P}} X$, there exists $N > 0$ such that $n \geq N$ implies that $\mathbb{P}(|X_n - X| > \delta) \leq \frac{\varepsilon}{2c}$. Thus $n \geq N$ implies that

$$|\mathbb{E}\, f(X_n) - \mathbb{E}\, f(X)| \leq \varepsilon + 2c\mathbb{P}(|X_n - X| > \delta) \leq 2\varepsilon. \qquad\qquad \square$$

We conclude this part with examples that show how reverse implications can fail in the above theorem.

Example 4.5. Let $U \sim \text{Uniform}(0, 1)$, and

$$X_1 = \mathbf{1}_{[0,\,1]}(U), \quad X_2 = \mathbf{1}_{[0,\,\frac{1}{2}]}(U), \quad X_3 = \mathbf{1}_{[\frac{1}{2},\,1]}(U), \quad X_4 = \mathbf{1}_{[0,\,\frac{1}{3}]}(U), \quad X_5 = \mathbf{1}_{[\frac{1}{3},\,\frac{2}{3}]}(U), \quad X_6 = \mathbf{1}_{[\frac{2}{3},\,1]}(U),$$

$$X_7 = \mathbf{1}_{[0,\,\frac{1}{4}]}(U), \quad X_8 = \mathbf{1}_{[\frac{1}{4},\,\frac{2}{4}]}(U), \quad X_9 = \mathbf{1}_{[\frac{2}{4},\,\frac{3}{4}]}(U), \quad X_{10} = \mathbf{1}_{[\frac{3}{4},\,1]}(U), \quad X_{11} = \mathbf{1}_{[0,\,\frac{1}{5}]}(U), \quad \ldots$$

This sequence converges to 0 in L^p, therefore in probability and weakly as well. (Just check the probability that $X_n \neq 0$.) However, there is always a later X_n with value 1, hence the sequence does not converge a.s.

Example 4.6. Let $U \sim \text{Uniform}(0, 1)$, and

- $X_n := U^n$. This converges to 0 in all senses.

- $X_n := nU^n$. This converges to 0 a.s., hence in probability and weakly as well. However,

$$\mathbb{E}\,|X_n - 0| = n\,\mathbb{E}\,U^n = n \cdot \frac{1}{n+1} \to 1 \neq 0,$$

  therefore L^1 convergence does not hold. Notice how both Monotone and Dominated convergence fail for X_n.

- We can take this to more extreme by $X_n := e^n\mathbf{1}_{[0,\,\frac{1}{n}]}(U)$. Again, this converges to 0 a.s. However, $||0 - X_n||_p = \frac{1}{n^{1/p}} \cdot e^n \to \infty$ for any $p > 0$.

Example 4.7. Given a sequence $0 \leq p_n \leq 1$, let $X_n \sim \text{Bernoulli}(p_n)$ and independent. Then

- by the definitions, $p_n \to 0$ is equivalent to each of L^p and in probability convergence to 0,

- by the two Borel-Cantelli lemmas, $\sum_n p_n < \infty$ is equivalent to a.s. convergence to 0.

The choice $p_n = \frac{1}{n}$ therefore gives L^p but not a.s. convergence.

We close this section by noting that, with the assumptions as stated there, Monotone convergence, Fatou's lemma and Dominated convergence hold if we require $X_n \to X$ a.s., or $X_n \xrightarrow{\mathbb{P}} X$ only, instead of convergence for all $\omega \in \Omega$.

5 Martingales, stopping times

Finally, all background is there to consider martingales. From here we mostly follow Williams [2]. Here are the definitions required.

Definition 5.1. A *filtered space* is $\big(\Omega,\, \mathcal{F},\, (\mathcal{F}_n)_{n=0}^{\infty},\, \mathbb{P}\big)$, where $(\Omega,\, \mathcal{F},\, \mathbb{P})$ is a probability space, and $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}$ are σ-algebras, jointly called a *filtration*. We also define $\mathcal{F}_\infty := \sigma\big(\bigcup_n \mathcal{F}_n\big) \subseteq \mathcal{F}$.

Definition 5.2. A process (sequence of random variables, that is) X_n is *adapted* to the filtration $(\mathcal{F}_n)_{n\geq 0}$, if for every n, the variable X_n is $\mathcal{F}_n$-measurable.

This is to say that $\mathcal{F}_n$ contains all information about X_n, in other words given $\mathcal{F}_n$, X_n is not random anymore. An often used scenario is defining $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ from a process $(X_n)_{n\geq 0}$. This is usually assumed when a filtration is not explicitly given.

Definition 5.3. A process $(M_n)_{n\geq 0}$ in a probability space $(\Omega,\, \mathcal{F},\, \mathbb{P})$ is a *martingale with respect to a filtration* $(\mathcal{F}_n)_{n\geq 0}$, if

- it is adapted to $(\mathcal{F}_n)_{n\geq 0}$;
- $\mathbb{E}\,|M_n| < \infty \;\; \forall n \geq 0$;
- $\mathbb{E}(M_{n+1}\,|\,\mathcal{F}_n) = M_n$ a.s., $\forall n \geq 0$.

If instead, in the last equality, we have $\leq$, then we call M a *supermartingale*, while if it is $\geq$ then it is a *submartingale*.

Notice that M_n is a (sub-, super-)martingale if and only if $M_n - M_0$ is. Also, tower rule checks that $\mathbb{E}\, M_n \overset{=}{\underset{\geq}{\leq}} \mathbb{E}\, M_0$ for the respective cases.

Example 5.4. Let X_k be independent random variables with $\mathbb{E}\, X_k = 0$ for each $k \geq 1$. Then $\sum_{k=1}^{n} X_k$ is a martingale. (For $n = 0$ we have an empty sum which we postulate to be zero.)

Example 5.5. Let X_k be independent random variables with $\mathbb{E}\, X_k = 1$ for each $k \geq 1$. Then $\prod_{k=1}^{n} X_k$ is a martingale. (For $n = 0$ we have an empty product which we postulate to be one.)

Example 5.6. Let $\xi \in L^1$ and $\mathcal{F}_n$ be a filtration. Then $M_n := \mathbb{E}(\xi\,|\,\mathcal{F}_n)$ is a martingale due to the Tower rule.

The first applications come from *stopping* a martingale at a time when something particular happens to a process. This is covered next.

Definition 5.7. A process $(C_n)_{n\geq 1}$ is *predictable* (also said *previsible*), if for every $n \geq 1$, C_n is $\mathcal{F}_{n-1}$-measurable.

Imagine a game in rounds that is based on random outcomes X_n. For example, X_n can be the share price of a certain stock. If a player possesses C_k of these stocks in the kth round of the game, then their joint value changes $C_k \cdot (X_k - X_{k-1})$ in this round. The total change in wealth is given by

$$(5.1) \qquad (C \bullet X)_n := \sum_{k=1}^{n} C_k \cdot (X_k - X_{k-1}).$$

Definition 5.8. The process $(C \bullet X)_n$ in (5.1) is called the *martingale transform of X by C*, or the *discrete stochastic integral of C by X*.

In this setup the player can change the amount C_k they own of the stock. However, as players do not foresee the future, it is natural to assume that C is predictable.

Theorem 5.9. *Let C be a predictable process.*

 (a) If there is a bound $K > 0$: $0 \leq C_n \leq K$ ($\forall n,\, \omega$), and X is (super)martingale, then so is $C \bullet X$.

 (b) If $|C_n| \leq K$ ($\forall n,\, \omega$), and X is martingale, then so is $C \bullet X$.

The bound on C_n can be relaxed to both C_n and X_n being in L^2 for each n.

Proof. That $C \bullet X$ is adapted follows easily. Consider

$$\mathbb{E}\big((C \bullet X)_{n+1} - (C \bullet X)_n\,|\,\mathcal{F}_n\big) = \mathbb{E}\big(C_{n+1} \cdot (X_{n+1} - X_n)\,|\,\mathcal{F}_n\big) = C_{n+1}\, \mathbb{E}\big((X_{n+1} - X_n)\,|\,\mathcal{F}_n\big).$$

The last step used predictability of C, and the bounds imposed on C or C and X. These latter are the necessary conditions for $C \bullet X \in L^1$ and for pulling C_{n+1} out of the expectation. The theorem follows from the (super)martingale property of X in the respective cases. $\qquad\square$

Definition 5.10. A random variable $T = 0, 1, 2, \ldots, \infty$ is a *stopping time*, if $\forall n \leq \infty$, the event $\{T \leq n\} \in \mathcal{F}_n$. For a process X_n, we call $X_n^T := X_{T \wedge n}$ the *stopped process*.

Here $i \wedge j$ means the smaller of the two numbers i or j. This definition expresses the property that we can decide whether the stopping time has arrived or not at time n just by looking at the history of the process(es) up to n. Notice that $n = \infty$ is included in the definition. As an elementary example, the time of the first Head in a sequence of coinflips is a stopping time, while the time one before the first Head is not.

Theorem 5.11. *Let T be a stopping time. If X is a (super)martingale, then so is X^T.*

Proof. Define
$$C_n := \mathbf{1}\{n \leq T\} = \mathbf{1}\{n - 1 < T\} = 1 - \mathbf{1}\{T \leq n - 1\},$$
which is $\mathcal{F}_{n-1}$-measurable by the right-hand side. In other words, C is predictable, which implies that $C \bullet X$ is a (super)martingale. On the other hand, the sum with this C is telescopic:
$$(C \bullet X)_n = \sum_{k=1}^{n} \mathbf{1}\{k \leq T\} \cdot (X_k - X_{k-1}) = \sum_{k=1}^{T \wedge n} (X_k - X_{k-1}) = X_{T \wedge n} - X_0 = X_n^T - X_0.$$
$\square$

Theorem 5.12 (Doob's optional stopping theorem)**.** *Let T be a stopping time and X a supermartingale. If either of the below holds:*

(i) T is bounded,

(ii) X is bounded and T is a.s. finite,

(iii) X is of bounded increments: $\exists K > 0: |X_n - X_{n-1}| \leq K \ \forall n \geq 1$, and $\mathbb{E}\, T < \infty$,

then $\mathbb{E}\, X_T \leq \mathbb{E}\, X_0$. If M is a martingale and either of (i), (ii), (iii) holds, then $\mathbb{E}\, M_T = \mathbb{E}\, M_0$.

Notice that one of the homework sheets (will) contain(s) an improved version of this theorem.

Proof. In all cases $T < \infty$ a.s., hence $X_{T \wedge n} \xrightarrow[n \to \infty]{} X_T$ a.s. We also know by Theorem 5.11 that $\mathbb{E}\, X_{T \wedge n} \leq \mathbb{E}\, X_0$. The question is whether we can pass this limit through the expectation. In the respective cases:

(i) for any n larger than the bound on T, $\mathbb{E}\, X_0 \geq \mathbb{E}\, X_{T \wedge n} = \mathbb{E}\, X_T$.

(ii) the bound on X allows to use Dominated convergence.

(iii) in this case we have, for every n,
$$|X_{T \wedge n} - X_0| = \left| \sum_{k=1}^{T \wedge n} (X_k - X_{k-1}) \right| \leq \sum_{k=1}^{T \wedge n} |X_k - X_{k-1}| \leq K \cdot (T \wedge n) \leq KT.$$

As the right-hand side is independent of n and has finite mean, it can act as the bounding Y variable in Dominated convergence, which then again allows to pass the limit through the expectation. The proof for a martingale M is analogous.
$\square$

Notice that we only used, hence it is sufficient to verify, the conditions on the stopped process X^T.

Corollary 5.13. *If M is a martingale of bounded increments, C is bounded and predictable, $T \in L^1$ is a stopping time, then $\mathbb{E}(C \bullet M)_T = 0$.*

When $X \geq 0$, we can apply Fatou's lemma instead of Dominated convergence to relax the conditions of optional stopping:

Theorem 5.14. *If $X \geq 0$ is a supermartingale and T is an a.s. finite stopping time, then $\mathbb{E}\, X_T \leq \mathbb{E}\, X_0$.*

Proof. In this case $X_{T \wedge n} \geq 0$ and converges to X_T a.s., hence by Fatou's lemma,
$$\mathbb{E}\, X_T = \mathbb{E} \lim_{n \to \infty} X_{T \wedge n} = \mathbb{E} \liminf_n X_{T \wedge n} \leq \liminf_n \mathbb{E}\, X_{T \wedge n} \leq \liminf_n \mathbb{E}\, X_0 = \mathbb{E}\, X_0.$$
$\square$

The next lemma is useful for checking if $\mathbb{E}\, T < \infty$ holds. It is enough to check that, no matter the history, stopping happens within a fix time interval with a probability bounded away from zero:

Lemma 5.15. *Let T be a stopping time, and assume $\exists N, \varepsilon > 0 \,:\, \mathbb{P}\{T \le n + N \,|\, \mathcal{F}_n\} \ge \varepsilon$ a.s. for all $n \ge 0$. Then $\mathbb{E}\, T < \infty$.*

Proof. For $k \ge 1$,

$$\mathbb{P}\{T > kN\} = \mathbb{P}\big\{\{T > kN\} \cap \{T > (k-1)N\}\big\} = \mathbb{P}\{T > kN \,|\, T > (k-1)N\}\mathbb{P}\{T > (k-1)N\}$$
$$\le (1 - \varepsilon)\mathbb{P}\{T > (k-1)N\}$$

by the assumption, which recursively gives $\mathbb{P}\{T > kN\} \le (1 - \varepsilon)^k$. The sequence $\mathbb{P}\{T > \ell\}$ is non-increasing, and its sum gives the expectation of T (check if you have not seen this before!). Hence

$$\mathbb{E}\, T = \sum_{\ell=0}^{\infty} \mathbb{P}\{T > \ell\} \le N \sum_{k=0}^{\infty} \mathbb{P}\{T > kN\} \le N \frac{1}{1 - (1 - \varepsilon)} = \frac{N}{\varepsilon} < \infty.$$

$\square$

We turn to a few applications.

Example 5.16. A monkey repeatedly types any of the 26 letters of the English alphabet independently with equal chance. Let T be the number of letters that have been typed when the entire word "ABRACADABRA" first appears. We are after $\mathbb{E}\, T$.

First notice that this is a stopping time and is finite due to Lemma 5.15: no matter what happened before, with probability $\varepsilon = 26^{-11}$ the stopping occurs in at most $N = 11$ steps.

To proceed, we add a gambling process to this problem. Before each hit of a new letter, a new gambler arrives and bets £1 on the letter "A". If the new letter is something else, the gambler loses their bet and exits the game. If the new letter is "A", then the gambler receives £26, which they all bet on the next letter being "B". If this next letter is something else, all bet is lost and the gambler exits the game. If this next letter is "B", then the gambler receives £26^2, which they all bet on the third letter being "R", and so on. If the gambler wins all the way down to "ABRACADABRA", the gambler exits the game with wealth £26^{11} and T becomes the total number of letters typed at this point. Otherwise the gambler loses all their money and exits the game. Also, new gamblers arrive at each step and start the same betting strategy. Denote by X_n the combined wealth of all gamblers who are in play after the nth letter has been typed. We claim that $M_n := X_n - n$ is a martingale.

To check this, notice that M_n is adapted to the natural filtration generated by the monkey, and is bounded for any given n, hence is in L^1. The martingale property is checked by first observing that the expected wealth of a gambler already in play does not change. That is because their bet is lost with probability $\frac{25}{26}$ and multiplied by 26 with probability $\frac{1}{26}$, hence the mean value stays. A gambler already out of play stays with their 0 wealth so that doesn't change either. However, a new gambler arrives and their expected wealth after a flip is $26 \cdot \frac{1}{26} = 1$ which adds to the conditional expectation of X_{n+1}. Therefore,

$$\mathbb{E}(M_{n+1} \,|\, \mathcal{F}_n) = \mathbb{E}(X_{n+1} \,|\, \mathcal{F}_n) - (n+1) = X_n + 1 - (n+1) = M_n.$$

Notice that $\mathbb{E}\, M_1 = \mathbb{E}\, X_1 - 1 = 0$, hence we can extend the definition by $M_0 = 0$.

The stopped martingale M^T is of bounded increments, and $\mathbb{E}\, T < \infty$, hence Optional stopping applies and gives

$$0 = M_0 = \mathbb{E}\, M_T = \mathbb{E}\, X_T - \mathbb{E}\, T.$$

It remains to calculate X_T. At the time of stopping, there are only three gamblers in play. The one who came at time $T - 11$ has wealth 26^{11}. Less lucky is the gambler who arrived at $T - 4$; they successfully bet on "ABRA" and have £26^4. Finally, the gambler who just arrived at T and bet on "A" has £26. Thus,

$$\mathbb{E}\, T = \mathbb{E}\, X_T = X_T = 26^{11} + 26^4 + 26 \simeq 3.7 \cdot 10^{15}.$$

Notice how repetitions in "ABRACADABRA" increase $\mathbb{E}\, T$ (why?).

Example 5.17. Let X_k be i.i.d. ± 1-valued with equal chance. The sum $S_n = \sum_{k=1}^{n} X_k$ is called *simple symmetric random walk* (SSRW). Empty sums are zero, hence $S_0 = 0$. Then S_n is a martingale, and

$$T := \inf\{n \,:\, S_n = 1\}$$

is a stopping time in the natural filtration of the walk. We will check in Example 5.18 that T is a.s. finite, and $S_T = 1$ a.s. Also, $\mathbb{E}\, S_{T \wedge n} = \mathbb{E}\, S_0 = 0$ from the martingale property, as seen in Theorem 5.11. However, the conditions of Optional stopping are not met, and

$$1 = \mathbb{E}\, S_T \ne \mathbb{E}\, S_0 = 0.$$

In particular this shows via case (iii) of Optional stopping that $\mathbb{E}\, T = \infty$ since S_n is of bounded increment.

Example 5.18. Consider the SSRW and the stopping time T as in Example 5.17. We show $T < \infty$ a.s., and even calculate its moment generating function. Finiteness of T implies *recurrence* of SSRW.

For any $\theta > 0$, the process

$$M_n^\theta := \frac{e^{\theta S_n}}{(\cosh \theta)^n}$$

is a martingale. It is clearly adapted and finite for each n, and the martingale property works out:

$$\mathbb{E}\big(M_{n+1}^\theta \,|\, \mathcal{F}_n\big) = \frac{\mathbb{E}\big(e^{\theta(S_n + X_{n+1})} \,|\, \mathcal{F}_n\big)}{(\cosh \theta)^{n+1}} = \frac{e^{\theta S_n}}{(\cosh \theta)^{n+1}} \cdot \frac{e^\theta + e^{-\theta}}{2} = \frac{e^{\theta S_n}}{(\cosh \theta)^n} = M_n^\theta.$$

We cannot rule out $T = \infty$. However, in this case we know $S_n \leq 0$ for each n, hence

$$0 \leq M_n^\theta \leq \frac{e^{\theta \cdot 0}}{(\cosh \theta)^n} \xrightarrow[n \to \infty]{} 0$$

by $\cosh \theta > 1$. On the event $\{T = \infty\}$ we therefore define $M_T^\theta = 0$ and have $M_{T \wedge n}^\theta = M_n^\theta \xrightarrow[n \to \infty]{} M_T^\theta$. We also have $M_{T \wedge n}^\theta \xrightarrow[n \to \infty]{} M_T^\theta$ on $\{T < \infty\}$.

By definition of T,

$$M_{T \wedge n}^\theta = \frac{e^{\theta S_{T \wedge n}}}{(\cosh \theta)^{T \wedge n}} \leq \frac{e^\theta}{(\cosh \theta)^{T \wedge n}} \leq e^\theta,$$

which is a bound independent of n, hence dominated convergence applies on the above limit:

$$1 = \lim_{n \to \infty} \mathbb{E} \, M_{T \wedge n}^\theta = \mathbb{E} \lim_{n \to \infty} M_{T \wedge n}^\theta = \mathbb{E} \, M_T^\theta = \mathbb{E}(M_T^\theta \,;\, T < \infty) = \mathbb{E}\Big(\frac{e^{\theta S_T}}{(\cosh \theta)^T} \,;\, T < \infty\Big)$$

$$= \mathbb{E}\Big(\frac{e^\theta}{(\cosh \theta)^T} \,;\, T < \infty\Big) = \mathbb{E}\Big(\frac{e^\theta}{(\cosh \theta)^T}\Big),$$

where the martingale property $\mathbb{E} \, M_{T \wedge n}^\theta = M_0^\theta = 1$, then $M_T^\theta = 0$ on $\{T = \infty\}$, then $S_T = 1$ on $\{T < \infty\}$, finally $\cosh \theta > 1$ were used. From this we conclude

(5.2)
$$\mathbb{E}\Big(\frac{1}{(\cosh \theta)^T}\Big) = e^{-\theta}.$$

Next we take $\theta \searrow 0$, which makes $\cosh \theta \searrow 1$. If $T = \infty$ then $\frac{1}{(\cosh \theta)^T} = 0$ for any $\theta > 0$, hence so is its limit. However, if $T < \infty$ then the limit is 1. Together,

$$\lim_{\theta \searrow 0} \frac{1}{(\cosh \theta)^T} = \mathbf{1}\{T < \infty\}.$$

As the bound $0 \leq \frac{1}{(\cosh \theta)^T} \leq 1$ holds no matter the value of T, we can apply dominated convergence on the above limit:

$$1 = \lim_{\theta \searrow 0} e^{-\theta} = \lim_{\theta \searrow 0} \mathbb{E}\Big(\frac{1}{(\cosh \theta)^T}\Big) = \mathbb{E}\Big(\lim_{\theta \searrow 0} \frac{1}{(\cosh \theta)^T}\Big) = \mathbb{E} \, \mathbf{1}\{T < \infty\} = \mathbb{P}\{T < \infty\},$$

which was our goal.

Substitute $0 < \alpha = \frac{1}{\cosh \theta} < 1$ into (5.2) and solve this substitution for $e^{-\theta}$ to obtain (check!) the generating function for the random variable T

$$\mathbb{E} \, \alpha^T = \frac{1 - \sqrt{1 - \alpha^2}}{\alpha}.$$

6 Martingale convergence

This section gives an introduction to yet another strength of the martingale property: convergence under rather general assumptions. We start with examining *upcrossings*:

Definition 6.1. Fix $a < b$ real numbers and let X_n be a process. The *upcrossing number until N* is defined as

$$U_N[a, b] := \max\{k \,:\, \exists 0 \leq s_1 < t_1 < s_2 < t_2 < \cdots < s_k < t_k \leq N \text{ with } X_{s_i} < a, \, X_{t_i} > b \text{ for all } 1 \leq i \leq k\}.$$

This definition counts what the name suggests: how many times the process increases from below level a to above level b before time N. To capture these increments, we also define

$$C_0 := 0, \qquad C_n := \mathbf{1}\{C_{n-1} = 1\}\mathbf{1}\{X_{n-1} \le b\} + \mathbf{1}\{C_{n-1} = 0\}\mathbf{1}\{X_{n-1} < a\} \quad (n \ge 1).$$

Think about C_n as an ON-OFF switch: once it is ON ($C_{n-1} = 1$), it stays on unless X exceeds level b. Once it is OFF ($C_{n-1} = 0$), it stays OFF until X descends below level a. Notice that if X_n is adapted, then this C_n is predictable. Let

$$Y_n := (C \bullet X)_n = \sum_{k=1}^{n}(X_k - X_{k-1})C_k.$$

This captures the increments of X during the ON periods of C. It then follows that

$$Y_N \ge (b - a) \cdot U_N[a, b] - (X_N - a)^-.$$

To see this notice that any upcrossing is of increment at least $b - a$, all of which is captured by Y_N. However, if X descends below a after the last upcrossing, that is also captured by Y_N and might contribute with a negative sign. This is taken care of by the last term.

Lemma 6.2 (Doob's upcrossing lemma). *If X_n is a supermartingale, then $(b - a)\,\mathbb{E}\,U_N[a, b] \le \mathbb{E}(X_N - a)^-$.*

Proof. By Theorem 5.9, Y_n is a supermartingale, hence $\mathbb{E}\,Y_N \le 0$. $\qquad\square$

Corollary 6.3. *Let $a < b$, and let X_n be a supermartingale that is bounded in L^1. (That is, $\sup_n \mathbb{E}\,|X_n| < \infty$.) Then*

$$(b - a)\,\mathbb{E}\,U_\infty[a, b] \le |a| + \sup_m \mathbb{E}\,|X_m| < \infty.$$

In particular, $U_\infty[a, b]$ is a.s. finite.

Proof. As $U_N[a, b] \ge 0$ is monotone in N, we have

$$(b - a)\,\mathbb{E}\,U_\infty[a, b] = (b - a)\,\mathbb{E}\lim_N U_N[a, b] = (b - a)\lim_N \mathbb{E}\,U_N[a, b] \le \lim_N \mathbb{E}(X_N - a)^-$$

by monotone convergence. To finish the proof,

$$\mathbb{E}(X_N - a)^- \le \mathbb{E}\,|X_N - a| \le \mathbb{E}\,|X_N| + |a| \le |a| + \sup_m \mathbb{E}\,|X_m|,$$

and the right-hand side is independent of N. $\qquad\square$

Theorem 6.4 (Doob's forward convergence theorem). *Let X_n be a supermartingale that is bounded in L^1. Then $X_\infty := \lim_n X_n$ exists a.s. and is finite.*

Proof.

$$\{X_n \text{ does not converge}\} = \{\liminf_n X_n < \limsup_n X_n\} = \bigcup_{\substack{a < b \\ a, b \in \mathbb{Q}}} \{\liminf_n X_n < a < b < \limsup_n X_n\}$$

$$\subseteq \bigcup_{\substack{a < b \\ a, b \in \mathbb{Q}}} \{U_\infty[a, b] = \infty\}.$$

The right-hand side is a countable union of zero probability events, has zero probability itself.

To see that the limit is a.s. finite, use Fatou's lemma:

$$\mathbb{E}\,|X_\infty| = \mathbb{E}\liminf_n |X_n| \le \liminf_n \mathbb{E}\,|X_n| \le \sup_n \mathbb{E}\,|X_n| < \infty,$$

which implies $|X_\infty| < \infty$ a.s. $\qquad\square$

Notice that if $X_n \ge 0$ is a supermartingale, then L^1-boundedness is automatic: $\mathbb{E}\,|X_n| = \mathbb{E}\,X_n \le \mathbb{E}\,X_0$.

Finally, we have a bit to say about L^2-martingales. The *scalar product* of L^2-random variables X and Y is defined by $\langle X, Y \rangle := \mathbb{E}(XY)$. Indeed check that this defines a scalar product. If M_n is an L^2-martingale, and $k \le m < n$, then $M_n - M_m$ is orthogonal to $\mathcal{F}_k$ of the filtration. Namely, for any Y $\mathcal{F}_k$-measurable random variable

$$\langle M_n - M_m, Y \rangle = \mathbb{E}\big((M_n - M_m)Y\big) = \mathbb{E}\,\mathbb{E}\big((M_n - M_m)Y \mid \mathcal{F}_k\big) = \mathbb{E}\big(Y\,\mathbb{E}(M_n - M_m \mid \mathcal{F}_k)\big) = \mathbb{E}\big(Y \cdot (M_k - M_k)\big) = 0.$$

In particular, increments of an L^2-martingale in disjoint time intervals are orthogonal: for $\ell < k \le m < n$, $\langle M_n - M_m, M_k - M_\ell \rangle = 0$. Another way of stating this is that these increments are *uncorrelated*. This is the main observation used in the next theorem.

Theorem 6.5. *An L^2 martingale M_n is bounded in L^2 if and only if $\sum_{k=1}^{\infty} \mathbb{E}(M_k - M_{k-1})^2 < \infty$. In this case $M_n \to M_\infty$ a.s. and in L^2.*

Proof. Due to the orthogonal increments as above, all cross terms of the square below have zero mean, and the following *Pythagorean theorem* holds:

$$\mathbb{E}\, M_n^2 = \mathbb{E}\Big(M_0 + \sum_{k=1}^{n}(M_k - M_{k-1})\Big)^2 = \mathbb{E}\, M_0^2 + \sum_{k=1}^{n} \mathbb{E}(M_k - M_{k-1})^2 \nearrow \mathbb{E}\, M_0^2 + \sum_{k=1}^{\infty} \mathbb{E}(M_k - M_{k-1})^2$$

as $n \to \infty$. This proves the first statement: the left-hand side is bounded iff the infinite sum is finite.

If M_n is bounded in L^2, then it is also bounded in L^1 by Ljapunov's inequality, and the forward convergence theorem provides the a.s. limit. To see the L^2 convergence, use Fatou's lemma and the Pythagorean theorem as

$$\mathbb{E}(M_\infty - M_n)^2 = \mathbb{E}\lim_{r \to \infty}(M_{n+r} - M_n)^2 \leq \liminf_r \mathbb{E}(M_{n+r} - M_n)^2 = \liminf_r \sum_{k=n+1}^{n+r} \mathbb{E}(M_k - M_{k-1})^2$$

$$= \sum_{k=n+1}^{\infty} \mathbb{E}(M_k - M_{k-1})^2 \xrightarrow[n\to\infty]{} 0$$

due to finiteness of the infinite sum. $\qquad\square$

7 Doob decomposition

We briefly cover a very useful technique of separating a martingale from a stochastic process. This is called Doob decomposition and its continuous-time analogue, called Doob-Meyer decomposition, is the basis of stochastic integration.

Theorem 7.1 (Doob decomposition)**.** *Let $(X_n)_{n\geq 0}$ be an adapted process in L^1. Then*

(a) there is

- *a martingale $(M_n)_{n\geq 0}$ with $M_0 = 0$,*
- *a predictable process $(A_n)_{n\geq 0}$ with $A_0 = 0$*

such that $X_n = X_0 + M_n + A_n$. This decomposition is almost everywhere unique in the sense that for any other pair $(\widehat{M}_n, \widehat{A}_n)_{n\geq 0}$ with the above properties we have $\mathbb{P}\{M_n = \widehat{M}_n \text{ and } A_n = \widehat{A}_n \text{ for all } n \geq 0\} = 1$.

(b) $(X_n)_{n\geq 0}$ is a submartingale if and only if $\mathbb{P}\{A_n \leq A_{n+1} \text{ for all } n \geq 0\} = 1$ in the above decomposition.

Assuming this decomposition works, we can actually guess what A should be via a next step analysis:

$$(7.1) \qquad \begin{aligned} \mathbb{E}(X_k - X_{k-1} \,|\, \mathcal{F}_{k-1}) &= \mathbb{E}(M_k - M_{k-1} \,|\, \mathcal{F}_{k-1}) + \mathbb{E}(A_k - A_{k-1} \,|\, \mathcal{F}_{k-1}) \\ &= M_{k-1} - M_{k-1} + \mathbb{E}(A_k - A_{k-1} \,|\, \mathcal{F}_{k-1}) = A_k - A_{k-1}. \end{aligned}$$

We start the proof by summing this display for the definition of A.

Proof. Define $A_0 = 0$ and for $n \geq 1$

$$A_n := \sum_{k=1}^{n} \mathbb{E}(X_k - X_{k-1} \,|\, \mathcal{F}_{k-1}).$$

This is predictable due to $k - 1 \leq n - 1$ in the above sum and the properties of the conditional expectation. Then let

$$M_n := X_n - X_0 - A_n.$$

$M_0 = 0$, $M_n \in L^1$, and M being adapted are clear and, by separating the last term in the sum,

$$\mathbb{E}(M_n \,|\, \mathcal{F}_{n-1}) = \mathbb{E}(X_n - X_0 - A_n \,|\, \mathcal{F}_{n-1}) = \mathbb{E}(X_n \,|\, \mathcal{F}_{n-1}) - X_0 - \sum_{k=1}^{n} \mathbb{E}(X_k - X_{k-1} \,|\, \mathcal{F}_{k-1})$$

$$= X_{n-1} - X_0 - \sum_{k=1}^{n-1} \mathbb{E}(X_k - X_{k-1} \,|\, \mathcal{F}_{k-1}) = X_{n-1} - X_0 - A_{n-1} = M_{n-1}.$$

Hence the above decomposition has the required properties. For a.e. uniqueness notice that any decomposition has to satisfy (7.1), which leaves no other choice of A, hence no other choice of M (up to zero probability sets). Part (b) is coming from $A_n = X_n - X_0 - M_n$:

$$A_{n+1} - A_n = \mathbb{E}(A_{n+1} \mid \mathcal{F}_n) - A_n = \mathbb{E}(X_{n+1} \mid \mathcal{F}_n) - X_0 - \mathbb{E}(M_{n+1} \mid \mathcal{F}_n) - X_n + X_0 + M_n = \mathbb{E}(X_{n+1} \mid \mathcal{F}_n) - X_n.$$

$\square$

Definition 7.2. Let $(M_n)_{n\geq 0}$ be an L^2-martingale with $M_0 = 0$. Then M^2 has the a.e. unique Doob decomposition into a martingale N and a predictable process A:

$$M^2 = N + A.$$

The process A is often denoted as $\langle M \rangle$, and is called the *brackets process* of M.

It is easy to check, and we will see later, that $(M_n^2)_{n\geq 0}$ is a submartingale. It therefore follows that A is a.s. non-decreasing, with an a.s. limit $A_\infty := \lim_{n\to\infty} A_n$. Since $\mathbb{E} M_n^2 = \mathbb{E} A_n$, we have that M is bounded in L^2 if and only if $\mathbb{E} A_\infty < \infty$.

We also note that

$$A_n - A_{n-1} = \mathbb{E}(M_n^2 - M_{n-1}^2 \mid \mathcal{F}_{n-1}) = \mathbb{E}\big((M_n - M_{n-1})^2 \mid \mathcal{F}_{n-1}\big).$$

8 Uniform integrability

We have seen before that L^p convergence is stronger than convergence in probability. However, there is a condition that allows to conclude L^p convergence from convergence in probability. This is explored below, and will be used later for martingales.

Notice that by monotone convergence,

(8.1) $$\lim_{c\to\infty} \mathbb{E}\big(|X|^p\,;\,|X| \geq c\big) = 0 \qquad \text{for any } X \in L^p.$$

This helps understanding the following definition.

Definition 8.1. A sequence X_n of random variables is p^{th} *power uniformly integrable*, if

$$\lim_{c\to\infty} \sup_n \mathbb{E}\big(|X_n|^p\,;\,|X_n| \geq c\big) = 0.$$

The following lemma will help exploiting uniform integrability.

Lemma 8.2. Let $X \in L^1$. Then $\forall \varepsilon > 0 \; \exists \delta > 0$ such that $\forall F \in \mathcal{F}$ event with $\mathbb{P}(F) < \delta$, $\mathbb{E}\big(|X|\,;\,F\big) < \varepsilon$ holds.

Proof. By contradiction, assume that there is an $\varepsilon > 0$ and a sequence F_n of events such that $\mathbb{P}(F_n) < 2^{-n}$, but $\mathbb{E}\big(|X|\,;\,F_n\big) \geq \varepsilon$. Denote $H := \limsup_n F_n$. Then, on one hand, Borel-Cantelli 1 implies that $\mathbb{P}(H) = 0$. On the other hand, an application of Fatou's lemma on $-|X| \cdot \mathbf{1}_{F_n}$ gives

$$\mathbb{E}\big(|X|\,;\,H\big) = \mathbb{E}\big(\limsup_n |X| \cdot \mathbf{1}_{F_n}\big) \geq \limsup_n \mathbb{E}\big(|X|\,;\,F_n\big) \geq \varepsilon,$$

which is a contradiction. $\square$

We can reprove (8.1) with this lemma. If $X \in L^p$, fix $Y = |X|^p \in L^1$, $\varepsilon > 0$, and δ for this Y as in Lemma 8.2. For this δ, via Markov's inequality, there is a large enough K, such that

$$\mathbb{P}\{|X| \geq K\} = \mathbb{P}\{|X|^p \geq K^p\} \leq \frac{\mathbb{E}\,|X|^p}{K^p} < \delta.$$

Then, with $F = \{|X| \geq K\}$, the lemma says

$$\mathbb{E}\big(|X|^p\,;\,|X| \geq K\big) = \mathbb{E}(Y\,;\,F) < \varepsilon.$$

That is, by picking large enough K, we could bring the expectation below ε. This is equivalent to (8.1).

With the help of the above, we can now go from in probability convergence and uniform integrability to L^p convergence.

Theorem 8.3. Let $p \geq 1$, suppose $X, X_1, X_2, \ldots \in L^p$, and $X_n \xrightarrow{\mathbb{P}} X$. Then $(iv) \Rightarrow (i) \Leftrightarrow (ii) \Leftrightarrow (iii) \Leftarrow (v)$, where

(i) $X_n \xrightarrow{L^p} X$;

(ii) X_n is p^{th} power uniformly integrable;

(iii) $\mathbb{E}\,|X_n|^p \to \mathbb{E}\,|X|^p$;

(iv) there exists a $p < q < \infty$, such that $\sup_n \mathbb{E}\,|X_n|^q < \infty$;

(v) there exists a $Y \in L^p$, such that $\forall n, \quad |X_n| < Y$.

Below partial proof is given to this statement.

Proof of $(i) \Rightarrow (ii)$, for $p = 1$. Fix $\varepsilon > 0$, we seek K such that $\forall n$, $\mathbb{E}\big(|X_n|\,;\,|X_n| \geq K\big) \leq \varepsilon$.

By the assumed L^1-convergence, there is an N such that $\mathbb{E}\,|X - X_n| < \frac{\varepsilon}{2}$ whenever $n > N$. Lemma 8.2 provides positive $\delta_0, \delta_1, \delta_2, \ldots, \delta_N$ such that

$$\forall F, \text{ if } \mathbb{P}(F) < \delta_0, \text{ then } \mathbb{E}\big(|X|\,;\,F\big) < \frac{\varepsilon}{2}, \qquad \forall F, \text{ if } \mathbb{P}(F) < \delta_n, \text{ then } \mathbb{E}\big(|X_n|\,;\,F\big) < \varepsilon$$

for $1 \leq n \leq N$. Set $\delta = \min\{\delta_0, \delta_1, \ldots, \delta_N\}$ and notice that this is still positive.

When $n \leq N$, define $F_n = \{|X_n| > K\}$, and pick K large enough that $\mathbb{P}(F_n) < \delta$ for each $1 \leq n \leq N$. Notice that this is possible due to only finitely many of the X_n's for this case. By the above, we then have $\mathbb{E}\big(|X_n|\,;\,|X_n| \geq K\big) < \varepsilon$.

When $n > N$, we argue as follows. The assumed L^1 convergence implies boundedness in L^1: $\sup_r \mathbb{E}\,|X_r| < \infty$. By increasing K if necessary, we can achieve

(8.2) $$\frac{\sup_r \mathbb{E}\,|X_r|}{K} < \delta.$$

A triangle inequality gives (with common factor $\mathbf{1}\{|X_n| \geq K\}$)

$$\mathbb{E}\big(|X_n|\,;\,|X_n| \geq K\big) \leq \mathbb{E}\big(|X|\,;\,|X_n| \geq K\big) + \mathbb{E}\big(|X_n - X|\,;\,|X_n| \geq K\big) \leq \mathbb{E}\big(|X|\,;\,|X_n| \geq K\big) + \mathbb{E}\,|X_n - X|.$$

For the first term, take $F_n = \{|X_n| \geq K\}$, and notice $\mathbb{P}(F_n) \leq \frac{\mathbb{E}\,|X_n|}{K} < \delta$ due to (8.2). The choice we made with $\delta \leq \delta_0$ then bounds this term by $\frac{\varepsilon}{2}$. The second term is also bounded by $\frac{\varepsilon}{2}$ due to our initial choice of N and $n > N$. $\qquad\square$

Proof of $(ii) \Rightarrow (i)$, for $p = 1$. Define the cutoff function

$$\varphi_K(x) = \begin{cases} K, & \text{if } x > K, \\ x, & \text{if } -K \leq x \leq K, \\ -K, & \text{if } x < -K. \end{cases}$$

Fix $\varepsilon > 0$ and notice that by the assumed uniform integrability and by (8.1), there is a K for which

$$\mathbb{E}\big|\varphi_K(X_n) - X_n\big| < \frac{\varepsilon}{3}, \qquad \text{and} \qquad \mathbb{E}\big|\varphi_K(X) - X\big| < \frac{\varepsilon}{3}.$$

Also, $\big|\varphi_K(x) - \varphi_K(y)\big| \leq |x - y|$ for any $x, y \in \mathbb{R}$, hence $\varphi_K(X_n) \xrightarrow{\mathbb{P}} \varphi(X)$ holds via $X_n \xrightarrow{\mathbb{P}} X$ (check!). As $|\varphi(\cdot)| \leq K$, Dominated convergence implies $\mathbb{E}\big|\varphi_K(X_n) - \varphi(X)\big| \to 0$, in particular this can be brought below $\frac{\varepsilon}{3}$ for large n's. Combining via triangle inequality,

$$\mathbb{E}\,|X_n - X| \leq \mathbb{E}\big|X_n - \varphi_K(X_n)\big| + \mathbb{E}\big|\varphi_K(X_n) - \varphi(X)\big| + \mathbb{E}\big|\varphi_K(X) - X\big| < \varepsilon.$$

$\qquad\square$

Proof of $(i) \Rightarrow (iii)$. This is just two triangle inequalities:

$$||X_n||_p = ||X_n - X + X||_p \leq ||X_n - X||_p + ||X||_p$$
$$||X||_p = ||X - X_n + X_n||_p \leq ||X - X_n||_p + ||X_n||_p, \text{ i.e.,}$$
$$||X_n||_p \geq ||X||_p - ||X - X_n||_p.$$

As we assumed L^p convergence, $||X - X_n||_p \to 0$. This results in

$$||X||_p \leq \liminf_n ||X_n||_p \leq \limsup_n ||X_n||_p \leq ||X||_p,$$

that is liminf and limsup agree and $||X_n||_p \to ||X||_p$. $\qquad\square$

$(iii) \Rightarrow (i)$ is called Sceffé's theorem, it will not be used later on and its proof is somewhat tedious, therefore it is skipped. Those interested can ask me about it on drop-in sessions.

Proof of $(iv) \Rightarrow (ii)$. Given $p < q$, we start with a little exercise in algebra. Set

$$p_{\mathrm{H}} = \frac{q}{p} > 1, \qquad q_{\mathrm{H}} = \frac{q}{q-p} > 1, \qquad \text{and check } \frac{1}{p_{\mathrm{H}}} + \frac{1}{q_{\mathrm{H}}} = 1.$$

Apply Hölder's inequality on the variables $|X_n|^p$ and $\mathbf{1}\{X_n \geq c\}$, with these p_{H} and q_{H} parameters:

$$\mathbb{E}\big(|X_n|^p \,;\, X_n \geq c\big) \leq \big(\mathbb{E}\,|X_n|^q\big)^{\frac{p}{q}} \cdot \big(\mathbb{E}\,\mathbf{1}\{X_n \geq c\}^{\frac{q}{q-p}}\big)^{\frac{q-p}{q}} = \big(\mathbb{E}\,|X_n|^q\big)^{\frac{p}{q}} \cdot \big(\mathbb{P}\{X_n \geq c\}\big)^{\frac{q-p}{q}}.$$

Here we also used that raising and indicator to a positive power does not change a thing. Markov's inequality gives $\mathbb{P}\{X_n \geq c\} \leq \mathbb{P}\{|X_n|^q \geq c^q\} \leq \frac{\mathbb{E}\,|X_n|^q}{c^q}$, which further bounds the above by

$$\big(\mathbb{E}\,|X_n|^q\big)^{\frac{p}{q}} \cdot \frac{\big(\mathbb{E}\,|X_n|^q\big)^{\frac{q-p}{q}}}{c^{q-p}} = \frac{\mathbb{E}\,|X_n|^q}{c^{q-p}}.$$

Under the assumptions of (iv), taking $\lim_{c \to \infty} \sup_n$ brings this to 0. $\square$

Proof of $(v) \Rightarrow (i)$. This is just an application of Dominated convergence. $\square$

9 Uniformly integrable martingales

Uniform integrability gives further powerful tools with martingales:

Theorem 9.1. *Let M_n be a uniformly integrable martingale. Then $M_\infty = \lim_{n \to \infty} M_n$ exists a.s. and in L^1, and $M_n = \mathbb{E}(M_\infty \,|\, \mathcal{F}_n)$ a.s.*

Proof. Due to uniform integrability, we can pick a $c > 0$ for which $\sup_n \mathbb{E}(|M_n| \,;\, |M_n| \geq c) \leq 1$. Then,

$$\mathbb{E}\,|M_n| = \mathbb{E}(|M_n| \,;\, |M_n| \geq c) + \mathbb{E}(|M_n| \,;\, |M_n| < c) \leq 1 + c,$$

where we used the simple algebraic fact that $|x| \cdot \mathbf{1}\{|x| < c\} \leq c$ for any real x. As the right-hand side is independent of n, it follows that M_n is bounded in L^1, thus converges a.s. to a finite limit by Doob's forward convergence. That implies convergence in probability, which in turn gives L^1 convergence when uniform integrability is added (Theorem 8.3).

For the last bit, fix $r > n > 0$, $F \in \mathcal{F}_n$:

$$|\mathbb{E}(M_n \,;\, F) - \mathbb{E}(M_\infty \,;\, F)| = |\mathbb{E}(M_r \,;\, F) - \mathbb{E}(M_\infty \,;\, F)| = |\mathbb{E}(M_r - M_\infty \,;\, F)| \leq \mathbb{E}\big(|M_r - M_\infty| \,;\, F\big) \leq \mathbb{E}\,|M_r - M_\infty|$$

a.s., where first the martingale property (check how!), then Jensen's inequality on the $|\cdot|$ function was used. By the L^1 convergence, the right-hand side goes to 0 as $r \to \infty$. As the left-hand side has no r in it, it is therefore zero: $\mathbb{E}(M_n \,;\, F) = \mathbb{E}(M_\infty \,;\, F)$. This proves $M_n = \mathbb{E}(M_\infty \,|\, \mathcal{F}_n)$ a.s. due to the Kolmogorov definition of conditional expectations and M_n being $\mathcal{F}_n$-measurable. $\square$

The next theorem is the reverse statement in some sense.

Theorem 9.2 (Lévy's upwards theorem). *Let $\xi \in L^1$, $(\mathcal{F}_n)_n$ be a filtration, and $M_n := \mathbb{E}(\xi \,|\, \mathcal{F}_n)$ (this is well defined almost everywhere). Then M_n is a uniformly integrable martingale, M_n converges a.s. and in L^1 to a limit M_∞, and $M_\infty = \mathbb{E}(\xi \,|\, \mathcal{F}_\infty)$ a.s.*

Proof. The proof has three parts.

1. That M_n is a martingale is a simple application of the Tower rule.

2. Next we show that M_n is uniformly integrable. Fix $\varepsilon > 0$ and $\delta > 0$ for ξ as in Lemma 8.2. Then pick $K > \frac{\mathbb{E}\,|\xi|}{\delta}$. An application of Markov's inequality, Jensen's inequality on the conditional expectation, then the Tower rule shows

$$\mathbb{P}\{|M_n| \geq K\} \leq \frac{\mathbb{E}\,|M_n|}{K} = \frac{\mathbb{E}\big|\mathbb{E}(\xi \,|\, \mathcal{F}_n)\big|}{K} \leq \frac{\mathbb{E}\,\mathbb{E}\big(|\xi| \,\big|\, \mathcal{F}_n\big)}{K} = \frac{\mathbb{E}\,|\xi|}{K} < \delta,$$

making $\{|M_n| \geq K\}$ a suitable event for Lemma 8.2. We apply this lemma in the last step below, following conditional Jensen again, the fact that $\{|M_n| \geq K\} \in \mathcal{F}_n$ and the Tower rule:

$$\begin{aligned}
\mathbb{E}[|M_n| \,;\, |M_n| \geq K] &= \mathbb{E}\big[\big|\mathbb{E}(\xi \,|\, \mathcal{F}_n)\big| \,;\, |M_n| \geq K\big] \\
&\leq \mathbb{E}\big[\mathbb{E}\big(|\xi| \,\big|\, \mathcal{F}_n\big) \,;\, |M_n| \geq K\big] \\
&= \mathbb{E}\,\mathbb{E}\big(|\xi| \,;\, |M_n| \geq K \,\big|\, \mathcal{F}_n\big) = \mathbb{E}\big(|\xi| \,;\, |M_n| \geq K\big) \leq \varepsilon.
\end{aligned}$$

The right-hand side has no n, hence uniform integrability is proved. This implies existence of the limit M_∞.

3. We need to show that this limit M_∞ a.s. coincides with $\eta := \mathbb{E}(\xi \mid \mathcal{F}_\infty)$. We do this for the case $\xi > 0$, otherwise the difference $\xi = \xi^+ - \xi^-$ then provides the general proof. For any set $F \in \mathcal{F}_\infty$, let

$$\mathbb{Q}_1(F) := \mathbb{E}(\eta\,;\, F) \qquad \text{and} \qquad \mathbb{Q}_2(F) := \mathbb{E}(M_\infty\,;\, F),$$

these are measures on $(\Omega, \mathcal{F}_\infty)$. If $F \in \mathcal{F}_n$, then Tower rules, measurability (recall $\mathcal{F}_n \subseteq \mathcal{F}_\infty$), and $M_n = \mathbb{E}(M_\infty \mid \mathcal{F}_n)$ from the previous theorem imply

$$\mathbb{Q}_1(F) = \mathbb{E}(\eta\,;\, F) = \mathbb{E}\big(\mathbb{E}(\xi \mid \mathcal{F}_\infty)\,;\, F\big) = \mathbb{E}\,\mathbb{E}(\xi\,;\, F \mid \mathcal{F}_\infty) = \mathbb{E}(\xi\,;\, F)$$
$$= \mathbb{E}\,\mathbb{E}(\xi\,;\, F \mid \mathcal{F}_n) = \mathbb{E}\big(\mathbb{E}(\xi \mid \mathcal{F}_n)\,;\, F\big) = \mathbb{E}(M_n\,;\, F)$$
$$= \mathbb{E}\big(\mathbb{E}(M_\infty \mid \mathcal{F}_n)\,;\, F\big) = \mathbb{E}\,\mathbb{E}(M_\infty\,;\, F \mid \mathcal{F}_n) = \mathbb{E}(M_\infty\,;\, F) = \mathbb{Q}_2(F).$$

This shows that $\mathbb{Q}_1$ agrees with $\mathbb{Q}_2$ on $\bigcup_n \mathcal{F}_n$, hence also on $\mathcal{F}_\infty = \sigma\big(\bigcup_n \mathcal{F}_n\big)$. Since $\{\eta > M_\infty\} \in \mathcal{F}_\infty$,

$$0 = \mathbb{Q}_1\{\eta > M_\infty\} - \mathbb{Q}_2\{\eta > M_\infty\} = \mathbb{E}(\eta - M_\infty\,;\, \eta > M_\infty),$$

therefore $\mathbb{P}\{\eta > M_\infty\} = 0$. In a similar way $\mathbb{P}\{\eta < M_\infty\} = 0$, which completes the proof. $\qquad\square$

Next we prove an important theorem in probability using the machinery built up so far.

Definition 9.3. Let $X_1, X_2, \ldots$ be random variables, and $\mathcal{T}_n := \sigma(X_{n+1}, X_{n+2}, \ldots)$. The *tail σ-algebra* is $\mathcal{T} := \bigcap_n \mathcal{T}_n$.

An event in $\mathcal{T}$ is $\mathcal{T}_n$-measurable for every n. In other words, *it does not depend on any finite number of changes in the sequence $X_1, X_2, \ldots$*.

Theorem 9.4 (Kolmogorov's 0-1 law)**.** *The tail σ-algebra of independent variables is trivial. That is, with the above definition, for any $F \in \mathcal{T}$ we have $\mathbb{P}(F) = 0$ or 1.*

It is important here that the random variables are independent.

Proof. Let, as before, $\mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n)$ and $F \in \mathcal{T}$. Define $\eta = \mathbf{1}_F$. As $F \in \mathcal{T} \subseteq \mathcal{F}_\infty$, η is $\mathcal{F}_\infty$-measurable, and is of course in L^1. Add Lévy's upwards theorem:

$$\eta = \mathbb{E}(\eta \mid \mathcal{F}_\infty) = \lim_{n \to \infty} \mathbb{E}(\eta \mid \mathcal{F}_n).$$

However, $F \in \mathcal{T} \subseteq \mathcal{T}_n$, which is independent of $\mathcal{F}_n$, as these are generated by a disjoint set of independent random variables. It follows that

$$\mathbf{1}_F = \eta = \lim_{n \to \infty} \mathbb{E}(\eta \mid \mathcal{F}_n) = \lim_{n \to \infty} \mathbb{E}\,\eta = \mathbb{E}\,\eta = \mathbb{P}(F),$$

which shows $\mathbb{P}(F)$ is either 0 or 1. $\qquad\square$

Example 9.5. For a SSRW $S_n = \sum_{k=1}^n X_k$, let $F := \{\frac{S_n}{n} \to v\}$ be the event that the walk has asymptotic velocity v. Changing any finite number of the i.i.d. X_k's does not influence the liminf, nor the limsup, of $\frac{S_n}{n}$, hence F is in the tail σ-algebra of the X_k's. It follows that F is trivial, has either probability 0 or 1. Indeed, the Strong law of large numbers states that $\mathbb{P}(F) = 1$ when $v = 0$, and zero in all other cases.

Recall that a filtration is an increasing system of σ-algebras and a natural interpretation is the expanding information collected from observing a process starting at time zero. In what follows we still have a filtration, but the σ-algebras are indexed up to time -1, rather than starting from time 0 as before. Accordingly, the limit is taken as the σ-algebras decrease, rather than increase which is what has been done so far. An application follows further down.

Theorem 9.6 (Lévy's downward theorem)**.** *Let $\mathcal{G}_{-n}$, $n \geq 1$, be σ-algebras with*

$$\mathcal{G}_{-n} \subseteq \mathcal{G}_{-n+1} \subseteq \mathcal{G}_{-n+2} \subseteq \ldots \subseteq \mathcal{G}_{-1}, \qquad \text{and} \qquad \mathcal{G}_{-\infty} := \bigcap_{n=1}^\infty \mathcal{G}_{-n}.$$

Let $\gamma \in L^1$, and $M_{-n} := \mathbb{E}(\gamma \mid \mathcal{G}_{-n})$. Then $M_{-\infty} := \lim_{n \to \infty} M_{-n}$ exists a.s. and in L^1, and $M_{-\infty} = \mathbb{E}(\gamma \mid \mathcal{G}_{-\infty})$ a.s.

Proof. The martingale property $\mathbb{E}(M_{-n+1} \,|\, \mathcal{G}_{-n}) = M_{-n}$ is checked the very same way as in Example 5.6, except the time index is negative.

Uniform integrability works the same way as in Lévy's upward theorem. This implies L^1-boundedness, and the upcrossing proof works as well to show a.s. convergence. Together with uniform integrability, L^1 convergence follows.

Finally, to show $M_{-\infty} = \mathbb{E}(\gamma \,|\, \mathcal{G}_{-\infty})$ a.s., fix $r > 0$ and an event $G \in \mathcal{G}_{-\infty} \subseteq \mathcal{G}_{-r}$. Then

$$\mathbb{E}(\gamma \,;\, G) = \mathbb{E}\,\mathbb{E}(\gamma \,;\, G \,|\, \mathcal{G}_{-r}) = \mathbb{E}\big(\mathbb{E}(\gamma \,|\, \mathcal{G}_{-r}) \,;\, G\big) = \mathbb{E}(M_{-r} \,;\, G).$$

Further, due to Jensen's inequality and the L^1 convergence,

$$\big|\, \mathbb{E}(M_{-r} \,;\, G) - \mathbb{E}(M_{-\infty} \,;\, G)\big| = \big|\mathbb{E}(M_{-r} - M_{-\infty} \,;\, G)\big| \leq \mathbb{E}\big(|M_{-r} - M_{-\infty}| \,;\, G\big) \leq \mathbb{E}\,|M_{-r} - M_{-\infty}| \xrightarrow[r \to \infty]{} 0.$$

Hence $\mathbb{E}(\gamma \,;\, G) = \mathbb{E}(M_{-\infty} \,;\, G)$, and $M_{-\infty} = \mathbb{E}(\gamma \,|\, \mathcal{G}_{-\infty})$ a.s. follows from $M_{-\infty}$ being $\mathcal{G}_{-\infty}$-measurable and the Kolmogorov definition of conditional expectations. $\qquad\square$

As an application, here is the proof of the Strong law of large numbers using martingales.

Theorem 9.7 (Strong law of large numbers). *Let X_k be i.i.d. random variables in L^1, and $S_n = \sum_{k=1}^{n} X_k$. Then $\frac{S_n}{n} \to \mathbb{E}\,X_1$ a.s. and in L^1.*

Proof. Let $\mathcal{G}_{-n} = \sigma(S_n, S_{n+1}, \dots)$, and notice that this system satisfies the conditions of the Downward theorem. Notice also that due to independence and the structure of S_n,

$$\mathcal{G}_{-n} = \sigma\big\{\sigma(S_n), \sigma(X_{n+1}, X_{n+2}, \dots)\big\},$$

where the σ-algebras $\sigma(S_n), \sigma(X_{n+1}, X_{n+2}, \dots)$ are independent. This is to say that the process of S_m's, $m \geq n$ from time n is determined by S_n and, independently, X_k's for $k > n$. It follows by symmetry (check!) that

$$M_{-n} := \mathbb{E}(X_1 \,|\, \mathcal{G}_{-n}) = \mathbb{E}\big(X_1 \,|\, \sigma(S_n)\big) = \frac{S_n}{n}.$$

Lévy's downward theorem applies on this martingale, and gives the existence of the a.s. and L^1 limit $M_{-\infty}$. All that is left is to identify what this limit is. To this order notice that $M_{-\infty} = \lim_{n \to \infty} M_{-n} = \lim_{n \to \infty} \frac{S_n}{n}$ a.s. (one can instead use limsup here to make sure it is defined surely) does not depend on changes made on any finitely many of the X_k's. It is therefore in the tail σ-algebra of the i.i.d. sequence, which is trivial by Kolmogorov's 0-1 law. For any $c \in \mathbb{R}$, $\mathbb{P}\{M_{-\infty} = c\}$ is therefore 0 or 1. However, it cannot be 0 for every c, there exists a non-random value which $M_{-\infty}$ takes a.s. and, due to $\mathbb{E}\,M_{-\infty} = \mathbb{E}\,M_{-1} = \mathbb{E}\,X_1$, this value can only be $\mathbb{E}\,X_1$. $\qquad\square$

10 Doob's submartingale inequality

We continue with yet another important property of (sub)martingales. Compare the below with Markov's inequality.

Theorem 10.1 (Doob's submartingale inequality). *Let $Z_n \geq 0$ be a submartingale. Then for every $c > 0$ real and $n > 0$ integer,*

$$\mathbb{P}\Big\{\sup_{k \leq n} Z_k \geq c\Big\} \leq \frac{\mathbb{E}\big(Z_n \,;\, \sup_{k \leq n} Z_k \geq c\big)}{c} \leq \frac{\mathbb{E}\,Z_n}{c}.$$

Since we are looking at a finite number of values, 'max' would be appropriate instead of 'sup'. However, 'sup' is the usual formulation as it also works for the continuous time version of the theorem.

Proof. If the event $F := \big\{\sup_{k \leq n} Z_k \geq c\big\}$ occurs, then there must be a first instance of k where $Z_k \geq c$. This is captured by the disjoint union

$$F = \bigcup_{k=0}^{n} F_k, \qquad \text{where} \qquad \begin{aligned} &F_0 := \{Z_0 \geq c\}, \\ &F_k := \bigcap_{i=0}^{k-1} \{Z_i < c\} \cap \{Z_k \geq c\} \qquad (k > 0). \end{aligned}$$

Notice that $F_k \in \mathcal{F}_k$, hence the submartingale property gives, for $k \leq n$,

$$\mathbb{E}(Z_n \,;\, F_k) = \mathbb{E}\,\mathbb{E}\big((Z_n \,;\, F_k) \,|\, \mathcal{F}_k\big) = \mathbb{E}\big(\mathbb{E}(Z_n \,|\, \mathcal{F}_k) \,;\, F_k\big) \geq \mathbb{E}(Z_k \,;\, F_k).$$

The event F_k implies $Z_k \geq c$, hence we can proceed by

$$\mathbb{E}(Z_n \,;\, F_k) \geq \mathbb{E}(Z_k \,;\, F_k) \geq c\mathbb{P}(F_k).$$

Summing this in k proves the theorem via the disjoint union above. $\qquad\square$

Submartingales occur more often than one would first think. Let M_n be a martingale, g a convex function, and assume $\mathbb{E}\,|g(M_n)| < \infty$ for each n. Then by Jensen's inequality,

$$\mathbb{E}\big(g(M_{n+1})\,|\,\mathcal{F}_n\big) \geq g\big(\mathbb{E}(M_{n+1}\,|\,\mathcal{F}_n)\big) = g(M_n),$$

hence $g(M_n)$ is a submartingale.

Example 10.2. A total of n people queue up to buy one ticket each for a small performance with n seats in the theatre. The price is one pound, and each person independently, with equal chance either has exact change, or a two pound coin in which case a one pound coin change is needed from the cashier. The cashier starts the service with m one pound coins. We seek an upper bound on the probability that the cashier runs out of one pound coins at some point while serving this queue.

Set X_i to be 1 if person i has one pound, and -1 if they have a two pound coin. Then $S_k := \sum_{i=1}^{k} X_i$ is the change in the number of one pound coins with the cashier due to serving the first k customers. It is a SSRW hence a martingale, and the cashier runs out of one pound coins iff $m + \min_{1 \leq k \leq n} S_k < 0$.

A non-negative submartingale is produced by taking square (a convex function) of S_k. Hence, the probability of the cashier running out of change is bounded by

$$\mathbb{P}\big\{ \min_{1 \leq k \leq n} S_k < -m \big\} \leq \mathbb{P}\big\{ \sup_{1 \leq k \leq n} S_k^2 \geq (m+1)^2 \big\} \leq \frac{\mathbb{E}\,S_n^2}{(m+1)^2} = \frac{n}{(m+1)^2}.$$

As an example, $n = 100$ and $m = 30$ already gives a bound of cca 10%.

11   A discrete Black-Scholes option pricing formula

As an application, a very simple version of the option pricing formula is presented, still following Williams [2]. First, the probability space that will govern the stock market is assumed in this form:

$$(11.1) \qquad \Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}, \text{ where } \omega_n = \begin{cases} +1, & \text{with probability } p, \\ -1, & \text{with probability } 1 - p, \end{cases}$$

independently for different n's. This generates the natural filtration $\mathcal{F}_n = \sigma\{\omega_1, \ldots, \omega_n\}$.

Stocks have value S_n, while *bonds* have value B_n per unit on day n, $0 \leq n \leq N$. We have A_n stocks, and V_n bonds in the morning of day n, so our total wealth is $A_n S_n + V_n B_n$. During the day, we are allowed to exchange these, and by the evening of day n we might have A_{n+1} stocks and V_{n+1} bonds. However, the total wealth during transactions must be conserved:

$$(11.2) \qquad X_n := A_n S_n + V_n B_n = A_{n+1} S_n + V_{n+1} B_n.$$

Overnight, the values change from S_n to S_{n+1} for stocks, and from B_n to B_{n+1} for bonds. Bonds are not very exciting, their value is deterministic, $B_n = (1 + r)^n B_0$ with a fixed $-1 < r < \infty$, which one can also write as

$$B_n - B_{n-1} = r B_{n-1}.$$

Stocks, on the other hand, will change randomly. With $-1 < a < r < b < \infty$ also fixed, the random rates are governed by the probability space as

$$(11.3) \qquad R_n := \frac{a+b}{2} + \frac{b-a}{2} \omega_n = \begin{cases} b, & \text{if } \omega_n = 1, \\ a, & \text{if } \omega_n = -1, \end{cases}$$

and we have

$$S_n - S_{n-1} = R_n S_{n-1}.$$

The *European option* is a contract made on day 0. It allows (but does not force) us to buy, at the end of day N, a stock at *striking price* K. Its value on day N is therefore $(S_N - K)^+$ (when $S_N < K$, we do not use it). But how much does it worth on day 0?

Definition 11.1. A *hedging strategy* for the above option with initial value x is a predictable process (A_n, V_n) such that for every $\omega \in \Omega$,

- $X_0 = x$,
- $X_n \geq 0$ for all $0 \leq n \leq N$,

- $X_N = (S_N - K)^+$.

(A_n and V_n can possibly go negative.)

If there exists a hedging strategy for exactly one value of x, then this should be the price of the European option (for anything cheaper, everybody would buy it, and any more expensive, people would rather do the hedging strategy). This is exactly the case:

Theorem 11.2. *A hedging strategy as above exists if and only if*

$$x = (r+1)^{-N} \cdot \mathbb{E}(S_N - K)^+$$

with respect to Ω (11.1) with $p = \frac{r-a}{b-a}$. This strategy is unique, and features $A_n \geq 0$ for all $0 \leq n \leq N$.

To prove this theorem, we need two lemmas. Define the martingale (with $Z_0 := 0$)

(11.4)
$$Z_n := \sum_{k=1}^{n} (\omega_k - 2p + 1).$$

The space Ω is simple enough to derive any other martingale from this one:

Lemma 11.3. *Let M_n be any martingale in $\big(\Omega, (\mathcal{F}_n)_{n \geq 0}\big)$. Then there is a unique predictable process H_n, such that*

$$M_n = M_0 + (H \bullet Z) = M_0 + \sum_{k=1}^{n} H_k (Z_k - Z_{k-1}).$$

Proof. The proof is by brute force. We can write all random variables as functions of sequences of ω_k's. Those $\mathcal{F}_n$-measurable will be functions of $\omega_1, \omega_2, \ldots, \omega_n$ only. Now let us reverse-engineer what H_n needs to be:

$$M_n - M_{n-1} = \sum_{k=1}^{n} H_k(Z_k - Z_{k-1}) - \sum_{k=1}^{n-1} H_k(Z_k - Z_{k-1}) = H_n(Z_n - Z_{n-1}) = H_n(\omega_n - 2p + 1)$$
$$= H_n(2 - 2p)\mathbf{1}\{\omega_n = 1\} - H_n 2p\mathbf{1}\{\omega_n = -1\}.$$

From here, checking the two cases $\omega_n = \pm 1$,

(11.5)
$$H_n(\omega_1, \ldots, \omega_{n-1}, 1) = \frac{M_n(\omega_1, \ldots, \omega_{n-1}, 1) - M_{n-1}(\omega_1, \ldots, \omega_{n-1})}{2 - 2p},$$
$$H_n(\omega_1, \ldots, \omega_{n-1}, -1) = \frac{M_{n-1}(\omega_1, \ldots, \omega_{n-1}) - M_n(\omega_1, \ldots, \omega_{n-1}, -1)}{2p}.$$

However, H_n needs to be predictable, it cannot depend on ω_n. In other words, the two lines of this display must agree. This is where the martingale property for M_n comes handy:

$$M_{n-1} = \mathbb{E}(M_n \mid \mathcal{F}_{n-1}) = pM_n(\omega_1, \ldots, \omega_{n-1}, 1) + (1-p)M_n(\omega_1, \ldots, \omega_{n-1}, -1)$$

Rearranging,

$$(1-p)\big(M_{n-1} - M_n(\omega_1, \ldots, \omega_{n-1}, -1)\big) = p\big(M_n(\omega_1, \ldots, \omega_{n-1}, 1) - M_{n-1}\big)$$

exactly saying that the two lines of (11.5) indeed agree, and H_n does not depend on ω_n. The choice

$$H_n(\omega_1, \ldots, \omega_{n-1}) = \frac{M_n(\omega_1, \ldots, \omega_{n-1}, 1) - M_{n-1}}{2 - 2p} = \frac{M_{n-1} - M_n(\omega_1, \ldots, \omega_{n-1}, -1)}{2p}$$

will then work, and, following the proof, is also unique. $\qquad\square$

Definition 11.4. Following a hedging strategy (A_n, V_n), the *discounted value of our wealth* is

(11.6) $\qquad Y_n := (1+r)^{-n} \cdot X_n = (1+r)^{-n} \cdot (A_n S_n + V_n B_n) = (1+r)^{-n} \cdot (A_{n+1}S_n + V_{n+1}B_n)$

(recall (11.2)).

This is the amount of bonds we would need to buy at time zero if we were to achieve wealth X_n at time n purely via the fixed bond interest rate r.

Lemma 11.5. *For any hedging strategy (A_n, V_n), the discounted wealth (11.6) is a martingale in the probability space (11.1) with $p := \frac{r-a}{b-a}$. It can be obtained by transforming the martingale (11.4) as $Y_n = Y_0 + (H \bullet Z)_n$ by the unique predictable process*

$$(11.7) \qquad\qquad H_n = \frac{b-a}{2}(1+r)^{-n} \cdot A_n S_{n-1}.$$

Proof. Rewrite the definition (11.6) using $B_n = (1+r)^n B_0$:

$$Y_n = (1+r)^{-n} \cdot A_n S_n + V_n B_0 = (1+r)^{-n} \cdot A_{n+1} S_n + V_{n+1} B_0.$$

Its increment can be written as (use the first expression for Y_n and the second one for Y_{n-1})

$$Y_n - Y_{n-1} = (1+r)^{-n} \cdot A_n \big(S_n - (1+r)S_{n-1}\big) = (1+r)^{-n} \cdot A_n (R_n - r) S_{n-1}.$$

To proceed, assume $p = \frac{r-a}{b-a}$, that is

$$r = (b-a)p + a = \frac{a+b}{2} + \frac{b-a}{2}(2p-1),$$

the mean of R_n (11.3). Hence

$$Y_n - Y_{n-1} = \frac{b-a}{2}(1+r)^{-n} \cdot A_n S_{n-1}(\omega_n - 2p + 1) = H_n \cdot (\omega_n - 2p + 1),$$

with the definition (11.7) of H_n. This is equivalent to $Y_n = Y_0 + (H \bullet Z)_n$ with the martingale Z_n from (11.4). This implies that Y is a martingale, and the previous lemma then assures that no other predictable H can provide $Y = Y_0 + H \bullet Z$. $\qquad\square$

Proof of Theorem 11.2. Define the martingale

$$Y_n := (r+1)^{-N} \cdot \mathbb{E}\big[(S_N - K)^+ \,|\, \mathcal{F}_n\big].$$

We construct the hedging strategy with this being its discounted wealth as in (11.6). To do that, use Lemma 11.3 to find the unique predictable process H_n for this martingale, from which the predictable strategy A_n can be read off via (11.7). (Identity (11.2) then produces V_n as well.) The wealth process with this strategy is $X_n = (1+r)^n Y_n$, and we check $X_0 = (1+r)^0 Y_0 = (r+1)^{-N} \cdot \mathbb{E}(S_N - K)^+$ since $\mathcal{F}_0$ is trivial; $X_n \geq 0$ is obvious, and $X_N = (1+r)^N (1+r)^{-N}(S_N - K)^+ = (S_N - K)^+$ since all random variables are $\mathcal{F}_N$-measurable. Hence the (A_n, V_n) created this way is a hedging strategy for the European option with initial value as stated in the theorem.

To see uniqueness of this strategy, assume there is another one (A'_n, V'_n) hedging the same European option. Its discounted wealth Y'_n is a martingale that satisfies

$$Y'_N = (1+r)^{-N} X'_N = (r+1)^{-N} \cdot (S_N - K)^+.$$

The martingale property then implies

$$Y'_n = \mathbb{E}(Y'_N \,|\, \mathcal{F}_n) = (r+1)^{-N} \cdot \mathbb{E}\big[(S_N - K)^+ \,|\, \mathcal{F}_n\big],$$

thus $Y'_n = Y_n$, which in turn implies that $A'_n = A_n$ for each n. Uniqueness of the martingale in particular implies uniqueness of the initial value $x = X_0 = Y_0 = (r+1)^{-N} \cdot \mathbb{E}(S_N - K)^+$ of the hedging strategy.

It remains to show that $A_n \geq 0$ for each n under this strategy. This is equivalent to $H_n \geq 0$, and from the lemma we have

$$\begin{aligned} H_n &= \frac{Y_n(\omega_1, \ldots, \omega_{n-1}, 1) - Y_{n-1}}{2 - 2p} \\ &= \frac{(1+r)^{-N}}{2 - 2p}\big[\mathbb{E}((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}, 1) - \mathbb{E}((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1})\big]. \end{aligned}$$

Hence we need to prove

$$\begin{aligned} \mathbb{E}\big((S_N - K)^+ \,&|\, \omega_1, \ldots, \omega_{n-1}, 1\big) \\ &\geq \mathbb{E}\big((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}\big) \\ &= p\,\mathbb{E}\big((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}, 1\big) + (1-p)\,\mathbb{E}\big((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}, -1\big) \end{aligned}$$

which happens if and only if

$$\mathbb{E}\big((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}, 1\big) \geq \mathbb{E}\big((S_N - K)^+ \,|\, \omega_1, \ldots, \omega_{n-1}, -1\big).$$

To see this, first notice that the function $(S_N - K)^+$ is non-decreasing in each of the variables $\omega_1, \ldots, \omega_N$. Calculating the above conditional expectations involves summing over $\omega_{n+1}, \ldots, \omega_N$ each taking values ± 1. For every such outcome,

$$(S_N - K)^+(\omega_1, \ldots, \omega_{n-1}, 1, \omega_{n+1}, \ldots, \omega_N) \geq (S_N - K)^+(\omega_1, \ldots, \omega_{n-1}, -1, \omega_{n+1}, \ldots, \omega_N),$$

and the inequality survives the summation for the conditional expectations. $\square$

References

[1] Albert Nikolaevich Shiryaev. *Probability (2nd Ed.)*. Springer, 1996.

[2] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.