Measurable sets and functions

Definition 1 (σ-algebras) *Let Ω be a set and $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ be a collection of subsets of Ω:*
i. $\mathcal{A}$ is an algebra *if $\varnothing \in \mathcal{A}$ and for $A, B \in \mathcal{A}$, $A^c = \Omega \setminus A \in \mathcal{A}$ and $A \cup B \in \mathcal{A}$.*
ii. $\mathcal{A}$ is a σ-algebra *if $\varnothing \in \mathcal{A}$, for $A \in \mathcal{A}$, $A^c \in \mathcal{A}$, and for (A_n) a sequence of sets in $\mathcal{A}$, $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.*

A collection of sets is an algebra subject to being closed under finite applications of the basic operators. The σ-algebra concept extends this slightly to infinite ones. Consider where this distinction is relevant?

Note that if we have $\{\mathcal{F}_i : i \in I\}$ are σ-algebras, then
$$\mathcal{F} = \bigcap_{i \in I} \mathcal{F}_i$$
is a σ-algebra. This allows us to consider the notion of a smallest σ-algebra containing a set (the σ-algebra 'generated' by a set). We write the σ-algebra generated by a collection of collections of sets $\mathfrak{A}$ as $\sigma(\mathfrak{A})$.

Definition 2 (Borel σ-algebra) *Let $(E, \mathcal{T})$ be a topological space. The σ-algebra generated by the open sets in E is called the* Borel σ-algebra *on E and is denoted $\mathcal{B}(E) = \sigma(\mathcal{T})$.*

Definition 3 *Suppose $(\Omega_i, \mathcal{F}_i)_{i \in I}$ are measurable spaces. With $\Omega = \prod_{i \in I} \Omega_i$, $\mathcal{F}$ the σ-algebra generated by $A = \prod_{i \in I} A_i$ where $A_i \in \mathcal{F}_i$ for all $i \in I$ and for all but finitely many $i \in I$, $A_i = \Omega_i$: $(\Omega, \mathcal{F})$ is the product space.*

This space is measurable, and $\mathcal{F}$ is a σ-algebra.

Definition 4 (π and λ-systems) *A collection of sets $\mathcal{A}$ is called a π-system if it is closed under intersections.*

A collection of sets $\mathcal{M}$ is called a λ-system if $\Omega \in \mathcal{M}$, if $A, B \in \mathcal{M}$, $A \subseteq B$, then $B \setminus A \in \mathcal{M}$, and if $(A_n) \subseteq \mathcal{M}$ with $A_n \subseteq A_{n+1}$ increasing then $\bigcup_{n \geq 1} A_n \in \mathcal{M}$.

A collection of sets is a σ-algebra if and only if it is both a π-system and a λ-system.

Lemma 1 (π-λ systems lemma) *Let $\mathcal{A}$ be a π-system and $\mathcal{M}$ a λ-system. Then if $\mathcal{A} \subseteq \mathcal{M}$ then $\sigma(\mathcal{A}) \subseteq \mathcal{M}$.*

We can use this with a convenient π-system to show that our λ-system contains more than is immediately obvious.

Let $\lambda(\mathcal{A})$ be the smallest λ-system containing $\mathcal{A}$. This is a subset of $\mathcal{M}$ and $\sigma(\mathcal{A})$, so we just need to show that $\lambda(\mathcal{A})$ is a σ-algebra (for which we just have to show that it is a π-system).

Definition 5 (Random variables) *With measurable spaces $(\Omega, \mathcal{F})$, $(E, \mathcal{E})$, a function $f : \Omega \to E$ is said to be an E-valued random variable (or a measurable function) if for all $A \in \mathcal{E}$, $f^{-1}(A) \in \mathcal{F}$.*

We get immediately that random variables can be composed as one would expect. We can also use random variables to define new σ-algebras. Note that $(\Omega, \{f^{-1}(A) : A \in \mathcal{E}\})$ is a σ-algebra.

Definition 6 *With $\{f_i : i \in I\}$ a family of functions $\Omega \to E$, $\sigma(f_i : i \in I)$ is the smallest σ-algebra on Ω for which all f_i are measurable.*

This is initially a slightly intimidating definition, but the intuition is just that we need our $\sigma(f_i : i \in I) = \sigma(f_i^{-1}(A) : A \in \mathcal{E}, i \in I)$.

Theorem 2 (Monotone Class Theorem) *Let $\mathcal{H}$ be a class of bounded functions from $\Omega \to \mathbb{R}$ such that*
- *$\mathcal{H}$ is a vector space over $\mathbb{R}$,*
- *the constant function $1 \in \mathcal{H}$,*
- *if $(f_n) \subseteq \mathcal{H}$, $f_n \to f$ monotonically increasing, then $f \in \mathcal{H}$,*

then if $\mathcal{C} \subseteq \mathcal{H}$, and $\mathcal{C}$ is closed under pointwise multiplication, then all bounded $\sigma(\mathcal{C})$-measurable functions are in $\mathcal{H}$.

To get an intuition for this, note that any $f \in \mathcal{C}$ is necessarily bounded and $\sigma(\mathcal{C})$-measurable, but the converse is not immediate. Thus we essentially get a statement of the λ-π systems lemma but for functions on analogous systems.

We can firstly see that $\mathcal{H}$ is closed in $\mathcal{F}_b(\Omega)$. Then, we can prove the statement for the special case of $\mathcal{C} = \{\chi_A : A \in \mathcal{A}\}$ for a π-system $\mathcal{A}$, then adding 1 to $\mathcal{C}$ without loss of generality we can make the proof more general (concretely, because $\sigma(\mathcal{C}) \subseteq \sigma(\mathcal{C} \cup \{1\})$).

It may allow this theorem to make more sense to note that λ-systems are sometimes referred to as 'monotone classes'. Thus the π-λ systems lemma can be seen as saying that for $\mathcal{A}$ a π-system, the (smallest) monotone class generated by $\mathcal{A}$ is $\sigma(\mathcal{A})$.

We can use the monotone class theorem to demonstrate that for $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is measurable, then fixing $\omega_1 \in \Omega_1$, $\omega_2 \mapsto f(\omega_1, \omega_2)$ is measurable.

Conditional Probability

Up until presently, we've considered the notion of event A conditioned on event B as having a fixed probability. This doesn't entirely capture what a conditional is however – we're conditioning on the amount of information we have, and therefore we want the conditional probability to change as a function of our information. In particular, we want our conditional probability to be a function of $\omega \in \Omega$, and in order to reflect conditioning as a reflection of information, we want to condition over events in a σ-algebra, rather than individual events.

Doing more algebra, we see that expectation is a more fitting operator, leading us to the following definition:
Definition 7 (Conditional expectation) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G} \subseteq \mathcal{F}$ a σ-algebra. A random variable $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbb{P})$ is (a version of) the conditional expectation of X given $\mathcal{G}$ if for $G \in \mathcal{G}$,*
$$\mathbb{E}\big[Y\chi_G\big] = \mathbb{E}\big[X\chi_G\big].$$
The key aspect of this statement can be rewritten as
$$\int_G \mathbb{E}\big[X \,|\, \mathcal{G}\big] \,\mathrm{d}\mathbb{P} = \int_G X \,\mathrm{d}\mathbb{P},$$
which allows us to carry over all of our normal integration properties to conditional expectations.

Theorem 3 *The conditional expectation of X given $\mathcal{G}$ exists, denoted $\mathbb{E}\big[X \,|\, \mathcal{G}\big]$, and if Z is also the conditional expectation of X given $\mathcal{G}$, then*
$$Z = \mathbb{E}\big[X \,|\, \mathcal{G}\big] \ \text{ a.s.}$$
Come back to the proof of this.

Note importantly how conditional expectations behave with respect to measurability. If we have X a $\mathcal{G}$-measurable random variable, then
$$\mathbb{E}\big[X \,|\, \mathcal{G}\big] \overset{\text{a.s.}}{=} X.$$
Meanwhile if $\sigma(X)$ and $\mathcal{G}$ are independent, then
$$\mathbb{E}\big[X \,|\, \mathcal{G}\big] \overset{\text{a.s.}}{=} \mathbb{E}\big[X\big].$$

Lemma 4 (Tower property) *Take $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}_1$, $\mathcal{F}_2$ both σ-algebras, satisfying $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}$. Then*
$$\mathbb{E}\Big[\mathbb{E}\big[X \,|\, \mathcal{F}_2\big] \,|\, \mathcal{F}_1\Big] \overset{\text{a.s.}}{=} \mathbb{E}\big[X \,|\, \mathcal{F}_1\big].$$
This should be relatively intuitive – $\mathbb{E}\big[X \,|\, \mathcal{F}_2\big]$ contains more information than can be represented in $\mathcal{F}_1$, but is fundamentally still expressing a reduced form of X, which can be reduced more to give $\mathbb{E}\big[X \,|\, \mathcal{F}_1\big]$.

One can also consider this as a commutativity statement: as $\mathbb{E}\big[X \,|\, \mathcal{F}_1\big]$ is $\mathcal{F}_2$-measurable, thus $\mathbb{E}\big[X \,|\, \mathcal{F}_1\big] = \mathbb{E}\big[\mathbb{E}\big[X \,|\, \mathcal{F}_1\big] \,|\, \mathcal{F}_2\big]$, so the tower property is stating that with $\mathcal{F}_1 \subseteq \mathcal{F}_2$:
$$\mathbb{E}\Big[\mathbb{E}\big[X \,|\, \mathcal{F}_1\big] \,|\, \mathcal{F}_2\Big] \overset{\text{a.s.}}{=} \mathbb{E}\Big[\mathbb{E}\big[X \,|\, \mathcal{F}_2\big] \,|\, \mathcal{F}_1\Big].$$

I'm tempted to claim the more general statement, that for $\mathcal{F}_1$, $\mathcal{F}_2$ both σ-algebras in $\mathcal{F}$:
$$\mathbb{E}\Big[\mathbb{E}\big[X \,|\, \mathcal{F}_1\big] \,|\, \mathcal{F}_2\Big] \overset{\text{a.s.}}{=} \mathbb{E}\big[X \,|\, \mathcal{F}_1 \cap \mathcal{F}_2\big].$$
This statement is true if $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent, because then both sides are equal to $\mathbb{E}[X]$, but it seems that there could be a 'middle-ground' between independence and containment for which commutativity stops holding.

On the other side, attempting to prove this statement, the tripping point is that it's unclear that the LHS is $\mathcal{F}_1$-measurable (although clearly it is $\mathcal{F}_2$-measurable).

Lemma 5 *Take X, Y random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with X, Y, and XY integrable. Then*
$$\mathbb{E}\big[XY \,|\, \sigma(Y)\big] \overset{\text{a.s.}}{=} Y\mathbb{E}\big[X \,|\, \sigma(Y)\big].$$
Ensure for yourself that it's clear why this implies the same holding for $\mathcal{G} \supseteq \sigma(Y)$ instead of $\sigma(Y)$.

Theorem 6 *Take $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, $\{\mathcal{F}_i : i \in I\}$ a family of σ-algebras in $\mathcal{F}$. Then $\big\{\mathbb{E}\big[X \,|\, \mathcal{F}_i\big] : i \in I\big\}$ is uniformly integrable.*

Using conditional expectation, we can introduce an inner product to $\mathcal{L}^2$, $\langle X, Y \rangle := \mathbb{E}[XY]$ ('introduce' is probably slightly strong, this inner product already exists for other purposes in functional analysis – although we usually use the Lebesgue measure). This gives us that $\mathcal{L}^2$ is a Hilbert space, and all of the corresponding results.

Measures on ℝ

Definition 8 *A measure space is a triple $(\Omega, \mathcal{F}, \mu)$ such that Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and $\mu : \mathcal{F} \to [0, \infty]$ is countably additive (μ is then a* measure *on $(\Omega, \mathcal{F})$).*

Definition 9 *Let μ be a probability measure on $\mathcal{B}(\mathbb{R})$. The distribution function of μ is $F_\mu(x) = \mu(-\infty, x]$, where we require that F_μ is non-decreasing, tends to 0 as $x \to -\infty$, to 1 as $x \to \infty$, and is right continuous.*

Definition 10 *We say that ν is absolutely continuous with respect to μ, $\nu \ll \mu$, if for any $A \in \mathcal{F}$, $\mu(A) = 0$ implies that $\nu(A) = 0$. Further, we say that μ and ν are equivalent, $\mu \sim \nu$, if $\mu \ll \nu$ and $\nu \ll \mu$.*

Extensions

For the most part, it's difficult to characterise a measure explicitly, due to σ-algebras being incredibly large in all but countable Ω. We therefore wish to characterise them in terms of their value on algebras.

Theorem 7 (Uniqueness of extension) *Let μ_1 and μ_2 be measures on a space $(\Omega, \mathcal{F})$, and $\mathcal{A} \subseteq \mathcal{F}$ is a π-system with $\sigma(\mathcal{A}) = \mathcal{F}$. Then if $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ and $\mu_1\big|_{\mathcal{A}} = \mu_2\big|_{\mathcal{A}}$, then $\mu_1 = \mu_2$.*

This follows immediately via the λ-π systems lemma.

Theorem 8 (Carathéodory Extension theorem) *Let Ω be a set and $\mathcal{A}$ an algebra on Ω, then with $\mu_0 : \mathcal{A} \to [0, \infty]$ a countably additive set function, there exists a measure $\mu : \sigma(\mathcal{A}) \to [0, \infty]$ such that $\mu\big|_{\mathcal{A}} = \mu_0$.*

One can derive this from defining the outer measure μ^* in terms of μ_0, and claiming that a set is measurable iff for all $E \subseteq \Omega$, $\mu^*(E) = \mu^*(E \cap B) + \mu^*(E \setminus B)$. We can then prove that this gives the smallest σ-algebra containing $\mathcal{A}$.

Definition 11 (Distribution function) *If a function $F : \mathbb{R} \to [0, 1]$ satisfies:*
i. F is non-decreasing;
ii. $F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$; and
iii. F is continuous from the right,
then F is a distribution function.

Theorem 9 *Let F be a distribution function. Then there exists a unique Borel probability measure μ on $\mathbb{R}$ such that $\mu(-\infty, x] = F(x)$. Further, every Borel probability measure on $\mathbb{R}$ defines a distribution function.*

A corollary of this is that there is a unique Borel measure such that for all $a < b \in \mathbb{R}$, $\mu(a, b] = b - a$.

This result demonstrates that there is a bijection between measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and distribution functions. In particular, we call these measures the Lebesgue-Stieltjes measures.

The proof of this theorem follows using both of the extension theorems. In particular, we use the algebra of left open right closed intervals.

Definition 12 (Pushforward measure) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X : \Omega \to E$, then for $A \in \mathcal{E}$,*
$$\mathbb{Q}(E) := \mathbb{P}(X^{-1}(E))$$

Thus we have a (very non-injective) map from random variables in E to probability measures on E. This is useful, on the basis that many properties of random variables will just be properties of the pushforward measure rather than the random variable itself.

Theorem 10 *For $\{(\Omega_i, \mathcal{F}_i, \mathbb{P}_i) : i \in \{1, \dots, n\}\}$ a set of probability measures, there is a unique measure $\mathbb{P}$ on $\left(\prod_{i=1}^n \Omega_i, \bigtimes_{i=1}^n \mathcal{F}_i\right)$ such that for $E_i \in \mathcal{F}_i$ with $i \in \{1, \dots, n\}$,*
$$\mathbb{P}\left(\prod_{i=1}^n E_i\right) = \prod_{i=1}^n \mathbb{P}_i(E_i).$$
It's hopefully natural here that one should aim an induction proof.

The theorem here then allows us to extend the matter to infinite products, although at this point we require that what we're dealing with are probability measures (to keep each term in $[0, 1]$ for convergence reasons), rather than just finite measures as could work with the previous statement of the theorem.

Independence

Definition 13 (Independence) *With $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $(\mathcal{G}_i)_{i=1}^n$ a collection of σ-algebras, these σ-algebras are independent if for $E_i \in \mathcal{G}_i$ for $i \in \{1, \dots, n\}$*
$$\mathbb{P}\left(\bigcap_{i=1}^n E_i\right) = \prod_{i=1}^n \mathbb{P}(E_i).$$
Further, an arbitrary collection $(G_i)_{i \in I}$ of σ-algebras is independent if any finite subset of the collection is independent.

Note that this means $\{\varnothing, \Omega\}$ is independent of anything else.

Additionally, we say that a set $(X_i)_{i \in I}$ of random variables is independent iff $(\sigma(X_i))_{i \in I}$ is independent.

This definition requires a bit of work to deal with properly. One of the best general results we can attain quickly gives a fairly applicable result:
Theorem 11 *With $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $(\mathcal{A}_i)_{i \in I}$ an arbitrary collection of π-systems, then $(\sigma(\mathcal{A}_i))_{i \in I}$ are independent iff for any finite $J \subseteq I$, $A_i \in \mathcal{A}_i$ for $i \in J$:*
$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$
We also have the result that for any independent set of σ-algebras, any subset is also independent.

It takes a small bit of proving, but from the above results we get the lemma:
Lemma 12 *With $(\Omega, \mathcal{F}, \mathbb{P})$, a family of independent random variables $X_i : \Omega \to E_i$, measurable functions $f_i : E_i \to \mathbb{R}$ for $i \in I$, then $(f(X_i))_{i \in I}$ are independent.*

Tail events

Definition 14 *For a sequence of random variables (X_n), the tail σ-algebra is defined as*
$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(\{X_k : k > n\})$$
The intuition here is that all events in the tail σ-algebra contain sample information distinguishing the results of functions of infinitely many sequence elements.

Theorem 13 (Kolmogorov's 0-1 Law) *Let (X_n) be a sequence of independent random variables. Then the tail σ-algebra of (X_n) contains only events with probability 0 or 1.*

To see this, we demonstrate that $\mathcal{T}$ is independent of a σ-algebra containing it, and therefore that all of its events are independent with themselves.

Borel-Cantelli lemmas

Definition 15 *With (A_n) a sequence of sets from $\mathcal{F}$:*
$$\begin{aligned} \limsup_{n\to\infty} A_n &= \bigcap_{n=1}^{\infty} \bigcup_{m \geq n} A_m \\ &= \{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\} \\ &= \{A_n \text{ infinitely often}\} \\ \text{and} \quad \liminf_{n\to\infty} A_n &= \bigcup_{n=1}^{\infty} \bigcap_{m \geq n} A_m \\ &= \{\omega \in \Omega : \omega \in A_n \text{ eventually }\} \\ &= \{A_n \text{ eventually}\} \end{aligned}$$

Lemma 14 (Fatou and Reverse Fatou for sets) *With (A_n) a sequence of sets in $\mathcal{F}$,*
$$\mathbb{P}(\liminf_{n\to\infty} A_n) \leq \liminf_{n\to\infty} \mathbb{P}(A_n)$$
$$\mathbb{P}(\limsup_{n\to\infty} A_n) \geq \limsup_{n\to\infty} \mathbb{P}(A_n).$$

Lemma 15 (First Borel-Cantelli lemma) *For (A_n) a sequence of events in $\mathcal{F}$, if*
$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$
then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Lemma 16 (Second Borel-Cantelli lemma) *For (A_n) a sequence of independent events in $\mathcal{F}$, if*
$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$
then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

By their nature, the BC lemmas are only informative in relation to almost sure events. While this may seem incredibly limited, by Kolmogorov's 0-1 Law, it turns out that many events of interest are in fact almost sure events.

Integration

As already covered in Part A Integration, we define integration as normal:
Definition 16 (Integral on simple functions) *For a measure space $(\Omega, \mathcal{F}, \mu)$, $f : \Omega \to [0, \infty]$ a non-negative simple function taking values $\{a_1, \dots, a_n\} \subseteq \mathbb{R}$:*
$$\int f \,\mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu\left(f^{-1}(\{a_i\})\right).$$

Definition 17 (Integral on non-negative functions) *For $f : \Omega \to [0, \infty]$ a non-negative measurable function:*
$$\int f = \sup\left\{\int g \,\mathrm{d}\mu : g \text{ simple}, 0 \leq g \leq f\right\}$$

Definition 18 (Integral) *For $f : \Omega \to \mathbb{R}$ a measurable function and $\int |f| \,\mathrm{d}\mu < \infty$, we write $f^+ = \max(f, 0)$, $f^- = -\min(f, 0)$, and*
$$\int f \,\mathrm{d}\mu = \int f^+ \,\mathrm{d}\mu - \int f^- \,\mathrm{d}\mu$$

Theorem 17 (Monotone convergence theorem) *For (f_n) a sequence of non-negative functions measurable on $(\Omega, \mathcal{F}, \mu)$, such that $f_n \to f$ monotonically. Then*
$$\int f_n \,\mathrm{d}\mu \to \int f \,\mathrm{d}\mu.$$

Theorem 18 (Fatou's lemma) *For (f_n) a sequence of non-negative functions measurable on $(\Omega, \mathcal{F}, \mu)$,*
$$\int \liminf_{n\to\infty} f_n \,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int f_n \,\mathrm{d}\mu$$

Lemma 19 (Reverse Fatou's lemma) *For (f_n) a sequence of non-negative functions measurable on $(\Omega, \mathcal{F}, \mu)$, assume that there is an integrable function g such that $f_n \leq g$ for $n \geq 1$. Then*
$$\int \limsup_{n\to\infty} f_n \,\mathrm{d}\mu \geq \limsup_{n\to\infty} \int f_n \,\mathrm{d}\mu$$

Theorem 20 (Dominated convergence theorem) *For (f_n) a sequence of functions measurable on $(\Omega, \mathcal{F}, \mu)$ with $f_n \to f$ pointwise. Assume that there is an integrable function g such that $|f_n| \leq g$ for $n \geq 1$. Then*
$$\int f_n \,\mathrm{d}\mu \to \int f \,\mathrm{d}\mu.$$

Theorem 21 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $X : \Omega \to E$, $g : E \to \mathbb{R}$ measurable on their respective spaces. Then g is $(\mathbb{P} \circ X^{-1})$-integrable iff $g \circ X$ is $\mathbb{P}$-integrable. Further,*
$$\int_E g(x) \,\mathrm{d}(\mathbb{P} \circ X^{-1})(x) = \int_\Omega g(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega).$$

Definition 19 (Expectation) *For $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$,*
$$\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\mathbb{P}.$$

We say that X admits an nth moment if $X \in \mathcal{L}^n(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 20 (Variance) *For $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$,*
$$\operatorname{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

Definition 21 (Standardised moment) *If $X \in \mathcal{L}^n(\Omega, \mathcal{F}, \mathbb{P})$, the nth standardised moment of X is*
$$\mathbb{E}\left[\left(\frac{X - \mathbb{E}[X]}{\sqrt{\operatorname{Var}(X)}}\right)^n\right]$$

Theorem 22 (Fubini-Tonelli) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the product of probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ for $i \in \{1, 2\}$, and $f : \Omega \to \mathbb{R}$ is a bounded measurable function. Then both*
$$x \mapsto \int_{\Omega_2} f(x, y) \,\mathrm{d}\mathbb{P}_2(y)$$
$$\text{and} \quad y \mapsto \int_{\Omega_1} f(x, y) \,\mathrm{d}\mathbb{P}_1(x)$$
are measurable (respectively in $\mathcal{F}_1$ and $\mathcal{F}_2$).

If either $f \geq 0$ or f is $\mathbb{P}$-integrable over Ω, then
$$\int_\Omega f \,\mathrm{d}\mathbb{P} = \int_{\Omega_2} \int_{\Omega_1} f(x, y) \,\mathrm{d}\mathbb{P}_1(x) \,\mathrm{d}\mathbb{P}_2(y) = \int_{\Omega_1} \int_{\Omega_2} f(x, y) \,\mathrm{d}\mathbb{P}_2(y) \,\mathrm{d}\mathbb{P}_1(x)$$

Radon-Nikodym theorem

Integration as we've defined it gives a canonical method of defining a measure on a space: as the integral of a fixed non-negative measurable function over the set being measured.

Concretely: with a measure space $(\Omega, \mathcal{F}, \mu)$, a measurable function $f : \Omega \to [0, \infty]$, $A \in \mathcal{F}$,
$$\nu(A) := \int_A f \,\mathrm{d}\mu$$
is a measure on $\mathcal{F}$ (via MCT).

We therefore want to characterise how often a measure can be characterised in this way (in particular, whether we can get this the case with respect to leb or the counting measure, both for which we have a wealth of tools).

Theorem 23 (Radon-Nikodym theorem) *Let μ, ν be two probability measures on a σ-algebra $(\Omega, \mathcal{F})$. Then $\nu \ll \mu$ if and only if there is a measurable function $f : \Omega \to [0, \infty]$ such that for $A \in \mathcal{F}$,*
$$\nu(A) = \int_A f \,\mathrm{d}\mu.$$
Further, $\nu \sim \mu$ if and only if $\mu(f^{-1}(\{0\})) = \nu(f^{-1}(\{0\})) = 0$.

We call f the radon-nikodym derivative of ν with respect to μ, $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. If $\nu \sim \mu$, then $\frac{1}{f} = \frac{\mathrm{d}\mu}{\mathrm{d}\nu}$.

This means that providing $\mathrm{leb}(A) = 0$ implies that $\nu(A) = 0$, we can construct ν in this way.

Using this theorem, we can define for $A, B \in \mathcal{F}$ the conditional distribution $\mathbb{P}(A \,|\, B)$, provided $\mathbb{P}(B) > 0$. If $\mathbb{P}(A) = 0$, then $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 0$, so there is some $f_B : \Omega \to [0, \infty]$ measurable such that
$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \int_A f_B(\omega) \,\mathrm{d}\mathbb{P}(\omega).$$

Uniform Integrability

We've now introduced the notions of convergence in $\mathcal{L}^1$, and seen that it implies convergence in probability. We'd quite like to go the other direction however, and find a sufficient condition for which the two are the same.

Definition 22 (Uniform integrability) *A collection $\mathcal{C}$ of random variables is called uniformly integrable (UI) if*
$$\lim_{N\to\infty} \sup_{X \in \mathcal{C}} \mathbb{E}[|X|\chi_{|X|>N}] = 0$$

Importantly, this is a property of collections, rather than individual random variables. The larger our collection is, the less likely it is to hold (and conversely, if $\mathcal{C}$ is UI, then $\mathcal{D} \subseteq \mathcal{C}$ is also UI). We can see immediately therefore that this is a property that can only hold for collections of random variables in $\mathcal{L}^1$, by considering the singleton sets which can be UI.

If we have a $Y \in \mathcal{L}^1$ such that for $X \in \mathcal{C}$, $|X| \leq Y$, then $\mathcal{C}$ is uniformly integrable. This is one of the most common methods for demonstrating a collection is UI.

Another useful characterisation is found by reframing $|X|\chi_{|X| \geq N}$. From the below fact, we can see that we can replace it with $(|X| - N)^+$ in the definition of uniform integrability without effect:
$$0 \leq (|X| - N)^+ \leq |X|\chi_{|X| \geq N} \leq 2(|X| - N/2)^+.$$
A common formulation that allows us to avoid having to determine the sets on which $|X| \geq N$ is given below:
Lemma 24 *Let $\mathcal{C}$ be a family of random variables. Then $\mathcal{C}$ is UI iff $\sup_{X \in \mathcal{C}} \mathbb{E}\big[|X|\big] < \infty$ and*
$$\sup\left\{\mathbb{E}\big[|X|\chi_A\big] : X \in \mathcal{C}, A \in \mathcal{F}, \mathbb{P}(A) \leq \delta\right\} \to 0$$
as $\delta \to 0$.

Theorem 25 (Vitali's convergence theorem) *Take (X_n) a sequence of integrable random variables which converge in probability to a random variable X. The following are equivalent:*
i. $\{X_n : n \geq 1\}$ is uniformly integrable.
ii. $X \in \mathcal{L}^1$ and $\mathbb{E}\big[|X_n - X|\big] \to 0$ as $n \to \infty$.
iii. $X \in \mathcal{L}^1$ and $\mathbb{E}\big[|X_n|\big] \to \mathbb{E}\big[|X|\big]$ as $n \to \infty$.

It's worth noting that (ii) is the definition of convergence in $\mathcal{L}^1$, from which we get (iii) just via the reverse triangle inequality. The difficult part of this proof is demonstrating that (iii) entails (i) which requires that characterisation of UI in terms of $(|X| - N)^+$ in conjunction with convergence in probability.

By a non-examinable result (Dunford-Pettis), a collection is UI iff its closure is compact in $\sigma(L^1, L^\infty)$, which is the weak topology on L^1.

Filtrations

Definition 23 (Filtrations) *For $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, a filtration $(\mathcal{F}_n)$ is an sequence of σ-algebras ($\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for $n \geq 1$). We then call $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ a filtered probability space.*

Definition 24 (Adapted stochastic process) *Take (X_n) a sequence of random variables, $(\mathcal{F}_n)$ a filtration. If for $n \geq 1$, X_n is $\mathcal{F}_n$ measurable, then (X_n) is said to be adapted to $(\mathcal{F}_n)$.*

Note that for a stochastic process (X_n), we get a natural filtration
$$\mathcal{F}_n := \sigma\left(\{X_k : k \leq n\}\right).$$
This is the smallest filtration possible.

It's worth noting that with $\mathcal{F}_\infty = \bigcup_{n \geq 1} \mathcal{F}_n$, we don't necessarily have that $\mathcal{F}_\infty = \mathcal{F}$. Nonetheless, once we're working primarily within the filtered space, this shouldn't make a significant amount of difference, because the remainder of events will be inaccessible to any adapted process.

Definition 25 (Stopping time) *Take $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ a filtered probability space. A random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ is a stopping time with respect to $(\mathcal{F}_n)$ if for $n \geq 1$, $\tau^{-1}(\{n\}) \in \mathcal{F}_n$.*

Immediate examples of stopping times can be given. Any constant is a stopping time, as are the maxima or minima of any two stopping times. Further, the first hitting time for an adapted stochastic process on any measurable set is a stopping time.

Definition 26 *Let τ be a stopping time on $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$. The σ-algebra of information at time τ is*
$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : \text{for } n \geq 1, A \cap \{\tau = n\} \in \mathcal{F}_n\}.$$
This definition isn't entirely straightforward to understand, but the key aspect is this: if we have the value of τ (say we know that $\omega \in \{\tau = n\}$), then we shouldn't be able to infer from $\omega \in A \in \mathcal{F}_\tau$ that $\omega \in B \in \mathcal{F}_\infty \setminus \mathcal{F}_n$. The above definition should then be clear as containing the events which are possible to have observed by time τ ('time' being defined in accordance to steps in the filtration).

We can then get our intuition that if for stopping times τ and ρ, $\tau \leq \rho$, then $\mathcal{F}_\tau \subseteq \mathcal{F}_\rho$ (we get more information if we have longer to discover it).

Note now that for (X_n) an adapted process, τ a stopping time, $X^\tau := (X_{\min(n,\tau)})$ is a 'stopped process', and is adapted to both the filtration $(\mathcal{F}_{\min(n,\tau)})$ and hence $(\mathcal{F}_n)$.

The filtration $(\mathcal{F}_{\min(n,\tau)})$ comes up often, and can be conceived essentially as the information given that you stop observing things ('exit the room') at time τ, whenever that happens to be. This is what we mean by τ being a *stopping* time, in that it stops the observations continuing.

Convergence

We now consider modes of convergence of random variables using our results in integration.

Definition 27 ($\mathcal{L}^p$ spaces) *For $p \geq 0$,*
$$\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P}) = \{X : \Omega \to \mathbb{R} \text{ measurable s.t. } \mathbb{E}\left[|X|^p\right] < \infty\}$$
In particular, $\mathcal{L}^0$ is the space of all random variables, and $\mathcal{L}^\infty$ is the space of random variables which are bounded almost surely.

In these notes I'll refer almost exclusively to $\mathcal{L}^p$, although it's worth noting that there's a nuance here: the well-behaved space we generally want to refer to for useful results is not $\mathcal{L}^p$, but rather $L^p := \mathcal{L}^p/\mathcal{N}$, where $\mathcal{N} := \{X \in \mathcal{L}^0 : \mathbb{E}\big[|X|\big] = 0\}$. At the same time, in this course it's not particularly desirable to be working with $X + \mathcal{N}$ constantly (not least because there are some instances where a property being 'just' almost sure *is* relevant), and therefore we principally use $\mathcal{L}^p$.

An exception to the comments above comes for $0 \leq p < 1$. L^p for these values is almost entirely useless, as it is not a normed space. $\mathcal{L}^p$ *does* have some use however, in particular for $p = 0$, in that it allows us to specify the set of measurable functions.

Definition 28 *For a sequence (X_n) of random variables over $(\Omega, \mathcal{F}, \mathbb{P})$, we say that X_n converges to X:*
i. almost surely ($X_n \overset{\text{a.s.}}{\to} X$ or $X_n \to X$ a.s.) if
$$\mathbb{P}(X_n \to X \text{ as } n \to \infty) = 1.$$
ii. in probability ($X_n \overset{\mathbb{P}}{\to} X$) if for all $\varepsilon > 0$
$$\mathbb{P}(|X_n - X| > \varepsilon) \to 0$$
as $n \to \infty$.
iii. in $\mathcal{L}^p$ ($X_n \overset{\mathcal{L}^p}{\to} X$) if $X_n \in \mathcal{L}^p$ for $n \geq 1$ and
$$\mathbb{E}[|X_n - X|^p] \to 0$$
as $n \to \infty$.
iv. weakly in $\mathcal{L}^1$ if $X_n \in \mathcal{L}^1$ for $n \geq 1$ and for all $Y \in \mathcal{L}^\infty$
$$\mathbb{E}[X_n Y] \to \mathbb{E}[XY]$$
as $n \to \infty$.
v. in distribution ($X_n \overset{d}{\to} X$) if for $x \in \mathbb{R}$ such that F_X is continuous,
$$F_{X_n}(x) \to F_X(x)$$
as $n \to \infty$.

Of these, notion (v) of convergence in distribution is the odd one out, as it is independent of any particular instance of a random variable with the same distribution. This is a strictly weaker property than convergence in probability.

Convergence in $\mathcal{L}^p$ is an identical notion to that in functional analysis, and (in a mathematical sense) leans very much into the measure-based notion of random variables. It is stronger both than weak convergence in $\mathcal{L}^1$ and convergence in probability.

Theorem 26 *For a sequence (X_n) of random variables,*
i. If $X_n \overset{\text{a.s.}}{\to} X$ then $X_n \overset{\mathbb{P}}{\to} X$.
ii. If $X_n \overset{\mathbb{P}}{\to} X$ then there is a subsequence (X_{n_k}) such that $X_{n_k} \overset{\text{a.s.}}{\to} X$.

Useful results

Lemma 27 (Markov's inequality) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X a non-negative random variable. Then for $\lambda > 0$,*
$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}$$

Corollary 1 (General Chebyshev's Inequality) *For a measurable set $A \subseteq \mathbb{R}$, $X : \Omega \to A$ a random variable, $\varphi : A \to [0, \infty]$ an increasing measurable function. For $\lambda \in A$ with $\varphi(\lambda) < \infty$ we have*
$$\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}[\varphi(X)]}{\varphi(\lambda)}.$$
This allows us to then demonstrate that for $p > 0$, $X_n \overset{\mathcal{L}^p}{\to} X$ implies $X_n \overset{\mathbb{P}}{\to} X$.

Further, we can also show the weak law of large numbers:
Corollary 2 *For (X_n) be a sequence of i.i.d. random variables with mean μ, variance σ^2, then*
$$\frac{1}{n}\sum_{k=1}^{n} X_k \to \mu$$
as $n \to \infty$.

Theorem 28 (Jensen's inequality) *Let $f : I \to \mathbb{R}$ be a convex function on an interval $I \subseteq \mathbb{R}$. If $X : \Omega \to I$ is an integrable random variable then*
$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

For considering the $\mathcal{L}^p$ spaces, we define $\|\cdot\|_p := (\mathbb{E}[|X|^p])^{1/p}$. We can note immediately that for $0 \leq p \leq q$, $\mathcal{L}^q \subseteq \mathcal{L}^p$.

Aside from this, note all the standard results regarding $\mathcal{L}^p$ spaces, in particular Hölder's inequality. A particular application of use is to note that for $1 < p, q < \infty$ with $1/p + 1/q = 1$, and for $x > 0$,
$$x\mathbb{P}(X \geq x) \leq \mathbb{E}[Y\chi_{X \geq x}],$$
then $\|X\|_p \leq q\|Y\|_p$.

In $\mathcal{L}^2$, we are able to introduce some geometry to the space, defining an inner product $\langle X, Y \rangle := \mathbb{E}[XY]$. This allows us to consider $\mathcal{L}^2$ as a Hilbert space (modulus random variables equal to 0 almost everywhere). Consequently we get all of the results that we normally have for Hilbert spaces within $\mathcal{L}^2$.

Martingales

We often consider notions of random walks in probability theory. The most interesting of these are random walks where the movement at each step has expectation 0, and this is a concept generalised by the notion of martingales. In this course we cover only the discrete case of martingales, although there is a rich theory concerning continuous martingales.

Definition 29 (Martingales) *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ be a filtered probability space. An integrable, $(\mathcal{F}_n)$-adapted stochastic process (X_n) is called*
i. a martingale if for $n \geq 0$, $\mathbb{E}\big[X_{n+1} \,|\, \mathcal{F}_n\big] = X_n$ a.s.;
ii. a submartingale if for $n \geq 0$, $\mathbb{E}\big[X_{n+1} \,|\, \mathcal{F}_n\big] \geq X_n$ a.s.; and
iii. a supermartingale if for $n \geq 0$, $\mathbb{E}\big[X_{n+1} \,|\, \mathcal{F}_n\big] \leq X_n$ a.s..

We can note immediately that the difference between a super- and submartingale is almost purely aesthetic – negate one and you get the other. Additionally, we can see that a sequence of random variables is a martingale iff it is both a super- and submartingale. Consequently, statements about submartingales are preferable for their generality to statements about martingales.

Under this definition, the key property to observe for a submartingale (X_n) is this: for $m, n \geq 0$,
$$\mathbb{E}\big[X_n \,|\, \mathcal{F}_m\big] \geq X_{\min(m,n)},$$
and the analogous relation follows for martingales and supermartingales. What this says is that a martingale is a sequence of random variables for which we expect no change on average. Meanwhile, a submartingale is a sequence which we expect to increase, and a supermartingale one we expect to decrease. To reflect this, we often refer to sequences (Y_n) satisfying $\mathbb{E}\big[Y_{n+1} \,|\, \mathcal{F}_n\big] = 0$ as 'martingale difference sequences'.

If a submartingale is adapted to a smaller filtration, then it is also a submartingale with respect to that filtration. This is because the only information a submartingale is using from the filtration is that in the natural filtration, and consequently the extra information is unnecessary. Note that this doesn't go the other direction though – it's straightforward to construct a large filtration (e.g. the constant filtration $(\mathcal{F})$) such that (X_n) is not a submartingale, despite potentially being a martingale in a smaller filtration.

A very general example of a martingale can be given just as a sum of independent random variables (Y_n) each with mean 0, with respect to the natural filtration. Another cute example of a martingale involves taking an integrable random variable X and an arbitrary filtration, then defining $X_n := \mathbb{E}\big[X \,|\, \mathcal{F}_n\big]$.

Lemma 29 *Let (X_n) be a martingale with respect to $(\mathcal{F}_n)$, and $f : \mathbb{R} \to \mathbb{R}$ convex. Then provided $(f(X_n))$ is a sequence of integrable random variables, then it is a submartingale.*

See this as an application of Jensen's inequality, and note that this has some very wide-reaching applications.

Definition 30 (Predictable process) *A sequence (V_n) of random variables is predictable with respect to $(\mathcal{F}_n)$ if $\sigma(V_n) \subseteq \mathcal{F}_{n-1}$ (V_n is $\mathcal{F}_{n-1}$-measurable) for $n \geq 1$.*

Theorem 30 *For $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ a filtered probability space, (Y_n) a martingale, (V_n) a predictable process, both with respect to $(\mathcal{F}_n)$, then defining for $n \geq 0$*
$$X_n = \sum_{k=1}^{n} V_k(Y_k - Y_{k-1}),$$
if each X_n is integrable then (X_n) is a martingale with respect to $(\mathcal{F}_n)$.

Note that the sequence (X_n) is a martingale transform, sometimes denoted $((V \circ Y)_n)$.

The definition of a submartingale prompts some questions about how exactly the sequence varies from being a martingale, and how exactly we might correct it back to a martingale.

Theorem 31 (Doob's Decomposition theorem) *Let $(\Omega, \mathcal{F}, (\mathcal{F})_n, \mathbb{P})$ be a filtered probability space, (X_n) an integrable adapted process.*
i. (X_n) has a Doob decomposition
$$X_n = X_0 + M_n + A_n$$
where (M_n) is a martingale, (A_n) is predictable with respect to $(\mathbb{F}_n)$, and $M_0 = A_0 = 0$.
ii. For any $(\widetilde{M}_n)$ a martingale, $(\widetilde{A}_n)$ predictable such that $X_n = X_0 + \widetilde{M}_n + \widetilde{A}_n$, $M_n = \widetilde{M}_n$ and $A_n = \widetilde{A}_n$ for all $n \geq 0$ almost surely.
iii. (X_n) is a sub(super)martingale iff (A_n) is non-decreasing (non-increasing) almost surely.

An important consequence of this is known as the angle bracket process. For (M_n) a martingale of random variables in $\mathcal{L}^2$, we have that (M_n^2) is a submartingale. We write the Doob decomposition as
$$M_n^2 = M_0^2 + N_n + \langle M \rangle_n$$
where (N_n) is a martingale, $(\langle M \rangle_n)$ an increasing predictable process. Using predictability, we see that
$$\langle M \rangle_{n+1} - \langle M \rangle_n = \mathbb{E}\big[M_{n+1}^2 - M_n^2 \,|\, \mathcal{F}_n\big] = \mathbb{E}\big[(M_{n+1} - M_n)^2 \,|\, \mathcal{F}_n\big].$$
This is the conditional variance of $(M_{n+1} - M_n)$.

Theorem 32 *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ be a filtered probability space, (X_n) a martingale, and τ a finite stopping time. Then $(X_{\min(n,\tau)})$ is a martingale with respect to $(\mathcal{F}_n)$ and hence $(\mathcal{F}_{\min(n,\tau)})$.*

This can be proven by noting that $(\chi_{n \leq \tau})$ is predictable, and thus we can write $(X_{\min(n,\tau)})$ in terms of both this and (X_n).

Theorem 33 (Doob's Optional Sampling theorem) *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$ be a filtered probability space, (X_n) a martingale, and τ, ρ two finite stopping times, $\tau \leq \rho$. Then*
$$\mathbb{E}\big[X_\rho \,|\, \mathcal{F}_\tau\big] \overset{\text{a.s.}}{=} X_\tau,$$
and in particular, $\mathbb{E}\big[X_\rho\big] = \mathbb{E}\big[X_\tau\big] = \mathbb{E}\big[X_0\big]$.

Similarly, if (X_n) is a submartingale, then $\mathbb{E}\big[X_\rho \,|\, \mathcal{F}_\tau\big] \geq X_\tau$ a.s..

We can prove this in two steps. Firstly, prove the case for ρ constant, and then consider the martingale $(X_{\min(n,\rho)} - X_{\min(n,\tau)})$ stopped at τ.

This theorem is quite an important one, as we essentially just expand the definition of a martingale (and, via decomposition, super- and submartingales). The intuition should be that if (X_n) is a process stopping at ρ, but we only observe it up to τ, the properties of the martingale should allow us to calculate the expected value of X_ρ as X_τ. Importantly, we also generalise that a martingale's expectation is constant, even if stopping at a random time (as long as the time is bounded).

We can also get a generalisation beyond just bounded stopping times:
Corollary 3 *Let (X_n) be a martingale, τ an a.s. finite stopping time. If either $\{X_n : n \geq 0\}$ is UI, or $\mathbb{E}[\tau] < \infty$ and the sequence $\left(\mathbb{E}\big[|X_{n+1} - X_n| \,\big|\, \mathcal{F}_n\big]\right)$ is bounded, then*
$$\mathbb{E}\big[X_\tau \chi_{\tau<\infty}\big] = \mathbb{E}\big[X_0\big].$$
We would like to keep characterising martingales further – in particular wishing to bound them.
Theorem 34 (Doob's maximal inequality) *Let (X_n) be a submartingale. Then, for $\lambda > 0$,*
$$Y_n^\lambda := (X_n - \lambda)\chi_{\bigcup_{k=1}^n \{X_k \geq \lambda\}}$$
is a submartingale, and for $n \geq 1$,
$$\lambda\mathbb{P}\left(\bigcup_{k=1}^{n} \{X_k \geq \lambda\}\right) \leq \mathbb{E}\big[X_n \chi_{\bigcup_{k=1}^n \{X_k \geq \lambda\}}\big] \leq \mathbb{E}\big[|X_n|\big].$$
Essentially, we expect (X_n) to grow, but only so quickly. This should (I think?) follow from considering the hitting time of $[\lambda, \infty)$. We then get that if (X_n) is a martingale in $\mathcal{L}^p$,
$$\mathbb{P}\left(\bigcup_{k=1}^{n} \{|X_k| \geq \lambda\}\right) \leq \frac{\mathbb{E}\big[|X_n|^p\big]}{\lambda^p}.$$
Theorem 35 (Doob's $\mathcal{L}^p$ inequality) *Let (X_n) be a non-negative submartingale in $\mathcal{L}^p$ for $p \geq 1$. Then $\max_{k \leq n} X_k \in \mathcal{L}^p$ and*
$$\mathbb{E}\big[X_n^p\big] \leq \mathbb{E}\left[\max_{k \leq n} X_k^p\right] \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}\big[X_n^p\big].$$
As a weaker result, we can also obtain for a supermartingale that
$$\lambda\mathbb{P}\left(\bigcup_{k=1}^{n} \{|X_k| \geq \lambda\}\right) \leq \mathbb{E}\big[X_0\big] + 2\mathbb{E}\big[X_n^-\big].$$

Martingale Convergence

Definition 31 (Upcrossings) *With a real sequence (x_n), $a, b \in \mathbb{R}$ with $a < b$, we can define the integer sequences (ρ_n), (τ_n) such that*
$$\rho_n = \inf\{m \geq \tau_{n-1} : x_m \leq a\}$$
$$\tau_n = \inf\{m \geq \rho_n : x_m \geq b\}.$$
We can then define the sequence $(U_n([a, b], (x_n)))$ as the number of upcrossings of $[a, b]$ by time n, $U([a, b], (x_n))$ the total number of upcrossings, by
$$U_n([a, b], (x_n)) = \sup\{k \geq 0 : \tau_k \leq n\}$$
$$U([a, b], (x_n)) = \sup\{k \geq 0 : \tau_k < \infty\}.$$

Lemma 36 (Doob's upcrossing lemma) *Let (X_n) be a supermartingale, $a, b \in \mathbb{R}$ with $a < b$. For $n \geq 0$,*
$$\mathbb{E}\big[U_n([a, b], (X_n))\big] \leq \frac{\mathbb{E}\big[(X_n - a)^-\big]}{b - a}.$$

To prove this, it's firstly useful to check that each ρ_n, τ_n as defined above are stopping times. We then define the predictable quantity
$$V_n = \sum_{k=1}^{\infty} \chi_{\rho_k < n \leq \tau_k},$$
which is 1 if at time n, (X_n) has begun but not finished an upcrossing, 0 otherwise. We can then define $(V \circ X)$ to represent, understanding (X_n) as the price of a stock, the profit made by time n if one employs a strategy of buying at each ρ_k, and selling at each τ_k. We can then take expectations of this quantity in order to derive the lemma.

Assuming one follows this hypothetical strategy, one can consider $(X_n - a)^-$ as the loss on the final buy. The upcrossing lemma indicates that in order to maintain the supermartingale property, for every upcrossing, we must spend a significant amount of time losing money on the initial investment.

Abnormally for this course, we now derive a lemma from real analysis:
Lemma 37 *A real sequence (x_n) converges in $[-\infty, \infty]$ iff $U([a, b], (x_n)) < \infty$ for all $a, b \in \mathbb{Q}$ with $a < b$.*

To make this easier when dealing with extended reals, one should work with the lim sup and lim inf to prove this.

Theorem 38 (Doob's forward convergence theorem) *Let (X_n) be a super- or submartingale. If (X_n) is bounded in $\mathcal{L}^1$ ($\mathbb{E}\big[|X_n|\big]$ is bounded), then (X_n) converges a.s. to some $X \in \mathcal{L}^1$.*

We can see this from applying the upcrossing lemma in conjunction with our previous lemma.

Definition 32 (Galton-Watson branching process) *Let $(X_{n,r})$ be an array of i.i.d. random variables. (Z_n) is a Galton-Watson branching process if $Z_0 = 1$, and for $n \geq 1$,*
$$Z_n = \sum_{k=1}^{Z_{n-1}} X_{n,k}.$$
If each $X_{n,r}$ is integrable, and $\mathbb{E}\big[X_{0,0}\big] = \mu \neq 0$, we can write $M_n = Z_n/\mu^n$ to get a martingale (M_n) with respect to $\mathcal{F}_n = \sigma(\{X_{k,r} : r \geq 1, k \leq n\})$, (Z_n) a sub-, super-, or normal martingale depending on whether μ is greater than, less than, or equal to 1. By the forward-convergence theorem, we see that (Z_n) can either converge or blow up.

Theorem 39 *Let (X_n) be a martingale on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n), \mathbb{P})$. The following are equivalent:*
i. (X_n) is uniformly integrable.
ii. There is some $\mathcal{F}_\infty$-measurable random variable X such that $X_n \to X$ almost surely and in $\mathcal{L}^1$.
iii. There is an integrable $\mathcal{F}_\infty$-measurable random variable X such that $X_n = \mathbb{E}\big[X \,|\, \mathcal{F}_n\big]$ a.s. for $n \geq 1$.
Further, if $X \in \mathcal{L}^p$ for $p \geq 1$, then the convergence $X_n \to X$ also holds in $\mathcal{L}^p$.

Note that these conditions require slightly more than the forward convergence theorem, because boundedness in $\mathcal{L}^1$ doesn't require that the X_n themselves have a uniform bound (only that their absolute expectation does). In the other direction, we can use the forward convergence theorem to prove (part of) this statement however, in conjunction with Vitali's convergence theorem.

We find that the optional sampling theorem, as well as the maximal and $\mathcal{L}^p$ inequalities all have stronger variants for UI martingales:
Theorem 40 *Let (X_n) be a UI martingale. Then for any stopping times $\tau \leq \rho$*
$$\mathbb{E}\big[X_\rho \,|\, \mathcal{F}_\tau\big] \overset{\text{a.s.}}{=} X_\tau$$
and in particular $\mathbb{E}\big[X_\tau\big] = \mathbb{E}\big[X_0\big]$.

For any $\lambda \geq 0$,
$$\lambda\mathbb{P}\left(\bigcup_{n \geq 1} \{|X_n| \geq \lambda\}\right) \leq \mathbb{E}\big[|X|\chi_{\bigcup_{n \geq 1}\{|X_n| \geq \lambda\}}\big].$$
Finally, if $X \in \mathcal{L}^p$ for some $p > 1$, then with $1/p + 1/q = 1$,
$$\|X\|_p \leq \|\max_{n \geq 1} |X_n|\|_p \leq q\|X\|.$$

Backwards Martingales

Definition 33 (Backwards martingales) *Let $(\mathcal{F}_{-n})$ with $\mathcal{F}_{-(n+1)} \subseteq \mathcal{F}_{-n} \subseteq \mathcal{F}$ for $n \geq 0$. A backwards martingale with respect to $(\mathcal{F}_{-n})$ is a sequence (X_{-n}) of integrable random variables such that X_{-n} is $\mathcal{F}_{-n}$-measurable, and*
$$\mathbb{E}\big[X_{-n+1} \,|\, \mathcal{F}_{-n}\big] = X_{-n}.$$

Since X_0 is integrable, we see that (X_{-n}) is automatically UI. We then need to adapt certain definitions such as with respect to upcrossings, but restore all of our normal results as expected, and in particular receive the following:
Theorem 41 *Let (X_{-n}) be a backwards martingale with respect to $(\mathcal{F}_{-n})$. Then X_{-n} converges a.s. and in $\mathcal{L}^1$ as $n \to \infty$ to $X_{-\infty} = \mathbb{E}\big[X_0 \,|\, \mathcal{F}_{-\infty}\big]$.*

Using this, we can prove the strong law of large numbers.
Theorem 42 (Kolmogorov's strong law of large numbers) *Let (X_n) be a sequence of i.i.d. random variables, each integrable with mean μ.*
$$\frac{1}{n}\sum_{k=1}^{n} X_k \to \mu$$
as $n \to \infty$ almost surely and in $\mathcal{L}^1$.

Prove, and finish non-examinable material in Chapter 9