PROBABILITY THEORY - PART 3
MARTINGALES

MANJUNATH KRISHNAPUR

Contents

PROGRESS OF LECTURES

- Jan 1: Introduction
- Jan 3: Conditional expectation, definition and existence
- Jan 6: Conditional expectation, properties
- Jan 8: Conditional probability and conditional distribution
- Jan 10: —Lecture cancelled—
- Jan 13: More on regular conditional distributions. Disintegration.
- Jan 15: —Sankranthi—
- Jan 17: Filtrations. Submartingales and martingales. Examples.
- Jan 20: New martingales out of old. Stopped martingale.
- Jan 22: Stopping time, stopped sigma-algebra, optional sampling
- Jan 24: Gambler's ruin. Waiting times for patterns in coin tossing.
- Jan 27: Discrete Dirichlet problem.
- Jan 29: Inequalities. Maximal inequality.
- Jan 31: Upcrossing inequality
- Feb 3: Proof of upcrossing inequality. Supermartingale and martingale convergence theorems.
- Feb 5: L^p-bounded martingales. Reverse martingale theorem.
- Feb 7: Lévy's forward and backward laws. SLLN via reverse martingales.
- Feb 10: Kolmogorov's 0-1 law. Mention of Hewitt-Savage. Bounded harmonic functions on $\mathbb{Z}^d$.
- Feb 12: Galton-Watson trees. Pólya's urn scheme.
- Feb 14: Pólya's urn scheme.
- Feb 17, 19, 21: Mid-term week.
- Feb 24:
- Feb 26:
- Feb 28:
-
-

1. Conditional expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub sigma algebra of $\mathcal{F}$. Let $X : \Omega \to \mathbb{R}$ be a real-valued integrable random variable, i.e., $\mathbf{E}[|X|] < \infty$. A random variable $Y : \Omega \to \mathbb{R}$ is said to be a conditional expectation of X given $\mathcal{G}$ if (a) Y is $\mathcal{G}$-measurable, (b) $\mathbf{E}[|Y|] < \infty$, and (c) $\int\limits_A Y d\mathbf{P} = \int\limits_A X d\mathbf{P}$ for all $A \in \mathcal{G}$.

We shall say that any such Y is a version of $\mathbf{E}[X \mid \mathcal{G}]$. The notation is justified, since we shall show shortly that such a random variable always exists and is unique up to $\mathbf{P}$-null sets in $\mathcal{G}$.

Example 1

Let $B, C \in \mathcal{F}$. Let $\mathcal{G} = \{\emptyset, B, B^c, \Omega\}$ and let $X = \mathbf{1}_C$. Since $\mathcal{G}$-measurable random variables must be constant on B and on B^c, we must take $Y = \alpha \mathbf{1}_B + \beta \mathbf{1}_{B^c}$. Writing the condition for equality of integrals of Y and X over B and over B^c, we get $\alpha \mathbf{P}(B) = \mathbf{P}(C \cap B)$, $\beta \mathbf{P}(B^c) = \mathbf{P}(C \cap B^c)$. It is easy to see that with then the equality also holds for integrals over $\emptyset$ and over Ω. Hence, the unique choice for conditional expectation of X given $\mathcal{G}$ is

$$Y(\omega) = \begin{cases} \mathbf{P}(C \cap B)/\mathbf{P}(B) & \text{if } \omega \in B, \\ \mathbf{P}(C \cap B^c)/\mathbf{P}(B^c) & \text{if } \omega \in B^c. \end{cases}$$

This agrees with the notion that we learned in basic probability classes. If we get to know that B happened, we update our probability of C to $\mathbf{P}(C \cap B)/\mathbf{P}(B)$ and if we get to know that B^c happened, we update it to $\mathbf{P}(C \cap B^c)/\mathbf{P}(B^c)$.

Exercise 1

Suppose $\Omega = \sqcup_{k=1}^n B_k$ is a partition of Ω where $B_k \in \mathcal{F}$ and $\mathbf{P}(B_k) > 0$ for each k. Then show that the unique conditional expectation of $\mathbf{1}_C$ given $\mathcal{G}$ is

$$\sum_{k=1}^n \frac{\mathbf{P}(C \cap B_k)}{\mathbf{P}(B_k)} \mathbf{1}_{B_k}.$$

Example 2

Suppose Z is $\mathbb{R}^d$-valued and (X, Z) has density $f(x, z)$ with respect to Lebesgue measure on $\mathbb{R} \times \mathbb{R}^d$. Let $\mathcal{G} = \sigma(Z)$. Then, show that a version of $\mathbf{E}[X \mid \mathcal{G}]$ is

$$Y(\omega) = \begin{cases} \frac{\int_{\mathbb{R}} x f(x, Z(\omega)) dx}{\int_{\mathbb{R}} f(x, Z(\omega)) dx} & \text{if the denominator is positive,} \\ 0 & \text{otherwise.} \end{cases}$$

Recall that $\mathcal{G}$-measurable random variables are precisely those of the form $h(Z)$ where $h : \mathbb{R}^d \to \mathbb{R}$ is a Borel measurable function. Here, it is clear that the set of ω for which $\int f(x, Z(\omega))dx$ is zero is a $\mathcal{G}$-measurable set. Hence, Y defined above is $\mathcal{G}$-measurable.
We leave it as an exercise to check that Y is a version of $\mathbf{E}[X \mid \mathcal{G}]$.

Example 3

This is really a class of examples. Assume that $\mathbf{E}[X^2] < \infty$. Then, we can show that existence of $\mathbf{E}[X \mid \mathcal{G}]$ by an elementary Hilbert space argument. Recall that $H = L^2(\Omega, \mathcal{F}, \mathbf{P})$ and $W = L^2(\Omega, \mathcal{G}, \mathbf{P})$ (here we write $\mathbf{P}$ again to mean $\mathbf{P}$ restricted to $\mathcal{G}$) are Hilbert spaces and W is a closed subspace of H.

Since elements of H and W are equivalence classes of random variables, let us write $[X]$ for the equivalence class of a random variables X (strictly, we should write $[X]_H$ and $[X]_W$, but who has the time?). By elementary Hilbert space theory, there exists a unique projection operator $P_W : H \to W$ such that $P_W v \in W$ and $v - P_W v \in W^{\perp}$ for each $v \in H$. In fact, $P_W v$ is the closest point in W to v.

Now let Y be any $\mathcal{G}$-measurable random variable such that $[Y] = P_W[X]$. We claim that Y is a version of $\mathbf{E}[X \mid \mathcal{G}]$. Indeed, if Z is $\mathcal{G}$-measurable and square integrable, then $\mathbf{E}[(X - Y)Z] = 0$ because $[X] - [Y] \in W^{\perp}$ and $[Z] \in W$. In particular, taking $Z = \mathbf{1}_A$ for any $A \in \mathcal{G}$, we get $\mathbf{E}[X\mathbf{1}_A] = \mathbf{E}[Y\mathbf{1}_A]$. This is the defining property of conditional expectation.

For later purpose, we note that the projection operator that occurs above has a special property (which does not even make sense for a general orthogonal projection in a Hilbert space).

Exercise 2

If $X \geq 0$ $a.s.$ and $\mathbf{E}[X^2] < \infty$, show that $P_W[X] \geq 0$ $a.s.$ **[Hint:** If $[Y] = P_W[X]$, then $\mathbf{E}[(X - Y_+)^2] \leq \mathbf{E}[(X - Y)^2]$ with equality if and only if $Y \geq 0$ $a.s.$]

Uniqueness of conditional expectation: Suppose Y_1, Y_2 are two versions of $\mathbf{E}[X \mid \mathcal{G}]$. Then $\int_A Y_1 d\mathbf{P} = \int_A Y_2 d\mathbf{P}$ for all $A \in \mathcal{G}$, since both are equal to $\int_A X d\mathbf{P}$. Let $A = \{\omega : Y_1(\omega) > Y_2(\omega)\}$. Then the equality $\int_A (Y_1 - Y_2) d\mathbf{P} = 0$ can hold if and only if $\mathbf{P}(A) = 0$ (since the integrand is positive on A). Similarly $\mathbf{P}\{Y_2 - Y_1 > 0\} = 0$. This, $Y_1 = Y_2$ $a.s.$ (which means that $\{Y_1 \neq Y_2\}$ is $\mathcal{G}$-measurable and has zero probability under $\mathbf{P}$).

Thus, conditional expectation, if it exists, is unique up to almost sure equality.

Existence of conditional expectation: There are two approaches to this question.

First approach: Radon-Nikodym theorem: Let $X \geq 0$ and $\mathbf{E}[X] < \infty$. Then consider the measure $\mathbb{Q} : \mathcal{G} \to [0,1]$ defined by $\mathbb{Q}(A) = \int_A X d\mathbf{P}$ (we assumed non-negativity so that $\mathbb{Q}(A) \geq 0$ for all $A \in \mathcal{G}$). Further, $\mathbf{P}$ is a probability measure when restricted to $\mathcal{G}$ (we continue to denote it by $\mathbf{P}$). It is clear that if $A \in \mathcal{G}$ and $\mathbf{P}(A) = 0$, then $\mathbb{Q}(A) = 0$. In other words, $\mathbb{Q}$ is absolutely continuous to $\mathbf{P}$ on $(\Omega, \mathcal{G})$. By the Radon-Nikodym theorem, there exists $Y \in L^1(\Omega, \mathcal{G}, \mathbf{P})$ such that $\mathbb{Q}(A) = \int_A Y d\mathbf{P}$ for all $A \in \mathcal{G}$. Thus, Y is $\mathcal{G}$-measurable and $\int_A Y d\mathbf{P} = \int_A X d\mathbf{P}$ (the right side is $\mathbb{Q}(A)$). Thus, Y is a version of $\mathbf{E}[X \mid \mathcal{G}]$.

For a general integrable random variable X, let $X = X_+ - X_-$ and let Y_+ and Y_- be versions of $\mathbf{E}[X_+ \mid \mathcal{G}]$ and $\mathbf{E}[X_- \mid \mathcal{G}]$, respectively. Then $Y = Y_+ - Y_-$ is a version of $\mathbf{E}[X \mid \mathcal{G}]$.

Remark 1

Where did we use the integrability of X in all this? When $X \geq 0$, we did not! In other words, for a non-negative random variable X (even if not integrable), there exists a Y taking values in $\mathbb{R}_+ \cup \{+\infty\}$ such that Y is $\mathcal{G}$-measurable and $\int_A Y d\mathbf{P} = \int_A X d\mathbf{P}$. However, it is worth noting that if X is integrable, so is Y.

In the more general case, if there is a set of positive measure on which both Y_+ and Y_- are both infinite, then $Y_+ - Y_-$ is ill-defined on that set. Therefore, it is best to assume that $\mathbf{E}[|X|] < \infty$ so that Y_+ and Y_- are finite $a.s.$

Second approach: Approximation by square integrable random variables: Let $X \geq 0$ be an integrable random variable. Let $X_n = X \wedge n$ so that X_n are square integrable (in fact bounded) and $X_n \uparrow X$. Let Y_n be versions of $\mathbf{E}[X_n \mid \mathcal{G}]$, defined by the projections $P_W[X_n]$ as discussed earlier.

Now, $X_{n+1} - X_n \geq 0$, hence by the exercise above $P_W[X_{n+1} - X_n] \geq 0$ $a.s.$, hence by the linearity of projection, $P_W[X_n] \leq P_W[X_{n+1}]$ $a.s.$ In other words, $Y_n(\omega) \leq Y_{n+1}(\omega)$ for all $\omega \in \Omega_n$ where $\Omega_n \in \mathcal{G}$ is such that $\mathbf{P}(\Omega_n) = 1$. Then, $\Omega' := \cap_n \Omega_n$ is in $\mathcal{G}$ and has probability 1, and for $\omega \in \Omega'$, the sequence $Y_n(\omega)$ is non-decreasing.

Define $Y(\omega) = \lim_n Y_n(\omega)$ if $\omega \in \Omega'$ and $Y(\omega) = 0$ for $\omega \notin \Omega'$. Then Y is $\mathcal{G}$-measurable. Further, for any $A \in \mathcal{G}$, by MCT we see that $\int_A Y_n d\mathbf{P} \uparrow \int_A Y d\mathbf{P}$ and $\int_A X_n d\mathbf{P} \uparrow X d\mathbf{P}$. If $A \in \mathcal{G}$, then $\int_A Y_n d\mathbf{P} = \int_A X_n d\mathbf{P}$. Thus, $\int_A Y d\mathbf{P} = \int_A X d\mathbf{P}$. This proves that Y is a conditional expectation of X given $\mathcal{G}$.

2. Conditional probability

Definition 1: Regular conditional probability

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G}$ be a sub sigma algebra of $\mathcal{F}$. By regular conditional probability of $\mathbf{P}$ given $\mathcal{G}$, we mean any function $Q : \Omega \times \mathcal{F} \to [0, 1]$ such that

(1) For $\mathbf{P}$-$a.e.$ $\omega \in \Omega$, the map $A \to Q(\omega, A)$ is a probability measure on $\mathcal{F}$.

(2) For each $A \in \mathcal{F}$, then map $\omega \to Q(\omega, A)$ is a version of $\mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]$.

The second condition of course means that for any $A \in \mathcal{F}$, the random variable $Q(\cdot, A)$ is $\mathcal{G}$-measurable and $\int_B Q(\omega, A) d\mathbf{P}(\omega) = \mathbf{P}(A \cap B)$ for all $B \in \mathcal{G}$.

It is clear that if it exists, it must be unique (in the sense that if Q' is another conditional probability, then $Q'(\omega, \cdot) = Q(\omega, \cdot)$ for $a.e.$ $\omega[\mathbf{P}]$. However, unlike conditional expectation, conditional probability does not necessarily exist[1]. Why is that?

Suppose we define $Q(\omega, B)$ to be a version of $\mathbf{E}[\mathbf{1}_B \mid \mathcal{G}]$ for each $B \in \mathcal{F}$. Can we not simply prove that Q is a conditional probability? The second property is satisfied by definition. But for $Q(\omega, \cdot)$ to be a probability measure, we require that for any $B_n \uparrow B$ it must hold that $Q(\omega, B_n) \uparrow Q(\omega, B)$. Although the conditional MCT assures us that this happens for $a.e.$ ω, the exceptional set where it fails depends on B and B_ns. As there are uncountably many such sequences (unless $\mathcal{F}$ is finite) it may well happen that for each ω, there is some sequence for which it fails (an uncountable union of zero probability sets may have probability one). This is why, the existence of conditional probability is not trivial. But it does exist in all cases of interest.

Theorem 1

Let M be a complete and separable metric space and let $\mathcal{B}_M$ be its Borel sigma algebra. Then, for any Borel probability measure $\mathbf{P}$ on $(M, \mathcal{B}_M)$ and any sub sigma algebra $\mathcal{G} \subseteq \mathcal{B}_M$, a regular conditional probability Q exists. It is unique in the sense that if Q' is another regular conditional probability, then $Q(\omega, \cdot) = Q'(\omega, \cdot)$ for $\mathbf{P}$-$a.e.$ $\omega \in M$.

In probability theory we generally do not ask for any structure on the probability space, but in this theorem we do. It is really a matter of language, since we always restrict our random variables to take values in complete and separable metric spaces. In that language, we can restate the above theorem as follows.

Theorem 2

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub sugma algebra. Let $\mathcal{F}' = \sigma(X)$ be any sub sigma algebra of $\mathcal{F}$ generated by a random variable $X : \Omega \mapsto M$ where M is a complete and separable metric space (endowed with its Borel sigma algebra). Then a

[1] So I have heard. If I ever saw a counterexample, I have forgotten it.

regular conditional probability for $\mathcal{G}$ exists on $\mathcal{F}'$. That is, there is a $Q : \Omega \times \mathcal{F}' \mapsto [0,1]$ such that $Q(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{F}', \mathbf{P})$ for each $\omega \in \Omega$ and $Q(\cdot, A)$ is a version of $\mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]$ for each $A \in \mathcal{F}'$.

We shall prove this for the special case when $\Omega = \mathbb{R}$. The same proof can be easily written for $\Omega = \mathbb{R}^d$, with only minor notational complication. The above general fact can be deduced from the following fact that we state without proof[2].

Theorem 3

Let (M, d) be a complete and separable metric space. Then, there is a Borel set $B \subseteq \mathbb{R}$ and a bijection $\varphi : M \to B$ such that φ and φ^{-1} are both Borel measurable (w.r.t the inherited Borel structure on B). In fact, if M is uncountable, we can take $B = [0,1]$.

With this fact, any question of measures on a complete, separable metric space can be transferred to the case of $[0,1]$. In particular, the existence of regular conditional probabilities can be deduced. Note that the above theorem applies to (say) $M = (0,1)$ although it is not complete in the usual metric. Indeed, one can put a complete metric on $(0,1)$ (how?) without changing the topology (and hence the Borel sigma algebra) and then apply the above theorem.

A topological space whose topology can be induced by a metric that makes it complete and separable is called a *Polish space* (named after Polish mathematicians who studied it, perhaps Ulam and others). In this language, Theorem 3 says that any Polish space is Borel isomorphic to a Borel subset of $\mathbb{R}$ and Theorem 1 says that regular conditional probabilities exist for any Borel probability measure on $\mathcal{B}$ with respect to an arbitrary sub sigma algebra thereof.

Proof of Theorem 1 when $M = \mathbb{R}$. We start with a Borel probability measure $\mathbf{P}$ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $\mathcal{G} \subseteq \mathcal{B}_{\mathbb{R}}$. For each $t \in \mathbb{Q}$, let Y_t be a version of $\mathbf{E}[\mathbf{1}_{(-\infty,t]} \mid \mathcal{G}]$. For any rational $t < t'$, we know that $Y_t(\omega) \leq Y_{t'}(\omega)$ for all $\omega \notin N_{t,t'}$ where $N_{t,t'}$ is a Borel set with $\mathbf{P}(N_{t,t'}) = 0$. Further, by the conditional MCT, there exists a Borel set N_* with $\mathbf{P}(N_*) = 0$ such that for $\omega \notin N_*$, we have $\lim_{t\to\infty} Y_t(\omega) = 1$ and $\lim_{t\to-\infty} Y_t = 0$ where the limits are taken through rationals only.

Let $N = \bigcup_{t,t'} N_{t,t'} \cup N_*$ so that $\mathbf{P}(N) = 0$ by countable additivity. For $\omega \notin N$, the function $t \to Y_t(\omega)$ from $\mathbb{Q}$ to $[0,1]$ is non-decreasing and has limits 1 and 0 at $+\infty$ and $-\infty$, respectively. Now define $F : \Omega \times \mathbb{R} \to [0,1]$ by

$$F(\omega, t) = \begin{cases} \inf\{Y_s(\omega) : s > t, s \in \mathbb{Q}\} & \text{if } \omega \notin N, \\ 0 & \text{if } \omega \in N. \end{cases}$$

[2] For a proof, see Chapter 13 of Dudley's book *Real analysis and probability* or this paper by B. V. Rao and S. M. Srivastava.

By exercise 3 below, for any $\omega \notin N$, we see that $F(\omega, \cdot)$ is the CDF of some probability measure μ_ω on $\mathbb{R}$, provided $\omega \notin N$. Define $Q : \Omega \times \mathcal{B}_\mathbb{R} \to [0,1]$ by $Q(\omega, A) = \mu_\omega(A)$. We claim that Q is a conditional probability of $\mathbf{P}$ given $\mathcal{G}$.

The first condition, that $Q(\omega, \cdot)$ be a probability measure on $\mathcal{B}_\mathbb{R}$ is satisfied by construction. We only need to prove the first condition. To this end, define

$$\mathcal{H} = \{A \in \mathcal{B}_\mathbb{R} : Q(\cdot, A) \text{ is a version of } \mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]\}.$$

First we claim that $\mathcal{H}$ is a λ-system. Indeed, if $A_n \uparrow A$ and $Q(\cdot, A_n)$ is a version of $\mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]$, then by the conditional MCT, $Q(\cdot, A)$ which is the increasing limit of $Q(\cdot, A_n)$, is a version of $\mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]$. Similarly, if $A \subseteq B$ and $Q(\cdot, A)$, $A(\cdot, B)$ are versions of $\mathbf{E}[one_A \mid \mathcal{G}]$ and $\mathbf{E}[one_B \mid \mathcal{G}]$, then by linearity of conditional expectations, $Q(\cdots, B \setminus A) = Q(\cdot, B) - Q(\cdot, A)$ is a version of $\mathbf{E}[\mathbf{1}_{B \setminus A} \mid \mathcal{G}]$.

Next, we claim that $\mathcal{H}$ contains the π-system of all intervals of the form $(-\infty, t]$ for some $t \in \mathbb{R}$. For fixed t, by definition $Q(\omega, (-\infty, t])$ is the decreasing limit of $Y_s(\omega) = \mathbf{E}[\mathbf{1}_{(-\infty,s]} \mid \mathcal{G}](\omega)$ as $s \downarrow t$, whenever $\omega \notin N$. By the conditional MCT it follows that $Q(\cdot, (-\infty, t])$ is a version of $\mathbf{E}[\mathbf{1}_{(-\infty,t]} \mid \mathcal{G}]$.

An application of the π-λ theorem shows that $\mathcal{H} = \mathcal{B}_\mathbb{R}$. This completes the proof. ■

The following exercise was used in the proof.

Exercise 3

Let $f : \mathbb{Q} \to [0,1]$ be a non-decreasing function such that $f(t)$ converges to 1 or 0 according as $t \to +\infty$ or $t \to -\infty$, respectively. Then define $F : \mathbb{R} \to [0,1]$ by $F(t) = \inf\{f(q) : t < q \in \mathbb{Q}\}$. Show that F is a CDF of a probability measure.

Proof of Theorem 1 for general M, assuming Theorem 3. Let $\varphi : M \to \mathbb{R}$ be a Borel isomorphism. That is φ is bijective and φ, φ^{-1} are both Borel measurable. We are given a probability measure $\mathbf{P}$ on $(M, \mathcal{B}_M)$ and a sigma algebra $\mathcal{G} \subseteq \mathcal{B}_M$. Let $\mathbf{P}' = \mathbf{P} \circ \varphi^{-1}$ be its pushforward probability measure on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Let $\mathcal{G}' = \{\varphi(A) : A \in \mathcal{G}\}$, clearly a sub sigma algebra of $\mathcal{B}_\mathbb{R}$.

From the already proved case, we get $Q' : \mathbb{R} \times \mathcal{B}_M \to [0,1]$, a conditional probability of $\mathbf{P}'$ given $\mathcal{G}'$. Define $Q : M \times \mathcal{B}_M \to [0,1]$ by $Q(\omega, A) = Q'(\varphi(\omega), \varphi(A))$. Check that Q' is a conditional probability of $\mathbf{P}$ given $\mathcal{G}$. ■

For those interested to go deeper into the subtleties of conditional probabilities, here are things I may expand on in the notes at some later time. You may safely skip all of this and increase the happiness in your life by a small amount.

Existence of regular conditional probabilities. Given $(\Omega, \mathcal{F}, \mathbf{P})$ and $\mathcal{G} \subseteq \mathcal{F}' \subseteq \mathcal{F}$, we want to know when a regular conditional probability $Q : \Omega \times \mathcal{F}' \mapsto [0,1]$ for conditioing with respect to $\mathcal{G}$, exists. If $\mathcal{F}'$ can be taken equal to $\mathcal{F}$, all the better! The strongest statements I know of are from a paper of Jirina[3], of which I state the friendlier statement (for the more general one, see the paper). We need the notion of a *perfect measure*, introduced by Gnedenko and Kolmogorov.

A probability measure $\mathbf{P}$ on $(\Omega, \mathcal{F})$ is said to be perfect if for every $f : \Omega \mapsto \mathbb{R}$ that is Borel measurable, there is a Borel set $B \subseteq f(\Omega)$ such that $\mathbf{P} \circ f^{-1}(B^c) = 0$. Since forward images of measurable sets need not be measurable, it is not always true that $f(\Omega)$ is a Borel set, hence this definition which settles for something a little less.

Theorem: Assume that $\mathbf{P}$ is a perfect measure on $(\Omega, \mathcal{F})$, and that $\mathcal{F}$ is countably generated. Then for any $\mathcal{G} \subseteq \mathcal{F}$, a regular conditional probability exists on $\mathcal{F}$.

One gets the following corollary for metric spaces.

Corollary: Let $(X, \mathcal{B})$ be a metric space with its Borel sigma algebra. Assume that $\mathbf{P}$ is an inner regular probability Borel probability measure (i.e., $\mathbf{P}(A) = \sup\{\mathbf{P}(K) : K \subseteq A,\ K \text{ compact}\}$ for any $A \in \mathcal{B}$). Then, for any sub-sigma algebra $\mathcal{G} \subseteq \mathcal{B}$, a regular conditional probability exists $Q : X \times \mathcal{B} \mapsto [0, 1]$ exists.

One of the fundamental facts about complete, separable metric spaces is that every Borel probability measure is inner regular. Hence, our earlier theorem that regular conditional probabilities exist when working on Polish spaces is a consequence of the above theorem.

Perfect probabilties were introduced in the 1950s when the foundations of weak convergence laid down by Prokhorov were still fresh. Over decades, the emphasis in probability has shifted to studying interesting models coming from various applications, and the setting of complete separable metric spaces has proved adequate for all purposes. Modern books in probability often don't mention this concept (even Kallenberg does not!). A good reference (if you still want to wade into it) for all this and more is the old book of K. R. Parthasarathy titled *Probability measures on metric spaces*.

Specifying a measure via conditional probabilities. A discrete time Markov chain on a state space S with a sigma algebra $\mathcal{S}$ is specified by two ingredients: A probability measure ν on S and a stochastic kernel $\kappa : S \times \mathcal{S} \mapsto [0, 1]$ such that $\kappa(\cdot, A)$ is measurable for all $A \in \mathcal{S}$ and $\kappa(x, \cdot)$ is a probability measure on $(S, \mathcal{S})$.

Then, a Markov chain with initial distribution ν and transition kernel κ is a collection of random variables $(X_n)_{n \geq 0}$ (on some probability space) such that $X_0 \sim \nu$ and the conditional distribution of X_{n+1} given $X_0, \ldots, X_n$ is $\kappa(X_n, \cdot)$.

[3] Jirina, Conditional probabilities on σ-algebras with countable basis. **Czech. Math. J.** 4 (79), 372-380 (1954) [Selected Transitions in Mathematical Statistics and Probability, vol. 2 (Providence: American Mathematical Society, 1962), pp. 79-86]

Does a Markov chain exist? It is easy to answer yes by defining probability measures μ_n on $(S^n, \mathcal{S}^{\otimes n})$ by

$$\mu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) = \int_{S^n} \nu(dx_1)\kappa(x_1, dx_2) \ldots \kappa(x_{n-3}, dx_{n-2})\kappa(x_{n-2}, dx_{n-1})$$

for $A_i \in \mathcal{S}$. This does define a probability measure on S^n, and further, these measures are consistent (the projection of μ_{n+1} to the first n co-ordinates gives μ_n). By Kolmogorov's consistency theorem, there is a measure μ on $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}})$ whose projection to the first n co-ordinates is μ_n. On $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, \mu)$, the co-ordinate random variables form a Markov chain with the given initial distribution and transition kernel. In fact, we could have allowed time-inhomogeneity and also gotten rid of Markov property by specifying stochastic kernels $\kappa_n : S^n \times \mathcal{S} \mapsto [0, 1]$ (supposed to specify the conditional distribution of X_n given $(X_0, \ldots, X_{n-1})$ by modifying the definition of the measures μ_n in the obvious way.

A more general question naturally suggested by the above discussion is the following question of great importance[4].

Let I be a countable set and Fix $\Omega = \{0, 1\}^I$, $\mathcal{G} = \mathcal{B}(\Omega)$. Suppose that for each finite $F \subseteq I$ we are given a stochastic kernel $\lambda_F : \Omega \times \mathcal{F} \mapsto [0, 1]$ such that

(1) $\lambda_F(x, \cdot)$ is a Borel probability measure on $\mathcal{G}$.

(2) $\lambda_F(\cdot, A)$ is measurable w.r.t $\mathcal{G}_F := \sigma\{\omega_j : j \notin F\}$.

(3) $\lambda_F(\cdot, A) = \mathbf{1}_A$ if $A \in \mathcal{G}_F$.

(4) If $F_1 \subseteq F_2$, then $\lambda_{F_2} \circ \lambda_{F_1} = \lambda_{F_2}$ where

$$\lambda_{F_2} \circ \lambda_{F_1}(x, A) := \int_\Omega \lambda_{F_1}(y, A)\lambda_{F_2}(x, dy).$$

A collection $\{\lambda_F\}$ satisfying these conditions is called a specification.

The question is whether there exists a measure μ on $(\Omega, \mathcal{G})$ such that the conditional distribution given $\mathcal{G}_F$ is λ_F, for any finite $F \subseteq I$. Such a measure μ is called a *Gibbs measure*. Equivalently, we may ask if there exist $\{0, 1\}$-valued random variables $(X_i)_{i \in I}$ (on some probability space) such that $\lambda_F(x, \cdot)$ is the conditional distribution of $(X_i)_{i \in F}$ given that $(X_i)_{i \in F^c} = (x_i)_{i \in F^c}$, for any finite $F \subseteq I$. Then μ is distribution of $(X_i)_{i \in I}$.

The following fundamental forms the basis of the probabilistic study of Gibbs measures coming from statistical physics. Ising model on an infinite graph is an example of such a measure specified by its conditional distributions.

Theorem: (Dobrushin-Lanford-Ruelle) Assume that a specification $\{\lambda_F\}$ satisfies the following conditions:

[4] The material below is taken from C. Preston, *Random Fields*, Springer, Berlin Heidelberg, 2006.

(1) There exists $x_0 \in \Omega$ such that given any finite $F \subseteq I$ and any $\epsilon > 0$, there is a probability measure ν on $\{0,1\}^F$ such that for any $A \subseteq \{0,1\}^F$ satisfying $\nu(A) < \delta$ and any finite $F' \supseteq F$, we have $\lambda_{F'}(x_0, A) < \epsilon$.

(2) For any finite dimensional cylinder set $A \in \mathcal{G}$ and any finite $F \subseteq I$ and any $\epsilon > 0$, there is a finite $F' \subseteq I$ and a function $f : \{0,1\}^{F'} \mapsto \mathbb{R}$ such that $|\lambda_F(x, A) - f(x_{F'})| < \epsilon$ for all $x \in \Omega$, where $x_{F'}$ is the projection of x to $\{0,1\}^{F'}$.

Then, a Gibbs measure exists for the given specification.

The question of uniqueness or non-uniqueness of Gibbs measure is one of the most fundamental questions in statistical physics, and underlies the mathematical study of phase transitions.

3. Relationship between conditional probability and conditional expectation

Let M be a complete and separable metric space (or in terms introduced earlier, a Polish space). Let $\mathbf{P}$ be a probability measure on $\mathcal{B}_M$ and let $\mathcal{G} \subseteq \mathcal{B}_M$ be a sub sigma algebra. Let Q be a regular conditional probability for $\mathbf{P}$ given $\mathcal{G}$ which exists, as discussed in the previous section. Let $X : M \to \mathbb{R}$ be a Borel measurable, integrable random variable. We defined the conditional expectation $\mathbf{E}[X \mid \mathcal{G}]$ in the first section. We now claim that the conditional expectation is actually the expectation with respect to the conditional probability measure. In other words, we claim that

$$\mathbf{E}[X \mid \mathcal{G}](\omega) = \int\limits_M X(\omega')dQ_\omega(\omega') \tag{1}$$

where $Q_\omega(\cdot)$ is a convenient notation probability measure $Q(\omega, \cdot)$ and $dQ_\omega(\omega')$ means that we use Lebesgue integral with respect to the probability measure Q_ω (thus ω' is a dummy variable which is integrated out).

To show this, it suffices to argue that the right hand side of (1) is $\mathcal{G}$-measurable, integrable and that its integral over $A \in \mathcal{G}$ is equal to $\int_A X d\mathbf{P}$.

Firstly, let $X = \mathbf{1}_B$ for some $B \in \mathcal{B}_M$. Then, the right hand side is equal to $Q_\omega(B) = Q(\omega, B)$. By definition, this is a version of $\mathbf{E}[\mathbf{1}_B \mid \mathcal{G}]$. By linearity, we see that (1) is valid whenever X is a simple random variable.

If X is a non-negative random variable, then we can find simple random variables $X_n \geq 0$ that increase to X. For each n

$$\mathbf{E}[X_n \mid \mathcal{G}](\omega) = \int\limits_M X_n(\omega')dQ_\omega(\omega') \;\; \text{a.e.}\omega[\mathbf{P}].$$

The left side increases to $\mathbf{E}[X \mid \mathcal{G}]$ for $a.e..$ ω by the conditional MCT. For fixed $\omega \notin N$, the right side is an ordinary Lebesgue integral with respect to a probability measure Q_ω and hence the usual MCT shows that it increases to $\int\limits_M X(\omega')dQ_\omega(\omega')$. Thus, we get (1) for non-negative random variables.

For a general integrable random variable X, write it as $X = X_+ - X_-$ and use (1) individually for $X_\pm$ and deduce the same for X.

Remark 2

Here we explain the reasons why we introduced conditional probability. In most books on martingales, only conditional expectation is introduced and is all that is needed. However, when conditional probability exists, conditional expectation becomes an actual expectation with respect to a probability measure. This makes it simpler to not have to prove many properties for conditional expectation as we shall see in the following section. Also, it is aesthetically pleasing to know that conditional probability exists in most circumstances of interest.

A more important point is that, for discussing Markov processes (as we shall do when we discuss Brownian motion), conditional probability is the more natural language in which to speak.

4. Properties of conditional expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. We write $\mathcal{G}, \mathcal{G}_i$ for sub sigma algebras of $\mathcal{F}$ and X, X_i for integrable $\mathcal{F}$-measurable random variables on Ω.

Proertied specific to conditional expectations:

(1) If X is $\mathcal{G}$-measurable, then $\mathbf{E}[X \mid \mathcal{G}] = X$ $a.s.$ In particular, this is true if $\mathcal{G} = \mathcal{F}$.

(2) If X is independent of $\mathcal{G}$, then $\mathbf{E}[X \mid \mathcal{G}] = \mathbf{E}[X]$. In particular, this is true if $\mathcal{G} = \{\emptyset, \Omega\}$.

(3) Tower property: If $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then $\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbf{E}[X \mid \mathcal{G}_1]$ $a.s.$ In particular (taking $\mathcal{G} = \{\emptyset, \Omega\}$), we get $\mathbf{E}[\mathbf{E}[X \mid \mathcal{G}]] = \mathbf{E}[X]$.

(4) $\mathcal{G}$-measurable random variables are like constants for conditional expectation: For any bounded $\mathcal{G}$-measurable random variable Z, we have $\mathbf{E}[XZ \mid \mathcal{G}] = Z\mathbf{E}[X \mid \mathcal{G}]$ $a.s.$

The first statement is easy, since X itself satisfies the properties required of the conditional expectation. The second is easy too, since the constant random variable $\mathbf{E}[X]$ is $\mathcal{G}$-measurable and for any $A \in \mathcal{G}$ we have $\mathbf{E}[X\mathbf{1}_A] = \mathbf{E}[X]\mathbf{E}[\mathbf{1}_A]$.

Property (4): First consider the last property. If $Z = \mathbf{1}_B$ for some $B \in \mathcal{G}$, it is the very definition of conditional expectation. From there, deduce the property when Z is a simple random variable, a non-negative random variable, a general integrable random variable. We leave the details as an exercise.

Property (3): Now consider the tower property which is of enormous importance to us. But the proof is straightforward. Let $Y_1 = \mathbf{E}[X \mid \mathcal{G}_1]$ and $Y_2 = \mathbf{E}[X \mid \mathcal{G}_2]$. If $A \in \mathcal{G}_1$, then by definition, $\int_A Y_1 d\mathbf{P} = \int_A X d\mathbf{P}$. Further, $\int_A Y_2 d\mathbf{P} = \int_A X d\mathbf{P}$ since $A \in \mathcal{G}_2$ too. This shows that $\int_A Y_1 d\mathbf{P} =$

$\int_A Y_2 d\mathbf{P}$ for all $A \in \mathcal{G}_1$. Further, Y_1 is $\mathcal{G}_1$-measurable. Hence, it follows that $Y_1 = \mathbf{E}[Y_2 \mid \mathcal{G}_1]$. This is what is claimed there.

Properties akin to expectation:

(1) Linearity: For $\alpha, \beta \in \mathbb{R}$, we have $\mathbf{E}[\alpha X_1 + \beta X_2 \mid \mathcal{G}] = \alpha \mathbf{E}[X_1 \mid \mathcal{G}] + \beta \mathbf{E}[X_2 \mid \mathcal{G}]$ $a.s.$

(2) Positivity: If $X \geq 0$ $a.s.$, then $\mathbf{E}[X \mid \mathcal{G}] \geq 0$ $a.s.$ and $\mathbf{E}[X \mid \mathcal{G}]$ is zero $a.s.$ if and only if $X = 0$ $a.s.$ As a corollary, if $X_1 \leq X_2$, then $\mathbf{E}[X_1 \mid \mathcal{G}] \leq \mathbf{E}[X_2 \mid \mathcal{G}]$.

(3) Conditional MCT: If $0 \leq X_n \uparrow X$ $a.s.$, then $\mathbf{E}[X_n \mid \mathcal{G}] \uparrow \mathbf{E}[X \mid \mathcal{G}]$ $a.s.$ Here either assume that X is integrable or make sense of the conclusion using Remark 1.

(4) Conditional Fatou's: Let $0 \leq X_n$. Then, $\mathbf{E}[\liminf X_n \mid \mathcal{G}] \leq \liminf \mathbf{E}[X_n \mid \mathcal{G}]$ $a.s.$

(5) Conditional DCT: Let $X_n \overset{a.s.}{\to} X$ and assume that $|X_n| \leq Y$ for some Y with finite expectation, then $\mathbf{E}[X_n \mid \mathcal{G}] \overset{a.s.}{\to} \mathbf{E}[X \mid \mathcal{G}]$.

(6) Conditional Jensen's inequality: If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and X and $\varphi(X)$ are integrable, then $\mathbf{E}[\varphi(X) \mid \mathcal{G}] \geq \varphi(\mathbf{E}[X \mid \mathcal{G}])$. In particular, of $\mathbf{E}[|X|^p] < \infty$ for some $p \geq 1$, then $\mathbf{E}[|X|^p \mid \mathcal{G}] \geq (\mathbf{E}[|X| \mid \mathcal{G}])^p$ of which the special cases $\mathbf{E}[|X| \mid \mathcal{G}] \geq |\mathbf{E}[X \mid \mathcal{G}]|$ and $\mathbf{E}[X^2 \mid \mathcal{G}] \geq (\mathbf{E}[X \mid \mathcal{G}])^2$ are particularly useful.

(7) Conditional Cauchy-Schwarz: If $\mathbf{E}[X^2], \mathbf{E}[Y^2] < \infty$, then $(\mathbf{E}[XY \mid \mathcal{G}])^2 \leq \mathbf{E}[X^2 \mid \mathcal{G}]\mathbf{E}[Y^2 \mid \mathcal{G}]$.

If we assume that Ω is a Polish space and $\mathcal{F}$ is its Borel sigma algebra, then no proofs are needed! Indeed, then a conditional probability exists, and conditional expectation is just expectation with respect to conditional probability measure. Thus, ω by ω, the properties above hold for conditional expectations[5].

But the assumption that conditional probability exists is not necessary for the above properties to hold. Recall that the difficulty with the existence of conditional probability was in choosing versions of conditional expectation for $\mathbf{1}_B$ for uncountably many B so that countable additivity of $B \mapsto \mathbf{E}[\mathbf{1}_B \mid \mathcal{G}](\omega)$ holds for each fixed ω. But if we restrict attention to countably many events or random variables, then we can find a common set of zero probability outside of which there is no problem. Since in all the properties stated above, we have only a finite or countable number of random variables, we can just consider a mapping of $\omega \mapsto (X_n(\omega))_n$ from Ω to $\mathbb{R}^{\mathbb{N}}$ and transfer the problem to the Polish space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$. We leave it as an exercise to work out the details and instead give direct arguments that amount to the same.

Proofs of properties of conditional expectations:

[5] You may complain that conditional MCT was used to show existence of conditional probability, then is it not circular reasoning to use conditional probability to prove conditional MCT? Indeed, at least a limited form of conditional MCT was already used. But the derivation of other properties using conditional probability is not circular.

(1) Let Y_i be versions of $\mathbf{E}[X_i \mid \mathcal{G}]$. Then for $A \in \mathcal{G}$,

$$\int_A (\alpha Y_1 + \alpha Y_2) d\mathbf{P} = \alpha \int_A Y_1 d\mathbf{P} + \beta \int_A Y_2 d\mathbf{P}$$
$$= \alpha \int_A X_1 d\mathbf{P} + \beta \int_A X_2 d\mathbf{P} = \int_A (\alpha X_1 + \beta X_2) d\mathbf{P}$$

which shows that $\alpha Y_1 + \beta Y_2$ is a version of $\mathbf{E}[\alpha X_1 + \beta X_2 \mid \mathcal{G}]$.

(2) This is clear if you go back to the proof of the existence of conditional expectation. Here is a more direct proof. Let Y be a version of $\mathbf{E}[X \mid \mathcal{G}]$ and set $A = \{Y < 0\} \in \mathcal{G}$. Then $\int_A Y d\mathbf{P} = \int_A X d\mathbf{P} \geq 0$ (as $X \geq 0$ $a.s.$) but $Y < 0$ on A, hence $\mathbf{P}(A) = 0$.

(3) Choose versions Y_n of $\mathbf{E}[X_n \mid \mathcal{G}]$. By redefining them on a zero probability set we may assume that $Y_1 \leq Y_2 \leq \ldots$, hence $Y = \lim Y_n$ exists. For any $A \in \mathcal{G}$, by the usual MCT we have $\mathbf{E}[Y_n \mathbf{1}_A] \uparrow \mathbf{E}[Y \mathbf{1}_A]$ and $\mathbf{E}[X_n \mathbf{1}_A] \uparrow \mathbf{E}[X \mathbf{1}_A]$. But also $\mathbf{E}[Y_n \mathbf{1}_A] = \mathbf{E}[X_n \mathbf{1}_A]$ for each n, hence $\mathbf{E}[Y \mathbf{1}_A] = \mathbf{E}[X \mathbf{1}_A]$. This is what was claimed.

(4) Since $Z := \liminf X_n$ is the increasing limit of $Z_n := \inf_{k \geq n} X_k$, for any $A \in \mathcal{G}$ by the conditional MCT we have $\mathbf{E}[Z_n \mid \mathcal{G}] \uparrow \mathbf{E}[Z \mid \mathcal{G}]$. But $X_n \geq Z_n$, hence $\mathbf{E}[Z_n \mid \mathcal{G}] \leq \mathbf{E}[X_n \mid \mathcal{G}]$. Putting these together, we see that $\liminf \mathbf{E}[X_n \mid \mathcal{G}] \geq \mathbf{E}[Z \mid \mathcal{G}]$ which is what we wanted.

(5) Apply the conditional Fatou's lemma to $Y - X_n$ and $Y + X_n$.

(6) Fix a version of $\mathbf{E}[X \mid \mathcal{G}]$ and $\mathbf{E}[\varphi(X) \mid \mathcal{G}]$. Write $\varphi(t) = \sum_{i \in I}(a_i + b_i t)$, where I is countable (e.g., supporting lines at all rationals). For each $i \in I$, we have $\mathbf{E}[\varphi(X) \mid \mathcal{G}] \geq \mathbf{E}[a_i + b_i X \mid \mathcal{G}] = a_i + b_i \mathbf{E}[X \mid \mathcal{G}]$. Take supremum over $i \in I$ to get $\varphi(\mathbf{E}[X \mid \mathcal{G}])$ on the right.

(7) Observe that $\mathbf{E}[(X - tY)^2 \mid \mathcal{G}] \geq 0$ $a.s.$ for any $t \in \mathbb{R}$. Hence $\mathbf{E}[X^2 \mid \mathcal{G}] + t^2 \mathbf{E}[Y^2 \mid \mathcal{G}] - 2t\mathbf{E}[XY \mid \mathcal{G}] \geq 0$ $a.s.$ The set of zero measure indicated by "$a.s.$" depends on t, but we can choose a single set of zero measure such that the above inequality holds for all $t \in \mathbb{Q}$, $a.s.$ (for a fixed version of $\mathbf{E}[X \mid \mathcal{G}]$ and $\mathbf{E}[Y \mid \mathcal{G}]$). By continuity in t, it holds for all $t \in \mathbb{R}$, $a.s.$ Optimize over t to get the conditional Cauchy-Schwarz.

5. Cautionary tales on conditional probability

Even when knows all the definitions in and out, it is easy to make mistakes with conditional probability. Extreme caution is advocated! Practising some explicit computations also helps. Two points are to be noted.

Always condition on a sigma-algebra: Always specify the experiment first and then the outcome of the experiment. From the nature of the experiment, we can work out the way probabilities and expectations are to be updated for every possible outcome of the experiment. Then we apply that to the outcome that actually occurs.

For example, suppose I tell you that the bus I caught today morning had a 4-digit registration number of which three of the digits were equal to 7, and ask you for the chance that the remaining digit is also a 7. You should refuse to answer that question, as it is not specified what experiment was conducted. Did I note down the first three digits and report them to you, or did I look for how many 7s there were and reported that to you? It is not enough to know what I observed, but also what else I could have observed.

Conditioning on zero probability events: If (X, Y) have a joint density $f(x, y)$, then $\mathbf{E}[X \mid Y] = \frac{\int xf(x,Y)dx}{\int f(x,Y)dx}$. If we set $Y = 0$ in this formula, we get $\mathbf{E}[X \mid Y = 0]$. However, since conditional expectation is only defined up to zero measure sets, we can also set $\mathbf{E}[X \mid Y = 0]$ to be any other value. Why this particular formula?

The point is the same as asking for the value of a measurable function at a point - changing the value at a point is of no consequence for most purposes. However, there may be some justification for choosing a particular value. For example, if 0 is a Lebesgue point of f, it makes sense to take $f(0)$ to be $\lim\limits_{\epsilon \downarrow 0} \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} f(x)dx$. This is true in particular if f is continuous at 0.

Similarly, if we have to specify a particular value for $\mathbf{E}[X \mid Y = 0]$, it has to be approached via some limits, for example we may define it as $\lim_{\epsilon \downarrow 0} \mathbf{E}[X \mid |Y| < \epsilon]$, if the limit exists (and $\mathbf{P}(|Y| < \epsilon) > 0$ for any $\epsilon > 0$). For instance, if the joint density $f(x, y)$ is continuous, this will limit will be equal to the formula we got earlier, i.e., $\frac{\int xf(x,0)dx}{\int f(x,0)dx}$.

Example 4

Here is an example that illustrates both the above points. Let (U, V) be uniform on $[0, 1]^2$. Consider the diagonal line segment $L = \{(u, v) \in [0, 1]^2 : u = v\}$. What is the expected value of U conditioned on the event that it lies on L? This question is ambiguous as the experiment is not specified and the event that (U, V) lies on L has zero probability. Here are two possible interpretations. See Figure 5.

(1) The experiment measured $Y = U - V$ and the outcome was 0. In this case we are conditioning on $\sigma\{Y\}$. If we take limits of $\mathbf{E}[U \mid |Y| < \epsilon]$ as $\epsilon \downarrow 0$, we get $\mathbf{E}[X \mid Y = 0] = \frac{1}{2}$.

(2) The experiment measured $Z = U/V$ and the outcome was 1. In this case we are conditioning on $\sigma\{Z\}$. If we take limits of $\mathbf{E}[U \mid |Z| < \epsilon]$ as $\epsilon \downarrow 0$, we get $\mathbf{E}[X \mid Z = 1] = \frac{2}{3}$ (do the calculation!).

Conditional probability is the largest graveyard of mistakes in probability, hence it is better to keep these cautions in mind[6]. There are also other kinds of mistakes such as mistaking $\mathbf{P}(A \mid B)$

[6] There are many probability puzzles or "paradoxes", where the gist is some mistake in conditioning of the kind stated above (the famous *Tuesday brother problem* is an example of conditioning on an event without telling what the measurement is). A real-life example: No less a probabilist than Yuval Peres told us of a mistake he made once: In

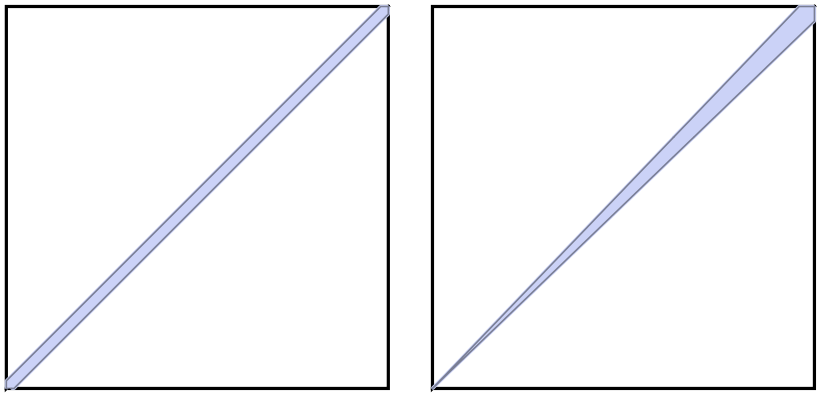


FIGURE 1. Two ways of conditioning a uniformly chosen point on the square to lie on the diagonal line. In the first case we condition on $|U-V|<\epsilon$ and in the second case on $|\frac{U}{V}-1|<\epsilon$ (a value of $\epsilon=0.02$ was taken). Under that conditioning, the point is uniform in the shaded regions. In the first conditioning, U is almost uniform on $[0,1]$ but not so in the second.

for $\mathbf{P}(B \mid A)$, a standard example being the Bayes' paradox (given that you tested positive for a rare disease, what is the chance that you actually have the disease?) that we talked about in more basic courses. This same thing is called *representational fallacy* by Kahnemann and Tversky in their study of psychology of probability (a person who is known to be a doctor or a mathematician is given to be intelligent, systematic, introverted and absent-minded. What is the person more likely to be - doctor or mathematician?).

Here is a mind-bender for your entertainment[7]. If you like this, you can find many other such questions on Gil Kalai's blog under the category Test your intuition.

Elchanan Mossel's amazing dice paradox: A fair die is thrown repeatedly until a **6** turns up. Given that all the throws showed up even numbers, what is the expected number of throws (including the last throw)?

One intuitive answer is that it is like throwing a die with only three faces, $2,4,6$, till a 6 turns up, hence the number of throws is a Geometric random variable with mean 3. This is wrong!

6. MARTINGALES

Let $(\Omega,\mathcal{F},\mathbf{P})$ be a probability space. Let $\mathcal{F}_\bullet=(\mathcal{F}_n)_{n\in\mathbb{N}}$ be a collection of sigma subalgebras of $\mathcal{F}$ indexed by natural numbers such that $\mathcal{F}_m\subseteq\mathcal{F}_n$ whenever $m<n$. Then we say that $\mathcal{F}_\bullet$ is a *filtration*. Instead of $\mathbb{N}$, a filtration may be indexed by other totally ordered sets like $\mathbb{R}_+$ or $\mathbb{Z}$ or

studying the random power series $f(z)=a_0+a_1z+a_2z^2+\ldots$ where a_k are i.i.d. $N(0,1)$, he got into a contradiction thinking that conditioning on $f(0)=0$ is the same as conditioning the zero set of f to contain the origin! The two conditionings are different for reasons similar to the example given in the text.

[7] Thanks to P. Vasanth, Manan Bhatia and Gaurang Sriramanan for bringing these to my attention!

$\{0, 1, \ldots, n\}$ etc. A sequence of random variables $X = (X_n)_{n\in\mathbb{N}}$ defined on $(\Omega, \mathcal{F}, \mathbf{P})$ is said to be *adapted* to the filtration $\mathcal{F}_\bullet$ if $X_n \in \mathcal{F}_n$ for each n.

Definition 2

In the above setting, let $X = (X_n)_{n\in\mathbb{N}}$ be adapted to $\mathcal{F}_\bullet$. We say that X is a *super-martingale* if $\mathbf{E}|X_n| < \infty$ for each $n \geq 0$, and $\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] \leq X_{n-1}$ a.s. for each $n \geq 1$.
We say that X is a *sub-martingale* if $-X$ is a super-martingale. If X is both a super-martingale and a sub-martingale, then we say that X is a *martingale*. When we want to explicitly mention the filtration, we write $\mathcal{F}_\bullet$-martingale or $\mathcal{F}_\bullet$-super-martingale etc.

Observe that from the definition of super-martingale, it follows that $\mathbf{E}[X_n \mid \mathcal{F}_m] \leq X_m$ for any $m < n$. If the index set is $\mathbb{R}_+$, then the right way to define a super-martingales is to ask for $\mathbf{E}[X_t \mid \mathcal{F}_s] \leq X_s$ for any $s < t$ (since, the "previous time point" $t-1$ does not make sense!).

Unlike say Markov chains, the definition of martingales does not appear to put too strong a restriction on the distributions of X_n, it is only on a few conditional expectations. Nevertheless, very power theorems can be proved at this level of generality, and there are any number of examples to justify making a definition whose meaning is not obvious on the surface. In this section we give classes of examples.

Example 5: Random walk

Let ξ_n be independent random variables with finite mean and let $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$ (so $\mathcal{F}_0 = \{\emptyset, \Omega\}$). Define $X_0 = 0$ and $X_n = \xi_1 + \ldots + \xi_n$ for $n \geq 1$. Then, X is $\mathcal{F}_\bullet$-adapted, X_n have finite mean, and

$$\begin{aligned}\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] &= \mathbf{E}[X_{n-1} + \xi_n \mid \mathcal{F}_{n-1}] \\ &= \mathbf{E}[X_{n-1} \mid \mathcal{F}_{n-1}] + \mathbf{E}[\xi_n \mid \mathcal{F}_{n-1}] \\ &= X_{n-1} + \mathbf{E}[\xi_n]\end{aligned}$$

since $X_{n-1} \in \mathcal{F}_{n-1}$ and ξ_n is independent of $\mathcal{F}_{n-1}$. Thus, if $\mathbf{E}[\xi_n]$ is positive for all n, then X is a sub-martingale; if $\mathbf{E}[\xi_n]$ is negative for all n, then X is a super-martingale; if $\mathbf{E}[\xi_n] = 0$ for all n, then X is a martingale.

Example 6: Product martingale

Let ξ_n be independent, non-negative random variables and let $X_n = \xi_1\xi_2\ldots\xi_n$ and $X_0 = 1$. Then, with $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$, we see that X is $\mathcal{F}_\bullet$-adapted and $\mathbf{E}[X_n]$ exists (equals the product of $\mathbf{E}[\xi_k]$, $k \leq n$). Lastly,

$$\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] = \mathbf{E}[X_{n-1}\xi_n \mid \mathcal{F}_{n-1}] = X_{n-1}\mu_n$$

where $\mu_n = \mathbf{E}[\xi_n]$. Hence, if $\mu_n \geq 1$ for all n, then X is a sub-martingale, if $\mu_n = 1$ for all n, then X is a martingale, and if $\mu_n \leq 1$ for all n, then X is a super-martingale.
In particular, replacing ξ_n by ξ_n/μ_n, we see that $Y_n := \frac{X_n}{\mu_1 \cdots \mu_n}$ is a martingale.

Example 7: Log-likelihood function

Let $S = \{1, 2, \ldots, m\}$ be a finite set with a probability mass function $p(i)$, $1 \leq i \leq m$. Suppose $X_1, X_2, \ldots$ are i.i.d. samples from this distribution. The likelihood-function of the first n samples is defined as

$$L_n = \prod_{k=1}^{n} p(X_k).$$

Its logarithm, $\ell_n := \log L_n = \sum_{k=1}^{n} \log p(X_k)$, is called the log-likelihood function. This is a sum of i.i.d. random variables $\log p(X_k)$, and they have finite mean $H := \mathbf{E}[\log p(X_k)] = \sum_{i=1}^{m} p(i) \log p(i)$ (if $p(i) = 0$ for some i, interpret $p(i)\log p(i)$ as zero). Hence $\ell_n - nH$ is a martingale (with respect to the filtration given by $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$), by the same logic as in the first example.

Example 8: Doob martingale

Here is a very general way in which any (integrable) random variable can be put at the end of a martingale sequence. Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\mathcal{F}_\bullet$ be any filtration. Let $X_n = \mathbf{E}[X \mid \mathcal{F}_n]$. Then, (X_n) is $\mathcal{F}_\bullet$-adapted, integrable and

$$\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] = \mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}_n\right] \mid \mathcal{F}_{n-1}\right] = \mathbf{E}[X \mid \mathcal{F}_{n-1}] = X_{n-1}$$

by the tower property of conditional expectation. Thus, (X_n) is a martingale. Such martingales got by conditioning one random variable w.r.t. an increasing family of sigma-algebras is called a *Doob martingale*[a].

Often $X = f(\xi_1, \ldots, \xi_m)$ is a function of independent random variables ξ_k, and we study X be sturying the evolution of $\mathbf{E}[X \mid \xi_1, \ldots, \xi_k]$, revealing the information of ξ_ks, one by one. This gives X as the end-point of a Doob martingale. The usefulness of this construction will be clear in a few lectures.

[a] J. Doob was the one who defined the notion of martingales and discovered most of the basic general theorems about them that we shall see. To give a preview, one fruitful question will be to ask if a given martingale sequence is in fact a Doob martingale.

Example 9: Increasing process

et A_n, $n \geq 0$, be a sequence of random variables such that $A_0 \leq A_1 \leq A_2 \leq \ldots a.s.$Assume that A_n are integrable. Then, if $\mathcal{F}_\bullet$ is any filtration to which A is adapted, then

$$\mathbf{E}[A_n \mid \mathcal{F}_{n-1}] - A_{n-1} = \mathbf{E}[A_n - A_{n-1} \mid \mathcal{F}_{n-1}] \geq 0$$

by positivity of conditional expectation. Thus, A is a sub-martingale. Similarly, a decreasing sequence of random variables is a super-martingale[a].

[a] An interesting fact that we shall see later is that any sub-martingale is a sum of a martingale and an increasing process. This seems reasonable since a sub-martingale increases on average while a martingale stays constant on average.

Example 10: Harmonic functions

Let $(V_n)_{n\geq 0}$ be a simple random walk on a graph G with a countable vertex set V where each vertex has finite degree. This means that V is a Markov chain with transition probabilities $p_{i,j} = \frac{1}{\deg(i)}$ if $j \sim i$, and $p_{i,j} = 0$ otherwise. Let $\varphi : V \mapsto \mathbb{R}$ be a harmonic function, i.e., $\varphi(i) = \frac{1}{\deg(i)} \sum_{j:j\sim i} \varphi(j)$, for all $i \in V$. Then, $X_n = \varphi(V_n)$ is a martingale. Indeed,

$$\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] = \frac{1}{\deg(V_{n-1})} \sum_{j:j\sim V_{n-1}} \varphi(j) = \varphi(V_{n-1}) = X_{n-1}.$$

Example 11: Branching process

Let $L_{n,k}$, $n \geq 1$, $k \geq 1$, be i.i.d. random variables taking values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. We define a sequence Z_n, $n \geq 0$ by setting $Z_0 = 1$ and

$$Z_n = \begin{cases} L_{n,1} + \ldots + L_{n,Z_{n-1}} & \text{if } Z_{n-1} \geq 1, \\ 0 & \text{if } Z_{n-1} = 0. \end{cases}$$

This is the formal definition of the generation sizes of a branching process.
Informally, a branching process is a random tree, also called a *Galton-Watson tree*, which has one individual in the 0th generation. That individual has $L_{1,1}$ offsprings all of who belong to the 1st generation. Each of them, independently, have offsprings (according to the distribution of L), and these individuals comprise the second generation. And so on, the process continues till some generation becomes empty or if that does not happen, it continues for ever. What we call Z_n is just the nth generation size, forgetting the tree structure. The basic question about branching processes is whether there is a positive probability for the tree to survive forever (we shall answer this later).

Returning to Z_n, let $\mathcal{F}_n = \sigma\{L_{m,k} : m \leq n, k \geq 1\}$ so that $Z_n \in \mathcal{F}_n$. Assume that $\mathbf{E}[L] = m < \infty$. Then, (see the exercise below to justify the steps)

$$\begin{aligned}\mathbf{E}[Z_n \mid \mathcal{F}_{n-1}] &= \mathbf{E}[\mathbf{1}_{Z_{n-1}\geq 1}(L_{n,1} + \ldots + L_{n,Z_{n-1}}) \mid \mathcal{F}_{n-1}] \\ &= \mathbf{1}_{Z_{n-1}\geq 1} Z_{n-1} m \\ &= Z_{n-1} m.\end{aligned}$$

Thus, $\frac{1}{m^n} Z_n$ is a martingale.

Exercise 4

If N is a $\mathbb{N}$-valued random variable independent of ξ_m, $m \geq 1$, and ξ_m are i.i.d. with mean μ, then $\mathbf{E}[\sum_{k=1}^{N} \xi_k \mid N] = \mu N$.

Example 12: Pólya's urn scheme

An urn has $b_0 > 0$ black balls and $w_0 > 0$ white balls to start with. A ball is drawn uniformly at random and returned to the urn with an additional new ball of the same colour. Draw a ball again and repeat. The process continues forever. A basic question about this process is what happens to the contents of the urn? Does one colour start dominating, or do the proportions of black and white equalize?

In precise notation, the above description may be captured as follows. Let U_n, $n \geq 1$, be i.i.d. Uniform$[0,1]$ random variables. Let $b_0 > 0$, $w_0 > 0$, be given. Then, define $B_0 = b_0$ and $W_0 = w_0$. For $n \geq 1$, define (inductively)

$$\xi_n = \mathbf{1}\left(U_n \leq \frac{B_{n-1}}{B_{n-1} + W_{n-1}}\right), \quad B_n \quad = B_{n-1} + \xi_n, \quad W_n = W_{n-1} + (1 - \xi_n).$$

Here, ξ_n is the indicator that the nth draw is a black, B_n and W_n stand for the number of black and white balls in the urn before the $(n+1)$st draw. It is easy to see that $B_n + W_n = b_0 + w_0 + n$ (since one ball is added after each draw).

Let $\mathcal{F}_n = \sigma\{U_1, \ldots, U_n\}$ so that ξ_n, B_n, W_n are all $\mathcal{F}_n$ measurable. Let $X_n = \frac{B_n}{B_n + W_n} = \frac{B_n}{b_0 + w_0 + n}$ be the proportion of balls after the nth draw (X_n is $\mathcal{F}_n$-measurable too). Observe that

$$\mathbf{E}[B_n \mid \mathcal{F}_{n-1}] = B_{n-1} + \mathbf{E}[\mathbf{1}_{U_n \leq X_{n-1}} \mid \mathcal{F}_{n-1}] = B_{n-1} + X_{n-1} = \frac{b_0 + w_0 + n}{b_0 + w_0 + n - 1} B_{n-1}.$$

Thus,

$$\begin{aligned}\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] &= \frac{1}{b_0 + w_0 + n} \mathbf{E}[B_n \mid \mathcal{F}_{n-1}] \\ &= \frac{1}{b_0 + w_0 + n - 1} B_{n-1} \\ &= X_{n-1}\end{aligned}$$

showing that (X_n) is a martingale.

New martingales out of old: Let $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ be a filtered probability space.

▶ Suppose $X = (X_n)_{n\ge 0}$ is a $\mathcal{F}_\bullet$-martingale and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function. If $\varphi(X_n)$ has finite expectation for each n, then $(\varphi(X_n))_{n\ge 0}$ is a sub-martingale. If X was a sub-martingale to start with, and if φ is increasing and convex, then $(\varphi(X_n))_{n\ge 0}$ is a sub-martingale.

Proof. $\mathbf{E}[\varphi(X_n) \mid \mathcal{F}_{n-1}] \ge \varphi(\mathbf{E}[X_n \mid \mathcal{F}_{n-1}])$ by conditional Jensen's inequality. If X is a martingale, then the right hand side is equal to $\varphi(X_{n-1})$ and we get the sub-martingale property for $(\varphi(X_n))_{n\ge 0}$.

If X was only a sub-martingale, then $\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] \ge X_{n-1}$ and hence the increasing property of φ is required to conclude that $\varphi(\mathbf{E}[X_n \mid \mathcal{F}_{n-1}]) \ge \varphi(X_{n-1})$. ■

▶ If $t_0 < t_1 < t_2 < \ldots$ is a subsequence of natural numbers, and X is a martingale/sub-martingale/super-martingale, then $X_{t_0}, X_{t_1}, X_{t_2}, \ldots$ is also a martingale/sub-martingale/super-martingale. This property is obvious. But it is a very interesting question that we shall ask later as to whether the same is true if t_i are random times.

If we had a continuos time-martingale $X = (X_t)_{t\ge 0}$, then again $X(t_i)$ would be a discrete time martingale for any $0 < t_1 < t_2 < \ldots$. Results about continuous time martingales can in fact be deduced from results about discrete parameter martingales using this observation and taking closely spaced points t_i. If we get to continuous-time martingales at the end of the course, we shall explain this fully.

▶ Let X be a martingale and let $H = (H_n)_{n\ge 1}$ be a predictable sequence. This just means that $H_n \in \mathcal{F}_{n-1}$ for all $n \ge 1$. Then, define $(H.X)_n = \sum_{k=1}^n H_k(X_k - X_{k-1})$. Assume that $(H.X)_n$ is integrable for each n (true for instance if H_n is a bounded random variable for each n). Then, $(H.X)$ is a martingale. If X was a sub-martingale to start with, then $(H.X)$ is a sub-martingale provided H_n are non-negative, in addition to being predictable.

Proof. $\mathbf{E}[(H.X)_n - (H.X)_{n-1} \mid \mathcal{F}_{n-1}] = \mathbf{E}[H_n(X_n - X_{n-1}) \mid \mathcal{F}_{n-1}] = H_n\mathbf{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}]$. If X is a martingale, the last term is zero. If X is a sub-martingale, then $\mathbf{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] \ge 0$ and because H_n is assumed to be non-negative, the sub-martingale property of $(H.X)$ is proved. ■

7. Stopping times

Let $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ be a filtered probability space. Let $T : \Omega \to \mathbb{N} \cup \{+\infty\}$ be a random variable. If $\{T \le n\} \in \mathcal{F}_n$ for each $n \in \mathbb{N}$, we say that T is a *stopping time*.

Equivalently we may ask for $\{T = n\} \in \mathcal{F}_n$ for each n. The equivalence with the definition above follows from the fact that $\{T = n\} = \{T \le n\} \setminus \{T \le n-1\}$ and $\{T \le n\} = \cup_{k=0}^n \{T = k\}$. The way we defined it, it also makes sense for continuous time. For example, if $(\mathcal{F}_t)_{t\ge 0}$ is a filtration and $T : \Omega \to [0, +\infty]$ is a random variable, then we say that T is a stopping time if $\{T \le t\} \in \mathcal{F}_t$ for all $t \ge 0$.

Example 13

Let X_k be random variables on a common probability space and let $\mathcal{F}^X$ be the natural filtration generated by them. If $A \in \mathcal{B}(\mathbb{R})$ and $\tau_A = \min\{n \ge 0 : X_n \in A\}$, then τ_A is a stopping time. Indeed, $\{\tau_A = n\} = \{X_0 \notin A, \dots, X_{n-1} \notin A, X_n \in A\}$ which clearly belongs to $\mathcal{F}_n$.

On the other hand, $\tau'_A := \max\{n : X_n \notin A\}$ is not a stopping time as it appears to require future knowledge. One way to make this precise is to consider $\omega_1, \omega_2 \in \Omega$ such that $\tau'_A(\omega_1) = 0 < \tau'_A(\omega_2)$ but $X_0(\omega_1) = X_0(\omega_2)$. I we can find such ω_1, ω_2, then any event in $\mathcal{F}_0$ contains both of them or neither. But $\{\tau'_A \le 0\}$ contains ω_1 but not ω_2, hence it cannot be in $\mathcal{F}_0$. In a general probability space we cannot guarantee the existence of ω_1, ω_2 (for example Ω may contain only one point or X_k may be constant random variables!), but in sufficiently rich spaces it is possible. See the exercise below.

Exercise 5

Let $\Omega = \mathbb{R}^{\mathbb{N}}$ with $\mathcal{F} = \mathcal{B}(\mathbb{R}^{\mathbb{N}})$ and let $\mathcal{F}_n = \sigma\{\Pi_0, \Pi_1, \dots, \Pi_n\}$ be generated by the projections $\Pi_k : \Omega \to \mathbb{R}$ defined by $\Pi_k(\omega) = \omega_k$ for $\omega \in \Omega$. Give an honest proof that τ'_A defined as above is not a stopping time (let A be a proper subset of $\mathbb{R}$).

Suppose T, S are two stopping times on a filtered probability space. Then $T \wedge S, T \vee S, T + S$ are all stopping times. However cT and $T - S$ need not be stopping times (even if they take values in $\mathbb{N}$). This is clear, since $\{T \wedge S \le n\} = \{T \le n\} \cup \{S \le n\}$ etc. More generally, if $\{T_m\}$ is a countable family of stopping times, then $\max_m T_m$ and $\min_m T_m$ are also stopping times.

Small digression into continuous time: We shall use filtrations and stopping times in the Brownian motion class too. There the index set is continuous and complications can arise. For example, let $\Omega = C[0, \infty)$, $\mathcal{F}$ its Borel sigma-algebra, $\mathcal{F}_t = \sigma\{\Pi_s : s \le t\}$. Now define $\tau, \tau' : C[0, \infty) \to [0, \infty)$ by $\tau(\omega) = \inf\{t \ge 0 : \omega(t) \ge 1\}$ and $\tau'(\omega) = \inf\{t \ge 0 : \omega(t) > 1\}$ where the infimum is interpreted to be $+\infty$ is the set is empty. In this case, τ is an $\mathcal{F}_\bullet$-stopping time but τ' is not (why?). In discrete time there is no analogue of this situation. When we discuss this in Brownian motion, we shall enlarge the sigma-algebra $\mathcal{F}_t$ slightly so that even τ' becomes a stopping time. This is one of the reasons why we do not always work with the natural filtration of a sequence of random variables.

The sigma algebra at a stopping time: If T is a stopping time for a filtration $\mathcal{F}_\bullet$, then we want to define a sigma-algebra $\mathcal{F}_T$ that contains all information up to and including the random time T.

To motivate the idea, assume that $\mathcal{F}_n = \sigma\{X_0, \dots, X_n\}$ for some sequence $(X_n)_{n\geq 0}$. One might be tempted to define $\mathcal{F}_T$ as $\sigma\{X_0, \dots, X_T\}$ but a moment's thought shows that this does not make sense as written since T itself depends on the sample point.

What one really means is to partition the sample space as $\Omega = \sqcup_{n\geq 0}\{T = n\}$ and on the portion $\{T = n\}$ we consider the sigma-algebra generated by $\{X_0, \dots, X_n\}$. Putting all these together we get a sigma-algebra that we call $\mathcal{F}_T$. To summarize, we say that $A \in \mathcal{F}_T$ if and only if $A \cap \{T = n\} \in \mathcal{F}_n$ for each $n \geq 0$. Observe that this condition is equivalent to asking for $A \cap \{T \leq n\} \in \mathcal{F}_n$ for each $n \geq 0$ (check!). Thus, we arrive at the definition

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T \leq n\} \in \mathcal{F}_n\} \quad \text{for each } n \geq 0.$$

Remark 3

When working in continuous time, the partition $\{T = t\}$ is uncountable and hence not a good one to work with. As defined, $\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \leq t\} \in \mathcal{F}_t\}$ for each $t \geq 0$.

We make some basic observations about $\mathcal{F}_T$.

(1) $\mathcal{F}_T$ is a sigma-algebra. Indeed,

$$A^c \cap \{T \leq n\} = \{T \leq n\} \setminus (A \cap \{T \leq n\}),$$
$$(\bigcup_{k\geq 1} A_k) \cap \{T \leq n\} = \bigcup_{k\geq 1} (A_k \cap \{T \leq n\}).$$

From these it follows that $\mathcal{F}_T$ is closed under complements and countable unions. As $\Omega \cap \{T \leq n\} = \{T \leq n\} \in \mathcal{F}_n$, we see that $\Omega \in \mathcal{F}_T$. Thus $\mathcal{F}_T$ is a sigma-algebra.

(2) The idea behind the definition of $\mathcal{F}_T$ is that it somehow encapsulates all the information we have up to the random time T. The following lemma is a sanity check that this intuition is captured in the definition (i.e., if the lemma were not true, we would have to change our definition!). Here and later, note that X_T is the random variable $\omega \mapsto X_{T(\omega)}(\omega)$. But this makes sense only if $T(\omega) < \infty$, hence we assume finiteness below. Alternately, we can fix some random variable X_∞ that is $\mathcal{F}$-measurable and use that to define X_T when $T = \infty$.

Lemma 4

Let $X = (X_n)_{n\geq 0}$ be adapted to the filtration $\mathcal{F}_\bullet$ and let T be a finite $\mathcal{F}_\bullet$-stopping time. Then X_T is $\mathcal{F}_T$-measurable.

Proof. $\{X_T \leq u\} \cap \{T \leq n\} = \{X_n \leq u\} \cap \{T \leq n\}$ which is in $\mathcal{F}_n$, since $\{X_n \leq u\}$ and $\{T \leq n\}$ both are. Therefore, $\{X_T \leq u\} \in \mathcal{F}_T$ for any $u \in \mathbb{R}$, meaning that X_T is $\mathcal{F}_T$-measurable. ■

(3) Another fact is that T is $\mathcal{F}_T$-measurable (again, it would be a strange definition if this was not true - after all, by time T we know that value of T). To show this we just need to show

that $\{T \le m\} \in \mathcal{F}_T$ for any $m \ge 0$. But that is true because for every $n \ge 0$ we have

$$\{T \le m\} \cap \{T \le n\} = \{T \le m \wedge n\} \in \mathcal{F}_{m\wedge n} \subseteq \mathcal{F}_n.$$

(4) If T, S are stopping times and $T \le S$ (caution! here we mean $T(\omega) \le S(\omega)$ for every $\omega \in \Omega$), then $\mathcal{F}_T \subseteq \mathcal{F}_S$. To see this, suppose $A \in \mathcal{F}_T$. Then $A \cap \{T \le n\} \in \mathcal{F}_n$ for each n.

Consider $A \cap \{S \le n\} = A \cap \{S \le n\} \cap \{T \le n\}$ since $\{S \le n\} \subseteq \{T \le n\}$. But $A \cap \{S \le n\} \cap \{T \le n\}$ can be written as $(A \cap \{T \le n\}) \cap \{S \le n\}$ which belongs to $\mathcal{F}_n$ since $A \cap \{T \le n\} \in \mathcal{F}_n$ and $\{S \le n\} \in \mathcal{F}_n$.

All these should make it clear that that the definition of $\mathcal{F}_T$ is sound and does indeed capture the notion of information up to time T. But if not, here is a last attempt to convince you!

Exercise 6

Let $(\Omega, \mathcal{F})$ be a measure space and let $\mathcal{F}_n = \sigma\{Y_1, \ldots, Y_n\}$, where Y_k are random variables (i.e., $\mathcal{F}$-measurable map into some measure space $(S_k, \mathcal{S}_k)$). Now suppose T be an $\mathcal{F}_\bullet$-stopping time. Show that $\mathcal{F}_T$ is the same as the sigma-algebra generated by the stopped process $Z = (Z_n)_{n\ge 0}$ where $Z_n = Y_{T\wedge n}$.

For the sake of completeness: In the last property stated above, suppose we only assumed that $T \le S$ $a.s.$ Can we still conclude that $\mathcal{F}_T \subseteq \mathcal{F}_S$? Let $C = \{T > S\}$ so that $C \in \mathcal{F}$ and $\mathbf{P}(C) = 0$. If we try to repeat the proof as before, we end up with

$$A \cap \{S \le n\} = [(A \cap \{T \le n\}) \cap \{S \le n\}] \ \cup \ (A \cap \{S \le n\} \cap C).$$

The first set belongs to $\mathcal{F}_n$ but there is no assurance that $A \cap C$ does, since we only know that $C \in \mathcal{F}$.

One way to get around this problem (and many similar ones) is to complete the sigma-algebras as follows. Let $\mathcal{N}$ be the collection of all null sets in $(\Omega, \mathcal{F}, \mathbf{P})$. That is,

$$\mathcal{N} = \{A \subseteq \Omega : \exists\, B \in \mathcal{F} \text{ such that } B \supseteq A \text{ and } \mathbf{P}(B) = 0\}.$$

Then define $\bar{\mathcal{F}}_n = \sigma\{\mathcal{F}_n \cup \mathcal{N}\}$. This gives a new filtration $\bar{\mathcal{F}}_\bullet = (\bar{\mathcal{F}}_n)_{n\ge 0}$ which we call the completion of the original filtration (strictly speaking, this completion depended on $\mathcal{F}$ and not merely on $\mathcal{F}_\bullet$. But we can usually assume without loss of generality that $\mathcal{F} = \sigma\{\cup_{n\ge 0}\mathcal{F}_n\}$ by decreasing $\mathcal{F}$ if necessary. In that case, it is legitimate to call $\bar{\mathcal{F}}_\bullet$ the completion of $\mathcal{F}_\bullet$ under $\mathbf{P}$).

It is to be noted that after enlargement, $\mathcal{F}_\bullet$-adapted processes remain adapted to $\bar{\mathcal{F}}_\bullet$, stopping times for $\mathcal{F}_\bullet$ remain stopping times for $\bar{\mathcal{F}}_\bullet$, etc. Since the enlargement is only by $\mathbf{P}$-null sets, it can be see that $\mathcal{F}_\bullet$-super-martingales remain $\bar{\mathcal{F}}_\bullet$-super-martingales, etc. Hence, there is no loss in working in the completed sigma algebras.

Henceforth we shall simply assume that our filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ is such that all $\mathbf{P}$-null sets in $(\Omega, \mathcal{F}, \mathbf{P})$ are contained in $\mathcal{F}_0$ (and hence in $\mathcal{F}_n$ for all n). Let us say that $\mathcal{F}_\bullet$ is complete to mean this.

Exercise 7

Let T, S be stopping times with respect to a complete filtration $\mathcal{F}_\bullet$. If $T \leq S$ $a.s$ (w.r.t. $\mathbf{P}$), show that $\mathcal{F}_T \subseteq \mathcal{F}_S$.

Exercise 8

Let $T_0 \leq T_1 \leq T_2 \leq \ldots$ $(a.s.)$ be stopping times for a complete filtration $\mathcal{F}_\bullet$. Is the filtration $(\mathcal{F}_{T_k})_{k \geq 0}$ also complete?

8. Optional stopping or sampling

Let $X = (X_n)_{n \geq 0}$ be a super-martingale on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$. We know that (1) $\mathbf{E}[X_n] \leq \mathbf{E}[X_0]$ for all $n \geq 0$ and (2) $(X_{n_k})_{k \geq 0}$ is a super-martingale for any subsequence $n_0 < n_1 < n_2 < \ldots$.

Optional stopping theorems are statements that assert that $\mathbf{E}[X_T] \leq \mathbf{E}[X_0]$ for a stopping time T. *Optional sampling theorems* are statements that assert that $(X_{T_k})_{k \geq 0}$ is a super-martingale for an increasing sequence of stopping times $T_0 \leq T_1 \leq T_2 \leq \ldots$. Usually one is not careful to make the distinction and OST could refer to either kind of result.

Neither of these statements is true without extra conditions on the stopping times. But they are true when the stopping times are bounded, as we shall prove in this section. In fact, it is best to remember only that case, and derive more general results whenever needed by writing a stopping time as a limit of bounded stopping times. For example, $T \wedge n$ are bounded stopping times and $T \wedge n \overset{a.s.}{\to} T$ as $n \to \infty$.

Now we state the precise results for bounded stopping times.

Theorem 5: Optional stopping theorem

Let $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ be a filtered probability space and let T be a stopping time for $\mathcal{F}_\bullet$. If $X = (X_n)_{n \geq 0}$ is a $\mathcal{F}_\bullet$-super-martingale, then $(X_{T \wedge n})_{n \geq 0}$ is a $\mathcal{F}_\bullet$-super-martingale. In particular, $\mathbf{E}[X_{T \wedge n}] \leq \mathbf{E}[X_0]$ for all $n \geq 0$.

Proof. Let $H_n = \mathbf{1}_{n \leq T}$. Then $H_n \in \mathcal{F}_{n-1}$ because $\{T \geq n\} = \{T \leq n-1\}^c$ belongs to $\mathcal{F}_{n-1}$. By the observation earlier, $(H.X)_n$, $n \geq 0$, is a super-martingale. But $(H.X)_n = X_{T \wedge n} - X_0$ and this proves that $(X_{T \wedge n})_{n \geq 0}$ is an $\mathcal{F}_\bullet$-super-martingale. Then of course $\mathbf{E}[X_{T \wedge n}] \leq \mathbf{E}[X_0]$. ■

Optional stopping theorem is a remarkably useful tool. The way it is applied is to strengthen the above statement to say that $\mathbf{E}[X_T] \leq \mathbf{E}[X_0]$ (equality if it is a martingale) for a stopping time T. This is not always true, for instance consider simple symmetric random walk on integers (which is a martingale) started at the origin and stopped at the first time T when the random walk visits the state 1 (it is well-known that $T < \infty \; a.s.$). Then $X_T = 1 \; a.s.$ but $X_0 = 0 \; a.s.$, hence the expectations do not match.

Theorem 6: Optional stopping theorem - an extension

Let $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ be a filtered probability space and let T be a finite stopping time for $\mathcal{F}_\bullet$. If $X = (X_n)_{n\geq 0}$ is a $\mathcal{F}_\bullet$-sub-martingale. Then $\mathbf{E}[X_T] \leq \mathbf{E}[X_0]$ if any one of the following conditions are met.

(1) $\{X_{T\wedge n}\}_{n\geq 1}$ is uniformly integrable.

(2) $\{X_{T\wedge n}\}_{n\geq 1}$ is uniformly bounded or dominated by an integrable random variable or bounded in L^2.

(3) T is uniformly bounded.

Proof. Since T is finite, $X_{T\wedge n} \overset{a.s.}{\to} X_T$. Hence, $\mathbf{E}[X_{T\wedge n}] \to \mathbf{E}[X_T]$ (and in fact $X_T \overset{L^1}{\to} X_T$) if and only if $\{X_{T\wedge n}\}$ is uniformly integrable. This proves the first statement.

Each of the three conditions in the second statement is sufficient for uniform integrability, hence the second follows from the first.

If $T \leq N \; a.s.$ then $|X_{T\wedge n}| \leq |X_0| + \ldots + |X_N|$ which is an integrable random variable. Therefore, the sequence $\{X_{T\wedge n}\}_{n\geq 1}$ is dominated and hence uniformly integrable. ■

Although the conditions given here may be worth remembering, it is much better practise to always write $\mathbf{E}[X_{T\wedge n}] \leq \mathbf{E}[X_0]$ and then think of ways in which to let $n \to \infty$ and get $\mathbf{E}[X_T] \leq \mathbf{E}[X_0]$. While uniform integrability is necessary and sufficient, it is hard to check, but there may be other situation-specific ways to interchange limit and expectation.

Needless to say, we just stated the result for super-martingales. From this, the reverse inequality holds for sub-martingales (by applying the above to $-X$) and hence equality holds for martingales.

In Theorem 5 we think of stopping a process at a stopping time. There is a variant where we sample the process at an increasing sequence of stopping times and the question is whether the observed process retains the martingale/super-martingale property. This can be thought of as a non-trivial extension of the trivial statement that if $(X_n)_n$ is a super-martingale w.r.t. $(\mathcal{F}_n)_n$, then for any $n_0 \leq n_1 \leq n_2 \leq \ldots$, the process $(X_{n_k})_k$ is a super-martingale w.r.t. $(\mathcal{F}_{n_k})_k$.

Theorem 7: Optional sampling theorem

Let $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ be a filtered probability space and let $X = (X_n)_{n\geq 0}$ be a $\mathcal{F}_\bullet$-super-martingale. Let T_n, $n \geq 0$ be bounded stopping times for $\mathcal{F}_\bullet$ such that $T_0 \leq T_1 \leq T_2 \leq \ldots$ Then, $(X_{T_k})_{k\geq 0}$ is a super-martingale with respect to the filtration $(\mathcal{F}_{T_k})_{k\geq 0}$.

If we only assume that $T_0 \le T_1 \le T_3 \le \dots$ *a.s.*, then the conclusion remains valid if we assume that the given filtration is complete.

The condition of boundedness of the stopping times can be replaced by the condition that $\{X_{T_k \wedge n}\}_{n \ge 0}$ is uniformly integrable for any k. The reasons are exactly the same as those that went into the proof of Theorem 6.

Proof. Since X is adapted to $\mathcal{F}_\bullet$, it follows that X_{T_k} is $\mathcal{F}_{T_k}$-measurable. Further, if $|T_k| \le N_k$ w.p.1. for a fixed number N_k, then $|X_{T_k}| \le |X_0| + \dots + |X_{N_k}|$ which shows the integrability of X_{T_k}. The theorem will be proved if we show that if $S \le T \le N$ where S, T are stopping times and N is a fixed number, then

$$\mathbf{E}[X_T \mid \mathcal{F}_S] \le X_S \;\; a.s. \tag{2}$$

Since X_S and $\mathbf{E}[X_T \mid \mathcal{F}_S]$ are both $\mathcal{F}_S$-measurable, (2) follows if we show that $\mathbf{E}[(X_T - X_S)\mathbf{1}_A] \le 0$ for every $A \in \mathcal{F}_S$.

Now fix any $A \in \mathcal{F}_S$ and define $H_k = \mathbf{1}_{S+1 \le k \le T}\mathbf{1}_A$. This is the indicator of the event $A \cap \{S \le k-1\} \cap \{T \ge k\}$. Since $A \in \mathcal{F}_S$ we see that $A \cap \{S \le k-1\} \in \mathcal{F}_{k-1}$ while $\{T \ge k\} = \{T \le k-1\}^c \in \mathcal{F}_{k-1}$. Thus, H is predictable. In words, this is the betting scheme where we bet 1 rupee on each game from time $S+1$ to time T, but only if A happens (which we know by time S). By the gambling lemma, we conclude that $\{(H.X)_n\}_{n \ge 1}$ is a super-martingale. But $(H.X)_n = (X_{T \wedge n} - X_{S \wedge n})\mathbf{1}_A$. Put $n = N$ and get $\mathbf{E}[(X_T - X_S)\mathbf{1}_A] \le 0$ since $(H.X)_0 = 0$. Thus (2) is proved. ■

An alternate proof of Theorem 7 is outlined below.

Second proof of Theorem 7. As in the first proof, it suffices to prove (2).

First assume that $S \le T \le S+1$ *a.s.* Let $A \in \mathcal{F}_S$. On the event $\{S = T\}$ we have $X_T - X_S = 0$. Therefore,

$$\begin{aligned}\int_A (X_T - X_S)dP &= \int\limits_{A \cap \{T = S+1\}} (X_{S+1} - X_S)dP \\ &= \sum_{k=0}^{N-1} \int\limits_{A \cap \{S=k\} \cap \{T=k+1\}} (X_{k+1} - X_k)dP. \end{aligned} \tag{3}$$

For fixed k, we see that $A \cap \{S = k\} \in \mathcal{F}_k$ since $A \in \mathcal{F}_S$ and $\{T = k+1\} = \{T \le k\}^c \cap \{S = k\} \in \mathcal{F}_k$ because $T \le S+1$. Therefore, $A \cap \{S = k\} \cap \{T = k+1\} \in \mathcal{F}_k$ and the super-martingale property of X implies that $\int_B (X_{k+1} - X_k)dP \le 0$ for any $B \in \mathcal{F}_k$. Thus, each term in (3) is non-positive. Hence $\int_A X_S dP \ge \int_A X_T dP$ for every $A \in \mathcal{F}_T$. This just means that $\mathbf{E}[X_S \mid \mathcal{F}_T] \le X_T$. This completes the proof assuming $S \le T \le S+1$.

In general, since $S \le T \le N$, let $S_0 = S$, $S_1 = T \wedge (S+1)$, $S_2 = T \wedge (S+2), \dots, S_N = T \wedge (S+N)$ so that each S_k is a stopping time, $S_N = T$, and for each k we have $S_k \le S_{k+1} \le S_k + 1$ *a.s.* Deduce from the previous case that $\mathbf{E}[X_T \mid \mathcal{F}_S] \le X_S$ *a.s.* ■

We end this section by giving an example to show that optional sampling theorems can fail if the stopping times are not bounded.

Example 14

Let ξ_i be i.i.d. $\text{Ber}_{\pm}(1/2)$ random variables and let $X_n = \xi_1 + \ldots + \xi_n$ (by definition $X_0 = 0$). Then X is a martingale. Let $T_1 = \min\{n \ge 1 : X_n = 1\}$.

A theorem of Pólya asserts that $T_1 < \infty$ w.p.1. But $X_{T_1} = 1$ $a.s.$ while $X_0 = 0$. Hence $\mathbf{E}[X_{T_1}] \neq \mathbf{E}[X_0]$, violating the optional stopping property (for bounded stopping times we would have had $\mathbf{E}[X_T] = \mathbf{E}[X_0]$). In gambling terminology, if you play till you make a profit of 1 rupee and stop, then your expected profit is 1 (an not zero as optional stopping theorem asserts).

If $T_j = \min\{n \ge 0 : X_n = j\}$ for $j = 1, 2, 3, \ldots$, then it again follows from Pólya's theorem that $T_j < \infty$ $a.s.$ and hence $X_{T_j} = j$ $a.s.$ Clearly $T_0 < T_1 < T_2 < \ldots$ but $X_{T_0}, X_{T_1}, X_{T_2}, \ldots$ is not a super-martingale (in fact, being increasing it is a sub-martingale!).

This example shows that applying optional sampling theorems blindly without checking conditions can cause trouble. But the boundedness assumption is by no means essential. Indeed, if the above example is tweaked a little, optional sampling is restored.

Example 15

In the previous example, let $-A < 0 < B$ be integers and let $T = \min\{n \ge 0 : X_n = -A \text{ or } X_n = B\}$. Then T is an unbounded stopping time. In gambling terminology, the gambler has capital A and the game is stopped when he/she makes a profit of B rupees or the gambler goes bankrupt. If we set $B = 1$ we are in a situation similar to before, but with the somewhat more realistic assumption that the gambler has finite capital.

By the optional sampling theorem $\mathbf{E}[X_{T \wedge n}] = 0$. By a simple argument (or Pólya's theorem) one can prove that $T < \infty$ w.p.1. Therefore, $X_{T \wedge n} \overset{a.s.}{\to} X_T$ as $n \to \infty$. Further, $|X_{T \wedge n}| \le B + A$ from which by DCT it follows that $\mathbf{E}[X_{T \wedge n}] \to \mathbf{E}[X_T]$. Therefore, $\mathbf{E}[X_T] = 0$. In other words optional stopping property is restored.

9. Applications of the optional stopping theorem

9.1. **Gambler's ruin problem.** Let $S_n = \xi_1 + \ldots + \xi_n$ be simple symmetric random walks, where ξ_i are i.i.d. $\text{Ber}_{\pm}(1/2)$. Fix $-a < 0 < b$. What is the probability that S hits b before $-a$? With $T = T_{-a} \wedge T_b$ where $T_x = \min\{n \ge 0 : X_n = x\}$ we know that $T < \infty$ $a.s.$[8] and hence $\mathbf{E}[X_{T \wedge n}] = 0$ for all n. Since $|X_{T \wedge n}| \le a + b$, we can let $n \to \infty$ and use DCT to conclude that $\mathbf{E}[X_T] = 0$. Hence,

[8] If you don't know this, here is a simple argument - Divide the coin tosses into disjoint blocks of length $\ell = a + b$, and observe that with probability $2^{-\ell}$, all tosses in a block are heads. Hence, there is some block which has all heads. If the random walk is not to the left of $-a$ at the beginning of this block, then it will be to the right of b at the end of the block.

if $\alpha = \mathbf{P}\{X_T = b\}$ then $1 - \alpha = \mathbf{P}\{X_T = -a\}$ and

$$0 = \mathbf{E}[X_T] = \alpha b - (1-\alpha)a$$

which gives $\alpha = \frac{a}{a+b}$.

Exercise 9

Let ξ_i be i.i.d. with $\xi_1 = +1$ w.p. p and $\xi_1 = -1$ w.p. $q = 1-p$. Let $X_n = \xi_1 + \ldots + \xi_n$. Find the probability that X hits B before $-A$ (for $A, B > 0$, of course).

One can get more information about the time T as follows. Recall that $\{S_n^2 - n\}$ is a martingale, hence $\mathbf{E}[S_{T\wedge n}^2 - (T \wedge n)] = 0$ for all n. To interchange expectation with limit as $n \to \infty$, we rewrite this as $\mathbf{E}[S_{T\wedge n}^2] = \mathbf{E}[T \wedge n]$. The left side converges to $\mathbf{E}[S_T^2]$ by DCT (as $|S_{T\wedge n}| \le a+b$) and the right side converges to $\mathbf{E}[T]$ (by MCT). Hence

$$E[T] = \mathbf{E}[S_T^2] = (-a)^2 \frac{b}{a+b} + b^2 \frac{a}{a+b} = ab.$$

In particular, when $a = b$, we get b^2, which makes sense in view of the fundamental fact that a random walk moves distance $\sqrt{t}$ in time t.

9.2. **Waiting times for patterns in coin tossing.** Let $\xi_1, \xi_2, \ldots$ be i.i.d. Ber$(1/2)$ variables (fair coin tosses). Let $\tau_{1011} = \min\{n \ge 1 : (\xi_{n-3}, \ldots, \xi_n) = (1,0,1,1)\}$ and similarly define τ_ϵ for any patter $\epsilon \in \{0,1\}^k$ for some k. Clearly these are stopping times for the filtration $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$. We would like to understand the distribution or the mean of these stopping times.

Clearly τ_1 and τ_0 are Geometric random variables with mean 2. Things are less simple for other patterns. Since this is written out in many places and was explained in class and is given as exercise to write out a proper proof, will skip the explanation here. The final answer depends on the overlaps in the pattern. If $\epsilon = (\epsilon_1, \ldots, \epsilon_k)$, then

$$\mathbf{E}[\tau_\epsilon] = \sum_{j=1}^{k} 2^j \mathbf{1}_{(\epsilon_1, \ldots, \epsilon_j) = (\epsilon_{k-j+1}, \ldots, \epsilon_k)}.$$

In particular, $\mathbf{E}[\tau_{11}] = 6$ while $\mathbf{E}[\tau_{10}] = 4$. You may remember that the usual proof of this involves setting up a gambling game where the kth gambler enters with 1 rupee in hand, just before the kth toss, and bets successively on the kth toss being ϵ_1 (at the same time the $(k-i)$th gambler, if still in the game, is betting on it being ϵ_{i+1}). If instead, the kth gambler come with k rupees in hand, one can find the second moment of τ_ϵ and so on. If the kth gambler comes with $e^{\theta k}$ rupees, where θ is sufficiently small, then the moment generating function of τ_ϵ can also be found.

Aside from the proof of this claim that uses optional stopping theorem, what is the reason for different waiting times for different patterns of the same length? This can be understood qualitatively in terms of the waiting time paradox.

The waiting time "paradox": If buses come regularly with inter-arrival times of one hour, but a person who has no watch goes at random to the bus stop, her expected waiting time is 30 minutes. However, if inter-arrival times are random with equal chance of being 90 minutes or 30 minutes (so one hour on average still), then the expected waiting time jumps to 37.5 minutes! The reason is that the person is 3 times more likely to have entered in a 90 minute interval than in a 30 minute interval.

What does this have to do with the waiting times in patterns. The "buses" $\underline{10}$ and $\underline{11}$ are equally frequent (chance $1/4$ at any $(n-1,n)$ slot), but 10 is more regularly spaced than $\underline{11}$. In fact $\underline{11}$ buses can crowd together as in the string $\underline{11111}$ which has 4 occurrences of $\underline{11}$. But to get four $\underline{10}$ buses we need at least 8 tosses. Thus, the waiting time for the less regular bus is more!

10. Random walks on graphs

Let $G = (V, E)$ be a graph with a countable vertex set V. We shall always assume that each vertex has finite degree and that the graph is connected. Simple random walk on G (usually written SRW) is a markov chain $X = (X_n)_{n\geq 0}$ with transition probabilities $p_{u,v} = \frac{1}{\deg(u)}$ for $v \sim u$, and $p_{u,v} = 0$ if v is not adjacent to u. Usually we fix X_0 to be some vertex w (in which case we write $\mathbf{P}_w$, $\mathbf{E}_w$ to indicate that).

Recall that a function $f : V \mapsto \mathbb{R}$ is said to be harmonic at a vertex u if $\frac{1}{\deg(u)}\sum_{v:v\sim u} f(v) = f(u)$ (this is called the *mean value property*). We saw that if f if harmonic on the whole of V, then $(f(X_n))_{n\geq 0}$ is a martingale. In such a situation, optional sampling theorem tells us that $(f(X_{\tau\wedge n}))_{n\geq 0}$ is also a martingale for any stopping time τ. Here is an extension of this statement.

Theorem 8

Let X be SRW on G. Let B be a proper subset of V and let τ denote the hitting time of B by the random walk. Suppose $f : V \mapsto \mathbb{R}$ is harmonic (or sub-harmonic) at all vertices of $V \setminus B$. Then, $(f(X_{\tau\wedge n}))_{n\geq 0}$ is a martingale (respectively, sub-martingale) with respect to the filtration $(\mathcal{F}_{\tau\wedge n})_{n\geq 0}$.

Note that this is not obvious and does not follow from the earlier statement. If we define $M_n = f(X_n)$, then M may not a martingale, since f need not harmonic on B. Therefore, $f(X_{\tau\wedge n})$ is not got by stopping a martingale (in which case OST would have implied the theorem), it is just that this stopped process is a martingale!

Proof. Let f be harmonic and set $M_n = f(X_{\tau\wedge n})$ and let $\mathcal{G}_n = \mathcal{F}_{\tau\wedge n}$. We want to show that $\mathbf{E}[M_{n+1} \mid \mathcal{G}_n] = M_n$. Clearly M_n is $\mathcal{G}_n$ measurable (since $X_{\tau\wedge n}$ is). Let $A \in \mathcal{G}_n$ and let $A' = A \cap \{\tau \leq n\}$ and $A'' = A \cap \{\tau > n\}$.

On A' we have $M_{n+1} = M_n = f(X_\tau)$ and hence $\mathbf{E}[M_{n+1}\mathbf{1}_{A'}] = \mathbf{E}[M_n\mathbf{1}_{A'}]$.

On A'', we have that X_{n+1} is a uniformly chosen random neighbour of X_n (independent of all the conditioning) and hence,

$$\mathbf{E}[M_{n+1}\mathbf{1}_{A''}] = \mathbf{1}_{A''}\frac{1}{\deg(X_n)}\sum_{v:v\sim X_n} f(u) = \mathbf{1}_{A''} f(X_n)$$

where the last equality holds because $X_n \notin B$ and f is harmonic there. But $f(X_n) = M_n$ on A'', (since $\tau > n$), and hence we see that $\mathbf{E}[M_{n+1}\mathbf{1}_{A''}] = \mathbf{E}[M_n\mathbf{1}_{A''}]$.

Adding the two we get $\mathbf{E}[M_{n+1}\mathbf{1}_A] = \mathbf{E}[M_n\mathbf{1}_A]$ for all $A \in \mathcal{G}_n$, hence $\mathbf{E}[M_{n+1} \mid \mathcal{G}_n] = M_n$. ■

Remark 4: Reversible Markov chains

Can the discussions of this section be carried over to general Markov chains? Not quite, but it can be to *reversible* Markov chains. Let X be a Markov chain on a countable state space S with transition matrix P. We shall assume that the chain is irreducible. Recall that the chain is said to be reversible if there is a $\pi = (\pi_i)_{i\in S}$ on S (called the stationary measure) such that $\pi(i)p_{i,j} = \pi(j)p_{j,i}$ for all $i, j \in S$.

If the chain is reversible, we can make a graph G with vertex set S and edges $i \sim j$ whenever $p_{i,j} > 0$ (reversibility forces $p_{j,i} > 0$, hence the graph is undirected). For any $i \sim j$, define the conductance of the corresponding edge as $C_{i,j} = \pi(i)p_{i,j}$. By reversibility, $C_{j,i} = C_{i,j}$, hence the conductance is associated to the edge, not the direction. Then the given Markov chain is a random walk on this graph, except that the transitions are not uniform. They are given by

$$p_{i,j} = \begin{cases} \frac{C_{i,j}}{C_{i,\cdot}} & \text{if } j \sim i, \\ 0 & \text{otherwise} \end{cases}$$

where $C_{i,\cdot} = \sum_{k:k\sim i} C_{i,k}$. Conversely, for any graph $G = (V, E)$ with specified conductances on edges, if we define transition probabilities as above, we get a reversible Markov chain.

All the discussions in the section can be taken over to general reversible chains, with appropriate modifications. If a chain is not reversible, for example suppose there are two states i, j such that $p_{i,j} > 0$ but $p_{j,i} = 0$, are quite different.

10.1. **Discrete Dirichlet problem and gambler's ruin.** Let $G = (V, E)$ be a connected graph with vertex set V and edge set E and every vertex having finite degrees. Let X denote the simple random walk on G. We consider two problems.

Gambler's ruin problem: Let A, C be disjoint proper subsets of V. Find $\mathbf{P}_x\{\tau_A < \tau_C\}$ for any $x \in V$. Here τ_A is the hitting time of the set A by the SRW X.

Discrete Dirichlet problem: Let B be a proper subset of V. Fix a function $\varphi : B \mapsto \mathbb{R}$. Find a function $f : V \mapsto \mathbb{R}$ such that (a) $f(x) = \varphi(x)$ for all $x \in B$, (b) f is harmonic on $V \setminus B$. This is a system of linear equations, one for each $v \in V \setminus B$, and in the variables $f(x)$, $x \in V \setminus B$.

These two problems are intimately related. To convey the main ideas without distractions, we restrict ourselves to finite graphs now.

(1) Observe that the solution to the Dirichlet problem, if it exists, is unique. Indeed, if f, g are two solutions, then $h = f - g$ is harmonic on $V \setminus B$ and $h = 0$ on B. Now let x_0 be a point where h attains its maximum (here finiteness of the graph is used). If $x_0 \notin B$, then $h(x_0)$ is the average of the values of h at the neighbours of x_0, hence each of those values must be equal to $h(x_0)$. Iterating this, we get a point $x \in B$ such that $h(x) = h(x_0)$ (connectedness of the graph is used here). Therefore, the maximum of h is zero. Similarly the minimum is zero and we get $f = g$.

(2) Let $f(x) = P_x\{\tau_A < \tau_B\}$ in the gambler's ruin problem. We claim that f is harmonic at every $x \notin B := A \cup C$. Indeed, for any $x \notin B$, condition on the first step of the Markov chain to see that

$$f(x) = \mathbf{E}_x[\mathbf{P}\{\tau_A < \tau_B \mid X_1\}] = \mathbf{E}_x[\mathbf{P}_{X_1}\{\tau_A < \tau_B\}] = \frac{1}{\deg(x)} \sum_{y:y\sim x} f(y).$$

Further, f is 1 on A and 0 on C. Hence f is just the solution to the discrete Dirichlet problem with $B = A \cup C$ and $\varphi = \mathbf{1}_A$. rst

(3) Conversely, suppose a set B is given and for every $x \in B$ we solve the gambler's ruin problem with $A = \{x\}$ and $C = B \setminus \{x\}$. Let $\mu_x(y) = \mathbf{P}_y\{\tau_x = \tau_B\}$ denote the solution. Then, given any $\varphi : B \mapsto \mathbb{R}$, it is easy to see that $f(\cdot) = \sum_{x\in B} \varphi(x)\mu_x(\cdot)$ is a solution to the discrete Dirichlet problem (linear combinations of harmonic functions is harmonic).

(4) The solution in the previous point may be rewritten as (with $M_n = f(X_{\tau\wedge n})$)

$$f(y) = \sum_{x\in B} \varphi(x)\mathbf{P}_y\{\tau_B = \tau_x\} = \sum_{x\in B} \varphi(x)\mathbf{P}_y\{M_\tau = x\} = \mathbf{E}_y[M_\tau].$$

(5) Here is another way to see that the solution f to the Dirichlet problem must be given like this. From Theorem 8 we know that M_n is a martingale. Hence $\mathbf{E}[f(X_{\tau\wedge n})] = \mathbf{E}[f(X_0)]$, in particular, if $X_0 = v$ then $\mathbf{E}_v[f(X_{\tau\wedge n})] = f(v)$. Let $n \to \infty$ and DCT ($f(X_{\tau\wedge n})$ is of course uniformly bounded) to conclude that $f(v) = \mathbf{E}[f(X_\tau)] = \mathbf{E}[M_\tau]$.

To summarize, we have shown the existence and uniqueness of the solution to the discrete Dirichlet problem, and related it to the solution to the gambler's ruin problem. This can be summarized as follows.

Theorem 9

Let $G = (V, E)$ be a finite connected graph and let B be a proper subset of vertices. Given $\varphi : B \mapsto \mathbb{R}$, the unique solution to the discrete Dirichlet problem with boundary data φ is given by $f(x) = \mathbf{E}_x[\varphi(X_\tau)]$ where X is the simple random walk on B and τ is its first hitting time of the set B.

Electrical networks: With the above discussion, we have related the gambler's ruin problem to the Dirichlet problem, without being able to solve either of them! Indeed, in general it is hopeless to expect an explicit solution. However, it is worth noting that the discrete Dirichlet problem arises in a different area that looks unrelated, namely that of electrical networks (a more sophisticated name is discrete potential theory). This will not bring any miracles, but the intuition from electrical networks can be of use in studying random walks and vice versa. Now we describe the electrical network formulation.

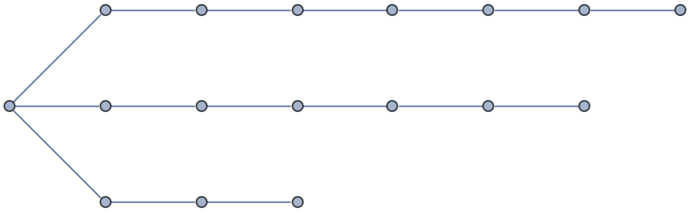
Imagine that G is an electric network where each edge is replaced by a unit resistor. The vertices in A are connected to batteries and the voltages at these points are maintained at $\varphi(x)$, $x \in A$. Then, electric current flows through the network and at each vertex a voltage is established. According to Kirchoff's law, the voltages at the vertices are precisely the solution to the discrete Dirichlet problem. Here is an example where we use knowledge of reduction in electrical networks to find the voltage at one vertex. This reduction is very special, and for general graphs there is not much one can do.

Example 16

Let G be the tree shown in the picture below. It is a tree with one vertex of degree 3 which we call the root O, and three leaves A, B, C. Let the distance (the number of edges between the root and the leaf) to the three leaves be a, b, c respectively. What is the probability that a simple random walk starting from O hits A before $\{B, C\}$?

As we have seen, the answer is given by $f(O)$, where $f : V \mapsto \mathbb{R}$ is a function that is harmonic except at the leaves and $f(A) = 1$, $f(B) = f(C) = 0$. As discussed above, this is the same problem in electrical networks (with each edge replaced by a unit resistor) of finding the voltage function when batteries are connected so that A is maintained at voltage 1 and B, C at voltage 0. In high school, we have seen ruled for resistors in series and parallel, so this is the same problem as a graph with four vertices O', A', B', C', where $O'A', OB', OC'$ have resistances $a, b, c,$, respectively. Then the effective resistance between A' and $\{B', C'\}$ is $a + \frac{1}{\frac{1}{b}+\frac{1}{c}}$, hence the effective current is the reciprocal of this. Therefore, the voltage at O' is $\frac{a^{-1}}{a^{-1}+b^{-1}+c^{-1}}$.

Exercise: Show this by solving for the harmonic function on the tree (without using this network reduction business!).



Variational principles: Using the terminology of electrical networks will not really help solve any problem. What do help are variational principles. Here is an easy exercise.

Exercise 10

Given a finite graph G and a proper subset of vertices B and a function $\varphi : B \mapsto \mathbb{R}$, consider the functional $\mathcal{L}[f] := \sum_{u \sim v}(f(u) - f(v))^2$ on $\mathcal{H} = \{f : V \mapsto \mathbb{R} : f(x) = \varphi(x) \text{ for all } x \in B\}$. Then the unique minimizer of L on $\mathcal{H}$ is the solution to the discrete Dirichlet problem with boundary data φ.

To illustrate the point, we now go to infinite graphs $G = (V, E)$ (again V is countable, each vertex has finite degree and G is connected). Recall that simple random walk X on G is recurrent if $\mathbf{P}_v\{\tau_v^+ < \infty\} = 1$ for some v (in which case it follows that $\mathbf{P}_w\{\tau_u < \infty\} = 1$ for all $w \neq u \in V$) where $\tau_v^+ = \min\{n \geq 1 : X_n = v\}$ (observe the condition $n \geq 1$, as opposed to $n \geq 0$ in the definition of τ_v). If not recurrent, it is called transient.

Again, fixing v and consider $f(x) = \mathbf{P}_x\{\tau_v < \infty\}$. If the graph is recurrent, then $f = 1$ everywhere, whereas if it is transient, we may prove that $f(x) \to 0$ as $x \to \infty$ (i.e., given $\epsilon > 0$, there is a finite $F \subseteq V$ such that $|f(x)| < \epsilon$ for all $x \notin F$). This way, one may expect to prove a theorem (this statement is not quite true as stated) that the graph is transient if and only if there is a harmonic function $f : V \mapsto \mathbb{R}$ such that $f(v) = 1$, $f(x) \to 0$ as $x \to \infty$ and f is harmonic on $V \setminus \{v\}$. But this is still hard to use, because finding harmonic functions may be delicate. This is where the variational principle is useful. We state the following theorem without proof[9]. For a fixed $v \in V$, a cut-set is any collection of edges such that every infinite simple path starting from v must use one of the edges in Π.

Theorem 10

Let G be an infinite connected network and let v be a fixed vertex. The following are equivalent.

(1) SRW on G is transient.

(2) There exists $W : E \mapsto \mathbb{R}_+$ such that $\sum_{e \in \Pi} W(e) \geq 1$ for every cut-set Π and $\sum_{e \in E} W(e)^2 < \infty$.

To illustrate the usefulness of this theorem, let us prove Pólya's theorem for random walks on $\mathbb{Z}^d$. Let us fix the vertex 0 and consider the existence of a W as required in the theorem.

$d = 1$: Any pair of edges $\{[n, n+1], [-m-1, -m]\}$ where $n, m > 0$, is a cut-set. From that it is easy to see that $W([n, n+1]) \geq 1$ for infintiely many n (in fact for all positive n or for all negative n or both). But then $\sum W(e)^2 = \infty$, showing that the random walk must be recurrent.

[9] Chapter 2 of the book *Probability on trees and networks* by Lyons and Peres is an excellent resource for this subject. Another important resource is the paper *The extremal length of a network* by R. J. Duffin.

$d = 2$: For any n, let $B(n) = \{-n, \ldots, n\}^2$. Let Π_n be the collection of edges that are in $B(n+1)$ but not in $B(n)$. There are $4(n+1)$ edges in $\Pi(n)$, and if the sum $\sum_{e \in \Pi_n} W(e) \geq 1$, then $\sum_{e \in \Pi_n} W(e)^2 \geq \frac{1}{4(n+1)}$ by Cauchy-Schwarz. As Π_ns are pairwise disjoint, this shows that $\sum_e W(e)^2 = \infty$.

$d \geq 3$. Define $W(e) = \frac{1}{|e|}$ where $|e|$ is the Euclidean distance from the origin to the mid-point of e. There are about n^{d-1} edges having $|e| \in [n, n+1]$, so the total sum of squares is like $\sum_n \frac{1}{n^{d-1}}$ which is finite. But is the condition $\sum_{e \in \Pi} W(e) \geq 1$ satisfied? For cut-sets of the form $B(n+1) \setminus B(n)$ where $B(n) = \{-n, \ldots, n\}^d$, this is clear. We leave the general case as an exercise.

The power of this theorem is in its robustness (as opposed to criteria such as $\sum_n p_{u,u}^{(n)} < \infty$ that we see in Markov chain class). If finitely many edges are added to the graph, it does not make a difference to the existence of W (also for finitely many deletions, provided it does not disconnect v from infinity) and hence to the question of recurrence or transience!

11. Maximal inequality

Kolmogorov's proof of his famous inequality was perhaps the first proof using martingales, although the term did not exist then!

Lemma 11: Kolmogorov's maximal inequality

Let ξ_k be independent random variables with zero means and finite variances. Let $S_n = \xi_1 + \ldots + \xi_n$. Then,

$$\mathbf{P}\left\{\max_{k \leq n} |S_k| \geq t\right\} \leq \frac{1}{t^2}\text{Var}(S_n).$$

Proof. We know that $(S_k)_{k \geq 0}$ is a martingale and $(S_k^2)_{k \geq 0}$ is a sub-martingale. Let $T = \min\{k : |S_k| \geq t\} \wedge n$ (i.e., T is equal to n or to the first time S exits $(-t, t)$, whichever is earlier). Then T is a bounded stopping time and $T \leq n$. By OST, $\{S_T^2, S_n^2\}$ is a sub-martingale and thus $\mathbf{E}[S_T^2] \leq \mathbf{E}[S_n^2]$. By Chebyshev's inequality,

$$\mathbf{P}\left\{\max_{k \leq n} |S_k| \geq t\right\} = \mathbf{P}\{S_T^2 \geq t^2\} \leq \frac{1}{t^2}\mathbf{E}[S_T^2] \leq \frac{1}{t^2}\mathbf{E}[S_n^2].$$

Thus the inequality follows. ■

This is an amazing inequality that controls the supremum of the entire path $S_0, S_1, \ldots, S_n$ in terms of the end-point alone! It takes a little thought to realize that the inequality $\mathbf{E}[S_T^2] \leq \mathbf{E}[S_n^2]$ is not a paradox. One way to understand it is to realize that if the path goes beyond $(-t, t)$, then there is a significant probability for the end point to be also large. This intuition is more clear in certain precursors to Kolmogorov's maximal inequality. In the following exercise you will prove one such, for symmetric, but not necessarily integrable, random variables.

Exercise 11

Let ξ_k be independent symmetric random variables and let $S_k = \xi_1 + \ldots + \xi_k$. Then for $t > 0$, we have

$$\mathbf{P}\left\{\max_{k\le n} S_k \ge t\right\} \le 2\mathbf{P}\left\{S_n \ge t\right\}.$$

Hint: Let T be the first time k when $S_k \ge t$. Given everything up to time $T = k$, consider the two possible future paths formed by $(\xi_{k+1}, \ldots, \xi_n)$ and $(-\xi_{k+1}, \ldots, -\xi_n)$. If $S_T \ge t$, then clearly for at least one of these two continuations, we must have $S_n \ge t$. Can you make this reasoning precise and deduce the inequality?

For a general super-martingale or sub-martingale, we can write similar inequalities that control the running maximum of the martingale in terms of the end-point.

Lemma 12: Doob's inequalities

Let X be a super-martingale. Then for any $t > 0$ and any $n \ge 1$,

(1) $\mathbf{P}\{\max_{k\le n} X_k \ge t\} \le \frac{1}{t}\{\mathbf{E}[X_0] + \mathbf{E}[(X_n)_-]\}$,

(2) $\mathbf{P}\{\min_{k\le n} X_k \le -t\} \le \frac{1}{t}\mathbf{E}[(X_n)_-]$.

Proof. Let $T = \min\{k : X_k \ge t\} \wedge n$. By OST $\{X_0, X_T\}$ is a super-martingale and hence $\mathbf{E}[X_T] \le \mathbf{E}[X_0]$. But

$$\begin{aligned}\mathbf{E}[X_T] &= \mathbf{E}[X_T \mathbf{1}_{X_T\ge t}] + \mathbf{E}[X_T \mathbf{1}_{X_T<t}] \\ &= \mathbf{E}[X_T \mathbf{1}_{X_T\ge t}] + \mathbf{E}[X_n \mathbf{1}_{X_T<t}] \\ &\ge \mathbf{E}[X_T \mathbf{1}_{X_T\ge t}] - \mathbf{E}[(X_n)_-]\end{aligned}$$

since $\mathbf{E}[X_n\mathbf{1}_A] \ge -\mathbf{E}[(X_n)_-]$ for any A. Thus, $\mathbf{E}[X_T\mathbf{1}_{X_T\ge t}] \le \mathbf{E}[X_0] + \mathbf{E}[(X_n)_-]$. Now use Chebyshev's inequality to write $\mathbf{P}\{X_T \ge t\} \le \frac{1}{t}\mathbf{E}[X_T\mathbf{1}_{X_T\ge t}]$ to get the first inequality.

For the second inequality, define $T = \min\{k : X_k \le -t\} \wedge n$. By OST $\{(X_T), (X_n)\}$ is a super-martingale and hence $\mathbf{E}[X_T] \ge \mathbf{E}[X_n]$. But

$$\begin{aligned}\mathbf{E}[X_T] &= \mathbf{E}[X_T \mathbf{1}_{X_T\le -t}] + \mathbf{E}[X_n \mathbf{1}_{X_T>-t}] \\ &\le -t\mathbf{P}\{X_T \le -t\} + \mathbf{E}[(X_n)_+].\end{aligned}$$

Hence $\mathbf{P}\{X_T \le -t\} \le \frac{1}{t}\{\mathbf{E}[(X_n)_+] - \mathbf{E}[X_n]\} = \frac{1}{t}\mathbf{E}[(X_n)_-]$. ■

For convenience, let us write down the corresponding inequalities for sub-martingales (which of course follow by applying Lemma 12 to $-X$): If $X_0, \ldots, X_n$ is a sub-martingale, then for any

$t > 0$ we have

$$\mathbf{P}\{\max_{k\le n} X_k \ge t\} \le \frac{1}{t}\mathbf{E}[(X_n)_+], \tag{4}$$

$$\mathbf{P}\{\min_{k\le n} X_k \le -t\} \le \frac{1}{t}\left\{-\mathbf{E}[X_0] + \mathbf{E}[(X_n)_+]\right\}. \tag{5}$$

If ξ_i are independent with zero mean and finite variances and $S_n = \xi_1 + \ldots + \xi_n$ is the corresponding random walk, then the above inequality when applied to the sub-martingale S_k^2 reduces to Kolmogorov's maximal inequality.

Maximal inequalities are useful in proving the Cauchy property of partial sums of a random series with independent summands. Here is an exercise.

Exercise 12

Let ξ_n be independent random variables with zero means. Assume that $\sum_n \text{Var}(\xi_n) < \infty$. Show that $\sum_k \xi_k$ converges almost surely. [Extra: If interested, extend this to independent ξ_ks taking values in a separable Hilbert space H such that $\mathbf{E}[\langle \xi_k, u\rangle] = 0$ for all $u \in H$ and such that $\sum_n \mathbf{E}[\|\xi_n\|^2] < \infty$.]

12. Doob's up-crossing inequality

For a real sequence $x_0, x_1, \ldots, x_n$ and any $a < b$, define the number of up-crossings of the sequence over the interval $[a, b]$ as the maximum number k for which there exist indices $0 \le i_1 < j_1 < i_2 < j_2 < \ldots < i_k < j_k \le n$ such that $x_{i_r} \le a$ and $x_{j_r} \ge b$ for all $r = 1, 2, \ldots, k$. Intuitively it is the number of times the sequence crosses the interval in the upward direction. Similarly we can define the number of down-crossings of the sequence (same as the number of up-crossings of the sequence $(-x_k)_{0\le k\le n}$ over the interval $[-b, -a]$).

Lemma 13: Doob's up-crossing inequality

Let $X_0, \ldots, X_n$ be a sub-martingale. Let $U_n[a, b]$ denote the number of up-crossings of the sequence $X_0, \ldots, X_n$ over the interval $[a, b]$. Then,

$$\mathbf{E}[U_n[a,b] \mid \mathcal{F}_0] \le \frac{\mathbf{E}[(X_n - a)_+ \mid \mathcal{F}_0] - (X_0 - a)_+}{b - a}.$$

What is the importance of this inequality? It will be in showing the convergence of martingales or super-martingales under some mild conditions. In continuous time, it will yield regularity of paths of martingales (existence of right and left limits).

The basic point is that a real sequence $(x_n)_n$ converges if and only if the number of up-crossings of the sequence over any interval is finite. Indeed, if $\liminf x_n < a < b < \limsup x_n$, then the sequence has infinitely many up-crossings and down-crossings over $[a, b]$. Conversely, if $\lim x_n$ exists, then the sequence is Cauchy and hence over any interval $[a, b]$ with $a < b$, there can be only finitely many up-crossings. In these statements the limit could be $\pm\infty$.

Proof. First assume that $X_n \geq 0$ for all n and that $a = 0$. Let $T_0 = 0$ and define the stopping times

$$T_1 := \min\{k \geq T_0 : X_k = 0\},\ T_3 := \min\{k \geq T_2 : X_k = 0\}, \quad \ldots$$

$$T_2 := \min\{k \geq T_1 : X_k \geq b\},\ T_4 := \min\{k \geq T_3 : X_k \geq b\}, \quad \ldots$$

where the minimum of an empty set is defined to be n. T_i are strictly increasing up to a point when T_k becomes equal to n and then the later ones are also equal to n. In what follows we only need T_k for $k \leq n$ (thus all the sums are finite sums).

$$\begin{aligned} X_n - X_0 &= \sum_{k\geq 0} X(T_{2k+1}) - X(T_{2k}) + \sum_{k \geq 1} X(T_{2k}) - X(T_{2k-1}) \\ &\geq \sum_{k \geq 0}(X(T_{2k+1}) - X(T_{2k})) + bU_n[0,b]. \end{aligned}$$

The last inequality is because for each k for which $X(T_{2k}) \geq b$, we get one up-crossing and the corresponding increment $X(T_{2k}) - X(T_{2k-1}) \geq b$.

Now, by the optional sampling theorem (since $T_{2k+1} \geq T_{2k}$ are both bounded stopping times), we see that

$$\mathbf{E}[X(T_{2k+1}) - X(T_{2k}) \mid \mathcal{F}_0] = \mathbf{E}\left[\mathbf{E}[X(T_{2k+1}) - X(T_{2k}) \mid \mathcal{F}_{T_{2k}}] \mid \mathcal{F}_0\right] \geq 0.$$

Therefore, $\mathbf{E}[X_n - X_0 \mid \mathcal{F}_0] \geq b\mathbf{E}[U_n[0,b] \mid \mathcal{F}_0]$. This gives the up-crossing inequality when $a = 0$ and $X_n \geq 0$.

In the general situation, just apply the derived inequality to the sub-martingale $(X_k - a)_+$ (which crosses $[0, b-a]$ whenever X crosses $[a,b]$) to get

$$\mathbf{E}[(X_n - a)_+ \mid \mathcal{F}_0] - (X_0 - a)_+ \geq (b-a)\mathbf{E}[U_n[a,b] \mid \mathcal{F}_0]$$

which is what we claimed. ■

The break up of $X_n - X_0$ over up-crossing and down-crossings was okay, but how did the expectations of increments over down-crossings become non-negative? There is a distinct sense of something suspicious about this! The point is that $X(T_3) - X(T_2)$, for example, is not always non-negative. If X never goes below a after T_2, then it can be positive too. Indeed, the sub-martingale property somehow ensures that this positive part off sets the $-(b-a)$ increment that would occur if $X(T_3)$ did go below a.

We invoked OST in the proof. Optional sampling was in turn proved using the gambling lemma. It is an instructive exercise to write out the proof of the up-crossing inequality directly using the gambling lemma (start betting when below a, stop betting when reach above b, etc.).

13. Convergence theorem for super-martingales

Now we come to the most important part of the theory.

Theorem 14: Super-martingale convergence theorem

Let X be a super-martingale on $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$. Assume that $\sup_n \mathbf{E}[(X_n)_-] < \infty$.

(1) Then, $X_n \overset{a.s.}{\to} X_\infty$ for some integrable (hence finite) random variable X_∞.

(2) In addition, $X_n \to X_\infty$ in L^1 if and only if $\{X_n\}$ is uniformly integrable. If this happens, we also have $\mathbf{E}[X_\infty \mid \mathcal{F}_n] \le X_n$ for each n.

In other words, when a super-martingale does not explode to $-\infty$ (in the mild sense of $\mathbf{E}[(X_n)_-]$ being bounded), it must converge almost surely!

Proof. Fix $a < b$. Let $D_n[a,b]$ be the number of down-crossings of $X_0, \ldots, X_n$ over $[a,b]$. By applying the up-crossing inequality to the sub-martingale $-X$ and the interval $[-b,-a]$, and taking expectations, we get

$$\begin{aligned}\mathbf{E}[D_n[a,b]] &\le \frac{\mathbf{E}[(X_n-b)_-] - \mathbf{E}[(X_0-b)_-]}{b-a} \\ &\le \frac{1}{b-a}(\mathbf{E}[(X_n)_-] + |b|) \le \frac{1}{b-a}(M+|b|)\end{aligned}$$

where $M = \sup_n \mathbf{E}[(X_n)_-]$. Let $D[a,b]$ be the number of down-crossings of the whole sequence (X_n) over the interval $[a,b]$. Then $D_n[a,b] \uparrow D[a,b]$ and hence by MCT we see that $\mathbf{E}[D[a,b]] < \infty$. In particular, $D[a,b] < \infty$ w.p.1.

Consequently, $\mathbf{P}\{D[a,b] < \infty \text{ for all } a < b, \;\; a, b \in \mathbb{Q}\} = 1$. Thus, X_n converges w.p.1., and we define X_∞ as the limit (for ω in the zero probability set where the limit does not exist, define X_∞ as 0). Thus $X_n \overset{a.s.}{\to} X_\infty$.

We observe that $\mathbf{E}[|X_n|] = \mathbf{E}[X_n] + 2\mathbf{E}[(X_n)_-] \le \mathbf{E}[X_0] + 2M$. By Fatou's lemma, $\mathbf{E}[|X_\infty|] \le \liminf \mathbf{E}[|X_n|] \le 2M + \mathbf{E}[X_0]$. Thus X_∞ is integrable.

This proves the first part. The second part is very general - whenever $X_n \overset{a.s.}{\to} X$, we have L^1 convergence if and only if $\{X_n\}$ is uniformly integrable. Lastly, $\mathbf{E}[X_{n+m} \mid \mathcal{F}_n] \le X_n$ for any $n, m \ge 1$. Let $m \to \infty$ and use L^1 convergence of X_{n+m} to X_∞ to get $\mathbf{E}[X_\infty \mid \mathcal{F}_n] \le X_n$.

This completes the proof. ■

A direct corollary that is often used is

Corollary 15

A non-negative super-martingale converges almost surely to a finite random variable.

14. Convergence theorem for martingales

Now we deduce the consequences for martingales.

Theorem 16: Martingale convergence theorem

Let $X = (X_n)_{n\geq 0}$ be a martingale with respect to $\mathcal{F}_\bullet$. Assume that X is L^1-bounded.

(1) Then, $X_n \overset{a.s.}{\to} X_\infty$ for some integrable (in particular, finite) random variable X_∞.

(2) In addition, $X_n \overset{L^1}{\to} X_\infty$ if and only if X is uniformly integrable. In this case, $\mathbf{E}[X_\infty \mid \mathcal{F}_n] = X_n$ for all n.

(3) If X is L^p bounded for some $p > 1$, then $X_\infty \in L^p$ and $X_n \overset{L^p}{\to} X_\infty$.

Observe that for a martingale the condition of L^1-boundedness, $\sup_n \mathbf{E}[|X_n|] < \infty$, is equivalent to the weaker looking condition $\sup_n \mathbf{E}[(X_n)_-] < \infty$, since $\mathbf{E}[|X_n|] - 2\mathbf{E}[(X_n)_-] = \mathbf{E}[X_n] = \mathbf{E}[X_0]$ is a constant.

Proof. The first two parts of the proof are immediate since a martingale is also a super-martingale. To conclude $\mathbf{E}[X_\infty \mid \mathcal{F}_n] = X_n$, we apply the corresponding inequality in the super-martingale convergence theorem to both X and to $-X$.

For the third part, if X is L^p bounded, then it is uniformly integrable and hence $X_n \to X_\infty$ $a.s.$and in L^1. To get L^p convergence, consider the non-negative sub-martingale $\{|X_n|\}$ and let $X^* = \sup_n |X_n|$. From Lemma 17 we conclude that $X^* \in L^p$. Of course, X^* dominates $|X_n|$ and $|X_\infty|$. Hence,

$$|X_n - X_\infty|^p \leq 2^{p-1}(|X_n|^p + |X_\infty|^p) \leq 2^p(X^*)^p$$

by the inequality $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ by the convexity of $x \mapsto |x|^p$. Thus, $|X_n - X_\infty|^p \overset{a.s.}{\to} 0$ and the sequence is dominated by $2^p(X^*)^p$ which is integrable. Dominated convergence theorem shows that $\mathbf{E}[|X_n - X_\infty|^p] \to 0$. ■

We used the following lemma in the above proof.

Lemma 17

Let $(Y_n)_{n\geq 0}$ be an L^p-bounded non-negative sub-martingale. Then $Y^* := \sup_n Y_n$ is in L^p and in fact $\mathbf{E}[(Y^*)^p] \leq C_p \sup_n \mathbf{E}[Y_n^p]$ where $C_p = \left(\frac{p}{p-1}\right)^p$.

Proof. Let $Y_n^* = \max_{k\leq n} Y_k$. Fix $\lambda > 0$ and let $T = \min\{k \geq 0 : Y_k \geq \lambda\}$. By the optional sampling theorem, for any fixed n, the sequence of two random variables $\{Y_{T\wedge n}, Y_n\}$ is a sub-martingale. Hence, $\int_A Y_n dP \geq \int_A Y_{T\wedge n} dP$ for any $A \in \mathcal{F}_{T\wedge n}$. Let $A = \{Y_{T\wedge n} \geq \lambda\}$ so that $\mathbf{E}[Y_n \mathbf{1}_A] \geq \mathbf{E}[Y_{T\wedge n}\mathbf{1}_{Y_{T\wedge n}\geq\lambda}] \geq \lambda\mathbf{P}\{Y_n^* \geq \lambda\}$. On the other hand, $\mathbf{E}[Y_n\mathbf{1}_A] \leq \mathbf{E}[Y_n\mathbf{1}_{Y^*\geq\lambda}]$ since $Y_n^* \leq Y^*$. Thus, $\lambda\mathbf{P}\{Y_n^* > \lambda\} \leq \mathbf{E}[Y_n\mathbf{1}_{Y^*\geq\lambda}]$.

Let $n \to \infty$. Since $Y_n^* \uparrow Y^*$, we get

$$\lambda\mathbf{P}\{Y^* > \lambda\} \leq \limsup_{n\to\infty} \lambda\mathbf{P}\{Y_n^* \geq \lambda\} \leq \limsup_{n\to\infty} \mathbf{E}[Y_n\mathbf{1}_{Y^*\geq\lambda}] = \mathbf{E}[Y_\infty\mathbf{1}_{Y^*\geq\lambda}].$$

where Y_∞ is the a.s. and L^1 limit of Y_n (exists, because $\{Y_n\}$ is L^p bounded and hence uniformly integrable). To go from the tail bound to the bound on pth moment, we use the identity $\mathbf{E}[(Y^*)^p] = \int_0^\infty p\lambda^{p-1}\mathbf{P}\{Y^* \ge \lambda\}d\lambda$ valid for any non-negative random variable in place of Y^*. Using the tail bound, we get

$$\begin{aligned}\mathbf{E}[(Y^*)^p] &\le \int_0^\infty p\lambda^{p-2}\mathbf{E}[Y_\infty \mathbf{1}_{Y^*\ge\lambda}]d\lambda \le \mathbf{E}\left[\int_0^\infty p\lambda^{p-2}Y_\infty \mathbf{1}_{Y^*\ge\lambda}d\lambda\right] \quad \text{(by Fubini's)} \\ &= \frac{p}{p-1}\mathbf{E}[Y_\infty \cdot (Y^*)^{p-1}].\end{aligned}$$

Let q be such that $\frac{1}{q}+\frac{1}{p}=1$. By Hölder's inequality, $E[Y_\infty \cdot (Y^*)^{p-1}] \le \mathbf{E}[Y_\infty^p]^{\frac{1}{p}}\mathbf{E}[(Y^*)^{q(p-1)}]^{\frac{1}{q}}$. Since $q(p-1)=p$, this gives us $\mathbf{E}[(Y^*)^p]^{1-\frac{1}{q}} \le \frac{p}{p-1}\mathbf{E}[Y_\infty^p]^{\frac{1}{p}}$. Hence, $\mathbf{E}[(Y^*)^p] \le C_p\mathbf{E}[Y_\infty^p]$ with $C_p = (p/(1-p))^p$. By virtue of Fatou's lemma, $\mathbf{E}[Y_\infty^p] \le \liminf \mathbf{E}[Y_n^p] \le \sup_n \mathbf{E}[Y_n^p]$. Thus, $\mathbf{E}[(Y^*)^p] \le C_p \sup_n \mathbf{E}[Y_n^p]$. ■

Alternately, from the inequality $\lambda\mathbf{P}\{Y_n^* > \lambda\} \le \mathbf{E}[Y_n\mathbf{1}_{Y^*\ge\lambda}]$ we could have (by similar steps, but without letting $n\to\infty$) arrived at a bound of the form $\mathbf{E}[(Y_n^*)^p] \le C_p\mathbf{E}[Y_n^p]$. The right hand side is bounded by $C_p\sup_n \mathbf{E}[Y_n^p]$ while the left hand side increases to $\mathbf{E}[(Y^*)^p]$ by monotone convergence theorem. This is another way to complete the proof.

One way to think of the martingale convergence theorem is that we have extended the martingale from the index set $\mathbb{N}$ to $\mathbb{N}\cup\{+\infty\}$ retaining the martingale property. Indeed, the given martingale sequence is the Doob martingale given by the limit variable X_∞ with respect to the given filtration.

While almost sure convergence is remarkable, it is not strong enough to yield useful conclusions. Convergence in L^1 or L^p for some $p\ge 1$ are much more useful. In this context, it is important to note that L^1-bounded martingales do not necessarily converge in L^1.

Example 17: Critical branching process

Consider a Galton-Watson tree (branching process) with mean off-spring distribution equal to 1 (any non-degenerate distribution will do, eg., Poisson(1)). Then if Z_n denotes the number of individuals in the nth generation (we start with $Z_0 = 1$), then Z_n is a non-negative martingale, and $\mathbf{E}[Z_n]=1$, hence it is L^1-bounded. But $Z_\infty = 0$ (either recall this fact from previous classes, or prove it from the martingale convergence theorem!). Thus $\mathbf{E}[Z_n] \not\to \mathbf{E}[Z_\infty]$, showing that L^1-convergence fails.

15. Another approach to martingale convergence theorems

The proof of almost sure convergence via upcrossings is something unusual and fascinating about the proof of martingale convergence. It is worth pondering whether it can be reworded in a more familiar form by showing that X_n is almost surely a Cauchy sequence, for example. In this section we make an attempt by first showing it for L^2-bounded martingales and then approximating general L^1-bounded martingales by L^2-bounded ones. As written, this does not give a complete proof because of a gap in the argument indicated later. However, the part about square integrable martingales is correct and complete.

Theorem 18: Square integrable martingales

Let X be a martingale on $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ such that $\sup_n \mathbf{E}[X_n^2] < \infty$. Then there is an L^2 random variable X_∞ such that $X_n \to X_\infty$ $a.s.$ and in L^2.

First, a basic observation about a square integrable martingale X. Assume $\mathbf{E}[X_n^2] < \infty$ for each n (no need for a uniform bound). By the projection interprtation of conditional expectations, $X_{n+1} - X_n$ is orthogonal to $L^2(\Omega, \mathcal{F}_n, \mathbf{P})$. In particular, $\{X_{k+1} - X_k\}_{k \geq 0}$ is an orthogonal set in $L^2(\Omega, \mathcal{F}, \mathbf{P})$ and hence for any $m > n$, we have

$$\mathbf{E}[(X_m - X_n)^2] = \sum_{k=n}^{m-1} \mathbf{E}[(X_{k+1} - X_k)^2]. \tag{6}$$

Proof of Theorem 18. Applying (6) with $n = 0$ and letting $m \to \infty$, we see that $\mathbf{E}[(X_{k+1} - X_k)^2]$ is summable. Hence, as $m, n \to \infty$, we see that $\mathbf{E}[(X_m - X_n)^2] \to 0$, again by (6), since the right hand side is the tail of a convergent series. Thus, $\{X_n\}$ is a Cauchy sequence in L^2 and hence there is some $X_\infty \in L^2$ such that $X_n \to X_\infty$ in L^2.

We now show almost sure convergence. Indeed, if $m > n$, then by Doob's maximal inequality applied to the sub-martingale $\{|X_k - X_n|\}_{k \geq n}$, we get

$$\mathbf{P}\{\max_{n \leq k \leq m} |X_k - X_n| \geq \epsilon\} \leq \frac{\mathbf{E}[|X_m - X_n|]}{\epsilon} \leq \frac{1}{\epsilon}\sqrt{\mathbf{E}[(X_m - X_n)^2]}.$$

As the latter goes to zero as $m, n \to \infty$, we see that

$$\mathbf{P}\{|X_k - X_j| \geq \epsilon \text{ for some } k > j > N\} \to 0 \text{ as } N \to \infty.$$

Let $\epsilon = \frac{1}{\ell}$ and choose N_ℓ so that the probability of the event on the left is less than $\frac{1}{\ell^2}$. By Borel-Cantelli lemma, almost surely, only finitely many of these events occur. Therefore, the sequence $\{X_n\}$ is a Cauchy sequence, almost surely. Thus $X_n \overset{a.s.}{\to} X'_\infty$ for some X'_∞. However, the L^2 limit is X_∞, therefore $X_\infty = X'_\infty$ $a.s.$ ■

What we have done is to give a completely self-contained proof (using Doob's maximal inequality, but not the upcrossing inequality) that an L^2-bounded martingale converges almost surely and in L^2. Can this idea be used to prove the theorem for L^1-bounded martingales? By Doob's inequality we can write again

$$\mathbf{P}\{\max_{n \leq k \leq m} |X_k - X_n| \geq \epsilon\} \leq \frac{\mathbf{E}[|X_m - X_n|]}{\epsilon}$$

but the right hand side need not go to zero, as we do not know that $\{X_n\}$ is Cauchy in L^1. Indeed, there are L^1-bounded martingales that do not converge in L^1 (eg., the generation sizes in a critical branching process). In probability, often the easiest way to show that something is finite is to show that it has finite expectation. For the number of times that X changes by more than ϵ (i.e., maximum k such that there is some $n_0 < n_1 < \ldots < n_k$ such that $|X_{n_{i+1}} - X_{n_i}| \geq \epsilon$), the bound in terms of $|X_n - X_m|$ does not help. Perhaps this motivated Doob to consider a fixed interval and the number of times the martingale crosses it...

However, we can try to use the theorem for square integrable martingales to derive the theorem for L^1-bounded martingales by approximation or to use the technical term, *localization*, that is, by using stopping times such that the stopped process is L^2-bounded.

Proof of almost sure convergence of L^1-bounded martingales. Let X be an L^1-bounded martingale. For a positive integer M and let $\tau_M = \min\{k : |X_k| \geq M\}$. Then $\{X(\tau_M \wedge n)\}_{n \geq 0}$ is a martingale.

Can we say that $X(\tau_M \wedge n)$ is also L^2-bounded? The problem is that when X exits $[-M, M]$, it may do so by a large amount. This is the gap in the argument that we referred to earlier. If we assume that the differences $X(j+1) - X(j)$ are uniformly bounded, then it is clear that $X(\tau_M \wedge n)$ is uniformly bounded and hence bounded in L^2. The proof is okay for such cases, but if you figure out how to fix the proof in general, or that it cannot be fixed, please do let me know!

If we assume that $\{X(\tau_M \wedge n)\}$ is L^2-bounded, then by the theorem for square integrable martingales, there is some $Z_M \in L^2$ such that $X(\tau_M \wedge n) \to Z_M$ $a.s.$ and in L^2, as $n \to \infty$.

Further, applying Doob's maximal inequality, if $C = \sup_n \mathbf{E}[|X_n|]$, then

$$\mathbf{P}\{\tau_M < \infty\} = \lim_{n \to \infty} \mathbf{P}\{\tau_M \leq n\} \leq \frac{1}{M}\mathbf{E}[|X_n|] \leq \frac{C}{M}.$$

From this it follows that $A = \cup_M \{\tau_M = \infty\}$ has probability 1. Further, on the event $\{\tau_M = \infty\}$, it is clear that $Z_{M'} = Z_M$ for all $M' < M$. Therefore, it follows that we can consistently define a random variable Z by setting it equal to Z_M on the event $\{\tau_M = \infty\}$. It is then clear that $X_n \overset{a.s.}{\to} Z$ on the event A. Since $\mathbf{P}(A) = 1$, we have proved that $X_n \overset{a.s.}{\to} Z$.

The integrability of Z follows by Fatou's lemma and the remaining parts of the martingale convergence theorem (that uniform integrability implies L^1 convergence etc.) are general facts that follows once we have almost sure convergence. ■

16. Reverse martingales

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let $\mathcal{F}_i$, $i \in I$ be sub-sigma algebras of $\mathcal{F}$ indexed by a partially ordered set $(I, \leq)$ such that $\mathcal{F}_i \subseteq \mathcal{F}_j$ whenever $i \leq j$. Then, we may define a martingale or a sub-martingale etc., with respect to this "filtration" $(\mathcal{F}_i)_{i \in I}$. For example, a martingale is a collection of integrable random variables X_i, $i \in I$ such that $\mathbf{E}[X_j \mid \mathcal{F}_i] = X_i$ whenever $i \leq j$.

If the index set is $-\mathbb{N} = \{0, -1, -2, \ldots\}$ with the usual order, we say that X is a reverse martingale or a reverse sub-martingale etc.

What is different about reverse martingales as compared to martingales is that our questions will be about the behaviour as $n \to -\infty$, towards the direction of decreasing information. It turns out that the results are even cleaner than for martingales!

Theorem 19: Reverse martingale convergence theorem

Let $X = (X_n)_{n\leq 0}$ be a reverse martingale. Then $\{X_n\}$ is uniformly integrable. Further, there exists a random variable $X_{-\infty}$ such that $X_n \to X_{-\infty}$ almost surely and in L^1.

Proof. Since $X_n = \mathbf{E}[X_0 \mid \mathcal{F}_n]$ for all n, the uniform integrability follows from Exercise 13.

Let $U_n[a,b]$ be the number of down-crossings of $X_n, X_{n+1}, \ldots, X_0$ over $[a,b]$. The up-crossing inequality (applied to $X_n, \ldots, X_0$ over $[a,b]$) gives $\mathbf{E}[U_n[a,b]] \leq \frac{1}{b-a}\mathbf{E}[(X_0 - a)_+]$. Thus, the expected number of up-crossings $U_\infty[a,b]$ by the full sequence $(X_n)_{n\leq 0}$ has finite expectation, and hence is finite w.p.1.

As before, w.p.1., the number of down-crossings over any interval with rational end-points is finite. Hence, $\lim_{n\to-\infty} X_n$ exists almost surely. Call this $X_{-\infty}$. Uniform integrability shows that convergence also takes place in L^1. ■

The following exercise was used in the proof.

Exercise 13

Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbf{P})$. Then the collection $\{\mathbf{E}[X \mid \mathcal{G}] : \mathcal{G} \subseteq \mathcal{F}\}$ is uniformly integrable.

What about reverse super-martingales or reverse sub-martingales? Although we shall probably have no occasion to use this, here is the theorem which can be proved on the same lines.

Theorem 20

Let $(X_n)_{n\leq 0}$ be a reverse super-martingale. Assume that $\sup_n \mathbf{E}[X_n] < \infty$. Then $\{X_n\}$ is uniformly integrable and X_n converges almost surely and in L^1 to some random variable $X_{-\infty}$.

Proof. Exercise. ■

This covers almost all the general theory that we want to develop. The rest of the course will consist in milking these theorems to get many interesting consequences.

17. Application: Lévy's forward and backward laws

Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$.

Question 1: If $\mathcal{F}_n$, $n \geq 0$, is an increasing sequence of sigma-algebras, then what happens to the sequence $\mathbf{E}[X \mid \mathcal{F}_n]$ as $n \to \infty$?

Question 2: If $\mathcal{G}_n$, $n \geq 0$ is a decreasing sequence of sigma-algebras, then what happens to $\mathbf{E}[X \mid \mathcal{G}_n]$ as $n \to \infty$.

Note that the question here is different from conditional MCT. The random variable is fixed and the sigm-algebras are changing. A natural guess is that the limit might be $\mathbf{E}[X \mid \mathcal{F}_\infty]$ and $\mathbf{E}[X \mid \mathcal{G}_\infty]$ respectively, where $\mathcal{F}_\infty = \sigma\{\bigcup_n \mathcal{F}_n\}$ and $\mathcal{G}_\infty = \bigcap_n \mathcal{G}_n$. We shall prove that these guesses are correct.

Forward case: The sequence $X_n = \mathbf{E}[X \mid \mathcal{F}_n]$ is a martingale because of the tower property $\mathbf{E}[\mathbf{E}[X \mid \mathcal{F}_n] \mid \mathcal{F}_m] = \mathbf{E}[X \mid \mathcal{F}_m]$ for $m < n$. Recall that such martingales are called Doob martingales.

Being conditional expectations of a given X, the martingale is uniformly integrable and hence X_n converges $a.s.$and in L^1 to some X_∞. We claim that $X_\infty = \mathbf{E}[X \mid \mathcal{F}_\infty]$ $a.s.$.

Indeed, X_n is $\mathcal{F}_\infty$-measurable for each n and hence the limit X_∞ is $\mathcal{F}_\infty$-measurable (since the convergence is almost sure, there is a null set issue which might make it necessary to either complete the sigma-algebras, or you may interpret it as saying that X_∞ can be modified on a set of zero probability to make it $\mathcal{F}_\infty$-measurable).

Define the measure μ and ν on $\mathcal{F}_\infty$ by $\mu(A) = \int_A XdP$ and $\nu(A) = \int_A X_\infty dP$ for $A \in \mathcal{F}_\infty$. What we want to show is that $\mu(A) = \nu(A)$ for all $A \in \mathcal{F}_\infty$. If $A \in \mathcal{F}_m$, then for any $n > m$, we have

$$\int_A XdP = \int_A X_m P = \int_A X_n dP \overset{n\to\infty}{\longrightarrow} \int_A X_\infty dP.$$

The first inequality holds because $X_m = \mathbf{E}[X \mid \mathcal{F}_m]$ and the second equality holds because $X_m = \mathbf{E}[X_n \mid \mathcal{F}_m]$ for $n > m$. The last convergence holds because $X_n \to X$ in L^1. Comparing the first and last quantities in the above display, we see that $\mu(A) = \nu(A)$ for all $A \in \bigcup_m \mathcal{F}_m$.

Thus, $\bigcup_n \mathcal{F}_n$ is a π-system on which μ and ν agree. By the $\pi - \lambda$ theorem, they agree of $\mathcal{F}_\infty = \sigma\{\bigcup_n \mathcal{F}_n\}$. This completes the proof that $\mathbf{E}[X \mid \mathcal{F}_n] \overset{a.s.,\, L^1}{\longrightarrow} \mathbf{E}[X \mid \mathcal{F}_\infty]$.

Backward case: Write $X_{-n} = \mathbf{E}[X \mid \mathcal{G}_n]$ for $n \in \mathbb{N}$. Then X is a reverse martingale w.r.t the filtration $\mathcal{G}_{-n}$, $n \in \mathbb{N}$. By the reverse martingale convergence theorem, we get that X_n converges almost surely and in L^1 to some X_∞.

We claim that $X_\infty = \mathbf{E}[X \mid \mathcal{G}_\infty]$. Since X_∞ is $\mathcal{G}_n$ measurable for every n (being the limit of X_k, $k \geq n$), it follows that X_∞ is $\mathcal{G}_\infty$-measurable. Let $A \in \mathcal{G}_\infty$. Then $A \in \mathcal{G}_n$ for any n and hence $\int_A XdP = \int_A X_n dP$ which converges to $\int_A X_\infty dP$. Thus, $\int_A XdP = \int_A X_\infty dP$ for all $A \in \mathcal{F}_\infty$.

18. Kolmogorov's zero-one law

As a corollary of the forward law, we may prove Kolmogorov's zero-one law.

Theorem 21: Kolmogoro'v zero-one law

Let ξ_n, $n \geq 1$ be independent random variables and let $\mathcal{T} = \bigcap_n \sigma\{\xi_n, \xi_{n+1}, \ldots\}$ be the tail sigma-algebra of this sequence. Then $\mathbf{P}(A)$ is 0 or 1 for every $A \in \mathcal{T}$.

Proof. Let $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$. Then $\mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n] \to \mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_\infty]$ in L^1 and almost surely. But $\mathcal{F}_\infty = \sigma\{\xi_1, \xi_2, \ldots\}$. Thus if $A \in \mathcal{T} \subseteq \mathcal{F}_\infty$ then $\mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_\infty] = \mathbf{1}_A$ *a.s.* On the other hand, $A \in \sigma\{\xi_{n+1}, \xi_{n+2}, \ldots\}$ from which it follws that A is independent of $\mathcal{F}_n$ and hence $\mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n] = \mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$. The conclusion is that $\mathbf{1}_A = \mathbf{P}(A)$ *a.s.*, which is possible if and only if $\mathbf{P}(A)$ equals 0 or 1. ■

19. Strong law of large numbers

The strong law of large number under first moment condition is an easy consequence of the reverse martingale theorem.

Theorem 22

Let ξ_n, $n \geq 1$ be i.i.d. real-valued random variables with zero mean and let $S_n = \xi_1 + \ldots + \xi_n$. Then $\frac{1}{n} S_n \overset{a.s.}{\to} 0$.

Proof. Let $\mathcal{G}_n = \sigma\{S_n, S_{n+1}, \ldots\} = \sigma\{S_n, \xi_{n+1}, \xi_{n+2}, \ldots\}$, a decreasing sequence of sigma-algebras. Hence $M_{-n} := \mathbf{E}[\xi_1 \mid \mathcal{G}_n]$ is a reverse martingale and hence converges almost surely and in L^1 to some $M_{-\infty}$.

But $\mathbf{E}[\xi_1 \mid \mathcal{G}_n] = \frac{1}{n} S_n$ (why?). Thus, $\frac{1}{n} S_n \to M_{-\infty}$ almost surely and in L^1. But the limit of $\frac{1}{n} S_n$ is clearly a tail random variable of ξ_ns and hence must be constant. Thus, $M_{-\infty} = \mathbf{E}[M_{-\infty}] = \lim \frac{1}{n} \mathbf{E}[S_n] = 0$. In conclusion, $\frac{1}{n} S_n \overset{a.s.}{\to} 0$. ■

20. Critical branching process

Let Z_n, $n \geq 0$ be the generation sizes of a Galton-Watson tree with offspring distribution $p = (p_k)_{k \geq 0}$. If $m = \sum_k k p_k$ is the mean, then Z_n / m^n is a martingale (we saw this earlier).

If $m < 1$, then $\mathbf{P}\{Z_n \geq 1\} \leq \mathbf{E}[Z_n] = m^n \to 0$ and hence, the branching process becomes extinct w.p.1. For $m = 1$ this argument fails. We show using martingales that extinction happens even in this cases.

Theorem 23

If $m = 1$ and $p_1 \neq 1$, then the branching process becomes extinct almost surely.

Proof. If $m = 1$, then Z_n is a non-negative martingale and hence converges almost surely to a finite random variable Z_∞. But Z_n is integer-valued. Thus,

$$Z_\infty = j \Leftrightarrow Z_n = j \text{ for all } n \geq n_0 \text{ for some } n_0.$$

But if $j \neq 0$ and $p_1 < 1$, then it is easy to see that $\mathbf{P}\{Z_n = j \text{ for all } n \geq n_0\} = 0$ (since conditional on $\mathcal{F}_{n-1}$, there is a positive probability of p_0^j that $Z_n = 0$). Thus, $Z_n = 0$ eventually. ■

In the supercritical case we know that there is a positive probability of survival. If you do not know this, prove it using the second moment method as follows.

Exercise 14

By conditioning on $\mathcal{F}_{n-1}$ (or by conditioning on $\mathcal{F}_1$), show that (1) $\mathbf{E}[Z_n] = m^n$, (2) $\mathbf{E}[Z_n^2] \asymp (1+\sigma^2)m^{2n}$. Deduce that $\mathbf{P}\{Z_n > 0\}$ stays bounded away from zero. Conclude positive probability of survival.

We also have the martingale Z_n/m^n. By the martingale convergence theorem $W := \lim Z_n/m^n$ exists, $a.s.$ On the event of extinction, clearly $W = 0$. On the event of survival, is it necessarily the case that $W > 0$ $a.s.$? If yes, this means that whenever the branching process survives, it does so by growing exponentially, since $Z_n \sim W \cdot m^n$. The answer is given by the famous Kesten-Stigum theorem.

Theorem 24: Kesten-Stigum theorem

Assume that $\mathbf{E}[L] > 1$ and that $p_1 \neq 1$. Then, $W > 0$ almost surely on the event of survival if and only if $\mathbf{E}[L \log_+ L] < \infty$.

We now prove a weaker form of this, that if $\mathbf{E}[L^2] < \infty$, then $W > 0$ on the event of survival (this was in fact proved by Kolmogorov earlier).

Kesten-Stigum under finite variance condition. Assume $\sigma^2 = \mathbf{E}[L^2] < \infty$. Then by Exercise 14, $\frac{Z_n}{m^n}$ is an L^2 bounded martingale. Therefore it converges to W almost surely and in L^2. In particular, $\mathbf{P}(W = 0) < 1$. However, by conditioning on the first generation, we see that $q = \mathbf{P}\{W = 0\}$ satisfies the equation $q = \mathbf{E}[q^L]$ (if the first generation has L children, in each of the trees under these individuals, the corresponding $W_i = 0$ and these W_i are independent). But the usual proof of the extinction theorem shows that there are only two solutions to the equation $q = \mathbf{E}[q^L]$, namely 1 and the extinction probability of the tree. Since we have see that $q < 1$, it must be equal to the extinction probability. That is $W > 0$ $a.s.$ on the event of survival. ■

21. Pólya's urn scheme

Initially the urn contain b black and w white balls. Let B_n be the number of black balls after n steps. Then $W_n = b + w + n - B_n$. We have seen that $X_n := B_n/(B_n + W_n)$ is a martingale. Since $0 \le X_n \le 1$, uniform integrability is obvious and $X_n \to X_\infty$ almost surely and in L^1. Since X_n are bounded, the convergence is also in L^p for every p. In particular, $\mathbf{E}[X_n^k] \to \mathbf{E}[X_\infty^k]$ as $n \to \infty$ for each $k \ge 1$.

Theorem 25

X_∞ bas Beta(b, w) distribution.

Proof. Let V_k be the colour of the kth ball drawn. It takes values 1 (for black) and 0 (for white). It is an easy exercise to check that

$$\mathbf{P}\{V_1 = \epsilon_1, \ldots, V_m = \epsilon_m\} = \frac{b(b+1)\ldots(b+r-1)w(w+1)\ldots(w+s-1)}{(b+w)(b+w+1)\ldots(b+w+n-1)}$$

if $r = \epsilon_1 + \ldots + \epsilon_m$ and $s = n - r$. The key point is that the probability does not depend on the order of ϵ_is. In other words, any permutation of $(V_1, \ldots, V_n)$ has the same distribution as $(V_1, \ldots, V_n)$, a property called *exchangeability*.

From this, we see that for any $0 \le r \le n$, we have

$$\mathbf{P}\{X_n = \frac{b+r}{b+w+n}\} = \binom{n}{r} \frac{b(b+1)\ldots(b+r-1)w(w+1)\ldots(w+(n-r)-1)}{(b+w)(b+w+1)\ldots(b+w+n-1)}.$$

In the simplest case of $b = w = 1$, the right hand side is $\frac{1}{n+1}$. That is, X_n takes the values $\frac{r+1}{n+2}$, $0 \le r \le n$, with equal probabilities. Clearly then $X_n \overset{d}{\to}$ Unif$[0,1]$. Hence, $X_\infty \sim$ Unif$[0,1]$. In general, we leave it as an exercise to show that X_∞ has Beta(b, w) distribution. ■

Here is a possibly clever way to avoid computations in the last step.

Exercise 15

For each initial value of b, w, let $\mu_{b,w}$ be the distribution of X_∞ when the urn starts with b black and w white balls. Each $\mu_{b,w}$ is a probability measure on $[0,1]$.

(1) Show that $\mu_{b,w} = \frac{b}{b+w}\mu_{b+1,w} + \frac{w}{b+w}\mu_{b,w+1}$.

(2) Check that Beta(b, w) distributions satisfy the above recursions.

(3) Assuming $(b, w) \mapsto \mu_{b,w}$ is continuous, deduce that $\mu_{b,w} =$ Beta(b, w) is the only solution to the recursion.

One can introduce many variants of Pólya's urn scheme. For example, whenever a ball is picked, we may add r balls of the same color and q balls of the opposite color. That changes the behaviour of the urn greatly and in a typical case, the proportions of black balls converges to a constant.

Here is a muti-color version which shares all the features of Pólya's urn above.

Multi-color Pólya's urn scheme: We have ℓ colors denoted $1, 2, \ldots, \ell$. Initially an urn contains $b_k > 0$ balls of color k (b_k need not be integers). At each step of the process, a ball is drawn uniformly at random from the urn, its color noted, and returned to the urn with another ball of the same color. Let $B_k(n)$ be the number of balls of kth color after n draws. Let ξ_n be the color of the ball drawn in the nth draw.

Exercise 16

(1) Show that $\frac{1}{n+b_1+\ldots+b_\ell}(B_1(n),\ldots,B_\ell(n))$ converges almost surely (and in L^p for any p) to some random vector $(Q_1,\ldots,Q_\ell)$.

(2) Show that $\xi_1,\xi_2,\ldots$ is an exchangeable sequence.

(3) For $b_1=\ldots=b_\ell=1$, show that $(Q_1,\ldots,Q_\ell)$ has Dirichlet$(1,1,\ldots,1)$ distribution. In general, it has Dirichlet$(b_1,\ldots,b_\ell)$ distribution.

This means that $Q_1+\ldots+Q_\ell=1$ and $(Q_1,\ldots,Q_{\ell-1})$ has density

$$\frac{\Gamma(b_1+\ldots+b_\ell)}{\Gamma(b_1)\ldots\Gamma(b_\ell)}x_1^{b_1-1}\ldots x_{\ell-1}^{b_{\ell-1}-1}(1-x_1-\ldots-x_{\ell-1})^{b_\ell-1}$$

on $\Delta=\{(x_1,\ldots,x_{\ell-1}):x_i>0 \text{ for all } i \text{ and } x_1+\ldots+x_{\ell-1}<1\}$.

Blackwell-Macqueen urn scheme: Here is a generalization of Pólya's urn scheme to infinitely many colours. Start with the unit line segment $[0,1]$, each point of which is thought of as a distinct colour. Pick a uniform random variable V_1, after which we add a line segment of length 1 that has colour V_1. Now we have the original line segment and a new line segment, and we draw a point uniformly at random from the union of the two line segments. If it falls in the original segment at location V_2, a new line segment of colour V_2 is added and if it falls in the segment of colour V_1, then a new line segment of length 1 having colour V_1 is added. The process continues.

If one considers the situation after the first step, the colour V_1 is like the black in a Pólya's urn scheme with $b=1=w$. Hence the proportion of V_1 converges almost surely to $P_1\sim$ unif$[0,1]$. When the kth colour appears, it appears with a line segment of length 1 and the original line segment has length 1. If we ignore all the points that fall in the other coloured segments that have appeared before, then again we have a Pólya urn with $b=w=1$. This leads to the following conclusion: The proportions of the colours that appear, in the order of appearance, converges almost surely to $(P_1,P_2,\ldots)$ where $P_1=U_1$, $P_2=(1-U_1)U_2$, $P_3=(1-U_1)(1-U_2)U_3$, …where U_i are i.i.d. uniform random variables on $[0,1]$.

The random vector P has a distribution on the infinite simplex $\Delta=\{(p_1,p_2,\ldots):p_i\geq 0,\ \sum_i p_i=1\}$ that is known as a GEM distribution (for Griffiths-Engel-McCloskey) and random vector $P^{\downarrow}$ got from P by ranking the co-ordinates in decreasing order is said to have Poisson-Dirichlet distribution (on the ordered simplex $\Delta^{\downarrow}=\{(p_1,p_2,\ldots):p_1\geq p_2\geq\ldots\geq 0 \text{ and } \sum_i p_i=1\}$. If we allow the initial stick to have length $\theta>0$ (the segments added still have length 1), then the resulting distribution on Δ and $\Delta^{\downarrow}$ are called GEM$(0,\theta)$ and PD$(0,\theta)$ distributions.

22. Liouville's theorem

Recall that a harmonic function on $\mathbb{Z}^2$ is a function $f:\mathbb{Z}^2\to\mathbb{R}$ such that $f(x)=\frac{1}{4}\sum_{y:y\sim x}f(y)$ for all $x\in\mathbb{Z}^2$.

Theorem 26: Liouville's theorem

If f is a non-constant harmonic function on $\mathbb{Z}^2$, then $\sup f = +\infty$ and $\inf f = -\infty$.

Proof. If not, by negating and/or adding a constant we may assume that $f \geq 0$. Let X_n be simple random walk on $\mathbb{Z}^2$. Then $f(X_n)$ is a martingale. But a non-negative super-martingale converges almost surely. Hence $f(X_n)$ converges almost surely.

But Pólya's theorem says that X_n visits every vertex of $\mathbb{Z}^2$ infinitely often w.p.1. This contradicts the convergence of $f(X_n)$ unless f is a constant. ■

Observe that the proof shows that a non-constant super-harmonic function on $\mathbb{Z}^2$ cannot be bounded below. The proof uses recurrence of the random walk. But in fact the same theorem holds on $\mathbb{Z}^d$, $d \geq 3$, although the simple random walk is transient there.

For completeness, here is a quick proof of Pólya's theorem in two dimensions.

Exercise 17

Let S_n be simple symmetric random walk on $\mathbb{Z}^2$ started at $(0,0)$.

(1) Show that $\mathbf{P}\{S_{2n} = (0,0)\} = \frac{1}{4^{2n}} \sum_{k=0}^{n} \frac{(2n)!}{k!^2(n-k)!^2}$ and that this expression reduces to $\left(\frac{1}{2^{2n}}\binom{2n}{n}\right)^2$.

(2) Use Stirling's formula to show that $\sum_n \mathbf{P}\{S_{2n} = (0,0)\} = \infty$.

(3) Conclude that $\mathbf{P}\{S_n = (0,0) \text{ i.o.}\} = 1$.

The question of existence of bounded or positive harmonic functions on a graph (or in the continuous setting) is important. Here are two things that we may cover if we get time.

- There are no bounded harmonic functions on $\mathbb{Z}^d$ (Blackwell).
- Let μ be a probability measure on $\mathbb{R}$ and let f be a harmonic function for the random walk with step distribution μ. This just means that f is continuous and $\int_{\mathbb{R}} f(x+a)d\mu(x) = f(a)$. Is f necessarily constant? We shall discuss this later (under the heading "Choquet-Deny theorem").

23. Hewitt-Savage zero-one law

There are many zero-one laws in probability, asserting that a whole class of events are trivial. For a sequence of random variables, here are three important classes of such events.

Below, ξ_n, $n \geq 1$, are random variables on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and taking values in $(X, \mathcal{F})$. Then $\xi = (\xi_n)_{n\geq 1}$ is a random variable taking values in $(X^{\mathbb{N}}, \mathcal{F}^{\otimes \mathbb{N}})$. These definitions can be extended to two sided-sequences $(\xi_n)_{n\in\mathbb{Z}}$ easily.

(1) The *tail sigma-algebra* is defined as $\mathcal{T} = \cap_n \mathcal{T}_n$ where $\mathcal{T}_n = \sigma\{\xi_n, \xi_{n+1}, \ldots\}$.

(2) The *exchangeable sigma-algebra* $\mathcal{S}$ is the sigma-algebra of those events that are invariant under finite permutations.

More precisely, let G be the sub-group (under composition) of all bijections $\pi : \mathbb{N} \to \mathbb{N}$ such that $\pi(n) = n$ for all but finitely many n. It is clear how G acts on $X^{\mathbb{N}}$:

$$\pi((\omega_n)) = (\omega_{\pi(n)}).$$

Then

$$\mathcal{S} := \{\xi^{-1}(A) : A \in \mathcal{F}^{\otimes \mathbb{N}} \text{ and } \pi(A) = A \text{ for all } \pi \in G\}.$$

If G_n is the sub-group of $\pi \in G$ such that $\pi(k) = k$ for every $k > n$ and $\mathcal{S}_n := \{\xi^{-1}(A) : A \in \mathcal{F}^{\otimes \mathbb{N}}$ and $\pi(A) = A$ for all $\pi \in G_n\}$, then $\mathcal{S}_n$ are sigma-algebras that decrease to $\mathcal{S}$.

(3) The *translation-invariant sigma-algebra* $\mathcal{I}$ is the sigma-algebra of all events invariant under translations.

More precisely, let $\theta_n : X^{\mathbb{N}} \to X^{\mathbb{N}}$ be the translation map $[\theta_n(\omega)]_k = \omega_{n+k}$. Then, $\mathcal{I} = \{A \in \mathcal{F}^{\otimes \mathbb{N}} : \theta_n(A) = A \text{ for all } n \in \mathbb{N}\}$ (these are events invariant under the action of the semi-group $\mathbb{N}$).

Kolmogorov's zero-one law asserts that under and product measure $\mu_1 \otimes \mu_2 \otimes \ldots$, the tail sigma-algebra is trivial. Ergodicity is the statement that $\mathcal{I}$ is trivial and it is true for i.i.d. product measures $\mu^{\otimes \mathbb{N}}$. The exchangeable sigma-algebra is also trivial under i.i.d. product measure, which is the result we prove in this section. First an example.

Example 18

The event $A = \{\omega \in \mathbb{R}^{\mathbb{N}} : \lim \omega_n = 0\}$ is an invariant event. In fact, every tail event is an invariant event. But the converse is not true. For example,

$$A = \{\omega \in \mathbb{R}^{\mathbb{N}} : \lim_{n\to\infty}(\omega_1 + \ldots + \omega_n) \text{ exists and is at most } 0\}$$

is an invariant event but not a tail event. This is because $\omega = (-1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots)$ belongs to A and so does every finite permutation of ω as the sum does not change. But changing the first co-ordinate to 0 gives $\omega' = (0, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots)$, which is not in A.

Theorem 27: Hewitt-Savage 0-1 law

Let μ be a probability measure on $(X, \mathcal{F})$. Then the invariant sigma-algebra $\mathcal{S}$ is trivial under the product measure $\mu^{\otimes \mathbb{N}}$.

In terms of random variables, we may state this as follows: Let ξ_n be i.i.d. random variables taking values in X. Let $f : X^{\mathbb{N}} \mapsto \mathbb{R}$ be a measurable function such that $f \circ \pi = f$ for all $\pi \in G$. Then, $f(\xi_1, \xi_2, \ldots)$ is almost surely a constant.

We give a proof using reverse martingale theorem. There are also more direct proofs.

Proof. For any integrable Y (that is measurable w.r.t $\mathcal{F}^{\otimes \mathbb{N}}$), the sequence $\mathbf{E}[Y \mid \mathcal{S}_n]$ is a reverse martingale and hence $\mathbf{E}[Y \mid \mathcal{S}_n] \overset{a.s., L^1}{\longrightarrow} \mathbf{E}[Y \mid \mathcal{S}]$.

Now fix $k \geq 1$ and let $\varphi : X^k \to \mathbb{R}$ be a bounded measurable function. Take $Y = \varphi(X_1, \ldots, X_k)$. We claim that

$$\mathbf{E}[\varphi(X_1, \ldots, X_k) \mid \mathcal{S}_n] = \frac{1}{n(n-1)\ldots(n-k+1)} \sum_{\substack{1 \leq i_1, \ldots, i_k \leq n \\ \text{distinct}}} \varphi(X_{i_1}, \ldots, X_{i_k}).$$

To see this, observe that by symmetry (since $\mathcal{S}_n$ does no distinguish between $X_1, \ldots, X_n$), we have $\mathbf{E}[\varphi(X_{i_1}, \ldots, X_{i_k}) \mid \mathcal{S}_n]$ is the same for all distinct $i_1, \ldots, i_k \leq n$. When you add all these up, we get

$$\mathbf{E}\left[\sum_{\substack{1 \leq i_1, \ldots, i_k \leq n \\ \text{distinct}}} \varphi(X_{i_1}, \ldots, X_{i_k}) \,\Big|\, \mathcal{S}_n \right] = \sum_{\substack{1 \leq i_1, \ldots, i_k \leq n \\ \text{distinct}}} \varphi(X_{i_1}, \ldots, X_{i_k})$$

since the latter is clearly $\mathcal{S}_n$-measurable. There are $n(n-1)\ldots(n-k+1)$ terms on the left, each of which is equal to $\mathbf{E}[\varphi(X_1, \ldots, X_k) \mid \mathcal{S}_n]$. This proves the claim.

Together with the reverse martingale theorem, we have shown that

$$\frac{1}{n(n-1)\ldots(n-k+1)} \sum_{\substack{1 \leq i_1, \ldots, i_k \leq n \\ \text{distinct}}} \varphi(X_{i_1}, \ldots, X_{i_k}) \overset{a.s.,\, L^1}{\longrightarrow} \mathbf{E}[\varphi(X_1, \ldots, X_k) \mid \mathcal{S}].$$

The number of summands on the left in which X_1 participates is $k(n-1)(n-2)\ldots(n-k+1)$. If $|\varphi| \leq M_\varphi$, then the total contribution of all terms containing X_1 is at most

$$M_\varphi \frac{k(n-1)(n-2)\ldots(n-k+1)}{n(n-1)(n-2)\ldots(n-k+1)} \to 0$$

as $n \to \infty$. Thus, the limit is a function of $X_2, X_3, \ldots$. By a similar reasoning, the limit is a tail-random variable for the sequence $X_1, X_2, \ldots$. By Kolmogorov's zero-one law it must be a constant (then the constant must be its expectation). Hence,

$$\mathbf{E}[\varphi(X_1, \ldots, X_k) \mid \mathcal{S}] = \mathbf{E}[\varphi(X_1, \ldots, X_k)].$$

As this is true for every bounded measurable φ, we see that $\mathcal{S}$ is independent of $\sigma\{X_1, \ldots, X_k\}$. As this is true for every k, $\mathcal{S}$ is independent of $\sigma\{X_1, X_2, \ldots\}$. But $\mathcal{S} \subseteq \sigma\{X_1, X_2, \ldots\}$ and therefore $\mathcal{S}$ is independent of itself. This implies that for any $A \in \mathcal{S}$ we must have $\mathbf{P}(A) = \mathbf{P}(A \cap A) = \mathbf{P}(A)^2$ which implies that $\mathbf{P}(A)$ equals 0 or 1. ■

24. Exchangeable random variables

Let ξ_n, $n \geq 1$, be any sequence of random variables. Recall that this means that

$$(\xi_{\pi(1)}, \xi_{\pi(2)}, \ldots) \overset{d}{=} (\xi_1, \xi_2, \ldots)$$

for any bijection (permutation) $\pi : \mathbb{N} \mapsto \mathbb{N}$ that fixes all but finitely many elements. Since distribution of infinitely many random variables is nothing but the collection of all finite dimensional distributions, this is equivalent to saying that

$$(\xi_{i_1}, \ldots, \xi_{i_n}) \overset{d}{=} (\xi_1, \ldots, \xi_n)$$

for any $n \geq 1$ and any distinct $i_1, \dots, i_n$.

We have seen an example of an exchangeable sequence in Pólya's urn scheme, namely the successive colours drawn.

Example 19

If ξ_n are i.i.d., then they are exchangeable. More generally, consider finitely many probability measures $\mu_1, \dots, \mu_k$ on some $(\Omega, \mathcal{F})$ and let $p_1, \dots, p_k$ be positive numbers that add up to 1. Pick $L \in \{1, \dots, k\}$ with probabilities $p_1, \dots, p_k$, and conditional on L, pick an i.i.d. sequence $\xi_1, \xi_2, \dots$ from μ_L. Then (unconditionally) ξ_is are exchangeable but not independent.

The above example essentially covers everything, according to a fundamental theorem of de Finetti! Before stating it, let us recall the exchangeable sigma-algebra $\mathcal{S}$ of the collection of all sets in the product sigma-algebra $\mathcal{F}^{\otimes \mathbb{N}}$ that are invariant under finite permutations of co-ordinates. Let us also define $\mathcal{S}_n$ as the collection of all events invariant under the permutations of the first n co-ordinates. The $\mathcal{S} = \cap_n \mathcal{S}_n$.

Theorem 28: de Finetti

Let $\xi_1, \xi_2, \dots$ be an exchangeable sequence of random variables taking values in $(X, \mathcal{F})$. Then, they are i.i.d. conditional on $\mathcal{S}$. By this we mean that

$$\mathbf{E}\left[\varphi_1(\xi_1) \dots \varphi_k(\xi_k) \mid \mathcal{S}\right] = \prod_{j=1}^{k} \mathbf{E}[\varphi_j(\xi_1) \mid \mathcal{S}]$$

for any $k \geq 1$ and any bounded measurable $\varphi_j : X \mapsto \mathbb{R}$.

If the situation is nice enough that a regular conditional probability given $\mathcal{S}$ exists, then the statement is equivalent to saying that the conditional distribution is (almost surely) a product of identical probability distributions on $\mathcal{F}$.

Before proving this, let us prove a lemma very similar to the one we used in the proof of the Hewitt-Savage zero one law.

Lemma 29

Let $\xi_1, \xi_2, \dots$ be an exchangeable sequence taking values in $(X, \mathcal{F})$. Fix $k \geq 1$ and any bounded measurable $\psi : X^k \mapsto \mathbb{R}$. Then, as $n \to \infty$

$$\frac{1}{n^k} \sum_{1 \leq i_1, \dots, i_k \leq n} \psi(\xi_{i_1}, \dots, \xi_{i_k}) \overset{a.s.}{\to} \mathbf{E}\left[\psi(\xi_1, \dots, \xi_k) \mid \mathcal{S}\right].$$

Proof. We claim that

$$\mathbf{E}\left[\psi(\xi_1, \dots, \xi_k) \mid \mathcal{S}_n\right] = \frac{1}{n(n-1)\dots(n-k+1)} \sum_{i_1, \dots, i_k \leq n}' \psi(\xi_{i_1}, \dots, \xi_{i_k}) \tag{7}$$

where $\sum'_{i_1,\ldots,i_k\le n}$ denotes summation over distinct $i_1,\ldots,i_k \le n$. The reason is that the right hand side is clearly in $\mathcal{S}_n$ (since it is a symmetric function of $\xi_1,\ldots,\xi_n$). Further, if $Z = g(\xi_1,\ldots,\xi_n)$ where g is a symmetric measurable bounded function from X^n to $\mathbb{R}$, then for any permutation π of $[n]$,

$$\begin{aligned}\mathbf{E}[Z\psi(\xi_1,\ldots,\xi_k)] &= \mathbf{E}[g(\xi_{\pi(1)},\ldots,\xi_{\pi(n)})psi(\xi_{\pi(1)},\ldots,\xi_{\pi(k)})]\\ &= \mathbf{E}[g(\xi_1,\ldots,\xi_n)psi(\xi_{\pi(1)},\ldots,\xi_{\pi(k)})]\end{aligned}$$

where the first line used the exchangeability of ξ_is and the second used the symmetry of g. By such symmetric functions generate the sigma-algebra $\mathcal{S}_n$, hence this shows that $\mathbf{E}[psi(\xi_{\pi(1)},\ldots,\xi_{\pi(k)}) \mid \mathcal{S}_n]$ is the same for all permutations π of $[n]$. Therefore the expectation of the right hand side of (7) is also the same.

Now, by Lévy's backward law (or reverse martingale theorem) we know that $\mathbf{E}[\varphi(\xi_1,\ldots,\xi_k) \mid \mathcal{S}_n]$ converges to $\mathbf{E}[\varphi(\xi_1,\ldots,\xi_k) \mid \mathcal{S}]$. On the right hand side, we may replace $n(n-1)\ldots(n-k+1)$ by n^k (the ratio goes to 1 as $n\to\infty$) and extend the sum to all $i_1,\ldots,i_k$ since the number of terms with at least two equal indices is of order n^{k-1} and its contribution is at most $\|\psi\|_{\sup}$ (thus the contribution gets washed away when divided by n^k). ■

Now we prove de Finetti's theorem.

Proof of de Finetti's theorem. By the lemma applied to $\psi(x_1,\ldots,x_k)=\varphi_1(x_1)\ldots\varphi_k(x_k)$,

$$\frac{1}{n^k}\sum_{i_1,\ldots,i_k\le n}\varphi_1(\xi_{i_1})\ldots\varphi_k(\xi_{i_k}) \overset{a.s.}{\to} \mathbf{E}\left[\varphi_1(X_1)\ldots\varphi_k(X_k) \mid \mathcal{S}\right].$$

On the other hand, the left hand side factors into a product of $\frac{1}{n}\sum_{i=1}^n \varphi_\ell(x_i)$ over $\ell = 1,2,\ldots,k$, and again by the Lemma the ℓth factor converges almost surely to $\mathbf{E}[\varphi_\ell(\xi_1) \mid \mathcal{S}]$. This proves the theorem. ■

There are many alternate ways to state the thorem of de Finetti. One is to say that every exchangeable measure is a convex combination of i.i.d. product measures. Another way is this:

If $(\xi_n)_n$ is an exchangeable sequence of random variables taking values in a Polish space X, then there exists a Borel measurable function $f:[0,1]\times[0,1]\mapsto X$ such that

$$(\xi_1,\xi_2,\xi_3,\ldots) \overset{d}{=} (f(V,V_1),f(V,V_2),f(V,V_3),\ldots)$$

where $V,V_1,V_2,\ldots$ are i.i.d.uniform$[0,1]$ random variables. Here V represents the common information contained in $\mathcal{S}$, and conditional on that, the variables are i.i.d.

24.1. **About the exchangeable sigma algebra.** Suppose X_i are i.i.d. By the Hewitt-Savage zero-one law, the exchangeable sigma algebra $\mathcal{S}$ is trivial. What is it in the case of a general exchangeable sequence $(X_n)_n$? To get an idea, first consider the case where X_ns take values in a finite set A. Then, by the lemma above, $\frac{1}{n}\sum_{k=1}^n \mathbf{1}_{X_k=a}$ converges almost surely to some $\theta(a)$ for each $a\in A$. Then θ is a random probability vector on A. Further, for any fixed n, it is clear that $\mathcal{S}_n$ is precisely the sigma algebra generated by $\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{X_k=a}$, $a \in A$. This suggests that the exchangeable sigma-algebra $\mathcal{S}$ must be just the sigma-algebra generated by θ (i.e., by $\theta(a)$, $a \in A$). To fill up with a precise statement

This also gives a way to think of de Finetti's theorem (in fact this was implicit in the proof). Think of an exchangeable sequence of random variables taking values in a finite set A. Then when we condition on $\mathcal{S}_n$, we know the number of times each $a \in A$ appears among $X_1, \ldots, X_n$. In other words, we know the multi-set $\{X_1, \ldots, X_n\}$. By exchangeablity, the conditional distribution of $(X_1, \ldots, X_n)$ is uniform distribution on all sequences in A^n that are consistent with these frequencies. Put another way, from the multi-set $\{X_1, \ldots, X_n\}$, sample n times without replacement, and place the elements in the order that they are sampled. If we fix a k and consider $X_1, \ldots, X_k$, then for large n sampling without replacement and sampling with replacement are essentially the same, which is the statement that $X_1, \ldots, X_k$, given $\mathcal{S}_n$, are approximately i.i.d.

25. Absolute continuity and singularity of product measures

Consider a sequence of independent random variables X_n (they may take values in different spaces). We are told that either (1) $X_n \sim \mu_n$ for each n or (2) $X_n \sim \nu_n$ for each n.Here μ_n and ν_n are given probability distributions. From one realization of the sequence $(X_1, X_2, \ldots)$, can we tell whether the first situation happened or the second?

In measure theory terms, the question may be formulated as follows.

Question: Let μ_n, ν_n be probability measures on $(\Omega_n, \mathcal{G}_n)$. Let $\Omega = \times_n \Omega_n$, $\mathcal{F} = \otimes_n \mathcal{G}_n$ and $\mu = \otimes_n \mu_n$ and $\nu = \otimes \nu_n$. Then, μ, ν are probability measures on $(\Omega, \mathcal{F})$. Assume that $\nu_n \ll \mu_n$ for each n. Can we say whether (1) $\nu \ll \mu$, (2) $\nu \perp \mu$ or (3) neither of the previous two options?

Let us consider a concrete example where direct calculations settle the above question. It also serves to show that both $\nu \perp \mu$ and $\nu \ll \mu$ are possibilities.

Example 20

Let $\mu_n = \text{unif}[0,1]$ and $\nu_n = \text{unif}[0, 1+\delta_n]$. Then, $\nu[0,1]^{\mathbb{N}} = \prod_n \frac{1}{1+\delta_n}$. Thus, if $\prod_n (1+\delta_n) = \infty$, then $\mu[0,1]^{\mathbb{N}} = 1$ while $\nu[0,1]^{\mathbb{N}} = 0$. Thus, $\mu \perp \nu$.

On the other hand, if $\prod_n (1+\delta_n) < \infty$, then we claim that $\nu \ll \mu$. To see this, pick U_n, V_n be i.i.d. unif$[0,1]$. Define $X_n = (1+\delta_n)U_n \sim \nu_n$. Further, set

$$Y_n = \begin{cases} X_n & \text{if } X_n \leq 1, \\ V_n & \text{if } X_n > 1. \end{cases}$$

Check that V_n are i.i.d with uniform distribution on $[0,1]$. In short, $(X_1, X_2, \dots) \sim \nu$ and $(Y_1, Y_2, \dots) \sim \mu$. Now,

$$\begin{aligned} \mathbf{P}\{X_n = Y_n \text{ for all n}\} &= \mathbf{P}\{X_n \le (1+\delta_n)^{-1} \text{ for all n}\} \\ &= \prod_{n=1}^{\infty} \frac{1}{1+\delta_n} \end{aligned}$$

which is positive by assumption. Thus, there is a way to construct $X \sim \mu$ and $Y \sim \nu$ such that $X = Y$ with positive probability. Then we cannot possibly have $\mu \perp \nu$ (in itself this is not enough to say that $\nu \ll \mu$).

We used the special properties of uniform distribution to settle the above example. In general it is not that easy, but Kakutani provided a complete answer.

Theorem 30: Kakutani's theorem

Let μ_n, ν_n be probability measures on $(\Omega_n, \mathcal{F}_n)$ and assume that $\mu_n \ll \nu_n$ with Radon-Nikodym theorem f_n Let $\mu = \otimes_n \mu_n$ and $\nu = \otimes_n \nu_n$, probability measures on $\Omega = \times_n \Omega_n$ with the product sigma algebra. Let $a_n = \int_{\Omega_n} \sqrt{f_n} d\nu_n$. Then,$f(x) := \prod_{k=1}^{\infty} f_k(x_k)$ converges ν-almost surely

(1) If $\prod_{k=1}^{\infty} a_k > 0$, then $\mu \ll \nu$ and and $d\mu(x) = f(x)\, d\nu(x)$.

(2) If $\prod_{k=1}^{\infty} a_k = 0$, then $\mu \perp \nu$.

First we prove a general lemma about product martingales.

Lemma 31

Let ξ_n be independent positive random variables with mean 1 and let $X_n = \xi_1 \xi_2 \dots \xi_n$ be the corresponding product martingale. Let $a_n = \mathbf{E}[\sqrt{\xi_n}]$ and let X_∞ be the almost sure limit of X_ns. Then there are two possibilities.

(1) $\prod_n a_n > 0$. In this case, $\{X_n\}$ is uniformly integrable, $\mathbf{E}[X_\infty] = 1$. If $\xi_n > 0 \; a.s.$ for all n, then $X_\infty > 0 \; a.s.$

(2) $\prod_n a_n = 0$. In this case, $\{X_n\}$ is not uniformly integrable and $X_\infty = 0 \; a.s.$

Observe that $a_k \le \sqrt{\mathbf{E}[\xi_k]} = 1$ for all k. Hence the partial products $\prod_{j=1}^{n} a_j$ are decreasing in n and have a limit in $[0,1]$, which is what we mean by $\prod_n a_n$.

Proof. Let $Y_n = \prod_{j=1}^{n} \frac{\xi_j}{\sqrt{a_j}}$. Then X_n and Y_n are both martingales (w.r.t. $\mathcal{F}_n = \sigma\{\xi_1, \ldots, \xi_n\}$) and are related as $X_n = Y_n^2 a_1^2 \ldots a_n^2$. As they are non-negative and have mean 1, we also know that $X_n \overset{a.s.}{\to} X_\infty$ and $Y_n \overset{a.s.}{\to} Y_\infty$ where X_∞ and Y_∞ are integrable (hence finite almost surely).

Suppose $\prod\limits_n a_n > 0$. Then $\mathbf{E}[Y_n^2] = \frac{1}{a_1 \ldots a_n}$ is uniformly bounded above. Hence Y is an L^2-bounded martingale and therefore $Y_n \to Y_\infty$ in L^2. In particular, Y_n^2 converges to Y_∞^2 almost surely and in L^1, which implies that $\{Y_n^2\}$ must be uniformly integrable. But $X_n \le Y_n^2$ (as $a_j \le 1$ for all j), which means that X is uniformly integrable. In particular, we also have $\mathbf{E}[X_\infty] = \lim_{n\to\infty} \mathbf{E}[X_n] = 1$. In particular, $\mathbf{P}\{X_\infty > 0\} > 0$. But if ξ_ns are strictly positive, then the event $\{X_\infty > 0\}$ is a tail event of $(\xi_n)_n$, hence by Kolmogorov's zero one law it must have probability 1.

Suppose $\prod\limits_n a_n = 0$. Observe that $X_\infty = Y_\infty^2 \prod\limits_j a_j^2$ and Y_∞ is a finite random variable. Hence $X_\infty = 0 \; a.s.$. ■

Proof of Kakutani's theorem. Define $\xi_n(\omega) = f_n(\omega_n)$ for $\omega \in \Omega$. Under the measure ν, the ξ_n are independent random variables with mean 1. Now form the product martingales (with respect to the νmeasure) X_n and Y_n as in the proof of Lemma 31.

If $\prod\limits_n a_n > 0$, then $\{X_n\}$ is uniformly integrable and $\mathbf{E}[X_\infty] = 1$ by that Lemma. We also know that $\mathbf{E}[X_\infty \mid \mathcal{F}_k] = X_k$ for any k by the martingale convergence theorem (for u.i. martingales). Define the measure θ on $(\Omega, \mathcal{F})$ by $d\theta(\omega) = X_\infty(\omega) d\nu(\omega)$. Then if $A \in \mathcal{F}_k$ for some k, we have

$$\theta(A) = \int_A X_\infty d\nu = \int_A X_k d\nu = \mu(A).$$

Thus μ and θ are two probability measures that agree on the π-system $\cup_k \mathcal{F}_k$. Hence they agree on the generated sigma algebra $\mathcal{F}$. That is μ has Radon-Nikodym derivative X_∞ w.r.t. ν.

If $\prod\limits_n a_n = 0$, then by Lemma 31, we see that $X_\infty = 0 \; a.s.[\nu]$. We show that $\mu\{X_\infty = 0\} = 0$, which of course shows that $\mu \perp \nu$. To show this, fix any $\epsilon > 0$ and observe that for n, we have (since $\{X_n < \epsilon\} \in \mathcal{F}_n$ and X_n is the Radon-Nikodym derivative of $\mu|_{\mathcal{F}_n}$ w.r.t. $\nu|_{\mathcal{F}_n}$)

$$\mu\{X_n < \epsilon\} = \int \mathbf{1}_{X_n < \epsilon} X_n d\nu < \epsilon \nu\{X_n < \epsilon\} \le \epsilon.$$

From this it follows (how?) that $\mu\{X_\infty < \epsilon\} \le \epsilon$ for any ϵ and hence $\mu\{X_\infty = 0\} = 0$. ■

There is a more general question, which we did not cover in class. Proofs can be found in most books having a chapter on martingales.

Question': Let μ, ν be probability measures on $(\Omega, \mathcal{F})$. Suppose $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ are sub sigma-algebras of $\mathcal{F}$ such that $\sigma\{\cup_n \mathcal{F}_n\} = \mathcal{F}$. Let $\mu_n = \mu\big|_{\mathcal{F}_n}$ and $\nu_n = \nu\big|_{\mathcal{F}_n}$ be the restrictions of μ and ν to $\mathcal{F}_n$. Assume that $\nu_n \ll \mu$. Is $\nu \ll \mu$. If not are there conditions?

This subsumes the question of product measures by taking $\Omega = \times_n \Omega_n$ and $\mathcal{F}_n = \sigma\{\Pi_1, \ldots, \Pi_n\}$, the sigma algebra generated by the first n projections. The answer for this question is as follows.

Let X_n be the Radon-Nikodym derivative of μ_n w.r.t. ν_n. Then X_n is a ν-martingale and converges to some X_∞ $a.s.[\nu]$. Then $\mu = X_\infty d\nu + \mathbf{1}_{X_\infty=\infty}\mu$ gives the decomposition of μ into a part absolutely continuous to ν and a part singular to ν.

26. The Haar basis and almost sure convergence

Consider $L^2[0,1]$ with respect to the Lebesgue measure. Abstract Hilbert space theory says that L^2 is a Hilbert space, it has an orthonormal basis, and that for any orthonormal basis $\{\varphi_n\}$ and any $f \in L^2$, we have

$$f \overset{L^2}{=} \sum_n \langle f, \varphi_n \rangle \varphi_n$$

which means that the L^2-norm of the difference between the left side and the nth partial sum on the right side converges to zero as $n \to \infty$.

But since L^2 consists of functions, it is possible to ask for convergence in other senses. In general, there is no almost-sure convergence in the above series.

Theorem 32

Let $H_{n,k}$, $n \ge 1$, $0 \le k \le 2^n - 1$ be the Haar basis for L^2. Then, for any $f \in L^2$, the convergence holds almost surely.

Proof. On the probability space $([0,1], \mathcal{B}, \lambda)$, define the random variables

$$X_n(t) = \sum_{m \le n} \sum_{k \le 2^m - 1} \langle f, H_{m,k} \rangle H_{m,k}(t).$$

We claim that X_n is a martingale. Indeed, it is easy to see that if $\mathcal{F}_n := \sigma\{H_{n,0}, \ldots, H_{n,2^n-1}\}$ (which is the same as the sigma algebra generated by the intervals $[k/2^n, (k+1)/2^n)$, $0 \le k \le 2^n - 1$), then $X_n = \mathbf{E}[f \mid \mathcal{F}_n]$. Thus, $\{X_n\}$ is the Doob-martingale of f with respect to the filtration $\mathcal{F}_\cdot$.

Further, $\mathbf{E}[X_n^2] = \sum_{m \le n} \sum_{k \le 2^m - 1} |\langle f, H_{m,k} \rangle|^2 \le \|f\|_2^2$. Hence $\{X_n\}$ is an L^2-bounded martingale. It converges almost surely and in L^2. But in L^2 it converges to f. Hence $X_n \overset{a.s.}{\to} f$. ■

27. Probability measures on metric spaces

In basic probability we largely study real-valued random variables or at most $\mathbb{R}^d$-valued random variables. From the point of view of applications of probability, it is clear that there are more complex random objects. For example, consider the graph of the daily value of the rupee versus the dollar over a calendar year. For each year we get a different graph, and in some ways, the ups and downs appear to be random. While one can consider it as a vector of length 365, it may be more meaningful to think of it as defined at each time point. Hence we need the notion of a random function. There are situations where one may also want the notion of a discontinuous random function or random functions on the plane (eg., random surfaces), or random measures (eg., the length measure of the zero set of a random function from $\mathbb{R}^2$ to $\mathbb{R}$) or the set of locations of an epidemic, etc.

Probabilists have found that all applications of interest so far can be captured by allowing random variables to take values in a general complete and separable metric space. The distribution of such a random variables is a probability measure on the metric space. A key part of the theory is the notion of weak convergence of measures on such spaces. In this section, we summarize (mostly without proofs), the basic facts[10].

Let (X, d) be a complete and separable metric space. Let $\mathcal{B}_X$ denote the Borel sigma-algebra of X and let $\mathcal{P}(X)$ denote the set of all probability measures on $(X, \mathcal{B}_X)$. For $\mu, \nu \in \mathcal{P}(X)$, define

$$d(\mu, \nu) = \inf\{r > 0 : \mu(A_r) + r \geq \nu(A) \text{ and } \nu(A_r) + r \geq \mu(A) \text{ for all } A \in \mathcal{B}_X\}$$

where $A_r = \bigcup_{x \in A} B(x, r)$ is the r-neighbourhood of A (it is an open set, hence measurable).

Lemma 33: Prohorov metric

d defines a metric on $\mathcal{P}(X)$.

Observe that $x \mapsto \delta_x$ is an isometry from X to $\mathcal{P}(X)$, hence using the same letter d for the metric can be excused. If $d(\mu_n, \mu) \to 0$ for $\mu_n, \mu \in \mathcal{P}(X)$, we say that μ_n converges in distribution to μ and write $\mu_n \overset{d}{\to} \mu$.

Lemma 34: Portmanteau theorem

For $\mu_n, \mu \in \mathcal{P}(X)$, the following are equivalent.

(1) $\mu_n \overset{d}{\to} \mu$.

(2) $\int f d\mu_n \to \int f d\mu$ for all $f \in C_b(X)$.

(3) $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$ for all open $G \subseteq X$.

(4) $\limsup_{n\to\infty} \mu_n(F) \leq \mu(F)$ for all closed $F \subseteq X$.

(5) $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ for all $A \in \mathcal{B}_X$ satisfying $\mu(\partial A) = 0$.

Except for the use of distribution functions (which is not available on general metric spaces), the similarity to the situation in $\mathbb{R}$ is readily seen. The Prohorov metric also agrees with the Lévy-Prohorov distance that we had defined, except that the class of sets over which the infimum is taken was only right-closed intervals (in general metric spaces, many books take infimum only over closed sets).

Following the usual definition in metric spaces, a subset $\mathcal{A} \subseteq \mathcal{P}(X)$ is said to be relatively compact (or precompact) if every subsequence has a convergent subsequence. This is the same as saying that $\bar{\mathcal{A}}$ is compact in $(\mathcal{P}(X), d)$. The fundamental theorem is a characterization of relatively compact sets (analogous to Helly's theorem for probability measures on $\mathbb{R}$).

[10] Billingsley's book

em Convergence of probability measures or K. R. Parthasarathy's *Probability measures on metric spaces* are excellent sources to know more. Of course, Kallenberg's book has everything succinctly.

Definition 3: Tightness

We say that $\mathcal{A} \subseteq \mathcal{P}(X)$ is *tight* if, for any $\epsilon > 0$, there is a compact $K_\epsilon \subseteq X$ such that $\mu(K_\epsilon) \geq 1 - \epsilon$ for all $\mu \in \mathcal{A}$.

Theorem 35: Prokhorov's theorem

A subset $\mathcal{A} \subseteq \mathcal{P}(X)$ is relatively compact if and only if it is tight.

Corollary 36

If (X, d) is compact, then $(\mathcal{P}(X), d)$ is also compact. In general for any complete, separable (X, d), the metric space $(\mathcal{P}(X), d)$ is also complete and separable.

That completes all we want to know in general. When it comes to a specific metric space, a key thing is to be able to check tightness of a subset of measures, which involves understanding compact subsets on the metric space itself. We work out a couple of examples below and write out the conditions for checking tightness. But before that let us indicate another exceedingly useful approach to showing convergence in distribution that avoids having to know all this machinery.

Lemma 37

Let μ_n, μ belong to $\mathcal{P}(X)$. Suppose X_n, X are X-valued random variables on some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that $\mathbf{P} \circ X_n^{-1} = \mu_n$, $\mathbf{P} \circ X^{-1} = \mu$ and $X_n \to X$ $a.s.[\mathbf{P}]$. Then, $\mu_n \overset{d}{\to} \mu$

Skorokhod showed the converse, that whenever $\mu_n \overset{d}{\to} \mu$, there is a probability space and random variables X_n, X having these distributions such that $X_n \overset{d}{\to} X$. However, the useful part is the above direction, although the proof is trivial!

Proof. Let $f \in C_b(X)$. Then $f(X_n) \overset{a.s.}{\to} f(X)$ and these are bounded real-valued random variables. Hence by the dominated convergence theorem $\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]$ as $n \to \infty$. But $\mathbf{E}[f(X_n)] = \int f d\mu_n$ and $\mathbf{E}[f(X)] = \int f d\mu$, hence $\mu_n \overset{d}{\to} \mu$. ■

Observe that almost sure convergence also makes it trivial to say that for any continuous function $\varphi : X \mapsto \mathbb{R}$, we have $\varphi(X_n) \to \varphi(X)$ almost surely and hence also in distribution. Thus, various "features" of μ_n also converge in distribution to the corresponding feature of μ (i.e., $\mu_n \circ \varphi^{-1} \overset{d}{\to} \mu \circ \varphi^{-1}$, as probability measures on $\mathbb{R}$).

Example 21

Let $X = \mathbb{R}^{\mathbb{N}}$. This is a complete and separable metric space with the metric $d(x, y) = \sum_n 2^{-n}(1 \wedge |x_n - y_n|)$ for $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$.

Example 22

Let $X = C[0,1]$ with the sup-norm metric. Arzela-Ascoli theorem tell us that $K \subseteq C[0,1]$ is compact if and only if it is closed and there is an $M < \infty$ such that $|f(0)| \leq M$ for all $f \in K$ and for each $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - f(y)| \leq \epsilon$ for any $x, y \in [0,1]$ with $|x - y| \leq \delta$ and for any $f \in K$. The last condition of *equicontinuity* is the crucial one.