Martingales and Brownian motion

Manjunath Krishnapur

Contents

CHAPTER 1

Conditional probability and expectation

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 random variable. The goal is to find the closest $\mathcal{G}$-measurable random variable to X. For example, if $\mathcal{G} = \sigma(Z)$, then we want the function g so that $g(Z)$ is the closest to X. This problem of predicting one (perhaps not easily observable) random variable in terms of other (observable) random variables is one of the fundamental problems of Statistics.

To say anything, we must first decide on the sense of closeness.

Square-integrable case. If $\mathbf{E}[X^2] < \infty$, then we can ask for $\mathcal{G}$-measurable Y that minimizes $\mathbf{E}[|X - Y|^2]$. Why does is exist, and is it unique? For this, we move to equivalence classes of random variables and regard $W = L^2(\Omega, \mathcal{G}, \mathbf{P}|_{\mathcal{G}})$ as a closed subspace of the Hilbert space $H = L^2(\Omega, \mathcal{F}, \mathbf{P})$. Hilbert space theory tell us that there is a projection map $P : H \to W$ such that for any u, the unique closest vector in W is Pu. An equivalent way of stating this is that $Pu \in W$ and $\langle u, v\rangle = \langle Pu, v\rangle$ for all $v \in W$.

Thus, if $[X] \in L^2$ is the equivalence class containing X, and Y is any member of the equivalence class $P[X]$ (any two choices agree up to a $\mathbf{P}$-null $\mathcal{G}$-measurable set), then $\mathbf{E}[|X - Y|^2] \leq \mathbf{E}[|X - Z|^2]$ for any $\mathcal{G}$-measurable square integrable random variable Z. Equivalently,

(1) Y is $\mathcal{G}$-measurable and square integrable,

(2) $\mathbf{E}[XZ] = \mathbf{E}[YZ]$ for any $\mathcal{G}$-measurable, square integrable Z.

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 1

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 $Y \geq 0$ $a.s.$]

Integrable case. Now suppose we only assume that $\mathbf{E}[|X|] < \infty$. It is tempting to consider the closed subspace $W = L^1(\Omega, \mathcal{G}, \mathbf{P}|_{\mathcal{G}})$ of the Banach space $H = L^1(\Omega, \mathcal{F}, \mathbf{P})$. But there is no good projection theory in L^1, hence we cannot repeat what we did for the square integrable case.

Example 1

On $([-1,1],\mathcal{B},\lambda)$ (where λ is the uniform probability measure), let $\mathcal{G} = \sigma(x \mapsto |x|)$. Then $\mathcal{G}$-measurable random variables are precisely measurable even functions. If $X(t) = t$ and $Y = f(|t|)$, then $\mathbf{E}[|X-Y|] = \int_{-1}^{1} |t - f(|t|)| \ = \int_0^1 (|t-f(t)|+|t+f(t)|)dt$. The integrand is at least $2t$, and equality is achieved if $-t < f(t) < t$. There are infinitely many measurable f satisfying this for all $t \in [0,1]$, hence there is no unique closest $\mathcal{G}$-measurable random variable to X. Find example where existence fails

The correct analogy with the square integrable case is to think of a random variable X as acting on other random variables by $Y \mapsto \mathbf{E}[XY]$. If $X \in L^2$, the correct space of Y is L^2, while if $X \in L^1$, then the correct space of Y is L^∞. This leads us to the following definition.

Definition 1: Conditional expectation

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 and integrable, and (b) $\mathbf{E}[YZ] = \mathbf{E}[XZ]$ for all bounded $\mathcal{G}$-measurable random variables Z. Any such Y is denoted $\mathbf{E}[X \bigm| \mathcal{G}]$.

Some remarks are in order.

(1) It suffices to check the second condition for indicator variables. That is, $\mathbf{E}[Y\mathbf{1}_A] = \mathbf{E}[X\mathbf{1}_A]$ for all $A \in \mathcal{G}$. If this holds, then $\mathbf{E}[YZ] = \mathbf{E}[XZ]$ for simple $\mathcal{G}$-measurable Z. For general bounded $\mathcal{G}$-measurable Z, find simple functions Z_n such that $|Z_n| \leq |Z|$ and $Z_n \overset{a.s.}{\to} Z$. DCT applies on both sides of $\mathbf{E}[XZ_n] = \mathbf{E}[YZ_n]$ to show that $\mathbf{E}[XZ] = \mathbf{E}[YZ]$.

(2) The same reasoning as in the previous point shows that if $\mathbf{E}[|X|^p] < \infty$ for some $p > 0$, and $\mathbf{E}[X\mathbf{1}_A] = \mathbf{E}[Y\mathbf{1}_A]$ for all $A \in \mathcal{G}$, then $\mathbf{E}[XZ] = \mathbf{E}[YZ]$ for the larger class of $Z \in L^q(\Omega,\mathcal{G},\mathbf{P}|_{\mathcal{G}})$, where $\frac{1}{p} + \frac{1}{q} = 1$.

(3) Taking $p = 2$ in the previous point, we see that for square integrable X, the conditional expectation exists and is the closest $L^2(\Omega,\mathcal{G},\mathbf{P}|_{\mathcal{G}})$ random variable to X.

The main question is whether a conditional expectation exists for integrable X, and if it is unique. Yes and Yes. Before giving a proof, let us see some examples.

Example 2

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 the equality of integrals over $\emptyset$ and over Ω also hold. 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 2

Let $B_1, \ldots, B_n$ be a measurable partition of Ω. Assume that $\mathbf{P}(B_k) > 0$ for each k. 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 3

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, we claim that a version of $\mathbf{E}[X \mid \mathcal{G}]$ is given by

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

As Y is a function of Z, it is $\mathcal{G}$-measurable. 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 $\mathbf{E}[X \mid \mathcal{G}]$.

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: We give two proofs.

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 2: 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.

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). A concrete example where it does not exist is given at the end of the section. 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. Thus, another way to state the above theorem is that in a general probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and $\mathcal{G} \subseteq \mathcal{F}$, regular conditional probability w.r.t. $\mathcal{G}$ may not exist on $\mathcal{F}$, but it will exist on any sub sigma algebra $\mathcal{F}' \subseteq \mathcal{F}$ that is generated by a random variable taking values in a complete, separable metric space. We state this as a theorem.

Theorem 2

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub sigma 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 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 $A \in \mathcal{F}'$.

In this situation of $\mathcal{F}' = \sigma(X)$, one can push forward $Q(\omega, \cdot)$ to M and get probability measures $\nu_\omega = Q(\omega, \cdot) \circ X^{-1}$. Then ν_ω is called the regular conditional distribution of X given $\mathcal{G}$. For example, if $X = (X_1, \ldots, X_d)$ is $\mathbb{R}^d$-valued, then ν_ω is the measure with the distribution function

$$F_\omega(t_1, \ldots, t_d) = Q(\omega, \{X_1 \le t_1, \ldots, X_m \le t_d\}).$$

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 general fact can be deduced from the fact that any complete separable metric space M is isomorphic as a measure space to a Borel subset of the real line[1]. Of course, it should be noted that the metric plays little role, if the topology is preserved by changing the metric, we may do so. For example, $(0, 1)$ is not complete with the usual metric, but we can endow it with a complete metric.

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 $Y_t(\omega) \le Y_{t'}(\omega)$ 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

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

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 $\mathcal{G}$.

[1] 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.

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 $\mathbf{E}[\mathbf{1}_A \mid \mathcal{G}]$ and $\mathbf{E}[\mathbf{1}_B \mid \mathcal{G}]$, then by linearity of conditional expectations, $Q(\cdot, B \setminus A) = Q(\cdot, B) - Q(\cdot, A)$ 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. 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 $\mathcal{G}$. ■

Now we give an example where regular conditional probability does not exist.

Example 4

Consider $([0, 1], \mathcal{B}, \lambda)$. Let $E \subseteq [0, 1]$ be a non-measurable set with $\lambda^*(E) = 1 = \lambda^*(E^c)$. Let $\mathcal{F} = \sigma\{\mathcal{B}, E\} = \{(A \cap E) \sqcup (B \cap E^c) : A, B \in \mathcal{B}\}$. On $\mathcal{F}$, define a measure by

$$\mu((A \cap E) \sqcup (B \cap E^c)) = \frac{1}{2}\lambda(A) + \frac{1}{2}\lambda(B).$$

This is well defined because $A \cap E = A' \cap E$ implies that $A \Delta A' \subseteq E^c$ and measurable subsets of E^c have zero measure (why?). Thus, $\lambda(A) = \lambda(A')$. Similarly, $\lambda(B \cap E^c)$ does not depend on the choice of B. That μ is a measure is then easy to see.

Suppose regular conditional probability Q of μ w.r.t. $\mathcal{B}$ were to exist. Then $Q(\cdot, A) = \mathbf{1}_A$ a.e., for any $A \in \mathcal{B}$, in particular for $A = [0, x]$ for any $x \in [0, 1]$. Take intersection over $x \in \mathbb{Q}$ to see that $Q(\omega, [0, x]) = \mathbf{1}_{[0,x]}(\omega)$ for all $x \in \mathbb{Q} \cap [0, 1]$, for a.e.$\omega$. As $x \mapsto Q(\omega, [0, x])$ is a distribution function, we see that $Q(\omega, \cdot) = \delta_\omega$, for a.e. ω. But then, $Q(\omega, E) = 1$ if $\omega \in E$ and $Q(\omega, E) = 0$ if $\omega \in E^c$. But this means that $Q(\cdot, E) = \mathbf{1}_E$ is not $\mathcal{B}$ measurable, contradicting a requirement of conditional probability.

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

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

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

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

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 $\int_M X(\omega')dQ_\omega(\omega')$. 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 and psychologically satisfying 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. This is explained next.

3.1. Specifying measures by conditional probabilities. Think of the following familiar objects in probability: Markov chains, Branching processes, Pólya's urn scheme. The last one will be defined later in the course, but the point here is that in all three cases and many others, the verbal description is of the form: "do something, and depending on what the outcome is, do this or that, ...". The very description contains the idea of conditioning. We explain with the example of Markov chains.

Markov chains: 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 $\kappa(X_n, \cdot)$.

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

$$\mu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) = \int_{A_0} \ldots \int_{A_{n-1}} \nu(dx_0)\kappa(x_0, dx_1) \ldots \kappa(x_{n-3}, dx_{n-2})\kappa(x_{n-2}, dx_{n-1})$$

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.

Remark 3

In fact, the Markov property is not needed. Suppose a measure ν on $(S, \mathcal{S})$, and stochastic kernels $\kappa_n : S^n \times \mathcal{S} \mapsto [0,1]$ are given. Define

$$\begin{aligned}&\mu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) \\ &= \int_{A_0} \cdots \int_{A_{n-1}} \nu(dx_0)\kappa_1(x_0, dx_1)\kappa_2(x_0, x_1, dx_2) \ldots \kappa_{n-1}(x_0, \ldots x_{n-2}, dx_{n-1}).\end{aligned}$$

This is again a consistent family of distributions, and we can construct random variables $(X_k)_{k \geq 0}$ on a suitable probability space so that $(X_0, \ldots, X_{n-1}) \sim \mu_n$.

In that sequence, $X_0 \sim \nu$ and the conditional distribution of X_n given $(X_0, \ldots, X_{n-1}) = (x_0, \ldots, x_{n-1})$ is given by $\kappa_n(x_0, \ldots, x_{n-1}, dx)$. Thus, a sequence of random variables may be described by giving the distribution of X_0, and for each $n \geq 1$ specifying the distribution of X_n given the previous X_is.

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.

Properties 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, and a general integrable random variable (but we also need XZ to be integrable, which is implied if Z is bounded). 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[2].

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

[2] 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.

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:

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

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

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 $\varphi(t) = \sup_{i \in I}(a_i + b_i t)$, 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 $\lim\limits_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} f(x)dx$. 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 $\lim_{\varepsilon \downarrow 0} \mathbf{E}[X \mid |Y| < \varepsilon]$, if the limit exists (and $\mathbf{P}(|Y| < \varepsilon) > 0$ for any $\varepsilon > 0$). 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 5

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 $\mathbf{E}[U \bigm| |Y| < \varepsilon]$ as $\varepsilon \downarrow 0$, we get $\mathbf{E}[X \bigm| Y = 0] = \frac{1}{2}$.

(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 $\mathbf{E}[U \bigm| |Z| < \varepsilon]$ as $\varepsilon \downarrow 0$, we get $\mathbf{E}[X \bigm| Z = 1] = \frac{2}{3}$ (do the calculation!).

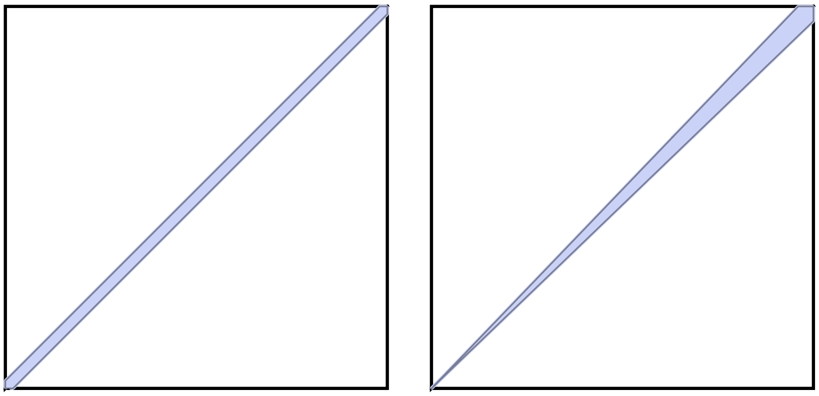


FIGURE 1. Two ways of conditioning a uniformly chosen point on the square to lie on the diagonal line. In the first case we condition on $|U - V| < \varepsilon$ and in the second case on $|\frac{U}{V} - 1| < \varepsilon$ (a value of $\varepsilon = 0.02$ was taken). Under that conditioning, the point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

Conditional probability is the largest graveyard of mistakes in probability, hence it is better to keep these cautions in mind[3]. There are also other kinds of mistakes such as mistaking $\mathbf{P}(A \bigm| B)$ for $\mathbf{P}(B \bigm| A)$, 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

[3] 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 studying the random power series $f(z) = a_0 + a_1 z + a_2 z^2 + \dots$ 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.

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[4]. 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. Finer aspects of conditional probabilities (omit on first, second, third and fourth readings)

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 non-negative amount. The two aspects touched upon here are not of equal value. The first one is somewhat esoteric and I don't know if anyone cares about it anymore. The second one is fundamental to statistical physics, which forms a large part of probability theory.

6.1. 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 conditioning 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[5], 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 $\mathcal{F}$.

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

[5] 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]

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 highly educational book of K. R. Parthasarathy titled *Probability measures on metric spaces*.

6.2. Specifying a measure via conditional probabilities. We already saw that the joint distribution of a sequence of random variables $X_1, X_2, \ldots$ may be specified by giving the marginal distribution of X_1 and the conditional distribution of X_n given $\sigma\{X_1, \ldots, X_{n-1}\}$ for $n \geq 2$.

What if the specifications are more complicated? For example, suppose we want $\{X_i : i \in I\}$, where the conditional distribution of $\{X_i : i \in F\}$ given $\{X_i : i \notin F\}$ are given for each finite set F. Can we construct such a collection?

It is clear that some consistency conditions are needed.

Example 6

Let X_1, X_2, X_3 be integer-valued random variables such that $\mathbf{P}\{(X_1, X_2, X_3) = (i, j, k)\} > 0$ for all $(i, j, k) \in \mathbb{Z}^3$ (to avoid worries about division by zero). Then

$$\sum_j \mathbf{P}\{X_1 = i \bigm| X_2 = j, X_3 = k\}\mathbf{P}\{X_2 = j \bigm| X_3 = k\} = \sum_j \mathbf{P}\{X_1 = i, X_2 = j \bigm| X_3 = k\}$$

which is a consistency requirement among the conditional distributions. You may object that the second factor in the sum on the left is not quite in the form of conditional distribution of $\{X_i : i \in F\}$ given $\{X_i : i \notin F\}$. No problem, rewrite the above as

$$\sum_{j,\ell} \mathbf{P}\{X_1 = i \bigm| X_2 = j, X_3 = k\}\mathbf{P}\{X_1 = \ell, X_2 = j \bigm| X_3 = k\} = \sum_j \mathbf{P}\{X_1 = i, X_2 = j \bigm| X_3 = k\}$$

so that all terms are of that form.

Let us now formulate the question taking into account this kind of consistency requirement. The problem is already very interesting and non-trivial if the random variables take only two values and I is countable[6].

Specification: Let I be a countable set and let $\Omega = \{0,1\}^I$ (a compact metric space) and $\mathcal{G} = \mathcal{B}(\Omega)$. 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

$$\lambda_{F_2} \circ \lambda_{F_1}(x, A) := \int_\Omega \lambda_{F_1}(y, A)\lambda_{F_2}(x, dy). \tag{2}$$

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

Gibbs measure: Given a specification as above, does there exists a measure μ on $(\Omega, \mathcal{G})$ such that the regular 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 result on existence of Gibbs measures: Unlike in the Kolmogorov consistency theorem, the obvious consistency conditions (2) are not sufficient to ensure the existence of Gibbs measures. We need more. The following fundamental forms the basis of the probabilistic study of Gibbs measures coming from statistical physics. The additional conditions imposed are not easy to interpret, but there are easy to check sufficient conditions that ensure they hold.

Theorem 3: Dobrushin-Lanford-Ruelle

Assume that the specification $\{\lambda_F\}$ satisfies the following conditions:

(1) There exists $x_0 \in \Omega$ such that given any finite $F \subseteq I$ and any $\varepsilon > 0$, 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 $\lambda_{F'}(x_0, A) < \varepsilon$.

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

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

CHAPTER 2

Martinagales: theory

1. Martingales

1.1. The setting. 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* and refer to the quadruple $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$ as a *filtered probability space*. We refer to n as "time", and the sigma-algebra at a given time represents the complete knowledge at that instant.

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 3: Martingales, Submartingales, Supermartingales

Let $X = (X_n)_{n \in \mathbb{N}}$ be an adapted process on a filtered probability space $(\Omega, \mathcal{F}, \mathcal{F}_\bullet, \mathbf{P})$. Assume that each X_n is integrable. We say that X is a

(1) *super-martingale* if $\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] \le X_{n-1}$ a.s. for each $n \ge 1$,

(2) *sub-martingale* if $-X$ is a super-martingale,

(3) *martingale* is it is both a super-martingale and a sub-martingale.

When we want to explicitly mention the filtration, we write $\mathcal{F}_\bullet$-martingale or $\mathcal{F}_\bullet$-super-martingale etc.

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.

1.2. Examples. In this section we give classes of examples.

Example 7: 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 $n \ge 1$. 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 8: Product martingale

Let ξ_n be independent, non-negative random variables and let $X_n = \xi_1\xi_2 \dots \xi_n$ and $X_0 = 1$. Then, with $\mathcal{F}_n = \sigma\{\xi_1, \dots, \xi_n\}$, we see that X is $\mathcal{F}_\bullet$-adapted and $\mathbf{E}[X_n]$ exists (equals the product of $\mathbf{E}[\xi_k]$, $k \le n$). 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 $\mu_n \ge 1$ for all n, then X is a sub-martingale, if $\mu_n = 1$ for all n, then X is a martingale, and if $\mu_n \le 1$ for all n, then X is a super-martingale.
In particular, replacing ξ_n by ξ_n/μ_n, we see that $Y_n := \frac{X_n}{\mu_1 \dots \mu_n}$ is a martingale.

Example 9: Log-likelihood function

Let $S = \{1, 2, \dots, m\}$ be a finite set with a probability mass function $p(i)$, $1 \le i \le m$. Suppose $X_1, X_2, \dots$ 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 $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$), by the same logic as in the first example.

Example 10: 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*.

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.

Example 11: Increasing process

Let 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 12: Harmonic functions

Let $R = (R_n)_{n\geq 0}$ be a simple random walk on a graph G with a countable vertex set V where each vertex has finite degree. This means that R 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,

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

We say that φ is subharmonic if $\varphi(i) \leq \frac{1}{\deg(i)} \sum_{j:j\sim i} \varphi(j)$, for all $i \in V$ and that φ is superharmonic if the inequality goes the other way. Correspondingly, X will be a submartingale or a supermartingale.

Example 13: Branching process

Consider a Galton-Watson process with offspring variable L with $\mathbf{P}\{L = k\} = p_k$ for $k \in \mathbb{N}$. Recall the informal description: At generation 0 there is one individual, who gives birth to a random number of offsprings according to the distribution of L. These offsprings belong to the first generation, and each of them give birth to offsprings according to the same distribution, independently of each other and their ancestors. And so on.

Let Z_n is the number of individuals in the nth generation. A precise construction of Z_ns can by setting $Z_0 = 1$ and

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

where $L_{n,k}$, $n, k \geq 1$, are i.i.d. copies of L.

The natural filtration to consider here is $\mathcal{F}_n = \sigma\{L_{m,k} : m \leq n, k \geq 1\}$. Clearly $Z = (Z_n)_{n\geq 0}$ is adapted to $\mathcal{F}_\bullet$. 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, $\frac{Z_n}{m^n}$ 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 14: 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

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

Thus,

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

showing that (X_n) is a martingale.

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

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

  Indeed, $\mathbf{E}[\varphi(X_n) \bigm| \mathcal{F}_{n-1}] \geq \varphi(\mathbf{E}[X_n \bigm| \mathcal{F}_{n-1}])$ 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 $(\varphi(X_n))_{n\geq 0}$.

  If X was only a sub-martingale, then $\mathbf{E}[X_n \bigm| \mathcal{F}_{n-1}] \geq X_{n-1}$ and hence the increasing property of φ is required to conclude that $\varphi(\mathbf{E}[X_n \bigm| \mathcal{F}_{n-1}]) \geq \varphi(X_{n-1})$.
- If $t_0 < t_1 < t_2 < \ldots$ is a subsequence of natural numbers, and X is a martingale (or sub-martingale or super-martingale), then $X_{t_0}, X_{t_1}, X_{t_2}, \ldots$ is also a martingale (respectively sub-martingale or super-martingale). Of course, we must take the new filtration

$\mathcal{F}_{t_0}, \mathcal{F}_{t_1}, \ldots$. This follows from the tower property of conditional expectation: If $n > m$,

$$\mathbf{E}[X_n \mid \mathcal{F}_m] = \mathbf{E}[\mathbf{E}[\ldots \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] \mid \mathcal{F}_{n-2}] \ldots \mid \mathcal{F}_m].$$

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 continuous 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 $H_n \ge 0$, the sub-martingale property of $(H.X)$ follows. ■

1.4. Continuous time martingales? For continuous time processes, we must change the setting. But most of what we said in this section goes through, with appropriate modifications.

- ▶ A filtration indexed by a totally ordered set I such as $\mathbb{R}_+ = [0, \infty)$ or $\mathbb{Z}$ or $\{0, 1, \ldots, n\}$ etc. is just a family of sub sigma algebras $\mathcal{F}_t$, $t \in I$ such that $\mathcal{F}_t \subseteq \mathcal{F}_s$ if $t \le s$.
- ▶ $X = (X_t)_{t \in I}$ is said to be adapted to $\mathcal{F}_\bullet = (\mathcal{F}_t)_{t \in I}$ if $X_t \in \mathcal{F}_t$ for all $t \in I$.
- ▶ Defining martingales may look tricky as there may be no "previous instant" $n - 1$. However, as we saw above, for a discrete time martingale, $\mathbf{E}[X_n \mid \mathcal{F}_m] = X_m$ for any $n > m$. This can be taken as the definition in general. That is $X = (X_t)_{t \in I}$ is defined to be a supermartingale if $\mathbf{E}[X_t \mid \mathcal{F}_s] \le X_s$ for any $s < t$.
- ▶ If $J \subseteq I$ and $(X_t)_{t \in I}$ is a submartingale w.r.t. $(\mathcal{F}_t)_{t \in I}$, then so is $(X_t)_{t \in J}$ w.r.t $(\mathcal{F}_t)_{t \in J}$.
- ▶ By the previous remark, if $(X_t)_{t \in \mathbb{R}_+}$ is a martingale, then so is any sequence $X_{q_1}, X_{q_2}, \ldots$ provided $q_1 < q_2 < \ldots$. Theorems that we prove for discrete time martingales apply to all such subsequences, and putting them together, one can get analogous theorems for the original process indexed by $\mathbb{R}_+$. This will be explained at the end. For now, this is just to say that studying discrete time martingales is sufficient.

Here are two examples of martingales in continuous time.

Exercise 5: Brownian motion

Let $W = (W_t)_{t\geq 0}$ be a standard Brownian motion. Recall that this means that for any $t_1 < \ldots < t_n$, the variables $W_{t_1}, W_{t_2} - W_{t_1} \ldots, W_{t_n} - W_{t_{n-1}}$ are independent Gaussians with zero means and variances $t_1, t_2 - t_1, \ldots, t_n - t_{n-1}$, respectively (the sample path continuity of $t \mapsto W_t$ is not needed for the following exercise).
Show that W_t and $W_t^2 - t$ are martingales.

Exercise 6: Poisson process

Let $N = (N_t)_{t\geq 0}$ be a homogenous Poisson process with intensity 1. Recall that this means that for any $t_1 < \ldots < t_n$, the variables $N_{t_1}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$ are independent Poisson random variables with parameters $t_1, t_2 - t_1, \ldots, t_n - t_{n-1}$ respectively.
Show that $N_t - t$ and $(N_t - t)^2 - t$ are martingales.

2. A short preview of things to come

Here are three distinct themes in the theory of martingales that we shall explore in some detail. The usefulness can only be appreciated when one sees the variety of problems that can be solved using these ideas.

(1) Several theorems that we have seen for sums of independent random variables go through for martingales, with a few modifications. Two examples are Hoeffding's inequality and Kolmogorov's maximal inequality. While the proofs are not all that different from the one for independent sums, the applicability is greatly expanded by this generalisation. With greater effort and imposing suitable conditions one can (we do not touch upon these) also prove central limit theorems, laws of iterated logarithm, etc.

(2) Optional stopping/sampling theorems. By definition of martingales, $\mathbf{E}[M_T] = \mathbf{E}[M_0]$ for any t, but the extension to certain kinds of random times T, known as stopping times, brings with it many amazing consequences. The closest things one might have seen are perhaps Wald's identities in sequential analysis (where the idea of stopping at a random time is the key point).

(3) Convergence theorems for martingales. Martingales stay constant on average. Turns out, each sample path of a martingale must either converge or oscillate wildly. Convergence theorems restrict the possibility of oscillating wildly to conclude convergence. The single most useful statement is that any martingale sequence that is uniformly integrable must converge almost surely and in L^1. This has innumerable consequences.

3. Hoeffding's inequality

Theorem 4: Hoeffding's inequality

Let $X = (X_0, \ldots, X_n)$ be a martingale. Assume that $|X_k - X_{k-1}| \le d_k$ a.s. for $1 \le k \le n$. Then $\mathbf{P}\{X_n - X_0 \ge t\} \le e^{-\frac{t^2}{2D^2}}$ for any $t > 0$, where $D^2 = d_1^2 + \ldots + d_n^2$.

Earlier we proved this for the martingale $X_k = \xi_1 + \ldots + \xi_k$, where ξ_i are independent random variables with $|\xi_k| \le d_k$. A key step in the proof is that for a zero mean random variable Y with $|Y| \le d$,

$$\mathbf{E}[e^{\theta Y}] \le e^{\frac{1}{2}\theta^2 d^2}. \tag{3}$$

Recall that this is proved by writing Y as the convex combination $\frac{Y+d}{2d}(-d) + \frac{d-Y}{2d}(d)$ and using convexity of exponential to get $\mathbf{E}[e^{\theta Y}] \le \frac{e^{\theta d}+e^{-\theta d}}{2}$. It is an elementary fact that the latter is

$$\sum_{k\ge 0} \frac{\theta^{2k} d^{2k}}{(2k)!} \le \sum_{k \ge 0} \frac{(\theta^2 d^2/2)^k}{k!} = e^{\frac{1}{2}\theta^2 d^2}.$$

PROOF OF HOEFFDING'S INEQUALITY. Fix $t > 0$ and $\theta > 0$ and use Markov's inequality:

$$\mathbf{P}\{X_n - X_0 \ge t\} = e^{-\theta t}\mathbf{E}[e^{\theta(X_n - X_0)}].$$

Conditioning on $\mathcal{F}_{n-1}$, the right side is just $\mathbf{E}[e^{\theta(X_{n-1}-X_0)}\mathbf{E}[e^{\theta(X_n - X_{n-1})} \bigm| \mathcal{F}_{n-1}]]$. Conditional on $\mathcal{F}_{n-1}$, the random variable $X_n - X_{n-1}$ has zero mean (martingale property) and is bounded by d_n anyway. By (3), the inner conditional expectation is at most $e^{\theta^2 d_n^2/2}$. Thus,

$$\mathbf{E}[e^{\theta(X_n - X_0)}] \le e^{\theta d_n^2/2}\mathbf{E}[e^{\theta(X_{n-1}-X_0)}].$$

Continuing, we get $\mathbf{E}[e^{\theta(X_n - X_0)}] \le e^{\frac{1}{2}\theta^2 D^2}$. Thus, $\mathbf{P}\{X_n - X_0 \ge t\} \le e^{-\theta t + \frac{1}{2}\theta^2 D^2}$. The optimal choice is $\theta = \frac{t}{D}$, which gives the claimed bound. ■

The only difference in the proof is that for independent random variables we factored $\mathbf{E}[e^{\theta S_n}]$ as a product of $\mathbf{E}[e^{\theta X_k}]$, but here we do it by conditioning on the previous step. While there is not much novelty in the proof of this extension, it greatly enhances the applicability.

3.1. Concentration for functions of independent random variables. Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ be a fixed function. Assume that $|f(x) - f(y)| \le d$ if $x, y \in \mathbb{R}^n$ differ in at most one co-ordinate (i.e., $x_i = y_i$ for all $i \ne j$, for some j).

Theorem 5: McDiarmid's inequality

In the above setting, let $Y = f(\xi_1, \ldots, \xi_n)$. Then

$$\mathbf{P}\{|Y - \mathbf{E}[Y]| \ge t\} \le 2e^{-\frac{t^2}{2nd^2}}.$$

PROOF. Create the Doob martingale $X_k = \mathbf{E}[Y \mid \mathcal{F}_k]$, where $\mathcal{F}_k = \sigma\{\xi_1, \ldots, \xi_k\}$. Then $X_n = Y$ while $X_0 = \mathbf{E}[Y]$. If we argue that $|X_k - X_{k-1}| \le d$ a.s., then by Hoeffding's inequality (apply to X and to $-X$ and add up the inequalities), we get the claimed result, since $D^2 = nd^2$.

To see the bound on $X_{k+1} - X_k$, consider ξ'_k, an independent copy of ξ_k. Then

$$X_k = \mathbf{E}[f(\xi_1, \ldots, \xi_{k-1}, \xi_k, \xi_{k+1}, \ldots, \xi_n) \mid \xi_1, \ldots, \xi_{k-1}, \xi_k],$$

$$X_{k-1} = \mathbf{E}[f(\xi_1, \ldots, \xi_{k-1}, \xi'_k, \xi_{k+1}, \ldots, \xi_n) \mid \xi_1, \ldots, \xi_{k-1}, \xi_k].$$

Note that the conditioning is on the same set in both cases, but f acts on vectors that differ only in the kth co-ordinate. Therefore,

$$|X_k - X_{k-1}| \le \mathbf{E}[|f(\xi_1, \ldots, \xi_{k-1}, \xi_k, \xi_{k+1}, \ldots, \xi_n) - f(\xi_1, \ldots, \xi_{k-1}, \xi'_k, \xi_{k+1}, \ldots, \xi_n)| \mid \mathcal{F}_k].$$

As the random variable is bounded by d, so is its conditional expectation. ■

Innumerable problems of probabilistic combinatorial optimization are of the form given here. For example,

(1) Let ξ be a uniformly sampled binary string. Let Y_n be the number of times the segment $s = 1011$ occurs in ξ. That is $Y_n = \sum_{i=1}^{n-3} \mathbf{1}_{(\xi_i, \ldots, \xi_{i+3}) = s}$. While we can compute the mean ($\mathbf{E}[Y_n] \sim n/4$) and variance of Y_n ($\text{Var}(Y_n) \asymp n$), the dependence causes difficulties in studying the variable. But Hoeffding's inequality tells us that

$$\mathbf{P}\{|Y_n - \mathbf{E}[Y_n]| \ge t\sqrt{n}\} \le e^{-ct^2}.$$

That is, Y_n has sub-Gaussian tails around its mean, on the scale of $\sqrt{n}$, the same behaviour (up to the constant c in the exponent) as if Y_n were a sum of i.i.d. random variables.

(2) Pick two independent binary strings U, V of length n uniformly and independently at random. Let

$$Y_n = \max\{k : \exists i_1 < \ldots < i_k,\ j_1 < \ldots < j_k, \text{ such that } U_{i_r} = V_{j_r} \text{ for all } r\},$$

be the longest length of a common subsequence. By some elementary arguments one can show that $\mathbf{E}[Y_n] \sim cn$ for some $0 < c < 1$, but the value of c is unknown. Remarkably, without any knowledge of the mean and variance, we can still say that $\mathbf{P}\{|Y_n - \mathbf{E}[Y_n]| \ge t\sqrt{n}\} \le 2e^{-t^2/2}$ (use Theorem 5 with $\xi_i = (U_i, V_i)$).

Of course, Hoeffding's gives a one-way inequality. The window length one gets in specific problems may not be optimal. The power is in the generality.

4. 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 15

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 7

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.

4.1. 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 $\mathcal{F}_n = \sigma\{X_0, \ldots, X_n\}$ for some sequence $(X_n)_{n\geq 0}$. One might be tempted to define $\mathcal{F}_T$ as $\sigma\{X_0, \ldots, X_T\}$ but a moment's thought shows that this does not make sense as written since T itself depends on the sample point. One way to fix this is to "freeze the process" at time T to get the *stopped process* $Y_n := X_{T\wedge n}$, $n \geq 0$ and define

$$\mathcal{F}_T := \sigma\{Y_n : n \geq 0\}. \tag{4}$$

This is well-defined, and it is clear that it captures all knowledge up to the stopping time T.

Another way to think of this 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 $\{X_0, \ldots, X_n\}$. 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

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T \leq n\} \in \mathcal{F}_n\} \quad \text{for each } n \geq 0. \tag{5}$$

The two definitions (4) and (5) are equivalent when the filtration is the one generated by X. We leave this as an exercise. For general filtrations, we take (5) as the definition. Some basic observations.

(1) It does not make a difference if we wrote $\{T = n\}$ instead of $\{T \leq n\}$ in (5). But in continuous time setting, it makes sense to define $\mathcal{F}_T = \{A \in \mathcal{F} : A \cap \{T \leq t\} \in \mathcal{F}_t \text{ for all } t\}$.

(2) $\mathcal{F}_T$ 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.

(3) 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 6

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. ■

(4) 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 $\{T \leq m\} \in \mathcal{F}_m$ for any $m \geq 0$. But that is true because for every $n \geq 0$ we have

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

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

If $A \in \mathcal{F}_T$, then $A \cap \{T \leq n\} \in \mathcal{F}_n$ and hence $(A \cap \{T \leq n\}) \cap \{S \leq n\} \in \mathcal{F}_n$, as S is a stopping time. But if $T \leq S$, then $A \cap \{S \leq n\} = A \cap \{S \leq n\} \cap \{T \leq n\}$ and hence $A \cap \{S \leq n\} \in \mathcal{F}_n$. This shows that $A \in \mathcal{F}_S$.

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.

For the sake of completeness: In the last property stated above, suppose we only assumed that $T \leq S$ $a.s.$ 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

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

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 $\bar{\mathcal{F}}_\bullet = (\bar{\mathcal{F}}_n)_{n\geq 0}$ 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 8

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

Exercise 9

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

5. Optional stopping or sampling

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

Optional stopping theorems are statements that assert that $\mathbf{E}[X_T] \le \mathbf{E}[X_0]$ for a stopping time T. *Optional sampling theorems* are statements that assert that $(X_{T_k})_{k\ge 0}$ is a super-martingale for an increasing sequence of stopping times $T_0 \le T_1 \le T_2 \le \ldots$. 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 7: 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 would follow if we could show that $\mathbf{E}[X_{T\wedge n}] \to \mathbf{E}[X_T]$ as $n \to \infty$. This seems reasonable as $X_{T\wedge n} \overset{a.s.}{\to} X_T$ for any finite stopping time T. However, it is not always true as the following example shows.

Example 16

Let (X_n) be a simple symmetric random walk on integers started at the origin and let T be the first time when the random walk visits the state 1 (it is well-known that $T < \infty$ $a.s.$). Then X is a martingale, $X_T = 1$ $a.s.$ but $X_0 = 0$ $a.s.$, hence the expectations do not match.

If (X_n) is a positive supermartingale, then for any finite stopping time T, almost sure convergence of $X_{T\wedge n}$ to X_T and Fatou's lemma imply that $\mathbf{E}[X_T] \leq \mathbf{E}[X_0]$. We give more conveniently applicable conditions in the theorem below (the conditions hold for X if and only if they hold for its negative, which is convenient to apply to martingales to get the conclusion $\mathbf{E}[X_T] = \mathbf{E}[X_0]$).

Theorem 8: 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 either (a) uniformly bounded or (b) dominated by an integrable random variable or (c) bounded in L^2.

(3) T is uniformly bounded.

(4) The differences $\{X_{n+1} - X_n\}$ are uniformly bounded by a constant, and $\mathbf{E}[T] < \infty$.

PROOF. Since T is finite, $X_{T\wedge n} \overset{a.s.}{\to} X_T$. Hence, $X_T \overset{L^1}{\to} X_T$ if and only if $\{X_{T\wedge n}\}$ is uniformly integrable. Convergence in L^1 implies convergence of expectations. 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 $T \le N$ $a.s.$ then $|X_{T\wedge n}| \le |X_0| + \ldots + |X_N|$ which is an integrable random variable. Therefore, the sequence $\{X_{T\wedge n}\}_{n\ge 1}$ is dominated and hence uniformly integrable.

If $|X_{n+1} - X_n| \le b$ a.s. for all n, then $|X_n| \le |X_0| + nb$. Hence $|X_{T\wedge n}| \le |X_0| + bT$. As $|X_0| + bT$ is integrable, the domination condition holds and thus $\{X_{T\wedge n}\}$ is uniformly integrable. ■

Although the conditions given here may be worth remembering, it is much better practise to always write $\mathbf{E}[X_{T\wedge n}] \le \mathbf{E}[X_0]$ and then think of ways in which to let $n \to \infty$ and get $\mathbf{E}[X_T] \le \mathbf{E}[X_0]$. 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 7 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 $n_0 \le n_1 \le n_2 \le \ldots$, the process $(X_{n_k})_k$ is a super-martingale w.r.t. $(\mathcal{F}_{n_k})_k$.

Theorem 9: Optional sampling theorem

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

If we only assume that $T_0 \le T_1 \le T_3 \le \ldots$ $a.s.$, 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 8.

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 $|X_{T_k}| \le |X_0| + \ldots + |X_{N_k}|$ 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

$$\mathbf{E}[X_T \mid \mathcal{F}_S] \le X_S \;\; a.s. \tag{6}$$

Since X_S and $\mathbf{E}[X_T \mid \mathcal{F}_S]$ are both $\mathcal{F}_S$-measurable, (6) 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 (6) is proved. ■

An alternate proof of Theorem 9 is outlined below.

SECOND PROOF OF THEOREM 9. As in the first proof, it suffices to prove (6).

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,

$$\begin{aligned}\int_A (X_T - X_S)dP &= \int_{A\cap\{T=S+1\}} (X_{S+1} - X_S)dP \\ &= \sum_{k=0}^{N-1} \int_{A\cap\{S=k\}\cap\{T=k+1\}} (X_{k+1} - X_k)dP. \qquad (7)\end{aligned}$$

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 (7) 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 $S_0 = S$, $S_1 = T \wedge (S+1)$, $S_2 = T \wedge (S+2), \ldots, S_N = T \wedge (S+N)$ 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 17

Let ξ_i be i.i.d. 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 $T_j = \min\{n \geq 0 : X_n = j\}$ 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 18

In the previous example, let $-A < 0 < B$ be integers and let $T = \min\{n \geq 0 : X_n = -A \text{ or } X_n = B\}$. 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, $|X_{T\wedge n}| \leq B+A$ 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.

6. Wald's identities

If $X_0, X_1, \ldots$ are i.i.d. random variables with finite mean, and T is an $\mathbb{N}$-valued random variable independent of X_is, then $\mathbf{E}[S_T] = \mathbf{E}[X_1]\mathbf{E}[T]$. To see this, write $S_T = \sum_n X_n \mathbf{1}_{T\geq n}$ and take expectations to get

$$\mathbf{E}[S_T] = \sum_{n=0}^{\infty} \mathbf{E}[X_n \mathbf{1}_{T\geq n}] = \mathbf{E}[X_1] \sum_{n=0}^{\infty} \mathbf{1}_{T\geq n} = \mathbf{E}[X_1]\mathbf{E}[T].$$

The application of Fubini's theorem to interchange of Expectation and sum in the first equality is justified by first working out the same kind of expression with $|X_n|$ in place of X_n.

Motivated by applications in statistics, Wald worked out conditions under which T could be allowed to depend on the sequence (X_n). These are known as Wald's identities. We give two of them.

Theorem 10: Wald's first identity

Let X_n be i.i.d. with finite expectation. Let T be a stopping time for the natural filtration of (X_n) with $\mathbf{E}[T] < \infty$. Then $\mathbf{E}[S_T] = \mathbf{E}[X_1]\mathbf{E}[T]$.

Theorem 11: Wald's second identity

Let X_n be i.i.d. with zero mean and finite variance σ^2. Let T be a stopping time for the natural filtration of (X_n) with $\mathbf{E}[T] < \infty$. Then $\mathbf{E}[S_T^2] = \sigma^2\mathbf{E}[T]$.

PROOF OF WALD'S FIRST IDENTITY. As $S_n - n\mathbf{E}[X_1]$ is a martingale, hence $\mathbf{E}[S_{T\wedge n} - (T \wedge n)\mathbf{E}[X_1]] = 0$ by optional stopping theorem. Write this as $\mathbf{E}[S_{T\wedge n}] = \mathbf{E}[T \wedge n]\mathbf{E}[X_1]$. By MCT, the right side converges to $\mathbf{E}[T]\mathbf{E}[X_1]$. On the left side, $S_{T\wedge n} \overset{a.s.}{\to} S_T$, and $S_{T\wedge n}$ is dominated by $Y = \sum_{k=0}^{\infty} |X_k|\mathbf{1}_{T\geq k}$. As $\{T \geq k\} = (\{T \leq k-1\})^c \in \mathcal{F}_{k-1}$ and X_k is independent of $\mathcal{F}_{k-1}$,

$$\mathbf{E}[|X_k|\mathbf{1}_{T\geq k}] = \mathbf{E}[\mathbf{E}[|X_k|\mathbf{1}_{T\geq k} \bigm| \mathcal{F}_{k-1}]] = \mathbf{E}[\mathbf{1}_{T\geq k}\mathbf{E}[X_k]] = \mathbf{E}[|X_k|]\mathbf{P}\{T \geq k\}.$$

Summing over k shows that $\mathbf{E}[Y] = \mathbf{E}[|X_1|]\mathbf{E}[T]$. ■

The application of optional stopping here could be avoided easily. The same argument that showed $\mathbf{E}[Y] = \mathbf{E}[|X_1|]\mathbf{E}[T]$ also shows that $\mathbf{E}[S_T] = \mathbf{E}[X_1]\mathbf{E}[T]$. Applying Fubini's now requires us to have already shown that $\mathbf{E}[Y] < \infty$.

PROOF OF WALD'S SECOND IDENTITY. We try to mimic the previous proof, replacing S_n by the martingale $S_n^2 - n\sigma^2$. Thus, $\mathbf{E}[S_{T\wedge n}^2] = \sigma^2\mathbf{E}[T \wedge n]$. By MCT, the right side converges to $\sigma^2\mathbf{E}[T]$. Now $S_{T\wedge n} \overset{a.s.}{\to} S_T$. If we show that $S_{T\wedge n}$ converges in L^2, the limit must be S_T again, and hence $\mathbf{E}[S_{T\wedge n}^2] \to \mathbf{E}[S_T^2]$, completing the proof. To show the required convergence in L^2, observe that for $m < n$

$$\begin{aligned}\mathbf{E}[|S_{T\wedge n} - S_{T\wedge m}|^2] &= \mathbf{E}\left[\left(\sum_{k=m+1}^{n} X_k\mathbf{1}_{T\geq k}\right)^2\right] \\ &= \sum_{k=m+1}^{n} \mathbf{E}[X_k^2\mathbf{1}_{T\geq k}] + 2\sum_{m+1\leq k<\ell\leq n} \mathbf{E}[X_kX_\ell\mathbf{1}_{T\geq \ell}].\end{aligned}$$

For $k < \ell$,

$$\mathbf{E}[X_kX_\ell\mathbf{1}_{T\geq\ell}] = \mathbf{E}[X_k\mathbf{1}_{T\geq\ell}\mathbf{E}[X_\ell \bigm| \mathcal{F}_{\ell-1}]] = 0$$

while

$$\mathbf{E}[X_k^2\mathbf{1}_{T\geq k}] = \mathbf{E}[\mathbf{1}_{T\geq k}\mathbf{E}[X_k^2 \bigm| \mathcal{F}_{k-1}]] = \mathbf{E}[X_1^2]\mathbf{P}\{T \geq k\}.$$

Thus, $\mathbf{E}[|S_{T\wedge n} - S_{T\wedge m}|^2] = \mathbf{E}[X_1]^2 \sum_{k=m+1}^{n} \mathbf{P}\{T \geq k\}$, which goes to zero as $m, n \to \infty$ (as $\mathbf{E}[T] < \infty$). This shows that $S_{T\wedge n}$ is Cauchy in L^2, and hence convergent. This completes the proof. ■

7. Applications of the optional stopping theorem

7.1. 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 $T_x = \min\{n \geq 0 : X_n = x\}$ we know that $T < \infty$ $a.s.$[1] and hence $\mathbf{E}[X_{T\wedge n}] = 0$ for all n. Since $|X_{T\wedge n}| \leq a + b$, we can let $n \to \infty$ and use DCT to conclude that $\mathbf{E}[X_T] = 0$. Hence, 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 10

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 $|S_{T\wedge n}| \leq a + b$) 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.

7.2. 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 $\tau_{1011} = \min\{n \geq 1 : (\xi_{n-3}, \ldots, \xi_n) = (1, 0, 1, 1)\}$ and similarly define τ_ε for any patter $\varepsilon \in \{0, 1\}^k$ 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

[1] 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.

exercise to write out a proper proof, will skip the explanation here. The final answer depends on the overlaps in the pattern. If $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_k)$, then

$$\mathbf{E}[\tau_\varepsilon] = \sum_{j=1}^{k} 2^j \mathbf{1}_{(\varepsilon_1,\ldots,\varepsilon_j)=(\varepsilon_{k-j+1},\ldots,\varepsilon_k)}.$$

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!

8. 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 12

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 $\mathbf{E}[M_{n+1} \bigm| \mathcal{G}_n] = M_n$. 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 $X_n \not\in B$ 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 $\mathbf{E}[M_{n+1} \bigm| \mathcal{G}_n] = M_n$. ■

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.

8.1. 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

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

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 12 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 13

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.

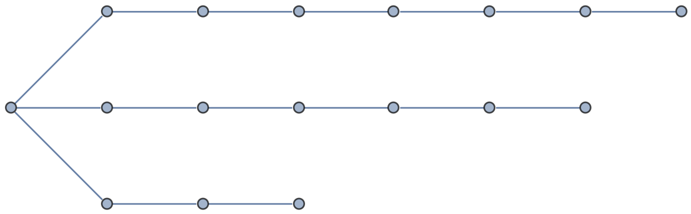
Imagine that G is an electric network where each edge is replaced by a unit resistor. The vertices in A are connected to batteries and the voltages at these points are maintained at $\varphi(x)$, $x \in A$. Then, electric current flows through the network and at each vertex a voltage is established. 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 19

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 11

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 $\varepsilon > 0$, there is a finite $F \subseteq V$ such that $|f(x)| < \varepsilon$ 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[2]. 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 14

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.

$d = 2$: 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 midpoint of e. There are about n^{d-1} edges having $|e| \in [n, n+1]$, so the total sum of squares is like

[2] 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.

$\sum_n \frac{1}{n^{d-1}}$ 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!

9. Maximal inequality

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

Lemma 15: 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 12

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

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

Hint: Let T be the first time k when $S_k \geq t$. 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 $S_T \geq t$, then clearly for at least one of these two continuations, we must have $S_n \geq t$. 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 16: Doob's inequalities

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

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

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

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

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

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

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

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

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

For convenience, let us write down the corresponding inequalities for sub-martingales (which of course follow by applying Lemma 16 to $-X$): If $X_0, \ldots, X_n$ is a sub-martingale, then for any $t > 0$ we have

$$\mathbf{P}\{\max_{k \leq n} X_k \geq t\} \leq \frac{1}{t}\mathbf{E}[(X_n)_+], \tag{8}$$

$$\mathbf{P}\{\min_{k \leq n} X_k \leq -t\} \leq \frac{1}{t}\left\{-\mathbf{E}[X_0] + \mathbf{E}[(X_n)_+]\right\}. \tag{9}$$

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 13

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 $\sum_n \mathbf{E}[\|\xi_n\|^2] < \infty$.]

10. 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 17: 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,

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

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

$$
\begin{aligned}
&T_1 := \min\{k \geq T_0 : X_k = 0\},\ T_3 := \min\{k \geq T_2 : X_k = 0\}, \quad \ldots \\
&T_2 := \min\{k \geq T_1 : X_k \geq b\},\ T_4 := \min\{k \geq T_3 : X_k \geq b\}, \quad \ldots
\end{aligned}
$$

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

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

Therefore, $\mathbf{E}[X_n - X_0 \bigm| \mathcal{F}_0] \geq b\mathbf{E}[U_n[0,b] \bigm| \mathcal{F}_0]$. 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

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

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.).

11. Convergence theorem for L^2-bounded martingales

One of the fundamental results about martingales is that if it is uniformly integrable, then it converges almost surely and in L^1. The almost sure convergence holds under the weaker assumption that the martingale is L^1-bounded, i.e., $\sup_n \mathbf{E}|X_n| < \infty$. However, it is the L^1 convergence, in particular the convergence of expectations, that is most useful.

In this section, we give a proof of the convergence under the stronger assumption of L^2-boundedness. The conclusion is also strengthened to convergence in L^2. In many applications, this is sufficient, but we also prove the more general theorem in a later section.

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 interpretation of conditional expectations, $X_{n+1} - X_n$ is an 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

$$\mathbf{E}[(X_m - X_n)^2] = \sum_{k=n}^{m-1} \mathbf{E}[(X_{k+1} - X_k)^2]. \tag{10}$$

PROOF OF THEOREM 18. Apply (10) with $n = 0$ and let $m \to \infty$ to see that

$$\sum_{k=0}^{\infty} \mathbf{E}[(X_{k+1} - X_k)^2] \leq \sup_m \mathbf{E}[(X_m - X_0)^2].$$

Under the L^2-boundedness assumption, the series on the left converges. Hence, $\mathbf{E}[(X_m - X_n)^2] \to 0$ as $m, n \to \infty$ by using (10) again, since the right 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. Applying Doob's maximal inequality to the sub-martingale $\{|X_k - X_n|\}_{k\geq n}$, we get for any $m > n$,

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

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

$$\mathbf{P}\{|X_k - X_j| \geq 2\varepsilon \text{ for some } k > j > n\} \to 0 \text{ as } n \to \infty.$$

Let $\varepsilon = \frac{1}{\ell}$ and choose N_ℓ so that for $n \geq N_\ell$, 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 $X_\infty = X'_\infty$ $a.s.$ ■

Can we deduce the martingale convergence theorem for L^1-bounded martingales, by approximating them with L^2-bounded martingales? This is a tempting approach, but the naive way of doing it will give the result only under additional restrictions.

Theorem 19: Convergence for uniformly integrable martingales with uniformly bounded differences

Let $X = (X_n)_{n\geq 0}$ be a uniformly integrable martingale. Assume that $|X_{n+1} - X_n| \leq b$ a.s., for all n for some $b < \infty$. Then, X_n converges almost surely and in L^1 to an integrable random variable X_∞.

PROOF. Fix 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. Further, $|X(\tau_M \wedge n)| \leq M + b$, since the jumps are bounded by b, and the martingale is within $[-M, M]$ at time $\tau_M - 1$. Thus, $\{X(\tau_M \wedge n)\}$ is an L^2-bounded martingale and hence by Theorem 18, 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

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

As $\tau_M \leq \tau_{M+1}$, 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 $M' > M$ (in fact, $X(\tau_M \wedge n) = X(\tau_{M'} \wedge n) = X(n)$ for all n). Therefore, we may 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 follow once we have almost sure convergence. ■

12. Convergence theorem for super-martingales

In this and the next section, we present the general results on martingale convergence. Unlike the square integrable case where we used only the maximal inequality, here the proofs use the upcrossing inequality.

Theorem 20: 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 $\mathbf{E}[X_\infty \mid \mathcal{F}_n] \leq X_n$ 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

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

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 $\mathbf{E}[|X_n|] = \mathbf{E}[X_n] + 2\mathbf{E}[(X_n)_-] \leq \mathbf{E}[X_0] + 2M$. By Fatou's lemma, $\mathbf{E}[|X_\infty|] \leq \liminf \mathbf{E}[|X_n|] \leq 2M + \mathbf{E}[X_0]$. 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, $\mathbf{E}[X_{n+m} \mid \mathcal{F}_n] \leq X_n$ for any $n, m \geq 1$. Let $m \to \infty$ and use L^1 convergence of X_{n+m} to X_∞ to get $\mathbf{E}[X_\infty \mid \mathcal{F}_n] \leq X_n$.

This completes the proof. ■

A direct corollary that is often used is

Corollary 21

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

13. Convergence theorem for martingales

Now we deduce the consequences for martingales.

Theorem 22: 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 supermartingale. To conclude $\mathbf{E}[X_\infty \mid \mathcal{F}_n] = X_n$, we apply the corresponding inequality in the supermartingale 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 23 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 last part of the above proof[3]. This lemma is similar in spirit and in its use to maximal inequalities in analysis, such as the famous one of Hardy and Littlewood.

[3] Arghydeep Chatterjee suggested the following argument that does not require this lemma to prove L^p-convergence in Theorem 22. We already know that $X_n \to X_\infty$ a.s. and in L^1 and that $X_n = \mathbf{E}[X_\infty \mid \mathcal{F}_n]$. From $|X_n|^p \overset{a.s.}{\to} |X_\infty|^p$ and Fatou's lemma and the assumption of L^p-boundedness, we see that $\mathbf{E}[|X_\infty|^p] < \infty$. By the conditional Jensen's inequality, $|X_n|^p \leq \mathbf{E}[|X_\infty|^p \mid \mathcal{F}_n]$, hence $\{|X_n|^p\}$ is a uniformly integrable sequence. As $|X_n - X_\infty|^p \leq 2^p(|X|^p_n + |X_\infty|^p)$, the sequence $\{|X_n - X_\infty|^p\}$ is also uniformly integrable. But $|X_n - X_\infty|^p \overset{a.s.}{\to} 0$, hence we also get $\mathbf{E}[|X_n - X_\infty|^p] \to 0$.

Lemma 23: A maximal inequality

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 $\mathbf{E}[(Y^*)^p] \leq C_p \sup\limits_n \mathbf{E}[Y_n^p]$ 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 $\mathbf{E}[(Y^*)^p] = \int_0^\infty p\lambda^{p-1}\mathbf{P}\{Y^* \geq \lambda\}d\lambda$ valid for any non-negative random variable in place of Y^*. Using the tail bound, we get

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

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

Alternately, from the inequality $\lambda\mathbf{P}\{Y_n^* > \lambda\} \leq \mathbf{E}[Y_n\mathbf{1}_{Y^*\geq\lambda}]$ we could have (by similar steps, but without letting $n\to\infty$) arrived at a bound of the form $\mathbf{E}[(Y_n^*)^p] \leq C_p\mathbf{E}[Y_n^p]$. 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 $p \geq 1$ 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 20: 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.

14. 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 indexed by $i \in I$ such that X_i is $\mathcal{F}_i$-measurable and $\mathbf{E}[X_j \mid \mathcal{F}_i] = X_i$ 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 24: 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 **??**.

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. ■

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 25

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.

CHAPTER 3

Martingales: applications

1. 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 $\mathbf{E}[X \bigm| \mathcal{F}_n]$ as $n \to \infty$?

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

Note that the question here is different from conditional MCT. The random variable is fixed and the sigma-algebras are changing. A natural guess is that the limit might be $\mathbf{E}[X \bigm| \mathcal{F}_\infty]$ and $\mathbf{E}[X \bigm| \mathcal{G}_\infty]$ 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 $X_n = \mathbf{E}[X \bigm| \mathcal{F}_n]$ is a martingale because of the tower property $\mathbf{E}[\mathbf{E}[X \bigm| \mathcal{F}_n] \bigm| \mathcal{F}_m] = \mathbf{E}[X \bigm| \mathcal{F}_m]$ 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 $X_\infty = \mathbf{E}[X \bigm| \mathcal{F}_\infty]$ $a.s.$.

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 can be dealt with by completing the sigma-algebras. Alternately, define $X_\infty(\omega) = \lim X_n(\omega)$ when the limit exists, and $X_\infty(\omega) = 0$ when $\lim X_n(\omega)$ does not exist. Then X_∞ is $\mathcal{F}_\infty$-measurable).

Define the measure μ and ν on $\mathcal{F}_\infty$ by $\mu(A) = \int_A X dP$ 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

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

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 $\int_A X dP = \int_A X_n dP$ which converges to $\int_A X_\infty dP$. Thus, $\int_A X dP = \int_A X_\infty dP$ for all $A \in \mathcal{F}_\infty$.

2. Kolmogorov's zero-one law

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

Theorem 26: Kolmogorov's 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. ■

3. 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 27

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 $\mathbf{E}[\xi_1 \bigm| \mathcal{G}_n] = \frac{1}{n}S_n$ (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, $\frac{1}{n}S_n \overset{a.s.}{\to} 0$. ■

4. 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 $m = \sum_k kp_k$ 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 28

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 $Z_n \sim W\ m^n$. The answer is given by the famous Kesten-Stigum theorem.

Theorem 29: 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. ■

5. 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 30

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

$$\mathbf{P}\{V_1 = \varepsilon_1, \ldots, V_m = \varepsilon_m\} = \frac{b(b+1)\ldots(b+r-1)w(w+1)\ldots(w+s-1)}{(b+w)(b+w+1)\ldots(b+w+n-1)}$$

if $r = \varepsilon_1 + \ldots + \varepsilon_m$ 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 $X_n \xrightarrow{d}$ 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 $\mu_{b,w} = \text{Beta}(b,w)$ 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 (you may imagine that the new segment is attached to the old one at the point 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 $P_1 \sim \text{unif}[0, 1]$. 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 $P_1 = U_1$, $P_2 = (1 - U_1)U_2$, $P_3 = (1 - U_1)(1 - U_2)U_3, \ldots$ 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 $\Delta^{\downarrow} = \{(p_1, p_2, \ldots) : p_1 \geq p_2 \geq \ldots \geq 0$ 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.

6. 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 31: 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").

To prove Blackwell's theorem, we first prove a lemma for general Markov chains. Let $P = (p_{i,j})_{i,j\in S}$ be a Markov transition matrix on a countable state space S. A function $f : S \to \mathbb{R}$ is said to be P-harmonic if $f(i) = \sum_{j\in S} p_{i,j} f(j)$ for all $i \in S$. If $X = (X_n)_{n\geq 0}$ is a Markov chain having transition P, then the P-harmonicity of f can be equivalently stated as $\mathbf{E}[f(X_{n+1}) \bigm| X_0, \ldots, X_n] = f(X_n)$. In other words, $(f(X_n))_n$ is a martingale.

Lemma 32

In the above setting, assume that for any $i, j \in S$, there is a coupling (X_n, Y_n) such that individually X and Y are Markov chains with transition matrix P and initial states i, j respectively, and such that $\tau = \min\{n : X_n = Y_n\} < \infty$ a.s. Then, any bounded P-harmonic function on S is constant.

Proof. Fix $i, j \in S$ and a coupling (X_n, Y_n) as assumed. If we define $Z_n = X_n$ for $n \geq \tau$ and $Z_n = Y_n$ for $n \leq \tau$, then $X_n, Z_n)$ is also a coupling with the same coupling time τ, but $X_n = Z_n$ for all $n \geq \tau$ (the two chains stick together at time τ).

Let f be a P-harmonic function such that $|f| \leq M$. As already observed, $f(X_n)$ and $f(Z_n)$ are both martingales and hence $\mathbf{E}[f(X_n) - f(Z_n)] = f(i) - f(j)$ for any $i, j \in S$. As $f(X_n) - f(Z_n) = 0$ if $n \geq \tau$, we deduce that $|f(i) - f(j)| \leq 2M\mathbf{P}\{\tau > n\} \to 0$ as $n \to \infty$. This shows that $f(i) = f(j)$ for any i, j, hence $f$must be constant. ■

Harmonic function on a graph is the same as the P-harmonic function where P is the transition for simple random walk on the graph. Observe that a P-harmonic function is also Q-harmonic if $Q = (P+I)/2$. The transition matrix Q is a *lazy version* of P, wherein at each step it stays put with probability $1/2$ and when it moves, it moves according to P.

Theorem 33: Blackwell: No bounded harmonic functions on $\mathbb{Z}^d$

If $f : \mathbb{Z}^d \to \mathbb{R}$ is bounded and harmonic, then f is constant.

PROOF. Let P be the transition matrix for the lazy version of simple symmetric random walk on $\mathbb{Z}^d$. Then f is P-harmonic. By Lemma 32, it suffices to show that for any $x, y \in \mathbb{Z}^d$, we can couple the chains starting at x, y so that they meet eventually. This is achieved by as follows.

Let $X_0 = x$ and $Y_0 = y$. Here is how the steps are coupled at time n. Pick $U \sim \text{Unif}\{1, \ldots, d\}$ and $\xi, \eta \sim \text{Unif}\{0, 1\}$, all independent and independent of the chains up to time n.

(1) If $U = i$ and $X_n(i) = Y_n(i)$, then let $(X_{n+1}, Y_{n+1}) - (X_n, Y_n)$ be equal to 0 if $\xi = 0$ and equal to (e_i, e_i) if $\xi = 1, \eta = 0$ and equal to $(-e_i, -e_i)$ if $\xi = 1, \eta = 1$.

(2) If $U = i$ and $X_n(i) < Y_n(i)$, then let $(X_{n+1}, Y_{n+1}) - (X_n, Y_n)$ be equal to $(e_i, 0)$ if $\xi = 0$ and equal to $(0, -e_i)$ if $\xi = 1$.

(3) If $U = i$ and $X_n(i) > Y_n(i)$, then let $(X_{n+1}, Y_{n+1}) - (X_n, Y_n)$ be equal to $(-e_i, 0)$ if $\xi = 0$ and equal to $(0, e_i)$ if $\xi = 1$.

We leave it as an exercise to check that X and Y are lazy simple symmetric random walks. Further, for any co-ordinate i, $X_n(i) - Y_n(i)$ is a lazy simple symmetric random walk in 1-dimension (with lazyness probability $1 - \frac{1}{2d}$) that gets absorbed at 0. Hence it will eventually get absorbed, in other words $X_n(i) = Y_n(i)$ for large enough n. This completes the proof of coupling. ■

7. 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, $\mathcal{I} = \{A \in \mathcal{F}^{\otimes \mathbb{N}} : \theta_n(A) = A$ 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 21

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 34: 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 $\mathbf{E}[Y \bigm| \mathcal{S}_n]$ is a reverse martingale and hence $\mathbf{E}[Y \bigm| \mathcal{S}_n] \overset{a.s., L^1}{\longrightarrow} \mathbf{E}[Y \bigm| \mathcal{S}]$.

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

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

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

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

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 $\mathbf{E}[\varphi(X_1, \ldots, X_k) \bigm| \mathcal{S}_n]$. This proves the claim.

Together with the reverse martingale theorem, we have shown that

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

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,

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

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. ■

8. 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 $i_1, \ldots, i_n$.

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

Example 22

If ξ_n are i.i.d., then they are exchangeable. More generally, consider finitely many probability measures $\mu_1, \ldots, \mu_k$ on some $(\Omega, \mathcal{F})$ and let $p_1, \ldots, p_k$ be positive numbers that add up to 1. Pick $L \in \{1, \ldots, k\}$ with probabilities $p_1, \ldots, p_k$, and conditional on L, pick an i.i.d. sequence $\xi_1, \xi_2, \ldots$ 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 35: de Finetti

Let $\xi_1, \xi_2, \ldots$ 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

$$\mathbf{E}\left[\varphi_1(\xi_1) \ldots \varphi_k(\xi_k) \,\middle|\, \mathcal{S}\right] = \prod_{j=1}^{k} \mathbf{E}[\varphi_j(\xi_1) \,\big|\, \mathcal{S}]$$

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 36

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 $n \to \infty$

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

PROOF. We claim that

$$\mathbf{E}\left[\psi(\xi_1, \dots, \xi_k) \,\middle|\, \mathcal{S}_n\right] = \frac{1}{n(n-1)\dots(n-k+1)} \sum_{i_1, \dots, i_k \leq n}^{\prime} \psi(\xi_{i_1}, \dots, \xi_{i_k}) \tag{11}$$

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

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

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 $\mathbf{E}[\psi(\xi_{\pi(1)}, \dots, \xi_{\pi(k)}) \mid \mathcal{S}_n]$ is the same for all permutations π of $[n]$. Therefore the expectation of the right hand side of (11) is also the same.

Now, by Lévy's backward law (or reverse martingale theorem) we know that $\mathbf{E}[\varphi(\xi_1, \dots, \xi_k) \mid \mathcal{S}_n]$ converges to $\mathbf{E}[\varphi(\xi_1, \dots, \xi_k) \mid \mathcal{S}]$. On the right hand side, we may replace $n(n-1)\dots(n-k+1)$ by n^k (the ratio goes to 1 as $n \to \infty$) and extend the sum to all $i_1, \dots, i_k$ 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 $\psi(x_1, \dots, x_k) = \varphi_1(x_1)\dots\varphi_k(x_k)$,

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

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 $\ell = 1, 2, \dots, k$, 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.

8.1. 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.

9. 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 23

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 $\nu[0,1]^{\mathbb{N}} = 0$. Hence, $\mu \perp \nu$.

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

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

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

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

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 37: 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 38

Let $\xi_n \geq 0$ be independent random variables with mean 1 and let $X_n = \xi_1\xi_2\ldots\xi_n$ 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 $a_k \leq \sqrt{\mathbf{E}[\xi_k]} = 1$ 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 OF LEMMA 38. 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).

(1) Suppose $\prod_n a_n > 0$. Then $\mathbf{E}[Y_n^2] = \frac{1}{a_1\ldots a_n}$ is uniformly bounded. As an L^2-bounded martingale, $Y_n \to Y_\infty$ 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 $X_n \leq Y_n^2$ (as $a_j \leq 1$ for all j), which means that $\{X_n\}$ is uniformly integrable. In particular, we also have $\mathbf{E}[X_\infty] = \lim \mathbf{E}[X_n] = 1$. 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.

(2) Suppose $\prod_n a_n = 0$. Observe that $X_\infty = Y_\infty^2 \prod_j a_j^2$ 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. Let $\mathcal{G}_n = \sigma\{\xi_1,\ldots,\xi_n\}$. Now form the product martingales (w.r.t. ν): $X_n = \xi_1\ldots\xi_n$ and $Y_n = \prod_{j=1}^n \frac{\xi_j}{\sqrt{a_j}}$ as in the proof of Lemma 38.

If $\prod_n a_n > 0$, then $\{X_n\}$ is uniformly integrable and $\mathbf{E}[X_\infty] = 1$ by that Lemma. We also know that $\mathbf{E}[X_\infty \mid \mathcal{G}_k] = X_k$ for any k by the martingale convergence theorem (for u.i. martingales). Define the measure θ on $(\Omega, \mathcal{G})$ by $d\theta(\omega) = X_\infty(\omega)d\nu(\omega)$. Then if $A \in \mathcal{G}_k$ 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 $\cup_k \mathcal{G}_k$. 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 38, we see that $X_\infty = 0$ $a.s.[\nu]$. We claim that $X_n \to +\infty$ a.s.$[\mu]$. Granting the claim, the sets $\{\lim X_n = 0\}$ and $\{\lim X_n = \infty\}$ provide a separation that proves that $\mu \perp \nu$.

To prove the claim, we first show that $\frac{1}{\sqrt{X_n}}$ is a μ-supermartingale. To see this, let $A \in \mathcal{G}_n$, and write $A = B \times \Omega_{n+1} \times \ldots$ for some $B \in \mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n$. Then,

$$\begin{aligned}\int_A \frac{1}{\sqrt{X_{n+1}}} d\mu &= \int_{\Omega_1 \times \ldots \times \Omega_{n+1}} \frac{\mathbf{1}_B(\omega_1, \ldots, \omega_n)}{\sqrt{f_1(\omega_1)} \ldots \sqrt{f_{n+1}(\omega_{n+1})}} \prod_{k=1}^{n+1} f_k(\omega_k)\, d\nu_1(\omega_1) \ldots d\nu_{n+1}(\omega_{n+1}) \\ &= \int_{\Omega_1 \times \ldots \times \Omega_n} \frac{\mathbf{1}_B(\omega_1, \ldots, \omega_n)}{\sqrt{f_1(\omega_1)} \ldots \sqrt{f_n(\omega_n)}} \prod_{k=1}^{n} f_k(\omega_k)\, d\nu_1(\omega_1) \ldots d\nu_n(\omega_n) \times \int_{\Omega_{n+1}} \sqrt{f_{n+1}} d\nu_{n+1} \\ &= \int_A \frac{1}{\sqrt{X_n}} d\mu \times \mathbf{E}[\sqrt{\xi_{n+1}}].\end{aligned}$$

The second factor is bounded by $\mathbf{E}[\xi_{n+1}] = 1$, hence we see that $\frac{1}{\sqrt{X_n}}$ is a μ-supermartingale. As a non-negative supermartingale, it converges a.s.$[\mu]$, say to a random variable Z. By Fatou's lemma,

$$\mathbf{E}_\mu[Z] \le \liminf \mathbf{E}_\mu[1/\sqrt{X_n}] \;=\; \liminf \mathbf{E}_\nu[\sqrt{X_n}] \;=\; \liminf \prod_{k=1}^{n} \mathbf{E}[\sqrt{\xi_k}] \;=\; \prod_{k=1}^{\infty} a_k$$

which is assumed to be zero. Hence $Z = 0$ a.s$[\mu]$, which means that $X_n \to +\infty$ a.s. $[\mu]$. ■

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.

10. 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 39

Let $H_{n,k}$, $n \geq 1$, $0 \leq k \leq 2^n - 1$ 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

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

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 $[k/2^n, (k+1)/2^n)$, $0 \leq k \leq 2^n - 1$), then $X_n = \mathbf{E}[f \mid \mathcal{F}_n]$. Thus, $\{X_n\}$ is the Doob-martingale of f with respect to the filtration $\mathcal{F}.$.

Further, $\mathbf{E}[X_n^2] = \sum_{m \leq n} \sum_{k \leq 2^m - 1} |\langle f, H_{m,k} \rangle|^2 \leq \|f\|_2^2$. 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 $X_n \overset{a.s.}{\to} f$. ■

11. Karlin-McGregor formula

Consider n independent simple random walks on $\mathbb{Z}$, each going up with probability p and down with probability $q = 1 - p$ at each step. If they start at locations $a_1, \ldots, a_n$ and time 0, the probability that they are at locations $b_1, \ldots, b_n$ at time t (in some order) is

$$\sum_{\sigma \in \mathcal{S}_n} \prod_{i=1}^{n} P_{a_i, b_{\sigma(i)}}(t) \tag{12}$$

where $P_{a,b}(t)$ is the probability that a simple random walk started at location a at time 0 is at location b at time t. Explicitly,

$$P_{a,b}(t) = \binom{t}{\frac{t+b-a}{2}} p^{\frac{t+b-a}{2}} q^{\frac{t-b+a}{2}} \tag{13}$$

where, for x a positive integer, $\binom{x}{y}$ is interpreted as zero unless y is an integer and $0 \leq y \leq x$. If instead we specify which random walk should end where, then we get only one term corresponding to the specified permutation.

Here is a different question. In the same setting, what is the probability that none of the random walks hit each other in the meantime? This is a much harder question, but there is an amazing explicit answer that turns out to be deep and important (we cannot explain the latter here). It holds in somewhat greater generality.

Definition 4

Let $\mu_0, \mu_1, \ldots$ be probability distributions on $\mathbb{Z}$. A random walk on $\mathbb{Z}$ with step distributions $(\mu_k)_{k\geq 0}$ is a sequence of random variables $\{S_0, S_1, \ldots\}$ such that $S_{k+1} - S_k \sim \mu_k$ for $k \geq 0$. If $S_0 = a$ w.p.1., we say that the random walk starts at a.

We say that the *skip-free* condition is satisfied if for independent random walks $S^{(1)}, \ldots, S^{(n)}$ started at $a_1 < \ldots < a_n$, if whenever $S^{(i)}(t) > S^{(j)}$ for some t, then there must be an $s \leq t$ such that $S^{(i)}(s) = S^{(j)}(s)$. Note that the skip-free condition depends only on the step-distributions and the initial states. Two cases where it is satisfied are:

(1) $\mu_k(-1) = 0$ for all k.

(2) $\mu_k(0) = 0$ for all k and $a_1, \ldots, a_n$ are all even or all odd.

We introduce the notation for the transition probabilities

$$P_{a,b}(s,t) = \mathbf{P}\{S_t = b \,\big|\, S_s = a\}$$

which can be written in terms of the μ_ks. Note that $P_{a,b}(s,t)$ depends on a, b only through $b - a$, and in the special case when μ_k does not depend on k (time-homogeneity), it depends on s, t only through $t - s$. When $\mu_k(+1) = p$ and $\mu_k(-1) = q$, this reduces to (13).

Theorem 40: Karlin–McGregor formula

Consider independent random walks $S^{(1)}, \ldots, S^{(n)}$ with common step distributions $(\mu_k)_{k\geq 0}$ and started at $a_1 < \ldots < a_n$. Assume that the skip-free condition is satisfied. Then, the probability that they end up at locations $b_1 < \ldots < b_n$ at time t without any two of them intersecting up to that time, is equal to

$$\sum_{\sigma\in\mathcal{S}_n} \mathrm{sgn}(\sigma) \prod_{i=1}^{n} P_{a_i, b_{\sigma(i)}}(0,t) = \det\left[P_{a_i,b_j}(0,t)\right]_{1\leq i,j\leq n}.$$

PROOF. Denote that random walks as $S^{(1)}, \ldots, S^{(n)}$, where $S_0^{(j)} = j$. For any $\sigma \in \mathcal{S}_n$, consider

$$M^{\sigma}(s) = \prod_{i=1}^{n} \mathbf{P}\left\{S^{(j)}(t) = b_{\sigma(j)} \text{ for each } j \,\big|\, \mathcal{F}_s\right\} \;=\; \mathbf{E}\left[\prod_{j=1}^{n} \mathbf{1}_{S^{(j)}} = b_{\sigma(j)} \,\big|\, \mathcal{F}_s\right]$$

where $\mathcal{F}_s = \sigma\{S^{(j)}(r) : 0 \le r \le s,\ 1 \le j \le n\}$ is the natural filtration. Clearly M^σ is an $\mathcal{F}_\bullet$ martingale, and hence so is

$$M(s) := \sum_{\sigma \in \mathcal{S}_s} \text{sgn}(\sigma) M^\sigma(s) \ = \ \mathbf{E}\left[\det\left(P_{S^{(j)}(s),b_k}(s,t)\right)_{1\le j,k\le n}\right].$$

Let $\tau = \inf\{s : S^{(j)}(s) = S^{(k)}(s) \text{ for some } j \neq k\}$ be the first time two of the random walks meet. The optional stopping theorem gives $\mathbf{E}[M(\tau \wedge t)] = \mathbf{E}[M(0)]$.

(1) $M(0)$ is precisely the quantity on the right side of the statement of the theorem.

(2) $M(\tau\wedge t) = 0$ if $\tau \le t$ (if $S^{(j)}(s) = S^{(i)}(s)$, then the jth and ith rows of $\left(P_{S^{(j)}(s),b_k}(s,t)\right)_{1\le j,k\le n}$ are equal). But if $\tau > t$, then $M(\tau \wedge t) = M(t) = \sum_\sigma \text{sgn}(\sigma) \prod_{i=1}^n \mathbf{1}_{S^{(j)}(t)=b_{\sigma(j)}}$, since $P_{a,b}(t,t) = \delta_{a,b}$. Thus, $\mathbf{E}[M(\tau \wedge t)] = \mathbf{P}\{\tau > t,\ \{S^{(1)}(t), \ldots, S^{(p)}(t)\} = \{b_1, \ldots, b_m\}\}$.

Thus we arrive at the identity

$$\sum_\sigma \text{sgn}(\sigma)\mathbf{P}\{\{S^{(j)}(t) = b_{\sigma(j)} \text{ for each } j, \ \text{ no intersection up to time } t\} = \det\left[P_{a_i,b_j}(0,t)\right]_{1\le i,j\le n}.$$

If the transitions are such that two paths cannot cross each other without touching, then only the term when σ is the identity permutation survives on the left and we arrive at

$$\mathbf{P}\{\{S^{(j)}(t) = b_j \text{ for each } j, \ \text{ no intersection up to time } t\} = \det\left[P_{a_i,b_j}(0,t)\right]_{1\le i,j\le n}.$$

That was the claim. ■

The Karlin-McGregor formula can be used to solve many counting problems such as the following.

A problem of counting lattice paths: On $\mathbb{Z}^2$, an oriented lattice path is one of the form $\ldots, \mathbf{u}_k, \mathbf{u}_{k+1} \ldots$ such that $\mathbf{u}_{k+1} - \mathbf{u}_k$ is $(1,0)$ or $(0,1)$.

Proposition 41

Let $(a_i, b_i), (c_i, d_i) \in \mathbb{Z}^2$ for $\le i \le n$, and assume that $a_i + b_i = 0$ and $c_i + d_i = L$ for all i and that $a_1 < \ldots < a_n$ and $c_1 < \ldots < c_n$. The number of packets of *non-intersecting* oriented lattice paths that lead from (a_i, b_i) to (c_i, d_i), $i \le n$, is

$$\det\left[\binom{L}{j + d_j - i - b_i}\right]_{i,j\le n}.$$

This follows directly from Karlin-McGregor, just rotate the lattice by 45° so that the starting points are on one vertical line and the ending points are on a parallel vertical line. After this rotation, lattice paths become simple random walk paths.

A generalization of the ballot problem: Suppose there are p candidates $C_1, \ldots, C_p$ in an election who get $N_1, \ldots, N_p$ votes respectively. As the votes are counted one by one, what is the chance that throughout the counting process, the true order of the candidates is maintained?

We may take $N_1 \geq N_2 \geq \ldots \geq N_p$ without loss of generality. The question is to find the chance that throughout the counting, C_1 leads, then C_2, then C_3 etc. When $p = 2$, this is the famous ballot problem, for which the answer (usually got by the reflection principle) is $\frac{N_1 - N_2 + 1}{N_1 + 1}$. The answer to the general question is

$$\det \left[\frac{N_j!}{(N_j + i - j)!} \right]_{1 \leq i,j \leq p} \tag{14}$$

and can be derived from the Karlin-McGregor formula, though the derivation is a little less obvious (try first!).

Proof. Consider random walk with Geo(r) steps, where $0 < r < 1$. Here Geo(r) takes the value k with probability $(1-r)^k r$, for $k \geq 0$. Consider p independent random walks $X^{(k)}$ started at $-k$, for $1 \leq k \leq p$. We claim that

$$\mathbf{P}\left\{ X^{(k)}(T) = N_k - k \text{ for each } k \text{ and they do not intersect ever} \;\Big|\; X^{(k)}(T) = N_k - k \text{ for each } k \right\}$$

$$= \lim_{Tr=1, r\downarrow 0} \frac{\det\left[(P_{-j, N_k - k}(T))_{j,k \leq p} \right]}{\prod_{k=1}^{p} P_{-k, N_k - k}(T)}.$$

This is because, as $r \to 0$ and $rT \to 1$, with probability converging to 1, all the jumps are of size at most 1 and no two random walks jump at the same time. That alloww us to apply the Karlin-McGregor formula (for fixed $r > 0$, skip-free condition is violated). By writing out the negative binomial coefficient or general Poisson convergence of rare events, it follows that

$$\lim_{Tr=1, r\downarrow 0} P_{a,b}(T) = \mathbf{P}\{\text{Pois}(1) = b - a\} = \frac{e^{-1}}{(b-a)!}.$$

This leads to the simplification

$$\lim_{Tr=1, r\downarrow 0} \frac{\det\left[(P_{-j, N_k - k}(T))_{j,k \leq p} \right]}{\prod_{k=1}^{p} P_{-k, N_k - k}(T)} = \frac{\det\left[\left(\frac{e^{-1}}{(N_k + j - k)!} \right)_{j,k \leq p} \right]}{\prod_{k=1}^{p} \frac{e^{-1}}{N_k!}} = \det \left[\frac{N_j!}{(N_j + i - j)!} \right]_{1 \leq i,j \leq p},$$

the expression that appears in (14).

But what does this have to do with the ballot problem? When all jumps happen at different times, conditional on then event $X^{(k)}(T) = N_k - k$, the walk $X^{(k)}$ jumps a total of N_k times. Further, the the $N_1 + \ldots + N_p$ jumps occur in a uniform random order, just as they should to correspond to counting votes at random. Lastly, the ballot problem asked for weak inequalities, hence we converted to strict inequality by starting $X^{(k)}$ at $-k$. ■

Remark 5

Karlin-McGregor formula can be stated for skip-free random walks in continuous time (and even continuous space, if sample paths are continuous, for example for Brownian motions), with essentially the same proof, except that we need optional stopping theorem for continuous time martingales. If we did that, the proof above can be stated more naturally by considering independent Poisson processes (which is essentially the scaling limit of the random walks with Geometric steps, as $r \to 0$ and $rT \to 1$).

12. Kahane's multiplicative cascade

A dyadic interval in $[0,1]$ is one of the form $I_{n,k} = [k2^{-n}, (k+1)2^{-n}]$ for some $n \geq 0$, $0 \leq k \leq 2^n - 1$. The interesting property of these intervals is that for any two of them, either one contains the other or the two have disjoint interiors. It is best to view this via the regular binary tree $\mathcal{T}$ that has root $\emptyset$ and where every vertex has two children.

We may index the vertices with the dyadic intervals $I_{n,k}$. The root vertex is indexed by $I_{0,0} = [0,1]$ and the vertex indexed by vertex $I_{n,k}$ has two children, namely $I_{n+1,2k}$ (left-child) and $I_{n+1,2k+1}$ (right child), the two dyadic intervals of the next generation that are contained in it.

Now let $W, W_{n,k}$, $n \geq 0$, $0 \leq k \leq 2^n - 1$, be i.i.d. strictly positive random variables with a distribution μ. We construct a sequence of random measures as follows: $dM_0(x) = dx$ and for $n \geq 0$ we set $dM_{n+1}(x) = g_{n+1}(x)dM_n(x)$, where $g_{n+1}(x) = g_n(x)W_{n+1,k}$ if $x \in I_{n+1,k}$. In other words, $dM_n(x) = f_n(x)dx$ where ($g_0(x) = 1$ and $k_0 = 0$ by definition)

$$f_n(x) = \prod_{j=0}^{n} g_j(x) = \prod_{m=0}^{n} W_{m,k_m} \quad \text{if } x \in I_{n,k_n} \subseteq I_{n-1,k_{n-1}} \subseteq \ldots \subseteq I_{1,k_0} \subseteq I_{0,0}.$$

Theorem 42: Kahane

The sequence of random measures M_n converge almost surely to a random measure M on $[0,1]$. Further, if $\mathbf{E}[W \log_2 W] < 1$, then $M \neq 0$ a.s. and M is almost surely a singular measure with no atoms.

Of course, the convergence here is in the Lévy metric on the space of probability measures on the line.

PROOF OF CONVERGENCE OF MEASURES. Fix $x \in [0,1]$. As it is a product of positive unit mean independent random variables, $f_n(x)$ is a positive martingale and hence converges a.s. Taking intersection over $x \in \mathbb{Q} \cap [0,1]$, we see that $f_n(x) \overset{a.s.}{\to} f(x)$ for all $x \in \mathbb{Q} \cap [0,1]$, almost surely, for some $f : \mathbb{Q} \cap [0,1] \to \mathbb{R}_+$.

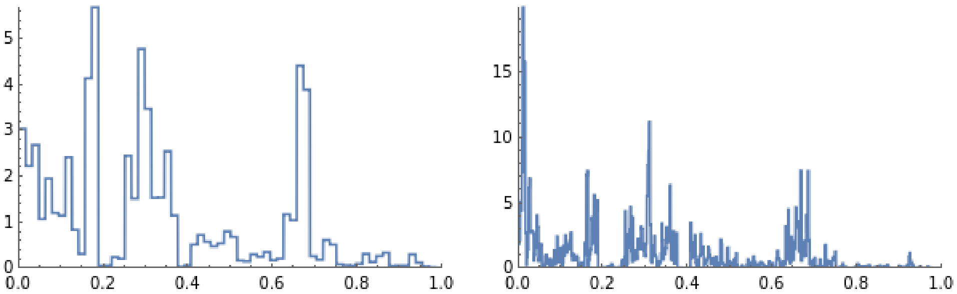


FIGURE 1. Figures of the densities f_7 and f_{10} when $W = e^{tZ-\frac{1}{2}t^2}$ where $Z \sim N(0,1)$ and $t = \frac{1}{2}$. It can be believed that the limit measure M is singular (note the markings on the Y-axis)

For the same reason, $M_n[0,x]$ is a martingale, as

$$\begin{aligned}\mathbf{E}[M_{n+1}[0,x] \,\big|\, \mathcal{F}_n] &= \mathbf{E}\left[\int_{[0,x]} f_{n+1}(t)dt \,\Big|\, \mathcal{F}_n\right] \\ &= \int_{[0,x]} \mathbf{E}[f_{n+1}(t) \,\big|\, \mathcal{F}_n]dt \\ &= \int_{[0,x]} f_n(t)dt \\ &= M_n[0,x].\end{aligned}$$

Therefore, $M_n[0,x]$is also a martingale for each x. Again taking intersection over $x \in \mathbb{Q} \cap [0,1]$, we see that $M_n[0,x] \to G(x)$ for all $x \in Q \cap [0,1]$ a.s., for some function $G : \mathbb{Q} \cap [0,1] \to \mathbb{R}_+$. Then $\bar{G}(x) = \inf\{G(u) : u > x\}$ defines a CDF of a measure on $[0,1]$, which we call M. It is now clear that $M_n \to M$, a.s. ■

For the remaining properties, to simplify the proofs we shall make the assumption that $2\mathbf{E}[W^2] < 1$. This is a stronger assumption, as $x \log_2 x \leq 2x^2$ for all $x > 0$. The proof under the assumption that $\mathbf{E}[W \log_2 W] < 1$ is along similar lines, but more involved.

STRICT POSITIVITY OF THE LIMITING MEASURE UNDER THE STRONGER ASSUMPTION. By conditioning on $W_{0,0}$, we see that

$$M_n[0,1] \overset{d}{=} W(M'_n[0,1] + M''_n[0,1])$$

where W, M'_n, M''_n are independent. Hence, $b_n = \mathbf{E}[M_n[0,1]^2]$ satisfies the recursion

$$\begin{aligned}b_{n+1} &= \mathbf{E}[W^2]\mathbf{E}[M'_n[0,1]^2 + M''_n[0,1]^2 + 2M'_n[0,1]M''_n[0,1]] \\ &= \mathbf{E}[W^2](2\mathbf{E}[M_n[0,1]^2 + 2\mathbf{E}[M'_n[0,1]]^2) \\ &= 2\mathbf{E}[W^2](1 + b_n)\end{aligned}$$

because $\mathbf{E}[M_n[0,1]] = 1$ by the martingale property. Writing $\beta = 2\mathbf{E}[W^2]$ and repeating, we see that

$$b_n = \beta + \beta^2 + \ldots + \beta^{n-1} + \beta^n b_0$$

which converges to $1/(1-\beta)$ if $\beta < 1$. Thus, the martingale $M_n[0,1]$ is L^2-bounded and hence converges in L^1. Thus $\mathbf{E}[M[0,1]] = 1$, showing that $\mathbf{P}\{M[0,1] > 0\} > 0$. As the event $M[0,1] > 0$ is clearly a tail event of the $W_{n,k}$s, it follows that $M[0,1] > 0$ a.s. ■

CONTINUITY PROPERTIES OF THE LIMITING MEASURE. ■

CHAPTER 4

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[1].

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 43: 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 $\mu_n \overset{d}{\to} \mu$.

[1] 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.

Lemma 44: 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 $\mathbb{R}$).

Definition 5: Tightness

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

Theorem 45: Prokhorov's theorem

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

Corollary 46

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 47

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 24

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 25

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 $\varepsilon > 0$ there is a $\delta > 0$ such that $|f(x) - f(y)| \leq \varepsilon$ 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.

CHAPTER 5

Brownian motion

1. Definition of Brownian motion and Wiener measure

Definition 6: Brownian motion

A collection of random variables $W = (W_t)_{t\ge 0}$ defined on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and satisfying the following properties.

(1) For any $n \ge 1$ and any $0 = t_0 < t_1 < \ldots < t_n$, the random variables $W_{t_k} - W_{t_{k-1}}$, $1 \le k \le n$, are independent.

(2) For any $s < t$ the distribution of $W_t - W_s$ is $N(0, t-s)$. Also, $W_0 = 0$, $a.s.$

(3) For $a.e.\ \omega \in \Omega$, the function $t \mapsto W_t(\omega)$ is continuous.

That such a collection of random variables exists requires proof. But first, why such a definition? We give some semi-historical and semi-motivational explanation in this section.

Einstein and the physical Brownian motion: In 1820s, the botanist Brown observed under water under a microscope and noticed certain particles buzzing about in an erratic manner. There was no explanation of this phenomenon till about 1905 when Einstein and Smoluchowski (independently of each other) came up with an explanation using statistical mechanics. More precisely, in Einstein's paper, he predicted that a small particle suspended in a liquid undergoes a random motion of a specific kind, and tentatively remarked that this could be the same motion that Brown observed.

We give a very cut-and dried (and half-understood) summary of the idea. Imagine a spherical particle inside water. The particle is assumed to be small in size but observable under a microscope, and hence much larger than the size of water molecules (which at the time of Einstein, was not yet universally accepted). According to the kinetic theory, at any temperature above absolute zero, molecules of water are in constant motion, colliding with each other, changing their direction, etc. (rather, it is this motion of molecules that defines the temperature). Now the suspended particle gets hit by agitating water molecules and hence gets pushed around. Each collision affects the particle very slightly (since it is much larger), but the number of collisions in a second (say), is very high. Hence, the total displacement of the particle in an interval of time is a sum of a large number of random and mutually independent small displacements. Then, letting W_t denote the displacement of the x-coordinate of the particle, we have the following conclusions.

(1) The displacements in two disjoint intervals of time are independent. This is the first condition in the definition of Brownian motion.

(2) The displacement in a given interval (provided it is long enough that the number of collisions with water molecules is large) must have Gaussian distribution. This is a consequence of the central limit theorem.

(3) If the liquid is homogeneous and isotropic and kept at constant temperature, then the displacement in a given interval of time must have zero mean and variance that depends only on the length of the time interval, say σ_t^2 for an interval of length t.

From the first and third conclusion, $\sigma_{t+s}^2 = \sigma_t^2 + \sigma_s^2$, which means that $\sigma_t^2 = D \cdot t$ for some constant D. If we set $D = 1$, we get the first two defining properties of Brownian motion. In his paper, Einstein wrote a formula for D in terms of the size of the suspended particle, the ambient temperature, some properties of the liquid (or water) and the Avogadro number N. All of these can be measured except N. By measuring the displacement of a particle over a unit interval of time many times, we can estimate $\mathbf{E}[W_1^2]$. Since $D = \mathbf{E}[W_1^2]$, this gives D and hence N. This was Einstein's proposal to calculate the Avogadro number by macroscopic observations and apparently this evidence convinced everyone of the reality of atoms.

Wiener and the mathematical Brownian motion: After the advent of measure theory in the first few years after 1900, mainly due to Borel and Lebesgue, mathematicians were aware of the Lebesgue measure and the Lebesgue integral on $\mathbb{R}^n$. The notion of abstract measure was also developed by Fréchet before 1915. Many analysts, particularly Gateaux, Lévy and Daniell and Wiener, pursued the question as to whether a theory of integration could be developed over infinite dimensional space[1]. One can always put an abstract measure on any space, but they were looking for something natural.

What is the difficulty? Consider an infinite dimensional Hilbert space such as ℓ^2, the space of square summable infinite sequences. Is there a translation invariant Borel measure on ℓ^2? Consider the unit ball B. There are infinitely many pairwise disjoint balls of radius 1 inside $\sqrt{2}B$ (for

[1] In 1924 or so, Wiener himself realized that dimension is irrelevant in measure theory. Indeed, in probability theory class we have see that once Lebesgue measure on $[0, 1]$ is constructed, one can just push it forward by appropriate maps to get all measures of interest such as Lebesgue measure on $[0, 1]^n$ and even product uniform measure on $[0, 1]^{\mathbb{N}}$. All these spaces are the same in measure theory, in sharp contrast to their distinctness in topology. Therefore, today no one talks of integration in infinite dimension anymore (I think!). We just think that Wiener measure is interesting.

example, take unit balls centered around each co-ordinate vector e_i, $i \geq 1$). Thus, if $\mu(B) > 0$, then by translation invariance, all these balls have the same measure and hence $\mu(\sqrt{2}B)$ must be infinite! This precludes the existence of any natural measure such as Lebesgue measure.

What else can one do? One of the things that was tried essentially amounted to thinking of a function $f : [0,1] \to \mathbb{R}$ as an infinite vector $f = (f_t)_{t\in[0,1]}$. In analogy with $\mathbb{R}^n$, where we have product measures, we can consider a product measure $\otimes_t \mu$ on $\mathbb{R}^{[0,1]}$ (the space of all functions from $[0,1]$ to $\mathbb{R}$) endowed with the product sigma-algebra. But this is very poor as a measure space as we have discussed in probability class. For example, the space $C[0,1]$ is not a measurable subset of $\mathbb{R}^{[0,1]}$, since sets in the product sigma-algebra are determined by countably many co-ordinates.

Norbert Wiener took inspiration from Einstein's theory to ask for the independence of *increments* of f rather than of independence of the *values* of f (which is what product measure does). And then, he showed that it is possible to put a Borel measure on $C[0,\infty)$ such that the increments are independent across disjoint intervals. This is why, his 1923 paper that introduced Brownian motion is titled *Differential space*, emphasizing that independence is at the level of differences of the function values.

2. The space of continuous functions

It is most appropriate to think of Brownian motion as a $C[0,\infty)$-valued random variable. Hence we recall the topology and measure structure on this space.

If X is a metric space, let $C_d(X)$ be the space of continuous functions from X to $\mathbb{R}^d$. If $d = 1$, we just write $C(X)$. Of particular interest to us are $C[0,\infty)$, $C[0,1]$. When discussing d-dimensional Brownian motion, we shall need $C_d[0,\infty)$ and $C_d[0,1]$.

On $C[0,1]$, define the norm $\|f\|_{\sup} = \max\{|f(t)| : t \in [0,1]\}$ and the metric $d(f,g) = \|f-g\|_{\sup}$. It is a fact that $C[0,1]$ is complete under this metric and hence, it is a Banach space. Obviously the sup-norm can be defined for $C[0,T]$ for any $T < \infty$, but not for $C[0,\infty)$, as the latter contains unbounded functions. The metric on $C[0,\infty)$ is defined by

$$d(f,g) = \sum_{n=1}^{\infty} \frac{1}{2^n} \frac{\|f-g\|_{\sup[0,n]}}{1+\|f-g\|_{\sup[0,n]}}.$$

The metric is irrelevant, what matters is the topology and the fact that the topology is metrizable. In fact, many other metrics such as $\tilde{d}(f,g) = \sum_{n=1}^{\infty} \frac{1}{n^2} \min\{1, \|f-g\|_{\sup[0,n]}\}$ induces the same topology on $C[0,\infty)$. In this topology, $f_n \to f$ if f_n converges to f uniformly on all compact sets of $\mathbb{R}_+ = [0,\infty)$. For $t \in [0,\infty)$, define the projection map $\Pi_t : C[0,\infty) \to \mathbb{R}$ by $\Pi_t(f) = f(t)$. The topology on $C[0,\infty)$ can also be described as the smallest topology in which all the projections are continuous (exercise!).

Once the topology is defined, we have the Borel σ-algebra $\mathcal{B}(C[0,\infty))$ which is, by definition, the smallest sigma-algebra containing all open sets. Alternately, we may say that the Borel σ-algebra is generated by the collection of projection maps. Sets of the form $(\Pi_{t_1},\ldots,\Pi_{t_n})^{-1}(B)$ for $n \geq 1$ and $t_1 < \ldots < t_n$ and $B \in \mathcal{B}(\mathbb{R}^n)$, are called (finite dimensional) cylinder sets. Cylinder sets form a π-system that generate the Borel sigma-algebra. Thus, by the $\pi - \lambda$ theorem, any two Borel probability measures that agree on cylinder sets agree on the entire Borel σ-algebra $\mathcal{B}(C[0,\infty))$. All these considerations apply if we restrict our attention to $C[0,1]$.

Definition 7: Wiener measure

Wiener measure is the Borel probability measure μ on $C[0,\infty)$ such that for any $n \geq 1$ and any $t_1 < \ldots < t_n$, the measure $\mu \circ (\Pi_{t_1},\ldots,\Pi_{t_n})^{-1}$ (a Borel probability measure on $\mathbb{R}^n$) is the multivariate Gaussian distribution with zero means and covariance matrix equal to $(t_i \wedge t_j)_{1\leq i,j\leq n}$.

It is not yet proved that Wiener measure exists. But if it exists, it must be unique, since any two such measures agree on all cylinder sets. In fact, Wiener measure and Brownian motion are two sides of the same coin, related to each other in the same way as a Gaussian random variable and the Gaussian measure are. In other words, Wiener measure is the distribution of Brownian motion, if it exists.

Exercise 18

(1) Suppose μ is the Wiener measure. Then, the collection of random variables $(\Pi_t)_{t\in\mathbb{R}_+}$ defined on the probability space $(C[0,\infty),\mathcal{B}(C[0,\infty)),\mu)$ is a Brownian motion.

(2) Suppose W is a Brownian motion on a probability space $(\Omega,\mathcal{F},\mathbf{P})$, then define the map $T:\Omega\to C[0,\infty)$ by

$$T(\omega) = \begin{cases} W_\bullet(\omega) & \text{if } t \mapsto W_t(\omega) \text{ is continuous,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the push-forward measure $\mu := \mathbf{P}\circ T^{-1}$ is the Wiener measure.

Remark 6

At first one might think it more natural to consider the space of all functions, $\mathbb{R}^{[0,1]}$, endowed with the cylinder sigma-algebra (the one generated by the projections $\Pi_t(f) = f(t)$). But the only events that are measurable in this sigma-algebra are those that are functions of countably many co-ordinates. In particular, sets such as $C[0,1]$ are not measurable subsets. In all of probability, when we talk of stochastic processes, it is usually on a space of functions with some continuity properties. Although $C[0,\infty)$ is restrictive for some purposes (eg., point processes, or events that happen in a time instant), in this course this will suffice for us. More generally one works with the space of right continuous functions having left limits (RCLL).

However, some books start by considering a measure on this space with the finite dimensional distributions of Brownian motion (such a measure exists by Kolmogorov consistency) and then show that that the outer measure of $C[0,1]$ is 1. From there, it becomes possible to get the measure to sit on $C[0,1]$ to get Brownian motion. I feel that this involves unnecessary technical digressions than the proof we give in the next section.

3. Chaining method and the first construction of Brownian motion

We want to construct random variables W_t, indexed by $t \in \mathbb{R}_+$, that are jointly Gaussian and such that $\mathbf{E}[W_t] = 0$ and $\mathbf{E}[W_t W_s] = t \wedge s$. Here is the sketch of how it is done by the so called chaining method of Kolmogorov and Centsov.

1. Let $D \subseteq [0,1]$ be a countable dense set. Because of countability, we know how to construct W_t, $t \in D$, on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$, having a joint Gaussian distribution with zero means and covariance $t \wedge s$.
2. We show that for $\mathbf{P} - a.e.\ \omega$, the function $t \mapsto W_t(\omega)$ is uniformly continuous. This is the key step.
3. By standard real analysis, this means that for each such ω, the function $t \mapsto W_t(\omega)$ extends to a continuous function on $[0,1]$.
4. Since limits of Gaussians are Gaussian, the resulting W_t, $t \in [0,1]$, have joint Gaussian distribution with the prescribed covariances.

Actually our construction will give more information about the continuity properties of Brownian motion. We start with some basic real analysis issues.

Let $D \subseteq [0,1]$ be a countable dense set and let $f : [0,1] \mapsto \mathbb{R}$ be given. We say that f extends continuously to $[0,1]$ if there exists $F \in C[0,1]$ such that $F \big|_D = f$. Clearly, a necessary condition for this to be possible is that f be uniformly continuous on D to start with. It is also sufficient. Indeed, a uniformly continuous function maps Cauchy sequences to Cauchy sequences, and hence, if $t_n \in D$ and $t_n \to t \in [0,1]$, then $(t_n)_n$ is Cauchy and hence $(f(t_n))_n$ is Cauchy and hence $\lim f(t_n)$ exists. Clearly, the limit is independent of the sequence $(t_n)_n$. Hence, we may define $F(t) = \lim\limits_{D \ni s \to t} f(s)$ and check that it is the required extension.

But we would like to prove a more quantitative version of this statement. Recall that the *modulus of continuity* of a function $f : [0,1] \to \mathbb{R}$ is defined as $w_f(\delta) = \sup\{|f(t) - f(s)| : |t-s| \le \delta\}$. Clearly, f is continuous if and only if $w_f(\delta) \downarrow 0$ as $\delta \downarrow 0$. The rate at which $w_f(\delta)$ decays to 0 quantifies the level of continuity of f. For example, if f is Lipschitz, then $w_f(\delta) \le C_f \delta$ and if f is Hölder(α) for some $0 < \alpha \le 1$, then $w_f(\delta) \le C_f \delta^{\alpha}$. For example, t^{α} is Hölder(α) (and not any better) on $[0,1]$.

Henceforth, we fix the countable dense set to be the set of dyadic rationals, i.e., $D = \bigcup_n D_n$ where $D_n = \{k2^{-n} : 0 \le k \le 2^n\}$.

Lemma 48: Kolmogorov-Centsov

Let $f : [0,1] \to \mathbb{R}$. Let Define $\Delta_n(f) = \max\{|f(\frac{k+1}{2^n}) - f(\frac{k}{2^n})| : 0 \le k \le 2^n - 1\}$. Assume that $\sum_n \Delta_n(f) < \infty$. Then, f extends to a continuous function on $[0,1]$ (we continue to denote it by f) and $w_f(\delta) \le 10 \sum_{n \ge m_\delta} \Delta_n(f)$ where $m_\delta = \lfloor \log_2(1/\delta) \rfloor$.

Assuming the lemma, we return to the construction of Brownian motion.

Construction of Brownian motion. First construct W_t, $t \in D$, that are jointly Gaussian with zero means and covariance $t \wedge s$. Then, $W(\frac{k+1}{2^n}) - W(\frac{k}{2^n})$, $0 \le k \le 2^n - 1$, are i.i.d. $N(0, 2^{-n})$. Hence, by the tail estimate of the Gaussian distribution,

$$\mathbf{P}\left\{\Delta_n(f) \ge 2\frac{\sqrt{n}}{\sqrt{2^n}}\right\} \;\le\; 2^n \mathbf{P}\left\{|\xi| \ge 2\sqrt{n}\right\} \;\le\; 2^n \exp\left\{-\frac{1}{2}(4n)\right\} \;\le\; 2^{-n}.$$

By the Borel-Cantelli lemma, it follows that $\Delta_n \le 2\frac{\sqrt{n}}{\sqrt{2^n}}$ for all $n \ge N$ for some random variable N that is finite w.p.1. If $N(\omega) < \infty$, then we can se a large constant $C(\omega)$ to take care of $\Delta_n(W_\bullet(\omega))$ for $n \le N(\omega)$ and write

$$\Delta_n(W_\bullet(\omega)) \le C(\omega)\frac{\sqrt{n}}{\sqrt{2^n}} \text{ for all } n \ge 1$$

for a random variable C that is finite w.p.1.

Fix any ω such that $C(\omega) < \infty$. Then, by the lemma, we see that $(W_t(\omega))_{t \in D}$ extends continuously to a function $(W_t(\omega))_{t \in [0,1]}$ and that the extension has modulus of continuity

$$w(\delta) \le \sum_{n \ge m_\delta} \frac{\sqrt{n}}{\sqrt{2^n}} \;\le\; 10C(\omega)\frac{\sqrt{m_\delta}}{\sqrt{2^{m_\delta}}} \;\le\; C'(\omega)\sqrt{\delta \log \frac{1}{\delta}}$$

using $m_\delta = \lfloor \log_2(1/\delta) \rfloor$. This shows that w.p.1., the extended function $t \mapsto W_t$ is not only uniformly continuous but has modulus of continuity $O(\sqrt{\delta}\sqrt{\log(1/\delta)})$.

It remains to check that the extended function has joint Gaussian distribution with the desired covariances. If $0 \le t_1 < \ldots < t_m \le 1$, then find $t_{i,n} \in D$ that converge to t_i, for $1 \le i \le m$. Then $(W_{t_{1,n},\ldots,W_{t_{m,n}}}) \overset{a.s.}{\to} (W_{t_1}, \ldots, W_{t_m})$. But $(W_{t_{1,n},\ldots,W_{t_{m,n}}})$ has joint Gaussian distribution. Hence, after taking limits, we see that $(W_{t_1}, \ldots, W_{t_m})$ has joint Gaussian distribution. In addition, the covariances converge, hence

$$\mathbf{E}[W_{t_1}W_{t_2}] = \lim_{n\to\infty} \mathbf{E}[W_{t_{1,n}}W_{t_{2,n}}] = \lim_{n\to\infty} t_{1,n} \wedge t_{2,n} = t_1 \wedge t_2.$$

Thus, W_t, $t \in [0,1]$ is the standard Brownian motion (indexed by $[0,1]$, extension to $[0,\infty)$ is simple and will be shown later). ■

It only remains to prove the lemma.

PROOF OF LEMMA 48. A function on D and its extension to $[0,1]$ have the same modulus of continuity. Hence, it suffices to show that $|f(t) - f(s)| \le 10\sum_{n\ge m_\delta} \Delta_n(f)$ for $t, s \in D$, $|t-s| \le \delta$.

Let $0 < t - s \le \delta$, $s, t \in D$. We write $I = [s,t]$ as a union of dyadic intervals using the following greedy algorithm. First we pick the largest dyadic interval (by this we mean an interval of the form $[k2^{-n}, (k+1)2^{-n}]$ for some n, k). contained in $[s,t]$. Call it, I_1 and observe that $|I_1| = 2^{-m}$ where $2^{-m} \le t - s \le 4.2^{-m}$. Then inside $I \setminus I_1$, pick the largest possible dyadic interval I_2. Then pick the largest possible dyadic interval in $I \setminus (I_1 \cup I_2)$ and so on. Since $t, s \in D_n$ for some n and hence, in a finite number of steps we end up with the empty set, i.e., we arrive at $I = I_1 \sqcup I_2 \sqcup \ldots \sqcup I_q$ for some positive integer q.

A little thought shows that for the lengths of I_j are non-increasing in j and that for any $n \ge m$, at most two of the intervals $I_1, \ldots, I_q$ can have length 2^{-n}. Write the intervals from left to right and express $f(t) - f(s)$ as a sum of the increments of f over these intervals to see that

$$|f(t) - f(s)| \le 2\sum_{n\ge m} \Delta_n(f).$$

Since $2^{-m} \le t - s$, we see that $m \ge \log_2 \frac{1}{t-s} \ge m_\delta$ and hence the conclusion in the statement of the lemma follows. ■

We put together the conclusions in the following theorem and extend the index set to $\mathbb{R}_+$.

Theorem 49

Standard Brownian motion $W = (W_t)_{t\in[0,\infty)}$ exists. Further, for any $\varepsilon > 0$ and any $T < \infty$, w.p.1., the sample paths $t \mapsto W_t$ are uniformly Hölder$\left(\frac{1}{2} - \varepsilon\right)$ on $[0, T]$.

PROOF. We used countably many i.i.d. standard Gaussians to construct standard Brownian motion on $[0,1]$. By using countably many such independent collections, we can construct (say on $([0,1], \mathcal{B}, \lambda)$) a collection of independent Brownian motions $W^{(k)} = (W^k(t))_{t\in[0,1]}$. Then for

$0 \le t < \infty$, define

$$W(t) = \sum_{k=1}^{m-1} W^{(k)}(1) + W^{(m)}(t-m)$$

if $m \le t < m+1$ for $m \in \mathbb{N}$. In words, we just append the Brownian motions successively to the previous ones.

We leave it for you to check that W is indeed a standard Brownian motion. Each $W^{(k)}$ has modulus of continuity $O(\sqrt{\delta}\sqrt{\log \frac{1}{\delta}})$ which is of course $O(\delta^{\frac{1}{2}-\varepsilon})$ for any $\varepsilon > 0$. For finite T, only finitely many $w_{W[0,T]}(\delta) \le 2 \max\{w_{W^{(k)}[0,1]}(\delta) : k \le T+1\}$. Hence, Hölder continuity holds on compact intervals. ■

4. Some insights from the proof

The proof of the construction can be used to extract valuable consequences.

Existence of continuous Gaussian processes with given covariance. Suppose $K : [0,1] \times [0,1] \mapsto \mathbb{R}$ is a postive semi-definite kernel. Do there exists random variables X_t, $t \in [0,1]$ having joint Gaussian distribution with zero means and covariance $\mathbf{E}[X_t X_s] = K(t,s)$? It is not difficult to see that continuity of K is a necessary condition (why?).

To get a sufficient condition, we may follow the same construction as before, and construct X_t, $t \in D$, having the prescribed joint distributions. How do we estimate Δ_n?

Set $h(\delta)^2 = \max\{K(t,t) + K(s,s) - 2K(t,s) : 0 \le t, s \le 1,\ |t-s| \le \delta\}$ (to understand what is happening, observe that if (X_t, X_s) has the prescribed bivariate Gaussian distribution, then $\mathbf{E}[(X_t - X_s)^2] = K(t,t) + K(s,s) - 2K(t,s)$). Then, each of $X(k+12^n) - X(\frac{k}{2^n})$ is Gaussian with standard deviation less than or equal to $h(2^{-n})$. By a union bound and the standard estimate for the Gaussian tail, we see that $\Delta_n \le \sqrt{10(1+\delta)}\sqrt{n}h(2^{-n})$, with probability $1 - 2^{-n}$ (observe that even though there is independence of increments in the Brownian case, we did not really use it in this step). Then the same steps as before show that X extends to a continuous function on $[0,1]$ provided $\sum_n \sqrt{n}h(2^{-n}) < \infty$.

In the case of Brownian motion, we had $h(\delta) = \sqrt{\delta}$. If $h(\delta) \le C\delta^p$ for any positive p, then $\sum_n \sqrt{n}h(2^{-n}) < \infty$. In fact, it suffices if $h(\delta) \le (\log(1/\delta))^p$ for a sufficiently large p.

Beyond Gaussians. Now suppose for every $k \ge 1$ and every $0 \le t_1 < t_2 < \ldots < t_k \le 1$, we are given a probability distribution $\mu_{t_1,\ldots,t_k}$ on $\mathbb{R}^n$ (in the Gaussian case it was enough to specify the means and covariances, but not in general). The question is whether there exist random variables X_t, $t \in [0,1]$, such that $(X(t_1), \ldots, X(t_k))$ has distribution $\mu_{t_1,\ldots,t_k}$ for every k and every $t_1 < \ldots < t_k$ and such that $t \mapsto X(t)$ is continuous *a.s.*? We shall of course need the consistency of the finite dimensional distributions, but that is not enough.

From the consistency, we can construct X_t, $t \in D$, as before. It remains to estimate Δ_n. The Gaussian distribution was used when we invoked the tail bound $\mathbf{P}\{Z > t\} \le e^{-t^2/2}$. Now that we do not have that, assume that $\mathbf{E}[(X_t - X_s)^\alpha] \le C|t-s|^{1+\beta}$ for some positive numbers C, α, β and for all $t, s \in [0,1]$. Observe that by $\mathbf{E}[|X_t - X_s|^\alpha]$ we mean the quantity $\int_{\mathbb{R}^2} |x-y|^\alpha d\mu_{t,s}(x,y)$. Then, it follows that

$$\mathbf{P}\left\{|X(\frac{k+1}{2^n}) - X(\frac{k}{2^n})| \ge u_n\right\} \le u_n^{-\alpha}\mathbf{E}[|X(\frac{k+1}{2^n}) - X(\frac{k}{2^n})|^\alpha] \le u_n^{-\alpha}2^{-n(1+\beta)}.$$

by the usual Chebyshev idea. Taking union over $0 \le k \le 2^n - 1$, we see that

$$\mathbf{P}\{\Delta_n \ge u_n\} \le C u_n^{-\alpha} 2^{-n\beta}.$$

which is summable if $u_n = 2^{-\gamma n}$ for some $0 < \gamma < \frac{\beta}{\alpha}$. Therefore, we get a process with continuous sample paths having modulus of continuity given by the series

$$\sum_{n \ge \log_2(1/|t-s|)} u_n \asymp 2^{-\gamma \log_2(1/|t-s|)} = |t-s|^\gamma.$$

The paths are Hölder continuous for any exponent smaller than β/α. This is the original form of the Kolmogorov-Centsov theorem.

Exercise 19

Deduce that Brownian motion is Hölder continuous with any exponent less than $\frac{1}{2}$.

The method of proof clearly cannot give any Hölder exponent larger than $1/2$. In fact by a little analysis, it is easy to see that Brownian motion is *not* uniformly Hölder of any exponent larger than $1/2$. We outline this in the exercise below. Later we shall show a much stronger fact, that Brownian motion has no Hölder points of exponent greater than $1/2$.

Exercise 20

If $Z \sim N(0,1)$, check that $\mathbf{P}\{|Z| < \varepsilon\} \le \varepsilon$ and hence deduce that $\mathbf{P}\{\Delta_n(W) \le C2^{-n\alpha}\}$ is summable for $\alpha > \frac{1}{2}$ and $C < \infty$. Deduce that Brownian motion on $[0,1]$ is not uniformly Hölder(α) for any $\alpha > \frac{1}{2}$, almost surely.

5. Lévy's construction of Brownian motion

Our first construction involved first defining W_t, $t \in D$, having the specified covariances, and then proving uniform continuity of the resulting function. For constructing W_t, $t \in D$, we showed in general that a countable collection of Gaussians with specified covariances can be constructed by choosing appropriate linear combinations of i.i.d. standard Gaussians.

In the following construction, due to Lévy and Cisielski, the special form of the Brownian covariance is exploited to make this construction very explicitly[2].

Lévy's construction of Brownian motion: As before, we construct it on time interval $[0,1]$. Let $\xi_{n,k}$, $k,n \geq 0$ be i.i.d. standard Gaussians. Let $F_0(t) = \xi_0 t$. For $n \geq 1$, define the random functions F_n by

$$F_n(t) = \begin{cases} \xi_{n,k} 2^{-\frac{1}{2}(n+1)} & \text{if } 0 \leq k \leq 2^n - 1 \text{ is odd,} \\ 0 & \text{if } 0 \leq k \leq 2^n - 1 \text{ is even,} \end{cases}$$

and such that F_n is linear on each dyadic interval $[\frac{k}{2^n}, \frac{k+1}{2^n}]$. Then define

$$W_n = F_0 + F_1 + \ldots + F_n.$$

In Figure 5, you may see the first few steps of the construction.

We claim that $\|F_n\|_{\sup} \leq 10\frac{\sqrt{n}}{\sqrt{2^n}}$ with probability $\geq 1 - \frac{1}{2^n}$. This is because F_n attains its maximum at $k2^{-n}$ for some odd k, and by definition, these values are independent Gaussians with mean zero and variance $1/2^{n+1}$. The usual estimate for the maximum of Gaussians gives the claim.

From this, it follows that $\sum_n \|F_n\|_{\sup} < \infty$ $a.s.$Therefore, w.p.1., the series $\sum_{n=0}^{\infty} F_n$ converges uniformly on $[0,1]$ and defines a random continuous function W. Further, at any dyadic rational $t \in D_m$, since $F_n(t) = 0$ for $n > m$, the series defining $W(t)$ is a finite sum of independent Gaussians. From this, we see that $W(t)$, $t \in D$ are jointly Gaussian.

We leave it as an exercise to check that $\mathbf{E}[W(t)W(s)] = t \wedge s$ (for $t, s \in D$). Since W is already continuous, and limits of Gaussians are Gaussian, conclude that the Gaussianity and covariance formulas are valid for all $t, s \in [0,1]$. Thus, W is standard Brownian motion on $[0,1]$.

Remark 7

Let $I_{n,k} = [\frac{k}{2^n}, \frac{k+1}{2^n}]$ for $0 \leq k \leq 2^n - 1$ and $n \geq 0$. Define $H_{n,k} : [0,1] \to \mathbb{R}$ by

$$H_{n,k}(x) = \begin{cases} +2^{-n/2} & \text{if } x \in [\frac{k}{2^n}, \frac{k+\frac{1}{2}}{2^n}), \\ -2^{-n/2} & \text{if } x \in [\frac{k+\frac{1}{2}}{2^n}, \frac{k+1}{2^n}], \\ 0 & \text{otherwise.} \end{cases}$$

[2] If the following description appears too brief, consult the book of Mörter and Peres where it is explained beautifully.

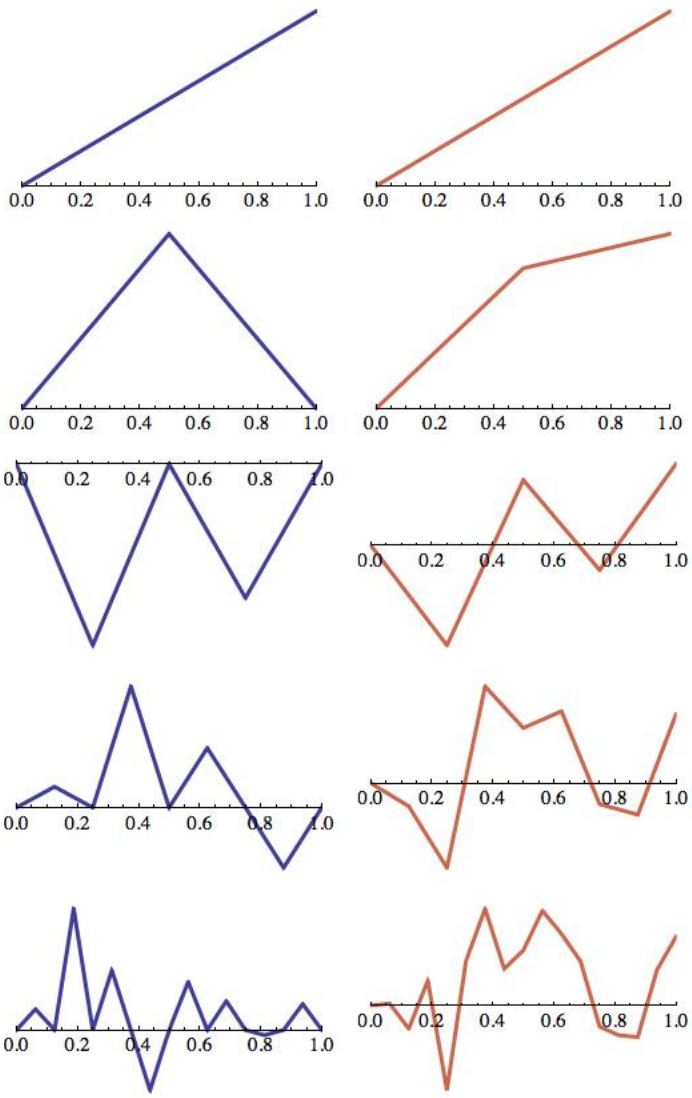


FIGURE 1. The first few steps in Lévy's construction. On the left are the functions F_n and on the right are the functions $F_0 + \ldots + F_n$, for $0 \leq n \leq 4$.

Then, together with the constant function $\mathbf{1}$, the collection $H_{n,k}$, $0 \leq k \leq 2^n - 1$, $0 \leq n$, form an orthonormal basis for $L^2[0, 1]$. It is easy to see that

$$F_{n+1}(t) = \sum_{k=0}^{2^n-1} \xi_{n+1,k} \int_0^t H_{n,k}(u)du.$$

Thus, the above construction gives the following "formula" for Brownian motion:

$$W(t) = \xi_0 \int_0^t \mathbf{1}(u)du + \sum_{n=0}^{\infty} \sum_{k=0}^{2^n-1} \xi_{n+1,k} \int_0^t H_{n,k}(u)du.$$

6. Series constructions of Brownian motion

Let us do some formal (i.e., non-rigorous) manipulations that sheds a light on the construction of Brownian motion. We start with the idea of "differential space" as Wiener termed it: If W is Brownian motion, the differentials $dW(t)$, $0 \le t \le 1$, are i.i.d. Gaussians (we can't say with what variance, because this is a formal statement without meaning!). Now take any orthonormal basis $\{\varphi_n\}$ for $L^2[0,1]$. We know that

$$\sum_n \langle f, \varphi_n\rangle\langle g, \varphi_n\rangle = \langle f, g\rangle \tag{15}$$

for any $f, g \in L^2[0,1]$. If we set $f = \delta_t$ and $g = \delta_s$, then formally we get $\sum_n \varphi_n(t)\varphi_n(s) = \langle \delta_t, \delta_s\rangle$, which is precisely the covariance structure we want for $dW(t)$. This suggests that we construct dW by setting $dW(t) = \sum_n X_n\varphi_n(t)$, where X_n are i.i.d. $N(0,1)$ (because when we compute $\mathbf{E}[dW(t)dW(s)]$, all terms with $m \neq n$ vanish and we get $\sum_n \varphi_n(t)\varphi_n(s)$. If so, since we want $W(0) = 0$, we must have

$$W(t) = \sum_n X_n \int_0^t \varphi_n(u)du \tag{16}$$

where X_n are i.i.d. standard Gaussians.

Now we can forget the means of derivation and consider the series on the right hand side of (16). If we can show that the series converges uniformly over $t \in [0,1]$ (with probability 1), then the resulting random function is continuous (since $t \mapsto \int_0^t \varphi_n$ is), and $W(t)$s will be jointly Gaussian with zero means. To compute their covariances, write $\int_0^t \varphi_n = \langle \varphi_n, \mathbf{1}_{[0,t]}\rangle$ and hence by taking limits of covariances of partial sums, we see that

$$\begin{aligned}\mathbf{E}[W(t)W(s)] &= \sum_{m,n} \mathbf{E}[X_m X_n]\langle \varphi_n, \mathbf{1}_{[0,t]}\rangle\langle \varphi_m, \mathbf{1}_{[0,s]}\rangle \;=\; \sum_n \langle \varphi_n, \mathbf{1}_{[0,t]}\rangle\langle \varphi_n, \mathbf{1}_{[0,s]}\rangle \\ &= \langle \mathbf{1}_{[0,t]}, \mathbf{1}_{[0,s]}\rangle \;=\; t \wedge s.\end{aligned}$$

In the first equality in the second line, we used (15).

This gives many new constructions (or new representations) of Brownian motion! The only remaining point is to show the uniform convergence. I do not know for what bases one gets uniform convergence, but here are a few important examples.

Haar basis: Consider the Haar basis, $\mathbf{1}, H_{0,0}, H_{1,0}, H_{1,1}, H_{2,0}, \ldots, H_{2,3}, \ldots$. In this case, it makes sense to index our i.i.d. Gaussian coefficients as $X, X_0, X_{1,0}, X_{1,1}, X_{2,0}, \ldots, X_{2,3}, \ldots$. The random function

$$\sum_{k=0}^{2^n-1} X_{n,k} \int_0^t H_{n,k}(u)du$$

is precisely what was called $F_{n+1}(t)$ in the previous section (see Remark 7). And it was shown that the series actually converges uniformly and has the correlations of the Brownian motion. What is special and helps here is that if t is a dyadic rational, then the series for $W(t)$ has only finitely many non-zero terms.

Trigonometric basis: 1, $\sqrt{2}\cos(2\pi nt)$, $\sqrt{2}\sin(2\pi nt)$, $n \geq 1$, form an orthonormal basis[3] for $L^2[0,1]$. In this case, the series form (16) becomes

$$W(t) = X_0 t + \sqrt{2}\sum_{n=1}^{\infty} \frac{1}{2\pi n}[X_n \sin(2\pi nt) + Y_n(1-\cos(2\pi nt))]$$

where X_n, Y_n are i.i.d. standard Gaussian random variables. In this case it is possible (but not trivial at all) to show that the series converges uniformly with probability 1, and that the resulting random function is Brownian motion.

Another trigonometric basis: The functions $\sqrt{2}\cos[\pi(n+\frac{1}{2})t]$, $n \geq 0$, form an orthonormal basis of $L^2[0,1]$. The series (16) then becomes

$$W(t) = \sqrt{2}\sum_{n\geq 0} X_n \frac{\sin[\pi(n+\frac{1}{2})t]}{\pi(n+\frac{1}{2})}. \tag{17}$$

Again, it can be shown that the series converges uniformly with probability 1, and gives back Brownian motion. This particular expansion is known as the Karhunen-Loeve expansion (it is an expansion first introduced by D. D. Kosambi. The orthonormal basis here are the eigenfunctions of the integral operator on $L^2[0,1]$ with kernel $K(t,s) = t \wedge s$).

Complex Brownian motion: By complex-valued Brownian motion we mean $W_{\mathbb{C}} = W(t) + iW'(t)$ where W, W' are i.i.d. Brownian motions on $[0,1]$. In the formal manipulation that we gave at the beginning of the section, if we allow complex valued functions and complex scalars, we end up with complex Brownian motion. In other words, the analogue of (16) is

$$W_{\mathbb{C}}(t) = \sum_n Z_n \int_0^t \varphi_n(u)du$$

where $\{\varphi_n\}$ is an orthonormal basis of $L^2[0,1]$ (now complex-valued functions) and Z_n are i.i.d. standard complex Gaussians (meaning that $\mathrm{Re}(Z_n)$ and $\mathrm{Im}(Z_n)$ are i.i.d. $N(0,1)$).

[3] You may have seen this in Fourier analysis class as an immediate consequence of Fejér's theorem. If not, consider the span of all these functions, and apply Stone-Weierstrass theorem to show that the span is dense in $C[0,1]$ with the sup-norm metric and hence in $L^2[0,1]$ with the L^2 metric.

Again, this may or may not be true for general orthonormal basis. We take the particular case of complex exponentials $\{e_n : n \in \mathbb{Z}\}$, where $e_n(t) = e^{2\pi int}$. Then the series becomes

$$W_{\mathbb{C}}(t) = Z_0 t + \sum_{n\neq 0} \frac{Z_n}{2\pi in} e^{2\pi int}.$$

The series converges uniformly (the proof of this assertion is nontrivial) with probability 1 and gives complex Brownian motion.

6.1. Ideas of proofs. In the last three examples, we did not present proofs. There are two stages: First prove that the series converges uniformly on $[0,1]$ with probability 1. Then show that the resulting random function has the right correlations. The first step is similar in all three examples, so let us consider the last one.

Lemma 50

The series $\sum_n \frac{Z_n}{2\pi in} e^{2\pi int}$ converges uniformly over $t \in [0,1]$, with probability 1.

If Z_n/n was absolutely summable with probability 1, then we would be done, but that is false! The main idea is to use cancellation between terms effectively by breaking the sum into appropriately large blocks. Another point worth noting is that for fixed t, the series converges almost surely, by Khinchine-Kolmogorov theorems on sums of independent random variables. One can adapt their proof to Hilbert-space valued random variables and show that the series converges in $L^2[0,1]$, with probability 1. The difficulty here is in getting uniform convergence.

PROOF OF LEMMA 50. For $n \geq 1$ define

$$F_n(t) = \sum_{k=2^{n-1}+1}^{2^n} \frac{Z_k}{k} e^{2\pi ikt}.$$

We aim to show that $\sum_n \|F_n\|_{\sup} < \infty$ with probability 1, which of course implies that $\sum_n F_n$ converges uniformly. That implies that the sum over $n \geq 1$ of $\frac{Z_n}{n} e^{2\pi int}$ converges uniformly with probability 1.

To control $\|F_n\|_{\sup}$, write $M = 2^{n-1}+1$ and $N = 2^n$ and observe that

$$\begin{aligned}
|F_n(t)|^2 &= \sum_{r=M-N+1}^{N-M-1} e^{2\pi irt} \sum_{k:M\leq k,k+r\leq N} \frac{\overline{Z}_k Z_{k+r}}{k(k+r)} \\
&\leq \frac{1}{M^2} \sum_{r=M-N+1}^{N-M-1} \left| \sum_{M\leq k,k+r\leq N} \overline{Z}_k Z_{k+r} \right|
\end{aligned}$$

and hence writing $\|F_n\|$ for the sup-norm of F_n on $[0,1]$, we have

$$\mathbf{E}[\|F_n\|^2] \le \frac{1}{M^2} \sum_{r=M-N+1}^{N-M-1} \mathbf{E}\left[\Big| \sum_{M\le k,k+r\le N} \overline{Z}_k Z_{k+r} \Big|\right].$$

Observe that $\mathbf{E}[\overline{Z}_k Z_\ell] = 2\delta_{k,\ell}$. Therefore, for $r=0$, the summand is $\mathbf{E}[\sum_{k=M}^{N}|Z_k|^2] = 2(N-M+1)$. For $r\neq 0$, we bound the summand by the square root of

$$\mathbf{E}\left[\Big| \sum_{M\le k,k+r\le N} \overline{Z}_k Z_{k+r} \Big|^2\right] = \mathbf{E}\left[\sum_{M\le k,k+r\le N} \sum_{M\le \ell,\ell+r\le N} \overline{Z}_k Z_{k+r} Z_\ell \overline{Z}_{\ell+r} \right] = 2(N-M+1)$$

because all terms with $k\neq \ell$ vanish. This shows that

$$\begin{aligned}\mathbf{E}[\|F_n\|^2] &\le \frac{1}{M^2}\left\{2(N-M+1) + 2(N-M)\sqrt{2(N-M+1)}\right\} \\ &\le 5\frac{N^{\frac{3}{2}}}{M^2} \;\le\; \frac{20}{2^{\frac{n}{2}}}.\end{aligned}$$

Therefore $\mathbf{E}[\|F_n\|] \le 5\times 2^{-n/4}$ which is summable, showing that $\sum_n \|F_n\| < \infty$ w.p.1. Hence the series converges uniformly with probability 1. ■

The proofs of uniform convergence is similar in the other cases.

7. Basic properties of Brownian motion

We have given two constructions of Brownian motion (and outlined one more). However, in our further study of Brownian motion, we would not like to use the specifics of this construction, but only the defining properties of Brownian motion. To this end, let us recall that standard Brownian motion is a collection of random variables $W = (W_t)_{t\in[0,\infty)}$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ such that

(1) $t\mapsto W_t(\omega)$ is continuous for $\mathbf{P}$-$a.e.\ \omega$,

(2) Increments over disjoint intervals are independent,

(3) $W_t - W_s \sim N(0, t-s)$ for any $s<t$.

Equivalently, we may define W as a $C[0,\infty)$-values random variable such that W_t, $t\ge 0$, are jointly Gaussian with mean zero and covariance $\mathbf{E}[W_tW_s] = t\wedge s$.

Symmetries of Brownian motion: Let W be standard Brownian motion and let μ_W denote the Wiener measure. By a symmetry, we mean a transformation $T: C[0,\infty)\to C[0,\infty)$ such that $\mu_W\circ T^{-1} = \mu_W$ or in the language of random variables, $T(W) \overset{d}{=} W$. Brownian motion has many symmetries, some of which we mention now.

- ▶ (Reflection symmetry). $T(f) = -f$. That is, if $X_t = -W_t$, then X is standard Brownian motion. To see this, observe that X is continuous w.p.1., X_t are jointly Gaussian and $X_t - X_s = -(W_t - W_s)$ has $N(0, t-s)$ distribution by the symmetry of mean zero Gaussian distribution.
- ▶ (Space-time scaling symmetry). Let $\alpha > 0$ and define $[T(f)](t) = \frac{1}{\sqrt{\alpha}} f(\alpha t)$. That is, if $X_t = \frac{1}{\sqrt{\alpha}} W_{\alpha t}$, then X is a standard Brownian motion.
- ▶ (Time-reversal symmetry) Let W be standard Brownian motion on $[0,1]$. Define $X(t) = W(1-t) - W(1)$ for $0 \le t \le 1$. Then X is standard Brownian motion on $[0,1]$.
- ▶ (Time-inversion symmetry). Define $X_t = tW_{1/t}$ for $t \in (0, \infty)$. Then X_t are jointly Gaussian, continuous in t w.p.1., and for $s < t$ we have
$$\mathbf{E}[X_t X_s] = ts\mathbf{E}\left[W\left(\frac{1}{t}\right)W\left(\frac{1}{s}\right)\right] = ts\frac{1}{t} = s.$$
Thus, $(X_s)_{s\in(0,infty)}$ has the same distribution as $(W_s)_{s\in(0,\infty)}$. In particular, if $M_\delta^X = \sup_{0<s\le\delta} X_s$ and $M_\delta^W = \sup_{0<s\le\delta} W_s$, then $(M_{1/k}^X)_{k\ge1}$ has the same distribution as $(M_{1/k}^W)_{k\ge1}$. But $\lim_{k\to\infty} M_{1/k}^W = 0$ w.p.1., and hence $\lim_{k\to\infty} M_{1/k}^X = 0$ w.p.1. But that precisely means that $\lim_{t\to0+} X(t) = 0$ w.p.1. The upshot is that if we set $X_0 = 0$, then X is standard Brownian motion.
- ▶ (Time-shift symmetry). Let $t_0 \ge 0$ and define $[Tf](t) = f(t+t_0) - f(t_0)$. That is, if $X_t = W_{t+t_0} - W_{t_0}$, then X is standard Brownian motion. Joint Gaussianity and continuity are clear. As for covariances, for $s < t$ we get
$$\begin{aligned}\mathbf{E}[X_t X_s] &= \mathbf{E}[W_{s+t_0} W_{t+t_0}] - \mathbf{E}[W_{t_0} W_{t+t_0}] - \mathbf{E}[W_{s+t_0} W_{t_0}] + \mathbf{E}[W_{t_0} W_{t_0}] \\ &= (s+t_0) - t_0 - t_0 + t_0 \\ &= s.\end{aligned}$$
Thus X is a standard Brownian motion. Whether the time-shift invariance holds at random times t_0 is an important question that we shall ask later.

8. Other processes from Brownian motion

Having constructed Brownian motion, we can use it to define various other processes with behaviour modified in many ways.

Brownian motion started at any location: If W is standard Brownian motion and $x \in \mathbb{R}$, the process X defined by $X_t = x + W_t$ for $t \ge 0$, is called Brownian motion started at x.

Brownian motion with drift and scaling: Let $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then define $X_t = \mu t + \sigma W_t$. This process X is called Brownian motion with drift μ and scale σ. More generally, we can consider the process $t \mapsto f(t) + \sigma W_t$ for some fixed function f as a noisy version of f (especially if σ is small). Brownian motion moves very randomly, these processes have a deterministic motion on which a layer of randomness is added.

Multi-dimensional Brownian motion: Brownian motion in $\mathbb{R}^d$, started at $x \in \mathbb{R}^d$, is defined as the stochastic process $W = (W(t))_{t \geq 0}$ where $W(t)$ are $\mathbb{R}^d$-valued random variables,(a) $W(0) = x$ $a.s.$, (b) for any $t_1 < \ldots < t_k$, the increments $W(t_1)$, $W(t_2) - W(t_1)$,... $W(t_k) - W(t_{k-1})$ are independent, (c) for any $s < t$ the distribution of $W(t) - W(s)$ is d-dimensional Gaussian with zero mean and covariance matrix $(t - s)I_d$, and (d) $t \mapsto W(t)$ is continuous with probability 1.

The existence of such a process need not be proved from scratch. Since we know that standard one-dimensional Brownian motion exists, we can find a probability space on which we have i.i.d. copies $W^{(k)}$, $k \geq 1$, of standard Brownian motion. Then define $W(t) = x + (W^{(1)}(t), \ldots, W^{(d)}(t))$. It is easy to check that this satisfies the properties stated above.

It is also worth noting that if we fix any orthonormal basis $v_1, \ldots, v_d$ of $\mathbb{R}^d$ and define $W(t) = x + W^{(1)}(t)v_1 + \ldots + W^{(d)}(t)v_d$, this also gives d-dimensional Brownian motion (check the properties!). Taking $x = 0$, this shows that standard Brownian motion W on $\mathbb{R}^d$ is invariant under orthogonal transformations, i.e., if $X(t) = PW(t)$ where P is a $d \times d$ orthogonal matrix, then $X \overset{d}{=} W$.

Ornstein-Uhlenbeck process: Is it possible to define Brownian motion indexed by $\mathbb{R}$ instead of $[0, \infty)$. An obvious thing is to take two independent standard Brownian motions and set $X(t) = W_1(t)$ for $t \geq 0$ and $X(t) = W_2(-t)$, then X may be called a two-sided Brownian motion. Somehow, it is not satisfactory, since the location 0 plays a special role (the variance of $X(t)$ increases on either side of it).

A better model is to set $X(t) = e^{-\frac{1}{2}t}W(e^t)$ for $t \in \mathbb{R}$. Then X is called Ornstein-Uhlenbeck process. It is a continuous process and X_t, $t \in \mathbb{R}$ are jointly Gaussian with zero means and covariances $\mathbf{E}[X_t X_s] = e^{-\frac{1}{2}(s+t)}\mathbf{E}[W(e^s)W(e^t)] = e^{-\frac{1}{2}|s-t|}$. Note that X does not have independent increments property. However, it has the interesting property of *stationarity* or *shift-invariance*: Fix $t_0 \in \mathbb{R}$ and define $Y(t) = X(t_0 + t)$. Then, check that Y has the same distribution of X (you may use space-time scale invariance of W). In other words, for the process X the origin is not a special time-point, it is just like any other point.

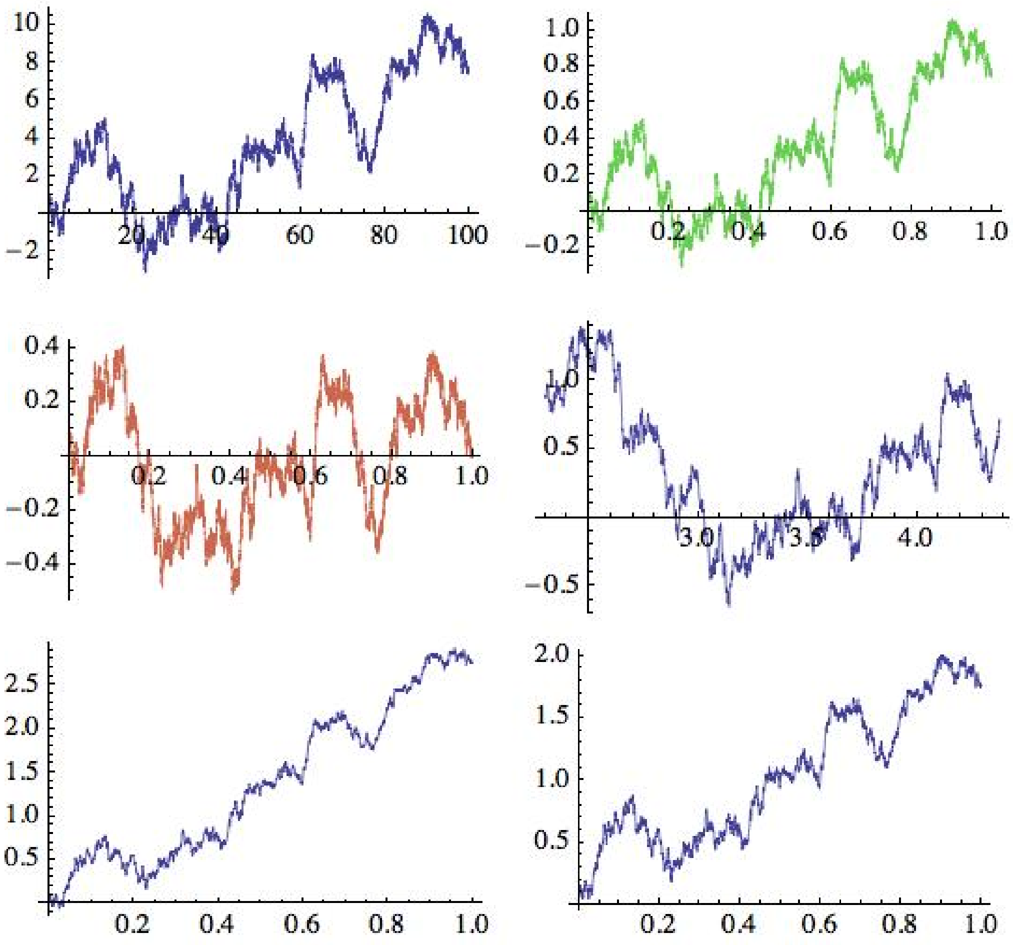


FIGURE 2. Top row left: Brownian motion run for time 10. Top row right: The same after a time-space scaling to time interval $[0,1]$. Middle row left: A Brownian bridge. Middle row right: An Ornstein-Uhlenbeck sample path. Bottom row left: Brownian motion with linear drift $2t$. Bottom row right: $W_t + \sqrt{2}t$. Take note of the markings on both axes.

Brownian bridge: Brownian bridge is the continuous Gaussian process $X = (X(t))_{t\in[0,1]}$ such that $\mathbf{E}[X_tX_s] = s(1-t)$ for $0 \le s < t \le 1$. Observe that $X(0) = X(1) = 0$ w.p.1. It arises in many situations, but for now we simply motivate it as a possible model for a random surface in $1+1$ dimensions (the graph of X is to be thought of as a surface) that is pinned down at both endpoints.

The existence of Brownian bride is easy t prove. Let W be a standard Brownian motion on $[0,1]$ and set $X(t) = W(t) - tW(1)$ for $0 \le t \le 1$. Check that X has the defining properties of Brownian bridge. This representation is also useful in working with Brownian bridge.

There is a third description of Brownian bridge. Consider standard Brownian motion $W = (W(t))_{t\in[0,1]}$ on some $(\Omega, \mathcal{F}, \mathcal{P})$. Let $\mathcal{G} = \sigma\{W(1)\}$. Then, a regular conditional distribution of

W given $\mathcal{G}$ exists. We may write it as $\mu(A, x)$, where $A \in \mathcal{B}(C[0,1])$ and $x \in \mathbb{R}$ (so $\mu(\cdot, x)$ is a probability measure that indicated the distribution of W given that $W(1) = x$). It can be checked that the conditional distributions are continuous in x. In fact, there is one measure μ_0 on $C[0,1]$ such that $\mu(A, x) = \mu_0\{g : t \mapsto g(t) + tx$ is in $A\}$. This is given in the homework and will be left as exercise.

Diffusions: Recall the physical motivation for Brownian motion as a particle in a fluid that is being bombarded on all sides by the molecules of the fluid. The mathematical definition that we have given assumes that the fluid is homogeneous (i.e., it is similar everywhere) and the motion is isotropic (there is no preferred direction of motion). If one imagines motion in a non-homogeneous medium, one arrives at the following kind of stochastic process.

For each $x \in \mathbb{R}^d$, let $m(x) \in \mathbb{R}^d$ and Σ_x be a positive definite $d \times d$ matrix. We want a $\mathbb{R}^d$-valued stochastic process $X = (X(t))_{t \geq 0}$ that has continuous sample paths, independent increments over disjoint intervals of time and such that conditional on $X(s)$, $s \leq t$, for small h, the distribution of $X(t+h) - X(t)$ is approximately Gaussian with mean vector $hm(X(t))$ and covariance matrix $h\Sigma_{X(t)}$. This last statement has to be interpreted in a suitable sense of $h \to 0$. Such a process is called a diffusion.

If $m(x) = 0$ and $\Sigma_x = I_d$, then we get back Brownian motion. If $m(x) = m$ (a constant) and $\Sigma_x = \Sigma$ (a constant matrix), then we can get such a process as $X(t) = tm + \Sigma^{\frac{1}{2}} W(t)$ where W is a standard d-dimensional Brownian motion. But more generally, it is not easy to show that such a process exists[4] and we shall not be able to touch upon this topic in this course.

9. Plan for the rest of the course

So far we have defined and constructed Brownian motion, and seen the most basic symmetries of it. We shall study the following aspects which cover only a small fraction (but reasonable enough for a first course) of things one could study about Brownian motion.

▶ Continuity properties of Brownian motion. The modulus of continuity is $O(\sqrt{\delta \log(1/\delta)})$ and hence it is Hölder($\frac{1}{2} - \varepsilon$) for any $\varepsilon > 0$. We shall see that W is nowhere Hölder($\frac{1}{2} + \varepsilon$) for any $\varepsilon > 0$.

▶ Markov property and martingales in Brownian motion. Brownian motion will be shown to have Markov and strong Markov property. We shall extract many martingales out of it. All this will be used to get substantial information about the maximum of a Brownian motion, the zero set, the

[4] One will have to either develop stochastic calculus first or a theory of general Markov processes and some existence theorems for Elliptic partial differential equations.

time to exit a given set, recurrence and transience, etc. If time permits, we shall see the relationship between multi-dimensional Brownian motion and harmonic functions and the Dirichlet problem.

► Brownian motion as a limiting object. We shall see that random walks converge to Brownian motion (Donsker's theorem). We shall use the connection between random walks and Brownian motion to deduce results about each from results about the other (eg., law of iterated logarithm, some arc-sine laws). If time permits we relate the difference between empirical distribution of an i.i.d. sample and the true distribution to a Brownian bridge.

► There are many other aspects we may not have time for. Some of them are the ideas of Wiener integral with respect to Brownian motion, Cameron-Martin formula, Hausdorff dimensions of random fractal sets coming from Brownian motion, stochastic Calculus ...

10. Further continuity properties of Brownian motion

Let W denote standard Brownian motion in $[0,1]$. We have see that W is Hölder($\frac{1}{2}-\varepsilon$) for any $\varepsilon > 0$ with probability 1. We shall show in this section that it is nowhere Hölder($\frac{1}{2}+\varepsilon$) for any $\varepsilon > 0$, in particular, the paths are nowhere differentiable.

If $f : [0,1] \to \mathbb{R}$ and $0 < \alpha \le 1$, we say that t is a Hölder(α) point for f if

$$\limsup_{h \downarrow 0} \frac{f(t+h)-f(t)}{h^{\alpha}} < \infty.$$

If the $\limsup$ on the left is less than or equal to c, then we say that t is a Hölder($\alpha; c$) point (then it is also a Hölder($\alpha; c'$) point for any $c' > c$). Observe that if f is differentiable at t, then t is a Hölder(1) point.

Theorem 51: Paley, Wiener, Zygmund

With probability 1, the following statements hold.

(1) Standard Brownian motion is nowhere differentiable.

(2) Standard Brownian motion is nowhere Hölder(α) for any $\alpha > \frac{1}{2}$.

(3) If $c < 0.3$, then Brownian motion has no Hölder($\frac{1}{2}; c$) points.

These statements are increasingly stronger, hence it suffices to prove the last one. The usual proof given in all books for the first two statements is a very elegant one due to Dvoretsky, Erdös and Kakutani. As far as I can see, that method cannot prove the third. I went back to the original proof of Paley, Wiener and Zygmund, and found that their proof, also very elegant, in fact gives the third statement! However, historically, it appears that such a statement only appeared much later in a paper of Dvoretsky, who proved the even stronger statement that Hölder($\frac{1}{2}; c$) points exist if and only if $c > 1$. I am a little confused but anyway...

PROOF OF NOWHERE DIFFERENTIABILITY DUE TO DVORETKSY, ERDÖS AND KAKUTANI. If f is differentiable at t, then $|f(s) - f(t)| \le C|s-t|$ for some $C < \infty$ and for all $s \in [0,1]$. Then, $|f(s) - f(u)| \le C(|s-t| + |u-t|)$ for all $s, u \in [0,1]$. In particular, for any $n \ge 0$ and any $0 \le k \le 2^n - 1$, this holds when we take $s = k2^{-n}$ and $u = (k+1)2^{-n}$. In particular, if ℓ is such that $[\ell 2^{-n}, (\ell+1)2^{-n}] \ni t$, then this holds for $k = \ell + j$, $j = 1,2,3$, or for $k = \ell - j$, $j = 1,2,3$ (if t is too close to 1, $\ell + 3$ may be greater than $2^n - 1$ and if t is too close to 0, $\ell - 3$ may be less than 0, hence we consider both possibilities). For such k, we get

$$\left| f\left(\frac{k+1}{2^n}\right) - f\left(\frac{k}{2^n}\right) \right| \le C\frac{10}{2^n} \tag{18}$$

since $k2^{-n}$ and $(k+1)2^{-n}$ are all within distance 5.2^{-n} of t. Thus, if we define

$$\mathcal{A} = \{f : f \text{ is differentiable at some } t \in [0,1]\},$$

$$\mathcal{A}_{n,C} = \{f : (18) \text{ holds for at least three consecutive } k \text{ in } 0, 1, \ldots, 2^n - 1\},$$

then what we have shown is that $\mathcal{A} \subseteq \bigcup\limits_{C=1}^{\infty} \bigcap\limits_{n=1}^{\infty} \mathcal{A}_{n,C}$.

We show for each fixed C that $\mathbf{P}\{W \in \mathcal{A}_{n,C}\} \to 0$ as $n \to \infty$. This implies[5] that $\mathbf{P}\{W \in \mathcal{A}\} = 0$. To show this,

$$\begin{aligned}
\mathbf{P}\{W \in \mathcal{A}_n\} &= \sum_{\ell=0}^{2^n-3} \mathbf{P}\{(18) \text{ holds for } f = W \text{ for } k = \ell, \ell+1, \ell+2\} \\
&\le (2^n - 2)\left(\mathbf{P}\left\{|\xi| \le \frac{10C}{\sqrt{2^n}}\right\}\right)^3 \\
&\le (2^n - 2)\left(\frac{1}{\sqrt{2\pi}}\frac{10C}{\sqrt{2^n}}\right)^3 \\
&\le 10^3 C^3 \frac{1}{\sqrt{2^n}}.
\end{aligned}$$

This proves the nowhere differentiability of Brownian motion. ■

By considering several increments in place of three, one can show that W has no Hölder($\frac{1}{2} + \varepsilon$) points.

Hölder($\frac{1}{2}; c$) points: Next we adapt the original proof of Paley, Wiener and Zygmund to show that there are no Hölder($\frac{1}{2}; c$) points if c is small. For convenience of notation, let $\Delta f(I) = f(b) - f(a)$

[5] One issue: Is $\mathcal{A}$ a Borel subset of $C[0,1]$?! It is, but we don't bother to prove it. Instead, let us always work with the completion of Wiener measure. In other words, if $\mathcal{A}_1 \subseteq \mathcal{A}_0 \subseteq \mathcal{A}_2$ and $\mathcal{A}_1$ and $\mathcal{A}_2$ are Borel and $\mathbf{P}\{W \in \mathcal{A}_1\} = \mathbf{P}\{W \in \mathcal{A}_2\}$, then the same is deemed to be the value of $\mathbf{P}\{W \in \mathcal{A}_0\}$.

for $f : [0,1] \mapsto \mathbb{R}$ and $I = [a,b]$ a subinterval of $[0,1]$. Also, let $I_{n,k} = [k2^{-n}, (k+1)2^{-n}]$ for $n \geq 0$ and $0 \leq k \leq 2^n - 1$.

A BRANCHING PROCESS PROOF DUE TO PALEY, WIENER AND ZYGMUND. Let t is a Hölder($\frac{1}{2}; c$) point, then there exists $M < \infty$ such that $|f(s) - f(t)| \leq c\sqrt{|s-t|}$ for all $s \in [t - 2^{-M}, t + 2^{-M}]$. In particular, if $n \geq M$ and $I_{n,k}$ is the dyadic interval containing t, then

$$|\Delta f(I)| \;\leq\; c\left\{\sqrt{(k+1)2^{-n} - t} + \sqrt{t - k2^{-n}}\right\} \;\leq\; \frac{\sqrt{2}c}{\sqrt{2^n}}. \tag{19}$$

In the last inequality we used the elementary fact that if $0 \leq x \leq a$, then $\sqrt{x} + \sqrt{a-x} \leq \sqrt{2a}$.

The collection of dyadic intervals carries a natural tree structure with $I_{0,0}$ being the root vertex and by declaring $I_{n+1,\ell}$ as a child of $I_{n,k}$ if $I_{n+1,\ell} \subseteq I_{n,k}$. This is a tree where each vertex has two children. Let us declare a dyadic interval $I_{n,k}$ to be alive if it satisfies $\Delta f(I_{n,k}) \leq c\sqrt{2}/\sqrt{2^n}$. Thus, if t is a Hölder($\frac{1}{2}; c$) point, then for some M, the tree beyond generation M has an infinite chain of descendents that are all alive (namely the dyadic intervals containing the point t).

The process of vertices alive is a Branching process that we shall prove will become extinct with probability 1. To do this, let $\mathcal{F}_n = \{\Delta W(I_{n,k}) : 0 \leq k \leq 2^n - 1\}$ so that these sigma-algebras are increasing. Whether an interval $I_{n,k}$ is alive or not is an event in $\mathcal{F}_n$. Condition on $\mathcal{F}_{n-1}$ and consider any live individual I in the $(n-1)$st generation. It has two children J, J' in the nth generation. Conditional on $\mathcal{F}_{n-1}$, we know the sum $\Delta W(J) + \Delta W(J') = \Delta W(I)$. From Exercise 10 we can write $\Delta W(J) = \frac{1}{2}\Delta W(I) + \frac{\xi}{\sqrt{2^{n+1}}}$ and $\Delta W(J') = \frac{1}{2}\Delta W(I) - \frac{\xi}{\sqrt{2^{n+1}}}$ where $\xi \sim N(0,1)$ is independent of $\mathcal{F}_{n-1}$. Now, J is alive if and only if $|\Delta W(J)| \leq \frac{c\sqrt{2}}{\sqrt{2^n}}$. This means that ξ must lie in an interval of length $4c$ centered at $\sqrt{2^{n-1}}\Delta W(I)$. By Exercise 10, irrespective of the value of $\Delta W(I)$, this probability is at most $4c/\sqrt{2\pi}$.

In summary, the expected number of offsprings of I is at most $\lambda = 8c/\sqrt{2\pi}$. If $c' < 1$, then the number of descendants of an interval $I_{M,k}$ in the generation $M + j$ is exactly λ^j. Thus the expected total number of live individuals live in the $M + j$ generation is $2^M\lambda^j$ which goes to zero as $j \to \infty$, provided $\lambda < 1$. Hence, for $c < \frac{\sqrt{2\pi}}{8} = 0.313\ldots$, the branching process goes extinct with probability 1.

Since this is true for every M, taking a countable union over positive integer M, it follows that for any $c < 0.31$, with probability 1, Brownian motion has no Hölder($\frac{1}{2}; c$) points. ■

We used two simple facts about Gaussian distribution in the proof. They are left as exercises.

Exercise 21

Let X, Y be i.i.d. $N(0,1)$. Then, the conditional distribution of (X,Y) given $X+Y=t$ is the same as the (unconditional) distribution of $(\frac{1}{2}t+\frac{1}{\sqrt{2}}\xi, \frac{1}{2}t-\frac{1}{\sqrt{2}}\xi)$ where $\xi \sim N(0,1)$.

Exercise 22

If $\xi \sim N(0,1)$, then $\sup\limits_{a\in\mathbb{R}} \mathbf{P}\{\xi \in [a-t,a+t]\} \le \frac{2t}{\sqrt{2\pi}}$.

11. Summary of continuity properties

Let W be standard Brownian motion on $[0,1]$. First and foremost is the point that $\mathbf{E}[|W_t - W_s|^2] = |t-s|$ from which we see that $W_{t+h}-W_t$ should behave like $\sqrt{h}$, typically. A summary of the basic continuity results is as follows.

(1) Almost surely $\limsup\limits_{h\downarrow 0} \max\limits_{t\in[0,1]} \frac{|W_t-W_s|}{\sqrt{h\log(1/h)}} < \infty$. We showed this (and if you follow our proof closely, you will see that the left hand side can be shown to be ≤ 10 w.p.1.).

We did not show Paul Lévy's sharp result that in fact

$$\max_{t\in[0,1]} \limsup_{h\downarrow 0} \frac{|W_t - W_s|}{\sqrt{h\log(1/h)}} = \sqrt{2} \;\; a.s.$$

(2) Almost surely W has no Hölder($\frac{1}{2}; c$) points for c sufficiently small. As a consequence, it is nowhere Hölder($\frac{1}{2}+\varepsilon$) and in particular, nowhere differentiable.

We showed this. We did not show the results of Dvoretzky (and Kahane?) that the sharp constant is 1. That is, for $c < 1$, there do not exist Hölder($\frac{1}{2}; c$) points while for $c > 1$, they do exist.

(3) We shall show later that at a fixed point, the continuity is faster than $\sqrt{h}$ and slower than $\sqrt{h\log(1/h)}$. This is the celebrated law of iterated logarithm which asserts that for any fixed $t \ge 0$,

$$\limsup_{h\downarrow 0} \frac{W(t+h)-W(t)}{\sqrt{2h\log\log(1/h)}} = 1 \;\; a.s.$$

In fact the set of limit points of $\frac{W(t+h)-W(t)}{\sqrt{2h\log\log(1/h)}}$ as $h \downarrow 0$ is almost surely equal to $[-1,1]$.

12. Blumenthal's zero-one law

We move towards the Markov property of Brownian motion and its consequences. To give a quick preview, standard Brownian motion turns out to be a strong Markov process, and we shall find many martingales hidden in it. These, together with optional sampling theorems applied to certain stopping times will allow us to study very fine properties of Brownian motion in depth. But as may be expected, certain technical matters will crop up. We start with one such.

Let W be a standard Brownian motion in 1-dimension, defined on some $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{F}_t := \sigma\{W_s : s \le t\}$ be the associated natural filtration. Define $\tau = \inf\{t : W(t) \ge 1\}$ and let $\tau' = \inf\{t : W(t) > 1\}$. It is easy to see that τ is a stopping time for the natural filtration but τ' is not (just find two paths that agree up to τ but that have different values for τ').

We would like τ' to also be a stopping time. This can be done by enlarging the filtration to $\mathcal{F}_t^+ := \bigcap\limits_{s>t} \mathcal{F}_s$. The filtration $\mathcal{F}_\bullet^+$ is called the right-continuous version of $\mathcal{F}_\bullet$ because $\bigcap\limits_{s>t} \mathcal{F}_s^+ = \mathcal{F}_t^+$ for every $t \ge 0$ or in other words $(\mathcal{F}_\bullet^+)^+ = \mathcal{F}_\bullet^+$. It is easy to see that τ' is indeed a stopping time with respect to $\mathcal{F}_\bullet^+$, since the event $\{\tau' \le t\} \in \mathcal{F}_s$ for each $s > t$.

Needless to say, τ remains a stopping time upon enlarging the filtration. What can go wrong with enlargement are Markov properties or martingale properties. For example, for any t we know that $W(\cdot + t) - W(t)$ is independent of $\mathcal{F}_t$. Does it remain true that $W(\cdot + t) - W(t)$ is independent of $\mathcal{F}_t^+$? If not, it is easy to imagine that the enlargement causes more difficulties than it solves.

The first and foremost task is to check that the enlargement is trivial - it adds only $\mathbf{P}$-null sets. This is indeed true.

Lemma 52: Blumenthal's zero-one law

If $A \in \mathcal{F}_0^+$, then $\mathbf{P}(A)$ equals 0 or 1.

PROOF. We know that $W^T := W(T + \cdot) - W(T)$ is independent of $(W_t)_{0 \le t \le T}$, for any $T > 0$. As $T \downarrow 0$, the sigma-algebra generated by $(W_t)_{0 \le t \le T}$ decreases to $\mathcal{F}_0^+$. Further, $\sigma(\cup_{T>0} \sigma(W^T)) = \sigma(W)$, since $W_T \to 0$ (to be more precise, because $W(t) = \lim_{T \downarrow 0} W^T(t)$). Therefore, $\sigma(W)$ is independent of $\mathcal{F}_0^+$. But $\mathcal{F}_0^+ \subseteq \sigma(W)$, hence $\mathcal{F}_0^+$ is independent of itself. Hence any $A \in \mathcal{F}_0^+$ must have probability 0 or 1. ■

In this proof, we used the following simple fact (observe that we could have worked with a sequence T_n decreasing to 0).

Exercise 23

Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots$ and $\mathcal{G}_1 \supseteq \mathcal{G}_2 \supseteq \ldots$ be sub-sigma algebras of $\mathcal{F}$ in $(\Omega, \mathcal{F}, \mathbf{P})$. If $\mathcal{F}_n$ is independent of $\mathcal{G}_n$ for each n, then $\sigma\left(\bigcup_n \mathcal{F}_n\right)$ is independent of $\bigcap_n \mathcal{G}_n$.

Remark 8

In the lectures, I indicated further enlargement by including all $\mathbf{P}$-null sets in each $\mathcal{F}_t$. More precisely, let $\mathcal{F}_\infty = \sigma\{W\} = \sigma\{\bigcup_{t\ge 0}\mathcal{F}_t\}$ and let

$$\mathcal{N} = \{A \subseteq \Omega : A \subseteq B \text{ for some } B \in \mathcal{F}_\infty \text{ with } \mathbf{P}(B) = 0\}.$$

Then define $\overline{\mathcal{F}}_t^+ = \sigma\{\mathcal{F}_t^+ \cup \mathcal{N}\}$. This is the completed, right-continuous filtration. All results stated below for the right-continuous filtration also hold for the completed right-continuous filtration.

13. Markov and strong Markov properties

Let W be a standard d-dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $\mathcal{F}_\bullet$ be the natural filtration generated by W and let $\mathcal{F}_\bullet^+$ be the right-continuous filtration defined by $\mathcal{F}_t^+ = \bigcap_{s>t}\mathcal{F}_s$.

Here is a naive way to state the Markov and strong Markov properties.

- (Markov property). Fix T and define $B(t) = W(T+t) - W(T)$ for $t \ge 0$. Then, B is a standard Brownian motion that is independent of $\mathcal{F}_t^+$.
- (Strong Martov property). Fix an $\mathcal{F}_\bullet^+$-stopping time τ and define $B(t) = W(t+\tau) - W(\tau)$ for $t \ge 0$. Then B is a standard Brownian motion independent of $\mathcal{F}_\tau^+$. Recall that $\mathcal{F}_\tau^+ = \{A \in \mathcal{F} : A \cap \{\tau \le t\} \in \mathcal{F}_t\}$.

We have already proved the Markov property when the filtration $\mathcal{F}_\bullet$ is used. By Blumenthal's zero-one law, $\mathcal{F}_t^+$ is got from $\mathcal{F}_t$ by augmenting some $\mathbf{P}$-null sets. Hence, independence of B from $\mathcal{F}_T$ is equivalent to independence of B from $\mathcal{F}_T^+$. Strong Markov property is slightly less obvious.

Proof of Strong Markov property. For simplicity we use the notation of 1-dimension. First assume that τ takes countably many values $s_0, s_1, s_2, \ldots$ for some $\delta > 0$. Fix any $A \in \mathcal{F}_\tau$, any $n \ge 1$ and $t_1, \ldots, t_n \ge 0$, and any $u_1 \ldots, u_n \in \mathbb{R}$. Let E be the event that $B(t_j) \le u_j$ for $1 \le j \le n$. Then,

$$\begin{aligned}\mathbf{P}\{E \cap A\} &= \sum_{m=0}^{\infty} \mathbf{P}\{E \cap A \cap \{\tau = s_m\}\} \\ &= \mathbf{P}\left\{\{B(s_m + t_j) - B(s_m) \le u_j \text{ for } j \le n\} \cap A \cap \{\tau = s_m\}\right\}.\end{aligned}$$

For fixed m, by Markov property and the fact that $A \cap \{\tau = s_m\} \in \mathcal{F}_m^+$, the mth summand above is equal to

$$\mathbf{P}\{W(t_j) \le u_j \text{ for } j \le n\}\mathbf{P}\{A \cap \{\tau = s_m\}\}.$$

Adding up and using $\mathbf{P}(A) = \sum_m \mathbf{P}\{A \cap \{\tau = s_m\}\}$ gives the identity $\mathbf{P}\{E \cap A\} = \mathbf{P}\{W(t_j) \leq u_j \text{ for } j \leq n\}\mathbf{P}\{A\}$. This shows that B is independent of $\mathcal{F}_\tau^+$ and that B has the same distribution as W.

Now consider a general stopping time τ. For $\ell \geq 1$ define $\tau_\ell = 2^{-\ell}\lceil 2^\ell \tau \rceil$. Then τ_ℓ is a stopping time, $\tau \leq \tau_\ell \leq \tau + 2^{-\ell}$. Thus $\tau_\ell \downarrow \tau$. Let $V = (W(\tau + t_1), \ldots, W(\tau + t_n))$ and $V_\ell = (W(\tau_\ell + t_1), \ldots, W(\tau_\ell + t_n))$ so that by continuity of Brownian motion, we have $V_\ell \overset{a.s.}{\to} V$. Thus, for most choices of $u_1, \ldots, u_n$ (we need u_j to be a continuity point of $V(j)$) we get

$$\begin{aligned}\mathbf{P}\{\{V(j) \leq u_j \text{ for } j \leq n\} \cap A\} &= \lim_{\ell \to \infty} \mathbf{P}\{\{V_\ell(j) \leq u_j \text{ for } j \leq n\} \cap A\} \\ &= \lim_{\ell \to \infty} \mathbf{P}\{W(t_j) \leq u_j \text{ for } j \leq n\}\mathbf{P}\{A\}\end{aligned}$$

where the last line used the strong Markov property for stopping times τ_ℓ that takes countably many values. ■

For our purposes this is sufficient. Observe that Markov property can be stated as saying that the conditional distribution of $W(T+t)$, $t \geq 0$, given $\mathcal{F}_T^+$ is the same as that of Brownian motion started at the point $W(T)$. Similarly, strong Markov property says that the conditional distribution of $W(\tau + t)$ given $\mathcal{F}_\tau^+$.

This is a better way of stating these properties. In case of Brownian motion, because of symmetries ($W + x$ is the same as Brownian motion conditioned on starting at x). In general, we consider a family of probability measure $\mathbf{P}_x$, $x \in \mathbb{R}$, on $C[0, \infty)$ such that $\mathbf{P}_x\{f : f(0) = x\} = 1$. This family is said to have (time-homogeneous) Markov property if:

Fix any $x \in \mathbb{R}^d$ and let $X = (X_t)_{t \geq 0} \sim \mathbf{P}_x$. Then, conditional on $\mathcal{F}_T^+$, the process $(X(T+t))_{t \geq 0}$ has the same distribution as $\mathbf{P}_{X(T)}$. Strong Markov property is stated in a similar way.

Example 26

Let $\mathbf{P}_x$ be the distribution of $(x + W_t + t)_t$ for $x \geq 0$ and the distribution of $(x + W_t - t)_{t \geq 0}$ for $x < 0$. Then $\mathbf{P}_x$ does not have Markov property.

Example 27

Let $\mathbf{P}_x$ be the distribution of $x + W$ for $x \neq 0$ and let $\mathbf{P}_0 = \delta_0$ be the Dirac measure at the constant function zero. Then, $\mathbf{P}$ satisfies Markov property but not the strong Markov property.

Indeed, if $x = 0$, then conditional on $\mathcal{F}_T$, the distribution of the future path $(W(T+t))_{t \geq 0}$ is degenerate at zero. If $x \neq 0$, ignoring the zero probability event $W_T = 0$, we see that the

future path $W(T+t)$ is that of Brownian motion stated at $W(T)$ (does not work if $W(T)=0$ but zero probability events may be ignored).
But if $\tau = \min\{t : W_t = 0\}$, then the conditional distribution of $(W(\tau+t))_{t\geq 0}$ is the standard Brownian motion, which is not the same as $\mathbf{P}_{W(\tau)} = \mathbf{P}_0$.

14. Zero set of Brownian motion

Let W be standard 1-dimensional Brownian motion and let $Z = \{t : W_t = 0\}$. Clearly Z is a random closed set of $\mathbb{R}_+$.

Theorem 53

Z has no isolated points, w.p.1.

PROOF. For $q \in \mathbb{Q}_+$, let $\tau_q = \min\{s > t : W(s) = 0\}$. By SMP, $W(\tau_q + t) - W(\tau_q) = W(\tau_q + t)$ is a standard Brownian motion. In particular, it has infinitely many zeros on any positive time interval $[0, \varepsilon)$. Hence, τ_q is an accumulation point (from the right) of Z, w.p.1. Take intersection over $q \in \mathbb{Q}_+$ to see that w.p.1., every τ_q, $q \in \mathbb{Q}$, is an accumulation point of Z.

Now, a zero $z \in Z$ is not of the form τ_q is and only if z is an accumulation point of Z from the left! Thus, all zeros of W are accumulation points. ■

15. Reflection principle

Let W be standard 1-dimensional Brownian motion. For $a > 0$ define the *running maximum* $M_t := \max\limits_{0\leq s\leq t} W_s$ and the *first passage time* $\tau_a := \min\{t \geq 0 : W(t) \geq a\}$. These are closely interconnected, since $M_t \geq a$ if and only if $\tau_a \leq t$.

Many questions can be asked: What is the distribution of M, of τ_a? Let T_* be the (unique) time in $[0,1]$ such that $W(T_*) = \max\limits_{s\leq 1} W(s)$. What is the distribution of T_*?

We shall answer all these questions. A basic tool is the reflection principle, a direct consequence of the strong Markov property.

Lemma 54: Reflection principle

Let W be standard 1-dimensional Brownian motion. Fix $a > 0$ and define

$$B(t) = \begin{cases} W(t) & \text{if } t \leq \tau_a, \\ 2W(\tau_a) - W(t) & \text{if } t > \tau_a. \end{cases}$$

Then, B is a standard Brownian motion.

PROOF. Let $X = (W_t)_{t \le \tau_a}$, $Y = (W_{t+\tau_a} - a)_{t \ge 0}$ and $Z = -Y$. Then $Y \overset{d}{=} Z$, X is independent of Y (by strong Markov property) and hence X is independent of Z. Hence $(X, Y) \overset{d}{=} (X, Z)$.

Concatenating X with Y gives W while concatenating X with Z gives B. Thus $B \overset{d}{=} W$. ■

Distribution of the running maximum: Let $a > 0$ and $t > 0$. Let W be a standard Brownian motion and let B be related to it as in the reflection principle. Then,

$$\begin{aligned}\{M_t > a\} &= \{M_t > a, W_t > a\} \sqcup \{M_t > a, W_t < a\} \\ &= \{W_t > a\} \sqcup \{B_t > a\}.\end{aligned}$$

Therefore, $\mathbf{P}\{M_t > a\} = 2\mathbf{P}\{W_t > a\} = 2\bar{\Phi}(a/\sqrt{t})$ where $\bar{\Phi}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$ is the tail of the standard normal distribution. Differentiating, we get the density of M_t to be

$$f_{M_t}(a) = -\frac{d}{dt}\mathbf{P}\{M_t > a\} = \frac{2}{\sqrt{2\pi}\sqrt{t}} e^{-\frac{1}{2t}a^2}.$$

Another way to say this is that *for each fixed t* we have $M_t \overset{d}{=} |W_t|$, since $\mathbf{P}\{|W_t| > a\} = 2\mathbf{P}\{W_t > a\}$ for any $a \ge 0$.

Distribution of the first passage times: As $\tau_a \le t$ if and only if $M_t \ge a$, we get $\mathbf{P}\{\tau_a \le t\} = 2\bar{\Phi}(a/\sqrt{t})$. The density of τ_a is

$$f_{\tau_a}(t) = \frac{d}{dt}\mathbf{P}\{\tau_a \le t\} = \frac{a}{\sqrt{2\pi}\, t^{\frac{3}{2}}} e^{-\frac{1}{2t}a^2}.$$

The density approaches zero at $t = 0$ and decayse like $t^{-3/2}$ as $t \to \infty$. Thus the tail is quite heavy.

Exercise 24

Deduce that $\mathbf{E}[\tau_a] = \infty$. In fact $\mathbf{E}[\tau_a^p] < \infty$ if and only if $p < \frac{1}{2}$.

Joint distribution of the Brownian motion and its running maximum: Fix $a > 0$ and $-\infty < b < a$. Then, by the definition of B in terms of W,

$$\{M_t > a \text{ and } W_t < b\} = \{B_t > 2a - b\}.$$

Since B is standard Brownian motion, we get $\mathbf{P}\{M_t > a \text{ and } W_t < b\} = \bar{\Phi}((2a-b)/\sqrt{t})$. Thus,

$$f_{(M_t, W_t)}(a, b) = -\frac{d^2}{da\, db}\bar{\Phi}((2a-b)/\sqrt{t}) = \frac{2(2a-b)}{\sqrt{2\pi}\, t^{\frac{3}{2}}} e^{-\frac{1}{2t}(2a-b)^2}.$$

Some distributional identities: The process $(|W_t|)_{t \ge 0}$ is called reflected Brownian motion. We have the following distributional identities.

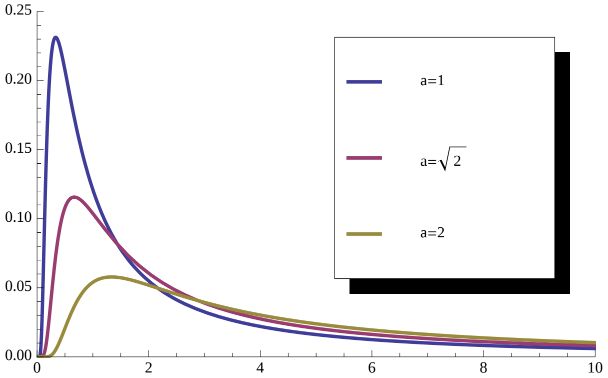


FIGURE 3. Densities of first passage time τ_a

- ▶ $M_t \overset{d}{=} |W_t|$ for each t. We already saw this.
- ▶ $M_t - W_t \overset{d}{=} |W_t|$ for each t. This can be computed from the joint intensity, but here is a computation-free proof[6]. The process $X_s = W_{t-s} - W_t$ for $0 \le s \le t$ is a standard Brownian motion. Observe that $M_t^X = M_t^W - W_t$. Hence, the distributional identity follows!

Do these equalities in distribution extend to those of the processes? The first one does not, since M is an increasing process while $|W|$ is not. But it is a non-trivial theorem of Lévy that the second one does, i.e., $M - W \overset{d}{=} |W|$. The key point is that $M - W$ is a Markov process. Once that is checked, the equality in distribution at fixed times easily extends to equality of finite dimensional distributions. Since both processes are continuous, this implies equality in distribution of the two processes.

Local times - a digression: Further, consider a probability space with two standard Brownian motions $W, \tilde{W}$ related such that $M^W - W = \tilde{W}$. Then, the process M^W is related to $\tilde{W}$ in a special way. Being an increasing function, M^W may be thought of as the distribution function of a random measure. Observe that M^W is constant on any interval (s,t) where $\tilde{W}$ has no zeros. This means that the random measure defined by M^W is supported on the zero set of $\tilde{W}$. It is called the *local time* of $\tilde{W}$, a clock that ticks only when the Brownian motion is at zero.

This is not entirely satisfactory. What we would like is to define a local time for the Brownian motion that we started with, in a canonical way. This is possible. Indeed, it can be shown that

$$L_t(0) := \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \text{Leb}\{s \le t : |W_s| \le \varepsilon\}$$

exists and defined the local time at 0. It is also possible to define $L_t(x)$ for $t > 0$ and $x \in \mathbb{R}$, simultaneously. But we shall not touch upon this matter in this course.

[6] Thanks to Arun Selvan for the nice proof!

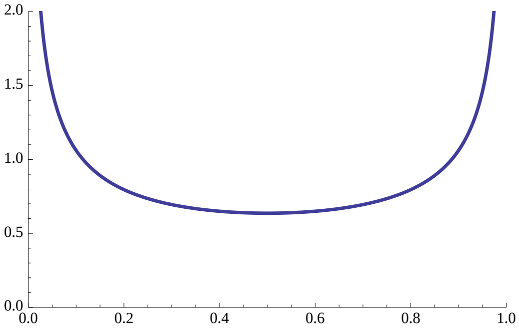


FIGURE 4. Arcsine density

16. Arcsine laws

Let $T^* = \arg\max_{0\le t\le 1} W_t$ be the location of the global maximum of Brownian motion in unit time. As all the values of local maxima are distinct, T^* is well-defined. Also, define $L := \max\{t \le 1 : W_t = 0\}$ be the time of last return to the origin.

Let us also recall the *arc-sine* distribution that has CDF $\frac{2}{\pi}\arcsin(\sqrt{t})$ and density $\frac{1}{\pi\sqrt{t(1-t)}}$, for $0 < t < 1$.

Theorem 55: Lévy

T^* and L have the arcsine distribution.

PROOF. (1) Fix $t \in (0,1)$. Then $T^* \le t$ if and only if $\max_{0\le s\le t} W_s \ge \max_{t\le s\le 1} W_s$ which is equivalent to $\max_{0\le s\le t} W_s - W_t \ge \max_{t\le s\le 1} W_s - W_t$.

If $\tilde{W}_{s-t} := W_s - W_t$ for $t \le s \le 1$, then $\tilde{W}$ is a standard Brownian motion (run for time $1-t$) that is independent of $(W_s)_{s\le t}$. Thus, putting everything together, we arrive at

$$\mathbf{P}\{T^* \ge t\} = \mathbf{P}\{M_t \ge \tilde{M}_{1-t}\}.$$

Because $M_t \overset{d}{=} |W_t|$, we may write $M_t = \sqrt{t}|X|$ and $\tilde{M}_{1-t} = \sqrt{1-t}|Y|$ where X, Y are i.i.d. standard Gaussians. Thus,

$$\mathbf{P}\{T^* \ge t\} = \mathbf{P}\{\sqrt{t}|X| \ge \sqrt{1-t}|Y|\} = \mathbf{P}\left\{\left|\frac{Y}{X}\right| \le \frac{\sqrt{t}}{\sqrt{1-t}}\right\}.$$

It is an easy exercise that Y/X has standard Cauchy distribution and hence the last probability is equal to $\frac{2}{\pi}\arctan(\sqrt{t}/\sqrt{1-t})$ which is equal to $\frac{2}{\pi}\arcsin(\sqrt{t})$. This shows that T^* has arcsine distribution.

(2) $L \ge t$ if and only if W hits zero somewhere in $[t,1]$. Let $\tilde{W}_s = W_{t+s} - W_t$ for $0 \le s \le 1-t$ which is a Brownian motion independent of $\mathcal{F}_t$.

Now, W hits zero in $[t,1]$ if and only if $\tilde{M}_{1-t} \geq |W_t|$ (if $W_t < 0$) or $\min_{s\leq 1-t} \tilde{W}_s \leq -|W_t|$ (if $W_t > 0$). Clearly either one has the same probability. Hence we arrive at

$$\mathbf{P}\{L \geq t\} = \mathbf{P}\{\tilde{M}_{1-t} \geq |W_t|\}.$$

But we may write $\tilde{M}_{1-t} = \sqrt{1-t}|X|$ and $|W_t| = \sqrt{t}|Y|$ where X, Y are i.i.d. standard Gaussians. Hence we return to the same calculation as for T^* and deduce that L must have arcsine distribution. ■

Lévy proved a third arcsine law. This is for $\{t \leq 1 : W_t > 0\}$, the proportion of time spent by the Brownian motion in the positive half-line. We shall prove this later.

Proof of Lévy's third arcsine law by chaos expansion:

Let γ be the standard Gaussian measure on $\mathbb{R}$. Applying Gram-Schmidt (without normalizing) procedure to $1, x, x^2, \ldots$ in $L^2(\gamma)$, we get a sequence of monic polynomials $H_0(x), H_1(x), \ldots$ that are orthogonal in $L^2(\gamma)$. Clearly H_n has degree n. These are known as *Hermite polynomials* and we can describe them explicitly in multiple ways:

(1) $H_n(x) = (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n} e^{-\frac{1}{2}x^2}$. This is clearly monic and has degree n. Hence it suffices to check that they are orthogonal in $L^2(\gamma)$, which follows by integrating by parts. If $m < n$,
$$\int H_n(x)H_m(x)d\gamma(x) = \int H_m(x)\frac{d^n}{dx^n}e^{-\frac{1}{2}x^2}\,dx \;=\; (-1)^n \int e^{-\frac{1}{2}x^2}\frac{d^n}{dx^n}H_m(x)dx \;=\; 0.$$

(2)

17. Martingales in Brownian motion

Using the strong Markov property, we found the distribution of the first passage times τ_a. It can be thought of as the exit time of a half-infinite interval. A natural question is to find the distribution of the exit time $\tau_{b,a}$ of a finite interval $[b,a]$ for $b < 0 < a$. In particular, since $\tau_{-a,a} \leq t$ if and only if $\max_{s\leq t}|W_s| \geq a$, this will also tell us the distribution of the running maximum of a reflected Brownian motion.

The tools we use are martingales inside Brownian motion. We know that W_t itself is a martingale. But W_t^2 is not. Indeed,

$$\begin{aligned}\mathbf{E}[W_t^2 \mid \mathcal{F}_s] &= \mathbf{E}[(W_s + (W_t - W_s))^2 \mid \mathcal{F}_s] \\ &= \mathbf{E}[W_s^2 + 2W_s(W_t - W_s) + (W_t - W_s)^2 \mid \mathcal{F}_s] \\ &= W_s^2 + (t-s).\end{aligned}$$

From this, we can deduce that $W_t^2 - t$ is a martingale. Similarly,

$$\begin{aligned}\mathbf{E}[W_t^3 \mid \mathcal{F}_s] &= \mathbf{E}[(W_s + (W_t - W_s))^3 \mid \mathcal{F}_s] \\ &= \mathbf{E}[W_s^3 + 3W_s^2(W_t - W_s) + 3W_s(W_t - W_s)^2 + (W_t - W_s)^3 \mid \mathcal{F}_s] \\ &= W_s^3 + 3W_s(t-s).\end{aligned}$$

From this, we deduce that $W_t^3 - 3tW_t$ is a martingale. Continuing, we find that $W_t^4 - 6tW_t^2 + 3t^2$ is a martingale. What is the general pattern?

Exponential martingales: Let $\lambda \in \mathbb{R}$ and define $M_\lambda(t) := e^{\lambda W_t - \frac{1}{2}\lambda^2 t}$. Then,

$$\mathbf{E}[M_\lambda(t) \mid \mathcal{F}_s^+] = e^{\lambda W_s - \frac{1}{2}\lambda^2 t}\mathbf{E}[e^{\lambda(W_t - W_s)}] = e^{\lambda W_s - \frac{1}{2}\lambda^2 t}e^{\frac{1}{2}\lambda^2(t-s)} = M_\lambda(s).$$

Thus, for each $\lambda \in \mathbb{R}$ we have a martingale $M_\lambda(t)$, $t \geq 0$.

Consider the power series expansion of function $e^{\lambda x - \frac{1}{2}\lambda^2} = \sum_{n=0}^\infty \frac{1}{n!}H_n(x)\lambda^n$ where

$$H_n(x) = \frac{d^n}{d\lambda^n}e^{\lambda x - \frac{1}{2}\lambda^2}\Big|_{\lambda=0}.$$

It is easy to see that $H_n(x)$ is a polynomial of degree n in x. These are called *Hermite polynomials*. By explicit computation one can see that $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, etc. The martingales that we got earlier are precisely $t^{n/2}H_n(W_t/\sqrt{t})$.

Exercise 25

Use differentiation under the integral sign and the fact that M_λ is a martingale, to show that $t^{n/2}H_n(W_t/\sqrt{t})$ is a martingale for every $n \geq 0$.

The usefulness of martingales is via the optional sampling theorem. We showed in class how to analyse the exit time of an interval by one-dimensional Brownian motion. And also how to find martingales for multi-dimensional Brownian motion. For instance, any $u : \mathbb{R}_+ \times \mathbb{R}^d$ that satisfies $\partial_t u(t,x) + \frac{1}{2}\Delta u(t,x) = 0$ and some growth conditions gives a martingale $u(t, W_t)$. In particular, harmonic functions v satisfying some growth conditions give the martingales $v(W_t)$.

We used these to prove recurrence and transience properties of Brownian motion. We may touch upon the Dirichlet problem in the last lecture (if we have time).

Read up on these in the books we have been referring to.

Like with discrete time martingales, optional stopping theorem is a great tool. We state a basic version.

Theorem 56: Optional stopping theorem

Let $X = (X_t)_{t \geq 0}$ be a martingale w.r.t. a filtration $\mathcal{F}_\bullet = (\mathcal{F}_t)_{t \geq 0}$. Assume that the sample paths of X are continuous. Let τ be a stopping time for $\mathcal{F}_\bullet$ and define $X^\tau(t) := X(\tau \wedge t)$. If X^τ is uniformly integrable, then $\mathbf{E}[X(\tau)] = \mathbf{E}[X(0)]$.

17.1. Recurrence and transience of Brownian motion.

18. Wiener's stochastic integral

Let W be standard Brownian motion on $(\Omega, \mathcal{F}, \mathbf{P})$. Let $f : [0,1] \to \mathbb{R}$. We want to make sense of $\int_0^1 f(t)dW(t)$ with extra conditions on f if necessary.

Let us first review what can be done in the non-random situation, where the integrating function is fixed.

- Let $\alpha \in C^1(\mathbb{R})$. Then for any $f \in C[0,1]$ we may define $\int_0^1 f(t)d\alpha(t)$ as $\int_0^1 f(t)\alpha'(t)dt$, the latter being the Riemann integral of a continuous function.
- More generally, if α is a function of bounded variation[7], then following ideas similar to that of Riemann intergral, Stieltjes showed that $\int_0^1 f(t)d\alpha(t)$ can be made sense of for any $f \in C[0,1]$.
- Suppose $\alpha \in C[0,1]$, not necessarily of bounded variation. Then it is no longer possible to define Stieltjes' integral. But for $f \in C[0,1]$, we can define
$$\int_0^1 f(t)d\alpha(t) := f(1)\alpha(1) - f(0)\alpha(0) - \int_0^1 \alpha(t)f'(t)dt.$$
The justification for this definition is that when α is of bounded variation, the expression on the right is equal to $\int_0^1 f d\alpha$, known as the integration by parts formula.

  This simple observation has considerable reach, and lies at the base of the theory of distributions in functional analysis. Any continuous function acts on smooth enough functions as above.

Now fix a sample path of Brownian motion. It is not of bounded variation, hence the first two approaches do not work. That is, we cannot make sense of $\int_0^1 f(t)dW(t)$ for all $f \in C[0,1]$. However, the sample path is indeed continuous, hence we can use the third approach and define $\int_0^1 f(t)dW(t)$ for $f \in C^1$ by the integration by parts formula.

But we can do more - we shall in fact define $\int_0^1 f(t)dW(t)$ for every $f \in L^2[0,1]$! This is done as follows.

[7] By definition, α is said to have bounded variation if $\sup \sum_{k=1}^n |\alpha(t_k) - \alpha(t_{k-1})|$ is finite, where the supremum is over all $0 = t_0 < t_1 < \ldots < t_n = 1$. It is a fact that a function is of bounded variation if and only if it can be written as a difference of two increasing functions. A

Step 1: Let $f : [0,1] \to \mathbb{R}$ be a step function, $f(t) = \sum_{k=1}^{n} \lambda_k \mathbf{1}_{[a_k,b_k]}(t)$ for some $0 \le a_1 < b_1 < a_2 < b_2 < \ldots < a_n < b_n$ for some $n \ge 1$. Then we define

$$I(f) = \sum_{k=1}^{n} \lambda_k (W(b_k) - W(a_k)).$$

If $\mathcal{S}$ denotes the collection of all step functions on $[0,1]$, then $\mathcal{S}$ is a dense subspace of $L^2[0,1]$. What we have defined is a function $I : \mathcal{S} \to L^2(\Omega, \mathcal{F}, \mathbf{P})$.

Step 2: We claim that $I : \mathcal{S} \to L^2(\Omega, \mathcal{F}, \mathbf{P})$ is a linear isometry. Further, $I(f)$ is a Gaussian random variable for each $f \in \mathcal{S}$.

Linearity is clear. To check isometry, by the independent increments property of W, we get

$$\|I(f)\|^2_{L^2(\mathbf{P})} = \mathbf{E}[|I(f)|^2] = \text{Var}(\sum_{k=1}^{n} \lambda_k (W(b_k) - W(a_k))) = \sum_{k=1}^{n} \lambda_k^2 (b_k - a_k) = \|f\|^2_{L^2[0,1]}.$$

That $I(f)$ is Gaussian is clear. Therefore it has $N(0, \|f\|^2_{L^2[0,1]})$ distribution.

Step 3: I maps Cauchy sequences in $\mathcal{S}$ to Cauchy sequences in $L^2(\mathbf{P})$. Hence, if $f_n \in \mathcal{S}$ and $f \in L^2[0,1]$ and $f_n \to f$ in L^2, then $\{f_n\}$ is Cauchy in $L^2[0,1]$ and therefore $\{I(f_n)\}$ is Cauchy in $L^2(\mathbf{P})$. By completeness of $L^2(\mathbf{P})$, $I(f_n)$ has a limit. Clearly this limit depends only on f and not on the sequence $\{f_n\}$. Therefore, we can unambiguously extend I to a linear isometry of $L^2[0,1]$ into $L^2(\mathbf{P})$.

This defines the stochastic integral and we usualy write $\int_0^1 f(t)dW(t)$ for $I(f)$. Since L^2-limits of Gaussians are Gaussians, it follows that for any $f, ginL^2[0,1]$, the distribution of $I(f)$ and $I(g)$ is bivariate Gaussian with zero means and $\text{Cov}(I(f), I(g)) = \int_0^1 fg$. In particular, $\text{Var}(I(f)) = \|f\|^2_{L^2[0,1]}$.

How was it possible to integrate every L^2 function? The point to remember is that when talking about Brownian motion, we are not talking of one function, but an entire ensemble of them. Therefore,

(1) For any given Brownian path, there is a function $f \in L^2[0,1]$ (even $f \in C[0,1]$) that cannot be integrated in any sense against the Brownian path.

(2) For a fixed $f \in L^2[0,1]$, this problem does not arise for almost every Brownian path and we can integrate f with respect to W.

(3) For almost every Brownian path, the integrals of all C^1 functions can be simultaneously defined (using the integration by parts formula).

In this sense, Brownian motion is better than a distribution, it can integrate functions with hardly any smoothness.

We shall not really use the Wiener integral to deduce any properties of Brownian motion. We discussed it to mention an important topic that we shall not touch upon in this course. This is the subject of *Ito integral* which makes sense of integrals of random functions such as $\int_0^t W_s dW_s$ (it turns out that this integral is $\frac{1}{2}W_t^2 - \frac{1}{2}t$ in contrast to C^1 functions α for which we always have $\int_0^t \alpha(s)d\alpha(s) = \frac{1}{2}\alpha(t)^2$). Here is an exercise.

Exercise 26

Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function such that $\int_0^t f^2 < \infty$ for all $t < \infty$. Then we may define $X_t = \int_0^t f(s)dW(s)$ exactly as above. Show that X is a martingale.

Stochastic summation: an analogy or more: Consider $\ell^2 = \{x = (x_n)_{n\in\mathbb{N}} : \sum_n x_n^2 < \infty\}$. Let $a = (a_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$. Two observations.

- If $a \in \ell^2$, then $\sum_n a_n x_n$ converges for every $x \in \ell^2$. This is the inner product in ℓ^2, well-defined because of the Cauchy-Schwarz inequality.
- Suppose $\sum_n a_n x_n$ converges for each $x \in \ell^2$. Then $a \in \ell^2$. To see this, define $L_m : \ell^2 \to \mathbb{R}$ by $L_m(x) = \sum_{k\le m} a_k x_k$. Then L_m is a bounded linear functional with $\|L_m\|^2 = \sum_{k\le m} a_k^2$. By the hypothesis, for ach $x \in \ell^2$, the sequence $\{L_m(x)\}_m$ is convergent in $\mathbb{R}$, and hence bounded in $\mathbb{R}$. By the uniform boundedness principle, $\{\|L_m\|\}$ is bounded. Thus $\sum_{k\le m} a_k^2 \le C$ for some C and for all m which implies that $a \in \ell^2$.

Now consider $\xi = (\xi_n)_{n\in\mathbb{N}}$, where ξ_n are i.i.d. $N(0,1)$ random variables on a common probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then, $\xi_n > 1$ infinitely often, w.p.1. and hence $\xi \notin \ell^2$, w.p.1. Thus, for almost every ω, there is an $x \in \ell^2$ such that $\sum_n \xi_n(\omega)x_n$ does not converge. However, for each $x \in \ell^2$, using standard results on sums of independent random variables, it follows that $\sum_n \xi_n$ converges w.p.1. But let us do it in a more roundabout way to bring out the analogy with the Wiener integral.

Step 1: Let $\mathcal{S} = \{x \in \ell^2 : x_n = 0 \text{ for all large } n\}$, a dense subspace of ℓ^2. For $x \in \mathcal{S}$, the sum $I(x) := \sum_n \xi_n x_n$ is a finite sum and therefore well-defined.

Step 2: $I : \mathcal{S} \mapsto L^2(\Omega, \mathcal{F}, \mathbf{P})$ is a linear isometry. In fact, for each $x \in \ell^2$, $I(x) \sim N(0, \|x\|^2)$. This is easy to see by computing $\mathbf{E}[I(x)^2]$.

Step 3: I extends as an isometry of ℓ^2 into $L^2(\Omega, \mathcal{F}, \mathbf{P})$. This step is carried out exactly the same way.

Interpret $I(x)$ as $\sum_n \xi_n x_n$ (in fact, the latter series converges almost surely, using standard theorems on sums of independent random variables). Thus, we have a very close analogy with the previous situation. To make the analogy even closer, you may want to define $S_n = \xi_1 + \ldots + \xi_n$ so that $\sum_n \xi_n x_n = \sum_n (S_n - S_{n-1}) x_n$ looks like "$\int x(n) dS(n)$".

19. A peek into Ito's integration theory

Let W be standard Brownian motion in one dimension, defined on some $(\Omega, \mathcal{F}, \mathbf{P})$. The Wiener integral made sense of $\int_0^1 f(t) dW(t)$ for $f \in L^2[0,1]$ as a random variable on the same probability space. Can we also integrate random integrands. In other words, suppose $X = (X(s))_{s \geq 0}$ is a stochastic process on the same probability space. Can we make sense of $\int_0^1 X(s) dW(s)$? If X is independent of W, then by conditioning on X, we reduce this to the situation of deterministic integrands. Hence we are more interested in the situation where X depends on W.

In the basic theory of Stieltjes' integral of f with respect to α, one starts out by considering partitions $0 = t_0 < t_1 < \ldots < t_N = 1$ and forming the Riemann sum $\sum_{k=0}^{N-1} f(t_k^*)(\alpha(t_{k+1}) - \alpha(t_k))$ where $t_k^* \in [t_k, t_{k+1}]$ are arbitrary points. If the limit of these sums exists, as the partitions get finer and over arbitrary choices of t_k^*s, then the limit value is defined to be $\int_0^1 f(t) d\alpha(t)$.

Taking inspiration from this (all integration must be limit of summation, after all), we can try to make sense of $\int_0^1 X(s) dW(s)$ by considering

$$\sum_{k=0}^{N-1} X(t_k^*)(W(t_{k+1}) - W(t_k)).$$

It turns out that the choice of t_k^* matters a lot! Henceforth, we consider the prototypical case of $X = W$. By taking t_k^* to be the left end-point or the right end-point or the mid-point of $[t_{k-1}, t_k]$, we get the following sums.

$$I_N := \sum_{k=0}^{N-1} W(\frac{k}{2^n})(W(\frac{k+1}{2^n}) - W(\frac{k}{2^n})),$$

$$J_N := \sum_{k=0}^{N-1} W(\frac{k+1}{2^n})(W(\frac{k+1}{2^n}) - W(\frac{k}{2^n})),$$

$$K_N := \sum_{k=0}^{N-1} W(\frac{k+\frac{1}{2}}{2^n})(W(\frac{k+1}{2^n}) - W(\frac{k}{2^n})).$$

Here is a simple calculation that shows that even if they have limits, the limits must be different.

$$J_N - I_N = \sum_{k=0}^{N-1} (W(\frac{k+1}{2^n}) - W(\frac{k}{2^n}))^2 \overset{a.s.}{\to} 1$$

where the almost sure convergence was a homework exercise (if we consider a general sequence of partitions, we only get convergence in probability, but that is not relevant for the point we are about to make). Thus, the limits of J_N and of I_N, if they exist, must differ by 1. In fact the limits exist because

$$J_N + I_N = \sum_{k=0}^{N-1}(W(\frac{k+1}{2^n})^2 - W(\frac{k}{2^n})^2) = W(1)^2.$$

Therefore, we deduce that

$$J_N \overset{a.s.}{\to} \frac{1}{2}W(1)^2 + \frac{1}{2}, \quad I_N \overset{a.s.}{\to} \frac{1}{2}W(1)^2 - \frac{1}{2}.$$

We leave it as an exercise to check that $K_N \overset{a.s.}{\to} \frac{1}{2}W(1)^2$. There is no difficulty in extending this for any fixed t and get three possible candidates $\frac{1}{2}W(t)^2 + \frac{1}{2}t$, $\frac{1}{2}W(t)^2 - \frac{1}{2}t$ and $\frac{1}{2}W(t)^2$ as possible definitions of $\int_0^t W(s)dW(s)$. Of course, choosing other t_k^*, one can get other candidates. Which is the right one?

It is tempting to choose the limit of K_N, since choosing the mid-point makes one seem less prejudiced to the temptations of left and right, and also because it agrees with the result for Stieltjes' integrals, $\int_0^t \alpha(t)d\alpha(t) = \frac{1}{2}\alpha(t)^2 - \frac{1}{2}\alpha(0)^2$. But when we consider the integral as a process in t (caution: We only showed that for fixed t, the limit exists, but let us suppose that the integral makes sense as a process in t), the leftist choice $\frac{1}{2}W(t)^2 - \frac{1}{2}t$ is more inviting, since it alone is a martingale! Why did it become a martingale? A hint is there already in the definition of I_N. If you consider the discrete-time martingale $W(k/2^n)$, $k = 0, 1, 2, \ldots$, then I_N is like enhancing the kth game by betting a *predictable* amount $W(k/2^n)$. The choices in J_N and K_N need knowledge of the future to make the bet, hence they fail to be martingales.

Ito integral: Let W be standard Brownian motion on $(\Omega, \mathcal{F}, \mathbf{P})$ and let X be a continuous stochastic process that is adapted to $\overline{\mathcal{F}_\bullet}^+$ (right-continuos, completed filtration). Let us assume that X is uniformly bounded (this can be relaxed if X is unlikely to be very large on bounded intervals). Then the limits

$$I_X(t) := \lim_{n\to\infty} \sum_{k=0}^{\lfloor 2^n t \rfloor} X(k/2^n)(W(\frac{k+1}{2^n}) - W(\frac{k}{2^n}))$$

exists as a process in t, and I_X is a continuous martingale. It is called the *Ito integral* of X with respect to W and denoted $I_X(t) = \int_0^t X(s)dW(s)$.

We have just stated this fact without proof. In specific cases, one can do this by hand. If nothing else, this may be thought of as a systematic way to obtain many martingales from Brownian motion!

Exercise 27

At least for a few small values of $p = 1, 2, \ldots$, show that

$$\int_0^t W(s)^p dW(s) = \frac{1}{p+1} W(t)^{p+1} - \frac{p}{2} \int_0^t W(s)^{p-1} ds.$$

20. Random walks and Brownian motion

Part of the original motivation for Brownian motion was that it was a kind of random walk in continuous time and continuous space. We shall now make this precise and show that random walks converge to Brownian motion in the sense of distribution. For this we need to see them both as objects in the same space.

Random walk as a continuous stochastic process: Let $x_1, x_2, \ldots, x_n$ be real numbers. Define the continuous W_n by

$$W_n(t) = \begin{cases} x_1 + \ldots + x_k & \text{if } t = \frac{k}{n} \text{ for some } 0 \le k \le n, \\ \text{linear in each interval } [\frac{k}{n}, \frac{k+1}{n}]. \end{cases}$$

For later purposes, let us introduce the notation $W_n = \mathcal{T}(x_1, \ldots, x_n)$, so that $\mathcal{T} : \mathbb{R}^n \mapsto C[0,1]$. If x_i are random variables, then W_n is a continuous stochastic process whose distribution is the push-forward of the distribution of T under $\mathcal{T}$. When x_i are i.i.d. random variables, W_n is essentially the random walk with these steps, except that we interpolate continuously to make it a continuous process of continuous time.

Weak convergence in $C[0,1]$: Suppose μ_n, μ are Borel probability measures on a complete, separable metric space (X, d). We say that $\mu_n \overset{d}{\to} \mu$ if $\int f d\mu_n \to \int f d\mu$ for all $f \in C_b(X)$ (the space of bounded continuous functions on X). If $X_n \sim \mu_n$ and $X \sim \mu$ are random variables (not necessarily on the same probability space), we abuse notation and write $X_n \overset{d}{\to} X$ to mean $\mu_n \overset{d}{\to} \mu$. This is the notion of convergence in distribution or weak convergence (that we have studied extensively when the metric space is Euclidean space). In particular, this applies to probability measures on $C[0,1]$.

For now, let us only make the observation that if $F : X \mapsto \mathbb{R}$ is a continuous function, then $\mu_n \overset{d}{\to} \mu$ implies that $\mu_n \circ F^{-1} \overset{d}{\to} \mu \circ F^{-1}$ (these are probability measures on $\mathbb{R}$). Thus, convergence in distribution of one sequence of measures on $C[0,1]$ encodes innumerable convergence in distribution statements on the real line (just by varying F). A convergence in distribution statement on $C[0,1]$ (or other such "large spaces") are often called a *functional limit theorem*.

With these definitions, we are ready to state one of the foundational theorems of probability theory. For one, it is a far-reaching generalization of the central limit theorem.

Theorem 57: Donsker's invariance principle

Let $X_1, X_2, \ldots$ be i.i.d. random variables with $\mathbf{E}[X_k] = 0$ and $\mathbf{E}[X_k^2] = 1$. Let $W_n = \mathcal{T}(\frac{X_1}{\sqrt{n}}, \ldots, \frac{X_n}{\sqrt{n}})$. Then $W_n \overset{d}{\to} W$, a standard Brownian motion on $[0,1]$.

How is this a generalization of the central limit theorem? Just consider the continuous function $F : C[0,1] \mapsto \mathbb{R}$ defined by $F(\varphi) = \varphi(1)$. As $F(W_n) = (X_1 + \ldots + X_n)/\sqrt{n}$ and $F(W) = W(1) \sim N(0,1)$, the standard central limit theorem follows. There are innumerable other functions one can use, and that gives us an amazing machinery to transfer results from random walks to Brownian motion or vice versa. Often from one particular random walk (e.g., simple symmetric random walk) to Brownian motion and hence to all other random walks with steps of zero mean and unit variance.

Why Donsker's theorem makes one break into a song: For example, let $F(\varphi) = \max_{0 \le t \le 1} \varphi(t)$. Then F is a continuous function from $C[0,1]$ to $\mathbb{R}$. From Donsker's theorem, it follows that $\max_{0\le t\le 1} W_n(t) \overset{d}{\to} \max_{0\le t\le 1} W(t)$. The left hand side is just the maximum of $\{\frac{1}{\sqrt{n}}S_0, \ldots, \frac{1}{\sqrt{n}}S_n\}$ and the right hand side is what we have been calling M_1. Since we worked out that $M_1 \overset{d}{=} |Z|$ where $Z \sim N(0,1)$, we now have limiting distribution of $\max\{S_0, \ldots, S_n\}/\sqrt{n}$. Observe that the special tricks that we used to compute the distribution of M_1, namely the reflection principle, is not available for general random walks, hence a direct proof of this statement about random walks may not be so easy. This shows how Brownian motion can be useful even for proving things about random walks!

To see the usefulness in the opposite direction, let us observe that the reflection principle is in fact available for the simple symmetric random walk (steps ±1 with equal probability), hence by some combinatorics (Feller's vol. 1, chapter 3 remains the best resource for this topic) one can actually show that $\frac{1}{\sqrt{n}} \max\{S_0, \ldots, S_n\} \overset{d}{\to} |Z|$. Now use Donsker's theorem to conclude that $M_1 \overset{d}{=} |Z|$. Furthermore, once you have it for Brownian motion, you have the result for random walk with any step distribution (with zero mean, unit variance). Thus, by proving it for one particular random walk, we can conclude it for all random walks, by passing through Brownian motion!

We shall use these ideas and revisit the three arcsine laws (we only proved two of them), and prove them for simple symmetric random walk (by combinatorics), for Brownian motion (by

Donsker's theorem) and hence for general random walks. We shall also prove the Khinchine-Hartman-Wintner law of iterated logarithm and a functional form of it, by first doing it for Brownian motion and then deducing it for random walks.

21. Proof of Donsker's invariance principle

The method of proof is even simpler than that of CLT, since we shall essentially get convergence in distribution by coupling random variables so that there is convergence in probability! Here is the precise statement.

Lemma 58

Let μ_n, μ be probability measures on a metric space (X, d). Assume that μ is *tight*, in the sense that given $\varepsilon > 0$, there exists a compact set $K_\varepsilon \subseteq X$ such that $\mu(K_\varepsilon) > 1 - \varepsilon$. Suppose we can construct X-valued random variables Y_n, Z_n on some probability space so that $Y_n \sim \mu_n$, $Z_n \sim \mu$, and $d(Y_n, Z_n) \overset{P}{\to} 0$. Then, $\mu_n \overset{d}{\to} \mu$.

PROOF. Let $f \in C_b(X)$. Then for any $\varepsilon > 0$, $\delta > 0$,

$$|\int_X f d\mu_n - \int_X f d\mu| \le \mathbf{E}[|f(Y_n) - f(Z_n)|]$$
$$\le \mathbf{E}[|f(Y_n) - f(Z_n)|\mathbf{1}_{Z_n \notin K_\varepsilon}] + \mathbf{E}[|f(Y_n) - f(Z_n)|\mathbf{1}_{d(Y_n,Z_n)\ge\delta}] + \mathbf{E}[|f(Y_n) - f(Z_n)|\mathbf{1}_{Z_n\in K_\varepsilon}\mathbf{1}_{d(Y_n,Z_n)<\delta}]$$
$$\le 2\|f\|_{\sup}\varepsilon + 2\|f\|_{\sup}\mathbf{P}\{d(Y_n, Z_n) \ge \delta\} + \mathbf{E}[|f(Y_n) - f(Z_n)|\mathbf{1}_{Z_n\in K_\varepsilon}\mathbf{1}_{d(Y_n,Z_n)<\delta}].$$

We observe that if δ is small enough, then $|f(y)-f(z)| < \varepsilon$ whenever $z \in K_\varepsilon$ and $d(y, z) < \delta$. If not, there would be $y_n \in X$, $z_n \in K_\varepsilon$ such that $d(y_n, z_n) \to 0$ and $|f(y_n) - f(z_n)| \ge \varepsilon$. By compactness we may assume $z_n \to z \in K_\varepsilon$, then $y_n \to z$ too, and by continuity $f(y_n) - f(z_n) \to 0$. Thus, the third term above may be bounded by ε.

Now let $n \to \infty$ and then $\varepsilon \to 0$ to see that $\int f d\mu_n \to \int f d\mu$. As this holds for any $f \in C_b(X)$, $\mu_n \overset{d}{\to} \mu$. ■

The argument above may be slightly extended to extend its applicability considerably.

Corollary 59

In the setting of the above lemma, the the conclusion $\int f d\mu_n \to \int f d\mu$ holds for any f that is continuous on a σ-compact set S_f such that $\mu(S_f) = 1$.

PROOF. Write $S = \cup_m L_m$ where $L_n$are compact and repeat the proof with K_ε replaced by $L_m \cap K_\varepsilon$. Then let $m \to \infty$ along with $n \to \infty$ and $\varepsilon \to 0$. ■

When μ is the Wiener measure, we know that it is tight (for example, take K_ε to be the set of all Hölder(1/4) functions with sufficiently large Hölder constant). Further, for any set $A \in \mathcal{B}_{C[0,1]}$, there are compact sets $L_m \subseteq A$ such that $\mu(L_m) \uparrow \mu(A)$.

Main step in the proof of Donsker's theorem. Let W be standard Brownian motion on $[0,\infty)$. By Skorokhod embedding theorem, there are stopping times $0 = \tau_0 \le \tau_1 \le \tau_2 \le \dots$ such that $(\tau_{i+1}-\tau_i, W(\tau_{i+1})-W(\tau_i))$, $i \ge 0$, are i.i.d., $W(\tau_{i+1})-W(\tau_i)$ are i.i.d. with the same distribution as X_1 and $\mathbf{E}[\tau_{i+1}-\tau_i] = 1$. In particular, $(W(\tau_0), W(\tau_1), \dots) \overset{d}{=} (S_0, S_1, \dots)$. Now define two $C[0,1]$-valued random variables (so $0 \le t \le 1$).

$$W_n(t) = \frac{1}{\sqrt{n}} W(nt), \qquad Y_n(t) = \begin{cases} \frac{W(\tau_k)}{\sqrt{n}} & \text{if } t = \frac{k}{n},\ 0 \le k \le n, \\ \text{linear in between.} \end{cases}$$

Then Y_n has the same distribution as the rescaled random walk and W_n is a standard Brownian motion. The key claim is that

$$\|W_n - Y_n\|_{\sup[0,1]} \xrightarrow{P} 0. \tag{20}$$

To show this, we observe that as Y is piecewise linear on each $[k/n, (k+1)/n]$,

$$\|W_n - Y_n\| = \left(\max_{0\le k\le n-1} \sup_{t\in[k/n,(k+1)/n]} |W_n(t) - Y_n(k/n)| \right) \vee \left(\max_{0\le k\le n-1} \sup_{t\in[k/n,(k+1)/n]} |W_n(t) - Y_n((k+1)/n)| \right).$$

For $\frac{k}{n} \le t \le \frac{k+1}{n}$, we have

$$W_n(t) - Y_n(k/n) = W_n(t) - W_n(\tau_k/n), \qquad W_n(t) - Y_n((k+1)/n) = W_n(t) - W_n(\tau_{k+1}/n)$$

hence if $\delta > \frac{1}{n}$, we have

$$\mathbf{P}\{\|W_n - Y_n\| > \varepsilon\} \le \mathbf{P}\{\omega_{W_n}(2\delta) > \varepsilon\} + \mathbf{P}\{\max_{0\le k\le n} |\frac{\tau_k}{n} - \frac{k}{n}| \ge \delta\}.$$

For fixed ε we can find δ so that the first probability is smaller than ε. As for the second, the probability goes to zero as $n \to \infty$ because $\max_{0\le k\le n} |\frac{\tau_k}{n} - \frac{k}{n} \overset{a.s.}{\to} 0$. This follows from the fact that if $\frac{x_n}{n} \to 1$, then $\max_{k\le n} |x_k - k|/n \to 0$. This completes the proof of (20).

Final touches to the proof of Donsker's theorem. From (20) and Lemma 58 it follows that Y_n converges in distribution to standard Brownian motion.

22. Lévy's arcsine laws

For $f \in C[0,1]$, we define

(1) $T(f) = \arg\max f$, the smallest t such that $f(t) = \max_{0\le s\le 1} f(s)$. We say smallest, to remove ambiguity in the definition.

(2) $L(f) = \max\{t \in [0,1] : f(t) = 0\}$ where the maximum is 1 if the set is empty.

(3) $A(f) = \text{Leb.}\{t \in [0,1] : f(t) \geq 0\}$.

Theorem 60: Lévy's arcsine laws

If W is standard Browninan motion on $\mathbb{R}$ run for unit time, $T(W)$, $L(W)$ and $A(W)$ have arcsine distribution having density $\frac{1}{\pi\sqrt{x(1-x)}}$.

The analogous asymptotic statements for random walks are as follows.

Theorem 61

Let $X_1, X_2, \ldots$ be i.i.d. random variables with zero mean and unit variance. Let $S_0 = 0$ and $S_n = X_1 + \ldots + X_n$. Define

(1) $T_n = \min\{0 \leq k \leq n : S_k = \max_{0 \leq j \leq n} S_j\}$,

(2) $L_n = \max\{0 \leq k \leq n : S_k S_{k+1} \leq 0\}$,

(3) $A_n = \max\{0 \leq k \leq n : S_k \geq 0\}$.

Then, $\frac{T_n}{n}$, $\frac{L_n}{n}$ and $\frac{A_n}{n}$ converge in distribution to the arcsine law.

One can make slight variations in the definitions of these random variables without changing the validity of the statement. For example, in the definition of L_n we can ask for strict inequality $S_k S_{k+1} < 0$ and similarly in A_n one can count strictly positive ones among $S_0, \ldots, S_n$. And in T_n one may take the last time S_k equals the global maximum.

First we deduce Theorem 61 from Theorem 60. This is a little less straightforward than what we discussed before (e.g., the maximum value), because none of T, W, A is a continuous function on $C_0[0,1]$. However, they are continuous a.e. on $C[0,1]$, with respect to Wiener measure.

Lemma 62

T, W, A are continuous $a.e.$ with respect to Wiener measure.

PROOF. To show this, define subsets of $C[0,1]$ that will be shown to have full Wiener measure and on which the corresponding functional will be shown to be continuous.

(1) $\mathcal{A}_1$: All f such that $L(f) < 1$ and for every $\varepsilon > 0$, there exist $s, t \in [L(f) - \varepsilon, L(f) + \varepsilon]$ such that $f(s) < 0 < f(t)$.

If f belongs to $\mathcal{A}_1$, fix $\varepsilon > 0$ and find $\delta > 0$ such that there exist $s, t \in [L(f) - \varepsilon, L(f) + \varepsilon]$ such that $f(s) < -\delta$ and $f(t) > \delta$ and such that $|f(u)| > \delta$ for all $u \geq L(f) + \varepsilon$. Then if $g \in C[0,1]$ and $\|g - f\| < \delta$, it is clear that g has no zeros in $[L(f) + \varepsilon, 1]$, while it does have a zero in $[L(f) - \varepsilon, L(f) + \varepsilon]$ (because $g(s) < 0 < g(t)$), hence $|L(g) - L(f)| \leq \varepsilon$.

Since $W(1) \neq 0$ and $W(0) = 0$ with probability 1, it follows that $L(W) < 1$ with probability 1. Since $L(f)$ is an accumulation point of the zero set of W (with probability 1), it follows that W has strict sign changes in $[L(W) - \varepsilon, L(W)]$ for any $\varepsilon > 0$, while it has no sign changes in $[L(W), 1]$. This implies that $W \in \mathcal{A}_1$ with probability 1.

(2) $\mathcal{A}_2$: All f for which $\{t : f(t) = \max_{0 \leq s \leq 1} f(s)\}$ is a singleton.

If $f \in \mathcal{A}_2$, then given $\varepsilon > 0$, there is a $\delta > 0$ such that the maximum of f outside $[T(f) - \delta, T(f) + \delta]$ is less than the global maximum by at least δ. Hence if $g \in C[0, 1]$ and $\|g - f\| < \delta$, then $T(g) \in [T(f) - \delta, T(f) + \delta]$.

Fix any $T \in [0, 1]$ and observe that $\max_{[0,T]} W - W(T)$ and $\max_{[T,1]} W - W(T)$ are independent random variables having the same distributions as $\sqrt{T}|Z_1|$ and $\sqrt{1-T}|Z_2|$, where Z_i are standard Gaussians. From this, it is clear that there is no chance that these two random variables are equal. But that is the same as saying that $\mathbf{P}\{\max_{[0,T]} W = \max_{[T,1]} W\} = 0$. As this is true for each fixed T, it is true for the union over all $T \in \mathbb{Q} \cap [0, 1]$. But if W has two distinct global maxima, then $\max_{[0,T]} W = \max_{[T,1]} W$ for any rational T for some rational T (any T between the two global maxima). Hence, $\mathbf{P}\{W \in \mathcal{A}_2\} = 1$.

(3) $\mathcal{A}_3$: All f such that $\{t : f(t) = 0\}$ has zero Lebesgue measure.

If $f \in \mathcal{A}_3$, then given $\varepsilon > 0$, there exists $\delta > 0$ such that the Lebesgue measure of $\{f \geq \delta\}$ and $\{f \leq -\delta\}$ are within ε of the Lebesgue measures of $\{f \geq 0\}$ and $\{f \leq 0\}$ respectively. If $g \in C[0, 1]$ and $\|f - g\| < \delta$, then $g > 0$ on $\{f \geq \delta\}$ and $g < 0$ on $\{f \leq -\delta\}$, from which it easily follows that $|\mathcal{A}(g) - \mathcal{A}(f)| \leq \varepsilon$.

We have already shown that the zero set of W has zero Lebesgue measure, hence $\mathbf{P}\{W \in \mathcal{A}_3\} = 1$.

All claims in the theorem are proved. ■

We need a slight extension of our earlier idea to functionals that are only continuous almost everywhere.

Lemma 63

Let μ_n, μ be Borel probability measures on a separable metric space (X, d) such that $\mu_n \overset{d}{\to} \mu$. If $F : X \mapsto \mathbb{R}$ is continuous $a.e.$ with respect to μ, then $\mu_n \circ F^{-1} \overset{d}{\to} \mu \circ F^{-1}$.

To prove this lemma, the easiest way is to use an idea of Skorokhod (called *Skorokhod's representation theorem*).

Lemma 64: Skorokhod's representation theorem

Let μ_n, μ be Borel probability measures on a metric space (X, d).

(1) Suppose $Y_n \sim \mu_n$ and $Y \sim \mu$ are random variables on a common probability space such that $Y_n \overset{a.s.}{\to} Y$, then $\mu_n \overset{d}{\to} \mu$.

(2) Assume $\mu_n \overset{d}{\to} \mu$. If X is separable, then there are random variables $Y_n \sim \mu_n$ and $Y \sim \mu$ are random variables on a common probability space such that $Y_n \overset{a.s.}{\to} Y$.

Assuming this, Lemma 63 is obvious. Get $Y_n \sim \mu_n$ and $Y \sim \mu$ such that $Y_n \overset{a.s.}{\to} Y$ and observe that $F(Y_n) \overset{a.s.}{\to} F(Y)$ (since Y falls inside the set of continuity of F, with probability 1) and hence $F(Y_n) \overset{d}{\to} F(Y)$. Thus it only remains to prove Skorokhod's representation theorem.

The first part is easy. Indeed, if $f \in C_b(X)$, then $f(Y_n) \overset{a.s.}{\to} f(Y)$ and by DCT it follows that $\mathbf{E}[f(Y_n)] \to \mathbf{E}[f(Y)]$, which is what it means to have $Y_n \overset{d}{\to} Y$. The other direction requires one to develop some machinery of weak convergence (e.g., that $\mu_n \overset{d}{\to} \mu$ if and only if $\mu_n(A) \to \mu(A)$ for all $A \in \mathcal{B}_X$ such that $\mu(\partial A) = 0$). You may see the proof in Dudley's book (Theorem 11.7.2). We shall skip it, but for our purposes there will be no gap owing to this, because one way we shall prove Donsker's theorem is by constructing W_n and W such that $W_n \overset{a.s.}{\to} W$. This means that Skorokhod representation will be proved for the special case of interest to us.

Proof of Theorem 61 from Theorem 60. Immediate from the two Lemmas above. ■

Now we turn to the proof of Theorem 60. We have already proved the claim for $L(W)$ and $T(W)$. Now we prove the statement for $A(W)$, by first proving an analogous statement for shall take a different approach, proving the analogous statements for simple symmetric random walk, using Donsker's theorem to deduce it for Brownian motion, and then again using Donsker's principle to deduce it for general random walks. The same idea can be carried out for L and T, but we leave that as exercises.

Lemma 65

Let $0 = S_0, S_1, \ldots$ be a simple symmetric random walk on $\mathbb{Z}$. Then

CHAPTER 6

Appendix: Miscellaneous background material

1. Gaussian random variables

Standard normal: A standard normal or Gaussian random variable is one with density $\varphi(x) := \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$. Its distribution function is $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$ and its tail distribution function is denoted $\bar{\Phi}(x) := 1 - \Phi(x)$. If X_i are i.i.d. standard normals, then $X = (X_1, \ldots, X_n)$ is called a standard normal vector in $\mathbb{R}^n$. It has density $\prod_{i=1}^{n} \varphi(x_i) = (2\pi)^{-n/2} \exp\{-|\mathbf{x}|^2/2\}$ and the distribution is denoted by γ_n, so that for every Borel set A in $\mathbb{R}^n$ we have $\gamma_n(A) = (2\pi)^{-n/2} \int\limits_A \exp\{-|\mathbf{x}|^2/2\}d\mathbf{x}$.

Exercise 28

[Rotation invariance] If $P_{n\times n}$ is an orthogonal matrix, then $\gamma_n P^{-1} = \gamma_n$ or equivalently, $PX \overset{d}{=} X$. Conversely, if a random vector with independent co-ordinates has a distribution invariant under orthogonal transformations, then it has the same distribution as cX for some (non-random) scalar c.

Multivariate normal: If $Y_{m\times 1} = \mu_{m\times 1} + B_{m\times n}X_{n\times 1}$ where $X_1, \ldots, X_n$ are i.i.d. standard normal, then we say that $Y \sim N_m(\mu, \Sigma)$ with $\Sigma = BB^t$. Implicit in this notation is the fact that the distribution of Y depends only on Σ and not on the way in which Y is expressed as a linear combination of standard normals (this follows from Exercise 36). It is a simple exercise that $\mu_i = \mathbf{E}[X_i]$ and $\sigma_{i,j} = \text{Cov}(X_i, X_j)$. Henceforth, for simplicity, we take the mean to be zero everywhere.

Since matrices of the form BB^t are precisely positive semi-definite matrices (defined as those $\Sigma_{m\times m}$ for which $\mathbf{v}^t\Sigma\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^m$), it is clear that covariance matrices of normal random vectors are precisely p.s.d. matrices. Clearly, if $Y \sim N_m(\mu, \Sigma)$ and $Z_{p\times 1} = C_{p\times m}Y + \theta_{p\times 1}$, then $Z \sim N_p(\theta + C\mu, C\Sigma C^t)$. Thus, affine linear transformations of normal random vectors are again normal.

Exercise 29

The random vector Y has density if and only if Σ is non-singular, and in that case the density is

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\exp\left\{-\frac{1}{2}\mathbf{y}^t\Sigma^{-1}\mathbf{y}\right\}.$$

If Σ is singular, then X takes values in a lower dimensional subspace in $\mathbb{R}^n$ and hence does not have density.

In particular, if $\mathbf{v} \in \mathbb{R}^m$, then $\mathbf{v}^tY$ is univariate normal with mean $\mathbf{v}^t\mu$ and variance $\mathbf{v}^t\Sigma\mathbf{v}$. The covariance of two different linear combinations $\mathbf{v}^tY$ and $\mathbf{u}^tY$ is $\mathbf{v}^t\Sigma\mathbf{u}$. The converse is also true. If $\mathbf{v}^tY$ is univariate Gaussian for every $\mathbf{v} \in \mathbb{R}^m$, then it is necessarily the case that Y is multi-variate Gaussian. You may prove this using characteristic functions, for example. The characteristic function of Gaussian distribution is given in the exercise below.

Exercise 30

Irrespective of whether Σ is non-singular or not, the characteristic function of Y is given by

$$\mathbf{E}\left[e^{i\langle\lambda,Y\rangle}\right] = e^{-\frac{1}{2}\lambda^t\Sigma\lambda}, \quad \text{for } \lambda \in \mathbb{R}^m.$$

In particular, if $X \sim N(0,\sigma^2)$, then its characteristic function is $\mathbf{E}[e^{i\lambda X}] = e^{-\frac{1}{2}\sigma^2\lambda^2}$ for $\lambda \in \mathbb{R}$.

Exercise 31

If $U_{k\times 1}$ and $V_{(m-k)\times 1}$ are such that $Y^t = (U^t, V^t)$, and we write $\mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ are partitioned accordingly, then

(1) $U \sim N_k(\mu_1, \Sigma_{11})$.

(2) $U\Big|_V \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1/2}V,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ (assume that Σ_{22} is invertible).

2. More about the univariate normal distribution

Tail of the standard Gaussian distribution: Recall the standard Gaussian density $\varphi(x)$. The corresponding cumulative distribution function is denoted by Φ and the tail is denoted by $\bar{\Phi}(x) := \int_x^\infty \varphi(t)dt$. The following estimates will be used very often.

Exercise 32

For all $x > 0$, we have

$$\frac{1}{\sqrt{2\pi}}\frac{x}{1+x^2}e^{-\frac{1}{2}x^2} \leq \bar{\Phi}(x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-\frac{1}{2}x^2}. \tag{21}$$

In particular[a], $\bar{\Phi}(x) \sim x^{-1}\varphi(x)$ as $x \to \infty$. Most often the following simpler bound, valid for $x \geq 1$, suffices.

$$\frac{1}{10x}e^{-\frac{1}{2}x^2} \leq \bar{\Phi}(x) \leq e^{-\frac{1}{2}x^2}. \tag{22}$$

[a] The notation $f(x) \sim g(x)$ means that $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

Maximum of independent standard Gaussians: Let $X_1, \ldots, X_n$ be (not necessarily independent) random variables with each having $N(0,1)$ distribution. Let $M_n = \max\{X_1, \ldots, X_n\}$. How big is M_n? In general, the maximum of correlated Gaussians is a very important question of great current interest. The i.i.d. case is a very special and easy case where we can extract the right answer easily.

Observe that $M_n \geq t$ if and only if $X_i \geq t$ for some $i \leq n$. Therefore,

$$\mathbf{P}\{M_n \geq t\} \leq \sum_{k=1}^{n} \mathbf{P}\{X_k \geq t\} = n\bar{\Phi}(t).$$

Using the upper bound in (22), and setting $t = \sqrt{2A\log n}$ with $A > 1$, we get (since $t \geq 1$ for $n \geq 2$),

$$\mathbf{P}\{M_n \geq \sqrt{2A\log n}\} \leq ne^{-A\log n} = \frac{1}{n^{A-1}}. \tag{23}$$

We shall use this quantitative bound many times in the lectures. In particular, for every $\delta > 0$, the above inequality implies that $\mathbf{P}\left\{\frac{1}{\sqrt{2\log n}}M_n \geq 1+\delta\right\} \to 0 \qquad$ as $n \to \infty$. This bound is actually tight if the random variables are independent.

Exercise 33: U

e the lower bound for the tail of the Normal distribution from (22), show that $\mathbf{P}\left\{\frac{1}{\sqrt{2\log n}}M_n \leq 1-\delta\right\} \to 0$ for any $\delta > 0$. Conclude that in this case $\frac{1}{\sqrt{2\log n}}M_n \overset{P}{\to} 1$.

Convergence and Gaussians: Distributional limits of Gaussians are Gaussians. In other words, if $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$, then $N(\mu_n, \sigma_n^2) \overset{d}{\to} N(\mu, \sigma^2)$. Conversely, if $N(\mu_n, \sigma_n^2) \overset{d}{\to} \nu$ for some probability measure ν, then $\nu = N(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$. If this is not clear, take it as an exercise!

Gaussian density and heat equation: For $t > 0$, let $p_t(x) := \frac{1}{\sqrt{t}}\varphi(x/\sqrt{t})$ be the $N(0,t)$ density. We interpret $p_0(x)dx$ as the degenerate measure at 0. These densities have the following interesting properties.

Exercise 34: S

ow that $p_t \star p_s = p_{t+s}$, i.e., $\int\limits_{\mathbb{R}} p_t(x-y)p_s(y)dy = p_{t+s}(x)$.

Exercise 35: S

ow that $p_t(x)$ satisfies the heat equation: $\frac{\partial}{\partial t}p_t(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p_t(x)$ for all $t > 0$ and $x \in \mathbb{R}$.

Remark 9: P

t together, these facts say that $p_t(x)$ is the *fundamental solution* to the heat equation. This just means that the heat equation $\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x)$ with the initial condition $u(0,x) = f(x)$ can be solved simply as $u(t,x) = (f \star p_t)(x) := \int_{\mathbb{R}} f(y)p_t(x-y)dy$. This works for reasonable f (say $f \in L^1(\mathbb{R})$).

3. Existence of countably many Gaussians with given covariances

Let $\Sigma = (\sigma_{i,j})_{i,j\geq 1}$ be a semi-infinite matrix. Do there exist random variables $X_1, X_2, \ldots$ that are jointly Gaussian (by which we mean that any finite sub-collection of them has joint Gaussian distribution) and such that $\mathbf{E}[X_i] = 0$ and $\mathbf{E}[X_iX_j] = \sigma_{i,j}$ for all $i,j \geq 1$?

A necessary condition is that Σ is (symmetric and) positive semi-definite. This means that $\sigma_{i,j} = \sigma_{j,i}$ for all i,j and $\sum_{i,j=1}^n u_iu_j\sigma_{i,j} \geq 0$ for all $n \geq 1$ and all $\mathbf{u} \in \mathbb{R}^n$. Symmetry is clearly necessary. As for the second condition, observe that

$$\sum_{i,j=1}^n u_iu_j\sigma_{i,j} = \mathbf{E}\left[\left(\sum_{i=1}^n u_iX_i\right)^2\right]$$

by expanding the square and interchanging expectation with the sum. From this, the p.s.d property is clear. Note that we did not require Gaussian property here - covariance matrix of any collection of random variables is p.s.d.

Claim 66

Let $\Sigma = (\sigma_{i,j})_{i,j\geq 1}$ be a symmetric p.s.d. matrix. Then, there exist random variables (on some probability space) X_i, $i \geq 1$, that are jointly Gaussian, have zero means and covariance matrix Σ.

Proof. Let ξ_n, $n \geq 1$, be i.i.d. $N(0,1)$ random variables (on your favourite probability space, for example, $([0,1], \mathcal{B}, \lambda)$). We shall define $X_n = a_{n,1}\xi_1 + \ldots + a_{n,n}\xi_n$, where the coefficients $a_{n,j}$, $1 \leq j \leq n$, will be chosen so as to satisfy the covariance conditions. That X_n, $n \geq 1$, have a joint Gaussian distribution is clear.

First, we define $a_{1,1} = \sqrt{\sigma_{1,1}}$ so that $X_1 \sim N(0,1)$. This definition is valid since p.s.d. property implies that $\sigma_{1,1} \geq 0$.

Next, from $\mathbf{E}[X_1X_2] = \sigma_{1,2}$ we get the equation $a_{1,2}\sqrt{\sigma_{1,1}} = \sigma_{1,2}$ and $a_{2,2}^2 + a_{2,1}^2 = \sigma_{2,2}$. As the 2×2 matrix $(\sigma_{i,j})_{i,j\leq 2}$ is p.s.d., we certainly have $\sigma_{1,1} \geq 0$ and $\sigma_{2,2}\sigma_{1,1} - \sigma_{1,2}^2 \geq 0$. If $\sigma_{1,1} > 0$, then the unique solutions are

$$a_{2,1} = \frac{\sigma_{1,2}}{\sqrt{\sigma_{1,1}}}, \quad a_{2,2} = \sqrt{\sigma_{2,2} - \frac{\sigma_{1,2}^2}{\sigma_{1,1}}}.$$

What if $\sigma_{1,1} = 0$. Then use p.s.d property to show that $\sigma_{1,i} = 0$ for all i (in general, if a diagonal entry vanishes, the entire row and column containing it must also vanish). But then the first equation is vacuous and we may set $a_{1,2} = 0$ (or anything else, it does not matter since X_1 is the zero random variable!) and $a_{2,2} = \sqrt{\sigma_{2,2}}$.

Now suppose we have solved for $a_{k,j}$, $1 \leq j \leq k \leq n-1$. We want to solve for $a_{n,j}$, $j \leq n$. Let us use matrix notation and write $B = (a_{k,j})_{j,k\leq n-1}$ (with $a_{k,j} = 0$ if $j > k$). Let $\mathbf{u}^t = (a_{n,1}, \ldots, a_{n,n-1})$ and let $\mathbf{v}^t = (\sigma_{n,1}, \ldots, \sigma_{n,n-1})$. Then, the equations that we must solve are $B\mathbf{u} = \mathbf{v}$ and $a_{n,n}^2 + \|u\|^2 = \sigma_{n,n}$. If $a_{k,k} > 0$ for $k \leq n-1$, then B is non-singular and we get the unique solutions $\mathbf{u} = B^{-1}\mathbf{v}$ and $a_{n,n} = \sqrt{\sigma_{n,n} - \|\mathbf{u}\|^2}$. The last square root makes sense because of the matrix theory fact that

$$\det \begin{bmatrix} X & \mathbf{v} \\ \mathbf{v}^t & c \end{bmatrix} = \det(X).(c - \mathbf{v}^t X^{-1}\mathbf{v})$$

whenever X is a non-singular matrix. Here we apply it with $X = (\sigma_{i,j})_{i,j\leq n-1}$, $\mathbf{v}$ as before and $c = \sigma_{n,n}$. Positive definiteness implies that both determinants are positive. Hence $c - \mathbf{v}^t X^{-1}\mathbf{v} > 0$ (in our case this is precisely $\sigma_{n,n} - \|u\|^2$.

All this is fine if Σ is strictly positive definite, for then $\det(\sigma_{i,j})_{i,j\leq n} > 0$ for every n. Hence, inductively, we see that $a_{n,n} > 0$ for all n and the above procedure continues without any difficulty. If $a_{n,n} = 0$ for some n, then we need to modify the procedure.

[Will write this, too tired now...] ■

4. Gaussian random variables

Standard normal: A standard normal or Gaussian random variable is one with density $\varphi(x) := \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$. Its distribution function is $\Phi(x) = \int_{-\infty}^{x} \varphi(t)dt$ and its tail distribution function is denoted $\bar{\Phi}(x) := 1 - \Phi(x)$. If X_i are i.i.d. standard normals, then $X = (X_1, \ldots, X_n)$ is called a standard normal vector in $\mathbb{R}^n$. It has density $\prod_{i=1}^n \varphi(x_i) = (2\pi)^{-n/2}\exp\{-|\mathbf{x}|^2/2\}$ and the distribution is denoted by γ_n, so that for every Borel set A in $\mathbb{R}^n$ we have $\gamma_n(A) = (2\pi)^{-n/2}\int\limits_A \exp\{-|\mathbf{x}|^2/2\}d\mathbf{x}$.

Exercise 36

[Rotation invariance] If $P_{n\times n}$ is an orthogonal matrix, then $\gamma_n P^{-1} = \gamma_n$ or equivalently, $PX \overset{d}{=} X$. Conversely, if a random vector with independent co-ordinates has a distribution invariant under orthogonal transformations, then it has the same distribution as cX for some (non-random) scalar c.

Multivariate normal: If $Y_{m\times 1} = \mu_{m\times 1} + B_{m\times n}X_{n\times 1}$ where $X_1, \ldots, X_n$ are i.i.d. standard normal, then we say that $Y \sim N_m(\mu, \Sigma)$ with $\Sigma = BB^t$. Implicit in this notation is the fact that the distribution of Y depends only on Σ and not on the way in which Y is expressed as a linear combination of standard normals (this follows from Exercise 36). It is a simple exercise that $\mu_i = \mathbf{E}[X_i]$ and $\sigma_{i,j} = \text{Cov}(X_i, X_j)$. Since matrices of the form BB^t are precisely positive semi-definite matrices (defined as those $\Sigma_{m\times m}$ for which $\mathbf{v}^t\Sigma\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^m$), it is clear that covariance matrices of normal random vectors are precisely p.s.d. matrices. Clearly, if $Y \sim N_m(\mu, \Sigma)$ and $Z_{p\times 1} = C_{p\times m}Y + \theta_{p\times 1}$, then $Z \sim N_p(\theta + C\mu, C\Sigma C^t)$. Thus, affine linear transformations of normal random vectors are again normal.

Exercise 37

The random vector Y has density if and only if Σ is non-singular, and in that case the density is

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}\mathbf{y}^t\Sigma^{-1}\mathbf{y}\right\}.$$

If Σ is singular, then X takes values in a lower dimensional subspace in $\mathbb{R}^n$ and hence does not have density.

Exercise 38

Irrespective of whether Σ is non-singular or not, the characteristic function of Y is given by

$$\mathbf{E}\left[e^{i\langle\lambda, Y\rangle}\right] = e^{-\frac{1}{2}\lambda^t\Sigma\lambda}, \quad \text{for } \lambda \in \mathbb{R}^m.$$

In particular, if $X \sim N(0, \sigma^2)$, then its characteristic function is $\mathbf{E}[e^{i\lambda X}] = e^{-\frac{1}{2}\sigma^2\lambda^2}$ for $\lambda \in \mathbb{R}$.

Exercise 39: I

$U_{k\times 1}$ and $V_{(m-k)\times 1}$ are such that $Y^t = (U^t, V^t)$, and we write $\mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ are partitioned accordingly, then

(1) $U \sim N_k(\mu_1, \Sigma_{11})$.

(2) $U\Big|_V \sim N_k(\mu_1 - \Sigma_{12}\Sigma_{22}^{-1/2}V,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ (assume that Σ_{22} is invertible).

Moments: All questions about a centered Gaussian random vector must be answerable in terms of the covariance matrix. In some cases, there are explicit answers.

Exercise 40

Prove the *Wick formula* (also called *Feynman diagram formula*) for moments of centered Gaussians.

(1) Let $X \sim N_n(0, \Sigma)$. Then, $\mathbf{E}[X_1 \ldots X_n] = \sum\limits_{M \in \mathcal{M}_n} \prod\limits_{\{i,j\} \in M} \sigma_{i,j}$, where $\mathcal{M}_n$ is the collection of all matchings of the set $[n]$ (thus $\mathcal{M}_n$ is empty if n is odd) and the product is over all matched pairs. For example, $\mathbf{E}[X_1X_2X_3X_4] = \sigma_{12}\sigma_{34} + \sigma_{13}\sigma_{24} + \sigma_{14}\sigma_{23}$.

(2) If $\xi \sim N(0,1)$, then $\mathbf{E}[\xi^{2n}] = (2n-1)(2n-3)\ldots(3)(1)$.

Cumulants: Let X be a real-valued random variable with $\mathbf{E}[e^{tX}] < \infty$ for t in a neighbourhood of 0. Then, we can write the power series expansions

$$\mathbf{E}[e^{i\lambda X}] = \sum_{k=0}^{\infty} m_n(X)\frac{\lambda^n}{n!}, \qquad \log \mathbf{E}[e^{i\lambda X}] = \sum_{k=1}^{\infty} \kappa_n[X]\frac{\lambda^n}{n!}.$$

Here $m_n[X] = \mathbf{E}[X^n]$ are the moments while $\kappa_n[X]$ is a linear combination of the first n moments ($\kappa_1 = m_1$, $\kappa_2 = m_2 - m_1^2$, etc). Then κ_n is called the nth cumulant of X. If X and Y are independent, then it is clear that $\kappa_n[X+Y] = \kappa_n[X] + \kappa_n[Y]$.

Exercise 41: (optional)

Prove the following relationship between moments and cumulants. The sums below are over partitions Π of the set $[n]$ and $\Pi_1, \ldots, \Pi_{\ell_\Pi}$ denote the blocks of Π.

$$m_n[X] = \sum_{\Pi}\prod_{i} \kappa_{|\Pi_i|}[X], \qquad \kappa_n[X] = \sum_{\Pi}(-1)^{\ell_\Pi - 1}\prod_{i} m_{|\Pi_i|}[X].$$

Thus $\kappa_1 = m_1$, $\kappa_2 = m_2 - m_1^2$,

Exercise 42

If $\xi \sim N(0,1)$, then $\kappa_1 = 0$, $\kappa_2 = 1$ and $\kappa_n = 0$ for all $n \geq 3$.

The converse of this result is also true and often useful in proving that a random variable is normal. For instance, the theorem below implies that to show that a sequence of random variables converges to normal, it suffices to show that cumulants $\kappa_m[X_n] \to 0$ for all $m \geq m_0$ for some m_0.

Result 67: Marcinkiewicz

If X is a random variable with finite moments of all orders and $\kappa_n[X] = 0$ for all $n \geq n_0$ for some n_0, then X is Gaussian.

Convergence and Gaussians:

Exercise 43

The family of distributions $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $0 \leq \sigma^2 < \infty$, is closed under convergence in distribution (for this statement to be valid we include $N(\mu, 0)$ which means δ_μ). Indeed, $N(\mu_n, \sigma_n^2) \overset{d}{\to} N(\mu, \sigma^2)$ if and only if $\mu_n \to \mu$ and $\sigma_n^2 \to \sigma^2$.

A vector space of Gaussian random variables: Let $Y \sim N_m(0, \Sigma)$ be a random vector in some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then, for every vector $\mathbf{v} \in \mathbb{R}^m$, define the random variable $Y_{\mathbf{v}} := \mathbf{v}^t Y$. Then, for any $\mathbf{v}_1, \ldots, \mathbf{v}_j$, the random variables $Y_{\mathbf{v}_1}, \ldots, Y_{\mathbf{v}_j}$ are jointly normal. The joint distribution of $\{Y_{\mathbf{v}}\}$ is fully specified by noting that $Y_{\mathbf{v}}$ have zero mean and $\mathbf{E}[Y_{\mathbf{v}} Y_{\mathbf{u}}] = \mathbf{v}^t \Sigma \mathbf{u}$.

We may interpret this as follows. If Σ is p.d. (p.s.d. and non-singular), then $(\mathbf{v}, \mathbf{u})_\Sigma := \mathbf{v}^t \Sigma \mathbf{u}$ defines an inner product on $\mathbb{R}^m$. On the other hand, the set $L_0^2(\Omega, \mathcal{F}, \mathbf{P})$ of real-valued random variables on Ω with zero mean and finite variance, is also an inner product space under the inner product $\langle U, V \rangle := \mathbf{E}[UV]$. The observation in the previous paragraph is that $\mathbf{v} \to Y_{\mathbf{v}}$ is an isomorphism of $(\mathbb{R}^m, (\cdot, \cdot)_\Sigma)$ into $L_0^2(\Omega, \mathcal{F}, \mathbf{P})$.

In other words, given any finite dimensional inner-product space $(V, \langle \cdot, \cdot \rangle)$, we can find a collection of Gaussian random variables on some probability space, such that this collection is isomorphic to the given inner-product space. Later we shall see the same for Hilbert spaces[1].

5. The Gaussian density

Recall the standard Gaussian density $\varphi(x)$. The corresponding cumulative distribution function is denoted by Φ and the tail is denoted by $\bar{\Phi}(x) := \int_x^\infty \varphi(t)dt$. The following estimate will be used very often.

[1] This may seem fairly pointless, but here is one thought-provoking question. Given a vector space of Gaussian random variables, we can multiply any two of them and thus get a larger vector space spanned by the given normal random variables and all pair-wise products of them. What does this new vector space correspond to in terms of the original $(V, \langle \cdot, \cdot \rangle)$?

Exercise 44

For all $x > 0$, we have $\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} e^{-\frac{1}{2}x^2} \le \bar{\Phi}(x) \le \frac{1}{\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2}x^2}$ In particular[a], $\bar{\Phi}(x) \sim x^{-1}\varphi(x)$ as $x \to \infty$. Most often the following simpler bound, valid for $x \ge 1$, suffices.

$$\frac{1}{10x} e^{-\frac{1}{2}x^2} \le \bar{\Phi}(x) \le e^{-\frac{1}{2}x^2}.$$

[a] The notation $f(x) \sim g(x)$ means that $\lim\limits_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

For $t > 0$, let $p_t(x) := \frac{1}{\sqrt{t}}\varphi(x/\sqrt{t})$ be the $N(0,t)$ density. We interpret $p_0(x)dx$ as the degenerate measure at 0. These densities have the following interesting properties.

Exercise 45

Show that $p_t \star p_s = p_{t+s}$, i.e., $\int\limits_{\mathbb{R}} p_t(x-y)p_s(y)dy = p_{t+s}(x)$.

Exercise 46

Show that $p_t(x)$ satisfies the heat equation: $\frac{\partial}{\partial t}p_t(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p_t(x)$ for all $t > 0$ and $x \in \mathbb{R}$.

Remark 10

Put together, these facts say that $p_t(x)$ is the *fundamental solution* to the heat equation. This just means that the heat equation $\frac{\partial}{\partial t}u(t,x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(t,x)$ with the initial condition $u(0,x) = f(x)$ can be solved simply as $u(t,x) = (f \star p_t)(x) := \int_{\mathbb{R}} f(y)p_t(x-y)dy$. This works for reasonable f (say $f \in L^1(\mathbb{R})$).

We shall have many occasions to use the following "integration by parts" formula.

Exercise 47

Let $X \sim N_n(0,\Sigma)$ and let $F : \mathbb{R}^n \to \mathbb{R}$. Under suitable conditions on F (state sufficient conditions), show that $\mathbf{E}[X_i F(X)] = \sum_{j=1}^n \sigma_{ij}\mathbf{E}[\partial_j F(X)]$. As a corollary, deduce the Wick formula of Exercise 40.

Stein's equation: Here we may revert to $t = 1$, thus $p_1 = \varphi$. Then, $\varphi'(x) = -x\varphi(x)$. Hence, for any $f \in C_b^1(\mathbb{R})$, we integrate by parts to get $\int f'(x)\varphi(x)dx = -\int f(x)\varphi'(x)dx = \int f(x)x\varphi(x)dx$. If $X \sim N(0,1)$, then we may write this as

$$\mathbf{E}[(Tf)(X)] = 0 \quad \text{for all } f \in C_b^1(\mathbb{R}), \text{ where } (Tf)(x) = f'(x) - xf(x). \tag{24}$$

The converse is also true. Suppose (24) holds for all $f \in C_b^1(\mathbb{R})$. Apply it to $f(x) = e^{i\lambda x}$ for any fixed $\lambda \in \mathbb{R}$ to get $\mathbf{E}[Xe^{i\lambda X}] = i\lambda\mathbf{E}[e^{i\lambda X}]$. Thus, if $\psi(\lambda) := \mathbf{E}[e^{i\lambda X}]$ is the characteristic function of X, then $\psi'(\lambda) = -\lambda\psi(\lambda)$ which has only one solution, $e^{-\lambda^2/2}$. Hence X must have standard normal distribution.

Digression - central limit theorem: One reason for the importance of normal distribution is of course the central limit theorem. The basic central limit theorem is for $W_n := (X_1 + \ldots + X_n)/\sqrt{n}$ where X_i are i.i.d. with zero mean and unit variance. Here is a sketch of how central limit theorem can be proved using Stein's method. Let $f \in C_b^1(\mathbb{R})$ and observe that $\mathbf{E}[W_n f(W_n)] = \sqrt{n}\mathbf{E}[X_1 f(W_n)]$. Next, write

$$f\left(\frac{X_1 + \ldots + X_n}{\sqrt{n}}\right) \approx f\left(\frac{X_2 + \ldots + X_n}{\sqrt{n}}\right) + \frac{X_1}{\sqrt{n}} f'\left(\frac{X_2 + \ldots + X_n}{\sqrt{n}}\right)$$

where we do not make precise the meaning of the approximation. Let $\hat{W}_n = \frac{X_2 + \ldots + X_n}{\sqrt{n}}$. Then,

$$\mathbf{E}[W_n f(W_n)] \approx \sqrt{n}\mathbf{E}[X_1]\mathbf{E}[f(\hat{W}_n)] + \mathbf{E}[X_1^2]\mathbf{E}[f'(\hat{W}_n)] = \mathbf{E}[f'(\hat{W}_n)].$$

Since $\hat{W}_n \approx W_n$, this shows that $\mathbf{E}[Tf(W_n)] \approx 0$. We conclude that $W_n \approx N(0,1)$.

There are missing pieces here, most important being the last statement - that if a random variable satisfies Stein's equation approximately, then it must be approximately normal. When included, one does get a proof of the standard CLT.