PROBABILITY THEORY 2 LECTURE NOTES

These lecture notes were written for MATH 6720 at Cornell University in the Spring semester of 2014. They were last revised in the Spring of 2016 and the schedule on the following page reflects that semester. These notes are for personal educational use only. Much of the material on martingales, Markov chains, and ergodic theory comes directly from the course text, *Probability: Theory and Examples* by Rick Durrett. Some of the discussion of Markov chains is also drawn from the book *Markov Chains and Mixing Times* by David Levin, Yuval Peres, and Elizabeth Wilmer. The sections on Brownian motion are based mainly on *Brownian Motion* by Peter Mörters and Yuval Peres. There are likely typos and mistakes in these notes. All such errors are the author's fault and corrections are greatly appreciated.

Day 1: Through Example 1.2
Day 2: Through Proposition 1.5
Day 3: Through definition of r.c.d.s
Day 4: Through Example 2.4
Day 5: Through Theorem 2.5
Day 6: Through Lemma 2.1
Day 7: Finished Section 2
Day 8: Through Theorem 3.3 (Skipped Polya's Urn)
Day 9: Through Theorem 4.3
Day 10: Through first observations about uniform integrability
Day 11: Through Theorem 4.9
Day 12: Through Theorem 5.2
Day 13: Finished Section 5
Day 14: Up to Example 6.1
Day 15: Through Example 6.6
Day 16: Finished Section 7
Day 17: Through Theorem 8.3
Day 18: Through Example 8.2
Day 19: Through Example 9.2
Day 20: Through Theorem 9.1
Day 21: Through Theorem 9.5
Day 22: Through Theorem 10.2
Day 23: Up to Theorem 10.3
Day 24: Student Presentations
Day 25: Through cut-off phenomenon
Day 26: Through Example 11.1
Day 27: Finished Section 11

(Skipped Sections 12 and 13)

Day 28: Through Theorem 14.1
Day 29: Through $B(d) = \sum_{i=0}^{\infty} F_i(d)$ for $d \in \mathcal{D}$.
Day 30: Through Theorem 15.1
Day 31: Finished Section 15
Day 32: Through Theorem 16.7
Day 33: Through Theorem 16.9
Day 34: Started Theorem 16.13
Day 35: Through Theorem 17.1
Day 36: Through Proposition 17.1 (and some discussion of remaining material)

Days 37-42: Student Presentations

1. Conditional Expectation

Let $(\Omega, \mathcal{F}, P)$ be a probability space and suppose that $A, B \in \mathcal{F}$ with $P(B) > 0$. In undergraduate probability, we learn that the probability of A conditional on B is defined as $P(A\,|B) = \frac{P(A \cap B)}{P(B)}$.

The idea is that if we learn that B has occurred, then the probability space must be updated to account for this new information. In particular, the sample space becomes B, the σ-algebra now includes only those events contained in B, $\mathcal{F}_B = \{E \cap B : E \in \mathcal{F}\}$, and the probability measure restricts to $\mathcal{F}_B$ and is normalized to account for this change. The formula for conditional probability is thus a description of how the probability function changes when additional information dictates that the sample space should shrink. ($P(\cdot\,|B)$ also defines a probability on $(\Omega, \mathcal{F})$, and it is often more convenient to adopt this perspective.)

When thinking about conditional probability, it can be instructive to take a step back and think of a second observer with access to partial information. Here we interpret $(\Omega, \mathcal{F}, P)$ as describing a random system whose chance of being in state $\omega \in \Omega$ is governed by P. $\mathcal{F}$ represents the possible conclusions that can be drawn about the state of the system: All that can be said is whether it lies in A for each $A \in \mathcal{F}$.
Now suppose that the observer has performed a measurement that tells her if B holds for some $B \in \mathcal{F}$ with $P(B) \in (0, 1)$. If she found out that B is true, her assessment of the probability of $A \in \mathcal{F}$ would be $P(A\,|B)$. If she found that B is false, she would evaluate the probability of A as $P\big(A\,|B^C\big)$. Thus, from our point of view, her description of the probability of A is given by the random variable

$$X_A(\omega) = \begin{cases} P(A\,|B), & \omega \in B \\ P\big(A\,|B^C\big), & \omega \notin B \end{cases}.$$

This is ultimately the kind of idea we are trying to capture with conditional expectation.

The typical development in elementary treatments of probability is to apply the definition of $P(A\,|B)$ to the events $\{X = x\}$ and $\{Y = y\}$ for discrete random variables X, Y in order to define the conditional mass function of X given that $Y = y$ as $p_X(x\,|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$. One then extrapolates to absolutely continuous X and Y by replacing mass functions with densities (which is problematic in that it treats pdfs as probabilities and raises issues concerning conditioning on null events). Finally, conditional expectation is defined in terms of integrating against the conditional pmfs/pdfs.

In what follows, we will need a more sophisticated theory of conditioning that avoids some of the pitfalls, paradoxes, and limitations of the framework sketched out above. Rather than try to arrive at the proper definition by way of more familiar concepts, we will begin with a formal definition and then work through a variety of examples and related results in order to provide motivation, build intuition, and make connections with ideas from elementary probability.

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ a random variable with $E\,|X| < \infty$, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. We define $E[X\,|\mathcal{G}]$, the *conditional expectation of X given $\mathcal{G}$*, to be any random variable Y satisfying

(i) $\quad Y \in \mathcal{G}$ (i.e. Y is measurable with respect to $\mathcal{G}$)
(ii) $\quad \int_A Y dP = \int_A X dP$ for all $A \in \mathcal{G}$

If Y satisfies (i) and (ii), we say that Y is a *version* of $E[X\,|\mathcal{G}]$.

Our most immediate order of business is to show that this definition makes good mathematical sense by proving existence and uniqueness theorems.

To streamline this task, we first take a moment to establish integrability for random variables which fit the definition so that we may manipulate various quantities of interest with impunity.

Lemma 1.1. *If Y satisfies conditions (i) and (ii) in the definition of $E[X\,|\mathcal{G}]$, then it is integrable.*

Proof. Letting $A = \{Y \geq 0\} \in \mathcal{G}$, condition (ii) implies

$$\int_A Y dP = \int_A X dP \leq \int_A |X|\, dP,$$
$$\int_{A^C} (-Y) dP = -\int_{A^C} Y dP = -\int_{A^C} X = \int_{A^C} (-X) dP \leq \int_{A^C} |X|\, dP.$$

It follows that

$$E\,|Y| = \int_A Y dP + \int_{A^C} (-Y) dP \leq \int_A |X|\, dP + \int_{A^C} |X|\, dP = E\,|X| < \infty. \qquad \square$$

Our next result makes use of a famous theorem from analysis whose proof can be found in any text on measure theory.

Theorem 1.1 (Radon-Nikodym)**.** *If μ and ν are σ-finite measures on $(S, \mathcal{S})$ with $\nu \ll \mu$, then there is a measurable function $f : S \to \mathbb{R}$ such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{S}$.*
f is called the Radon-Nikodym derivative of ν with respect to μ, written $f = \frac{d\nu}{d\mu}$.

The following existence proof gives an interpretation of conditional expectation in terms of Radon-Nikodym derivatives.

Theorem 1.2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ a random variable with $E\,|X| < \infty$, and $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra. There exists a random variable Y satisfying*

(i) $Y \in \mathcal{G}$
(ii) $\int_A Y dP = \int_A X dP$ *for all* $A \in \mathcal{G}$

Proof. First suppose that $X \geq 0$. Define $\nu(A) = \int_A X dP$ for $A \in \mathcal{G}$. Then $P|_{\mathcal{G}}$ and ν are finite measures on $(\Omega, \mathcal{G})$. (That ν is countably additive is an easy application of the DCT.) Moreover, ν is clearly absolutely continuous with respect to P. The Radon-Nikodym theorem therefore implies that there is a function $\frac{d\nu}{dP} \in \mathcal{G}$ such that

$$\int_A X dP = \nu(A) = \int_A \frac{d\nu}{dP} dP.$$

It follows that $Y = \frac{d\nu}{dP}$ is a version of $E[X\,|\mathcal{G}]$.

For general X, write $X = X^+ - X^-$ and let $Y_1 = E\,[X^+\,|\mathcal{G}]$, $Y_2 = E\,[X^-\,|\mathcal{G}]$. Then $Y = Y_1 - Y_2$ is integrable and $\mathcal{G}$-measurable, so for all $A \in \mathcal{G}$,

$$\int_A Y dP = \int_A Y_1 dP - \int_A Y_2 dP = \int_A X^+ dP - \int_A X^- dP = \int_A X dP. \qquad \square$$

* The proof of Theorem 1.2 is really a one-liner since the Radon-Nikodym theorem holds for signed measures and $\rho(A) = \int_A g d\mu$ defines a signed measure for μ-integrable g.

To conclude our discussion of the definition of conditional expectation, we record

Theorem 1.3. *Y is unique up to null sets.*

Proof. Suppose that Y' is also a version of $E[X\,|\mathcal{G}]$.

Condition (ii) implies that

$$\int_A Y'dP = \int_A XdP = \int_A YdP$$

for all $A \in \mathcal{G}$.

By condition (i), the event $A_\varepsilon = \{Y - Y' \geq \varepsilon\}$ is in $\mathcal{G}$ for all $\varepsilon > 0$, hence

$$0 = \int_{A_\varepsilon} YdP - \int_{A_\varepsilon} Y'dP = \int_{A_\varepsilon} (Y - Y')dP \geq \varepsilon P\left(Y - Y' \geq \varepsilon\right).$$

It follows that $Y \leq Y'$ a.s. Interchanging the roles of Y and Y' in the preceding argument shows that $Y' \leq Y$ a.s. as well, and the proof is complete. □

The following observation helps elucidate the sense in which uniqueness holds.

Proposition 1.1. *If Y is a version of $E[X\,|\mathcal{G}]$ and $Y' \in \mathcal{G}$ with $Y = Y'$ a.s., then Y' is also a version of $E[X|\mathcal{G}]$.*

Proof. Since Y and Y' are $\mathcal{G}$-measurable, $E = \{\omega : Y(\omega) \neq Y'(\omega)\} \in \mathcal{G}$. Since $P(E) = 0$, we see that for any $B \in \mathcal{G}$,

$$\begin{aligned}\int_B XdP = \int_B YdP &= \int_{B\cap E} YdP + \int_{B\setminus E} YdP = \int_{B\setminus E} YdP \\ &= \int_{B\setminus E} Y'dP = \int_{B\setminus E} Y'dP + \int_{B\cap E} Y'dP = \int_B Y'dP.\end{aligned}$$

□

Lemma 1.1, Theorem 1.3, and Proposition 1.1 combine to tell us that conditional expectation is unique as an element of $L^1(\Omega, \mathcal{G}, P)$. Just as elements of L^p spaces are really equivalence classes of functions (rather than specific functions) in classical analysis, conditional expectations are equivalence classes of random variables. Here versions play the role of specific functions.

Often we will omit the "almost sure" qualification when speaking of relations between conditional expectations, but it is important to keep this issue in mind.

In light of Proposition 1.1, we can often work with convenient versions of $E[X\,|\mathcal{G}]$ when we need to make use of pointwise results.

Examples.

Intuitively, sub-σ-algebras represent (potentially available) information – for each $A \in \mathcal{G}$ we can ask whether or not A has occurred. From this perspective, we can think of $E[X\,|\mathcal{G}]$ as giving the "best guess" for the value of X given the information in $\mathcal{G}$. The following examples are intended to clarify this view.

Example 1.1. If $X \in \mathcal{G}$, then our heuristic suggests that $E[X\,|\mathcal{G}] = X$ since if we know X, then our best guess is X itself. This clearly satisfies the definition as X always satisfies condition (ii) and condition (i) is met by assumption.
Since constants are measurable with respect to any σ-algebra, taking $X = c$ shows that $E[c\,|\mathcal{G}] = c$.

Example 1.2. At the other extreme, suppose that X is independent of $\mathcal{G}$ – that is, for all $A \in \mathcal{G}$, $B \in \mathcal{B}$, $\{X \in B\}$ and A are independent events. In this case, $\mathcal{G}$ tells us nothing about X, so our best guess is $E[X]$. As a constant, $E[X]$ automatically satisfies condition (i). To see that (ii) holds as well, note that for any $A \in \mathcal{G}$,

$$\int_A E[X]dP = E[X]P(A) = E[X]E[1_A] = E[X1_A] = \int_A XdP$$

by independence.
In particular, ordinary expectation corresponds to conditional expectation w.r.t. $\mathcal{G} = \{\Omega, \emptyset\}$.

Example 1.3. We now expand upon our introductory example: Suppose that $\Omega_1, \Omega_2, ...$ is a countable partition of Ω into disjoint measurable sets, each having positive probability (e.g. B and B^C). Let $\mathcal{G} = \sigma(\Omega_1, \Omega_2, ...)$. We claim that $E[X\,|\mathcal{G}] = P(\Omega_i)^{-1}E[X;\Omega_i]$ on Ω_i. The interpretation is that $\mathcal{G}$ tells us which Ω_i contains the outcome, and given that information, our best guess for X is its average over Ω_i.

To verify our claim, note that

$$E[X\,|\mathcal{G}](\omega) = \sum_i \frac{E[X;\Omega_i]}{P(\Omega_i)} 1_{\Omega_i}(\omega)$$

is $\mathcal{G}$-measurable since each Ω_i belongs to $\mathcal{G}$. Also, since each $A \in \mathcal{G}$ is a countable disjoint union of the Ω_i's, it suffices to check condition (ii) on the elements of the partition. But this is trivial as

$$\int_{\Omega_i} P(\Omega_i)^{-1}E[X;\Omega_i]dP = E[X;\Omega_i] = \int_{\Omega_i} XdP.$$

If we make the obvious definition $P(A\,|\mathcal{H}) = E[1_A\,|\mathcal{H}]$, then the above says that

$$P(A\,|\mathcal{G}) = P(\Omega_i)^{-1}\int_{\Omega_i} 1_A dP = \frac{P(A \cap \Omega_i)}{P(\Omega_i)} \text{ on } \Omega_i.$$

Example 1.4. Conditioning on a random variable can be seen as a special case of our definition by taking $E[X\,|Y\,] = E[X\,|\sigma(Y)\,]$. To see how this compares with the definition given in undergraduate probability, suppose that X and Y are discrete with joint pmf $p_{X,Y}$ and marginals p_X, p_Y. Then $\sigma(Y)$ is generated by the countable partition $\{Y = y\}_{y\in \text{Range}(Y)}$, so the previous example shows that if $E\,|X| < \infty$, then

$$E[X|Y] = P(Y=y)^{-1}E[X;\{Y=y\}] = \frac{1}{P(Y=y)}\sum_x xP(X=x,Y=y)$$
$$= \sum_x x\frac{p_{X,Y}(x,y)}{p_Y(y)}$$

on $\{Y = y\}$.

Example 1.5. Similarly, suppose that X and Y are jointly absolutely continuous with joint density $f_{X,Y}$ and marginals f_X, f_Y. Suppose for simplicity that $f_Y(y) > 0$ for all $y \in \mathbb{R}$. In this case, if $E\,|g(X)| < \infty$, then $E[g(X)\,|Y\,] = h(Y)$ where

$$h(y) = \int g(x)\frac{f_{X,Y}(x,y)}{f_Y(y)}dx.$$

The Doob-Dynkin lemma shows that $E[g(X)\,|Y\,] \in \sigma(Y)$. To see that the second criterion is satisfied as well, recall that every $A \in \sigma(Y)$ is of the form $A = \{Y \in B\}$ for some $B \in \mathcal{B}$. The change of variables formula shows that

$$\int_{\{Y\in B\}} h(Y)dP = \int_B h(y)f_Y(y)dy = \int 1_B(y)\left(\int g(x)\frac{f_{X,Y}(x,y)}{f_Y(y)}dx\right)f_Y(y)dy$$
$$= \int\int g(x)1_B(y)f_{X,Y}(x,y)dxdy = E[g(X)1_B(Y)] = \int_{\{Y\in B\}} g(X)dP.$$

Note that the condition $f_Y > 0$ is actually unnecessary since the above proof only needs h to satisfy

$$h(y)f_Y(y) = \int g(x)f_{X,Y}(x,y)dx,$$

so h can take on any value at those y with $f_Y(y) = 0$. (Since $f_Y(y) = \int f_{X,Y}(x,y)dx$ and $f_{X,Y} \geq 0$, the right-hand side of the above equation will also be 0 at such y.)

Example 1.6. Suppose that X and Y are independent and φ satisfies $E\,|\varphi(X,Y)| < \infty$. Then

$$E[\varphi(X,Y)\,|X\,] = g(X) \text{ where } g(x) = E[\varphi(x,Y)].$$

As in the previous example, condition (i) is satisfied by Doob-Dynkin, and condition (ii) can be verified by letting μ and ν denote the distributions of X and Y, respectively, and computing

$$\int_{\{X\in B\}} g(X)dP = \int_B g(x)d\mu(x) = \int 1_B(x)\left(\int \varphi(x,y)d\nu(y)\right)d\mu(x)$$
$$= \int\int 1_B(x)\varphi(x,y)d(\mu\times\nu)(x,y) = \int 1_B(X)\varphi(X,Y)dP = \int_{\{X\in B\}}\varphi(X,Y)dP.$$

Properties.

Many of the properties of ordinary expectation carry over to conditional expectation as they are ultimately facts about integrals:

Proposition 1.2 (Linearity). $E[aX + Y|\mathcal{G}] = aE[X|\mathcal{G}] + E[Y|\mathcal{G}]$

Proof. Sums and constant multiples of $\mathcal{G}$-measurable functions are $\mathcal{G}$-measurable, and for any $A \in \mathcal{G}$

$$\begin{aligned}\int_A (aE[X|\mathcal{G}] + E[Y|\mathcal{G}])\, dP &= a\int_A E[X|\mathcal{G}]\, dP + \int_A E[Y|\mathcal{G}]\, dP \\ &= a\int_A X dP + \int_A Y dP = \int_A (aX + Y) dP.\end{aligned}$$

□

Proposition 1.3 (Monotonicity). *If $X \leq Y$, then $E[X|\mathcal{G}] \leq E[Y|\mathcal{G}]$.*

Proof. By assumption, we have

$$\int_A E[X|\mathcal{G}] dP = \int_A X dP \leq \int_A Y dP = \int_A E[Y|\mathcal{G}] dP$$

for all $A \in \mathcal{G}$. For all $\varepsilon > 0$, $A_\varepsilon = \{\omega : E[X|\mathcal{G}] - E[Y|\mathcal{G}] \geq \varepsilon\} \in \mathcal{G}$, so

$$\varepsilon P(A_\varepsilon) \leq \int_{A_\varepsilon} (E[X|\mathcal{G}] - E[Y|\mathcal{G}])\, dP = \int_{A_\varepsilon} E[X|\mathcal{G}] dP - \int_{A_\varepsilon} E[Y|\mathcal{G}] dP \leq 0.$$

It follows that $E[X|\mathcal{G}] \leq E[Y|\mathcal{G}]$ a.s. □

Proposition 1.4 (Monotone Convergence Theorem). *If $X_n \geq 0$ and $X_n \nearrow X$, then $E[X_n|\mathcal{G}] \nearrow E[X|\mathcal{G}]$.*

Proof. By monotonicity, $0 \leq E[X_n|\mathcal{G}] \leq E[X_{n+1}|\mathcal{G}] \leq E[X|\mathcal{G}]$ for all n. (The inequalities are almost sure, but we can work with versions of the conditional expectations where they hold pointwise.) Since bounded nondecreasing sequences of reals converge to their limit superior, there is a random variable Y with $E[X_n|\mathcal{G}] \nearrow Y$.

Moreover, $Y \in \mathcal{G}$ as it is the limit of $\mathcal{G}$-measurable functions.

Finally, applying the ordinary MCT to $E[X_n|\mathcal{G}]1_B \nearrow Y1_B$, invoking the definition of conditional expectation, and then applying the MCT to $X_n 1_B \nearrow X1_B$ shows that

$$\int_B Y dP = \lim_{n\to\infty} \int_B E[X_n|\mathcal{G}] dP = \lim_{n\to\infty} \int_B X_n dP = \int_B X dP$$

for all $B \in \mathcal{G}$, hence Y is a version of $E[X|\mathcal{G}]$. □

Note that since we have established a conditional MCT, conditional versions of Fatou and dominated convergence follow from the usual arguments.

The final analogue we will consider is a conditional form of Jensen's inequality. It is fairly straightforward to derive conditional variants of other familiar theorems using these examples as templates.

Proposition 1.5 (Jensen)**.** *If φ is convex and $E|X|, E|\varphi(X)| < \infty$, then*

$$\varphi(E[X|\mathcal{G}]) \leq E[\varphi(X)|\mathcal{G}].$$

Proof. When we proved the original Jensen inequality, we established that if φ is convex, then for every $c \in \mathbb{R}$, there is a linear function $l_c(x) = a_c x + b_c$ such that $l_c(c) = \varphi(c)$ and $l_c(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.

Let $S = \{(a_r, b_r)\}_{r\in\mathbb{Q}}$. Then S is countable with $ax + b \leq \varphi(x)$ for all $x \in \mathbb{R}$, $(a, b) \in S$. Moreover, since $\mathbb{Q}$ is dense in $\mathbb{R}$ and convex functions are continuous, we have $\varphi(x) = \sup_{(a,b)\in S} ax + b$ for all $x \in \mathbb{R}$.

Monotonicity and linearity imply that

$$E[\varphi(X)|\mathcal{G}] \geq E[aX + b|\mathcal{G}] = aE[X|\mathcal{G}] + b \text{ a.s.}$$

whenever $(a, b) \in S$.

As S is countable, the event $A = \{E[\varphi(X)|\mathcal{G}] \geq aE[X|\mathcal{G}] + b \text{ for all } (a, b) \in S\}$ has full probability.

Thus with probability one, we have

$$E[\varphi(X)|\mathcal{G}] \geq \sup_{(a,b)\in S} aE[X|\mathcal{G}] + b = \varphi(E[X|\mathcal{G}]). \qquad \square$$

One use for conditional expectation is as an intermediary for computing ordinary expectations. This is justified by the "law of total expectation":

Proposition 1.6. $E[E[X|\mathcal{G}]] = E[X]$.

Proof. Taking $A = \Omega$ in the definition of $E[X|\mathcal{G}]$ yields

$$E[X] = \int_\Omega X dP = \int_\Omega E[X|\mathcal{G}] dP = E[E[X|\mathcal{G}]]. \qquad \square$$

As an example of the utility of the preceding observation, we prove

Proposition 1.7. *Conditional expectation is a contraction in L^p, $p \geq 1$.*

Proof. Since $\varphi(x) = |x|^p$ is convex, Proposition 1.5 implies that $|E[X|\mathcal{G}]|^p \leq E[|X|^p|\mathcal{G}]$. Taking expectations and appealing to Proposition 1.6 gives

$$E[|E[X|\mathcal{G}]|^p] \leq E[E[|X|^p|\mathcal{G}]] = E[|X|^p]. \qquad \square$$

Proposition 1.6 is actually a special case of the "tower property" of conditional expectation.

This result is one of the more useful theorems about conditional expectation and is often summarized as "The smaller σ-algebra always wins."

Theorem 1.4. *If $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then*

$$E\left[E[X\,|\mathcal{G}_1]\,|\mathcal{G}_2\right] = E\left[E[X\,|\mathcal{G}_2]\,|\mathcal{G}_1\right] = E[X\,|\mathcal{G}_1].$$

Proof.

Since $E[X\,|\mathcal{G}_1] \in \mathcal{G}_1 \subseteq \mathcal{G}_2$, Example 1.1 shows that $E\left[E[X\,|\mathcal{G}_1]\,|\mathcal{G}_2\right] = E[X\,|\mathcal{G}_1]$.

To see that $E\left[E[X\,|\mathcal{G}_2]\,|\mathcal{G}_1\right] = E[X\,|\mathcal{G}_1]$, we observe that $E[X\,|\mathcal{G}_1] \in \mathcal{G}_1$ and for any $A \in \mathcal{G}_1 \subseteq \mathcal{G}_2$,

$$\int_A E[X\,|\mathcal{G}_1]dP = \int_A X dP = \int_A E[X\,|\mathcal{G}_2]dP. \qquad \square$$

* Proposition 1.6 is the case $\mathcal{G}_1 = \{\Omega, \emptyset\}$, $\mathcal{G}_2 = \mathcal{G}$.

The second criterion in our definition of conditional expectation can be expressed in more probabilistic language as $E[Y1_A] = E[X1_A]$ for all $A \in \mathcal{G}$. One sometimes sees the alternative criterion $E[YZ] = E[XZ]$ for all bounded $Z \in \mathcal{G}$. The equivalence of the two conditions follows from the usual four-step procedure for building general integrals from integrals of indicators. We will stick with our original definition as it is easier to check.

The following theorem (which Durrett describes as saying that "for conditional expectation with respect to $\mathcal{G}$, random variables $Y \in \mathcal{G}$ are like constants [in that] they can be brought outside the integral") generalizes this alternative definition.

Theorem 1.5. *If $W \in \mathcal{G}$ and $E\,|X|\,, E\,|WX| < \infty$, then $E[WX\,|\mathcal{G}] = WE[X\,|\mathcal{G}]$.*

Proof. $WE[X\,|\mathcal{G}] \in \mathcal{G}$ by assumption, so we need only check the second criterion.

We first suppose that $W = 1_B$ for some $B \in \mathcal{G}$. Then for all $A \in \mathcal{G}$,

$$\begin{aligned}\int_A WE[X\,|\mathcal{G}]dP &= \int_A 1_B E[X\,|\mathcal{G}]dP = \int_{A\cap B} E[X\,|\mathcal{G}]dP \\ &= \int_{A\cap B} X dP = \int_A 1_B X dP = \int_A WX dP.\end{aligned}$$

By linearity, we see that the condition $\int_A WE[X\,|\mathcal{G}]dP = \int_A WX dP$ also holds when W is a simple function.

Now if $W, X \geq 0$, we can take a sequence of simple functions $W_n \nearrow W$ and use the MCT to conclude that

$$\begin{aligned}\int_A WE[X\,|\mathcal{G}]dP &= \lim_{n\to\infty} \int_A W_n E[X\,|\mathcal{G}]dP \\ &= \lim_{n\to\infty} \int_A W_n X dP = \int_A WX dP.\end{aligned}$$

The general result follows by splitting W and X into positive and negative parts. $\square$

Our final major result about conditional expectation gives a geometric interpretation in the case of square integrable X. Namely, noting that $L^2(\mathcal{F}) = \{Y \in \mathcal{F} : E[Y^2] < \infty\}$ is a Hilbert space and $L^2(\mathcal{G})$ is a closed subspace of $L^2(\mathcal{F})$, we will show that if $X \in L^2(\mathcal{F})$, then $E[X\,|\mathcal{G}]$ is the orthogonal projection of X onto $L^2(\mathcal{G})$.

Theorem 1.6. *If $E[X^2] < \infty$, then $E[X\,|\mathcal{G}]$ minimizes the mean square error $E[(X-Y)^2]$ amongst all $Y \in \mathcal{G}$.*

Proof. To begin, we note that if $Z \in L^2(\mathcal{G})$, then $E\,|ZX| < \infty$ by the Cauchy-Schwarz inequality, so Theorem 1.5 implies

$$ZE[X\,|\mathcal{G}] = E[ZX\,|\mathcal{G}].$$

Taking expected values gives

$$E\left[ZE[X\,|\mathcal{G}]\right] = E\left[E[ZX\,|\mathcal{G}]\right] = E[ZX],$$

showing that

$$E\left[Z\left(X - E[X\,|\mathcal{G}]\right)\right] = E[ZX] - E\left[ZE[X\,|\mathcal{G}]\right] = 0$$

for $Z \in L^2(\mathcal{G})$.

Thus for any $Y \in L^2(\mathcal{G})$, if we set $Z = E[X\,|\mathcal{G}] - Y$, then we have

$$\begin{aligned} E\left[(X-Y)^2\right] &= E\left[\left((X - E[X\,|\mathcal{G}]) + Z\right)^2\right] \\ &= E\left[\left(X - E[X\,|\mathcal{G}]\right)^2\right] + 2E\left[Z\left(X - E[X\,|\mathcal{G}]\right)\right] + E[Z^2] \\ &= E\left[\left(X - E[X\,|\mathcal{G}]\right)^2\right] + E[Z^2]. \end{aligned}$$

(Proposition 1.7 shows that $E[X\,|\mathcal{G}] \in L^2(\mathcal{G})$, so $Z = E[X\,|\mathcal{G}] - Y \in L^2(\mathcal{G})$ as well.)

It follows that $E\left[(X-Y)^2\right]$ is minimized over $L^2(\mathcal{G})$ when $E[X\,|\mathcal{G}] - Y = Z = 0$.

To see that $E[X\,|\mathcal{G}]$ minimizes the MSE over $L^0(\mathcal{G})$, we make use of the inequality

$$(a+b)^2 \le (a+b)^2 + (a-b)^2 = 2a^2 + 2b^2.$$

If $Y \in \mathcal{G}$ is such that $E[(X-Y)^2] = \infty$, then it certainly doesn't minimize the MSE since $E[X\,|\mathcal{G}] \in L^2(\mathcal{G})$ with

$$E\left[\left(X - E[X\,|\mathcal{G}]\right)^2\right] \le 2E[X^2] + 2E\left[E[X\,|\mathcal{G}]^2\right] < \infty,$$

and if $E[(X-Y)^2] < \infty$, then

$$E[Y^2] = E[((Y-X)+X)^2] \le 2E[(X-Y)^2] + 2E[X^2] < \infty. \qquad \square$$

In some treatments of conditional expectation, the Radon-Nikodym approach is bypassed entirely by first defining $E[X\,|\mathcal{G}]$ for $X \in L^2(\mathcal{F})$ in terms of projection onto $L^2(\mathcal{G})$, and then extending the definition to $X \in L^1(\mathcal{G})$ using approximating sequences of square integrable random variables. An upshot of this strategy is that one can then prove the Radon-Nikodym theorem using martingales!

Regular Conditional Distributions.

Before moving on to martingales, we pause briefly to discuss regular conditional distributions/probabilities.

Recall that if $(\Omega, \mathcal{F}, P)$ is a probability space, $(S, \mathcal{S})$ is a measurable space, and X is an $(S, \mathcal{S})$-valued random variable on $(\Omega, \mathcal{F}, P)$, then X induces the pushforward measure $P \circ X^{-1}$ on $(S, \mathcal{S})$. The theory of regular conditional distributions provides a conditional analogue of this construction.

To see how this works, suppose that $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra. For any set $A \in \mathcal{S}$, the conditional probability of $\{X \in A\}$ given $\mathcal{G}$ is given by the conditional expectation $\mu(\omega, A) = E[1_A(X) \,|\mathcal{G}](\omega)$. As the notation suggests, we are thinking of μ as a function $\mu : \Omega \times \mathcal{S} \to [0, 1]$.
We would like to know whether $\mu(\omega, \cdot)$ defines a probability measure on $(S, \mathcal{S})$ for P-a.e. $\omega \in \Omega$.

Note that for every $A \in \mathcal{S}$, $\mu(\cdot, A)$ is a.s. uniquely defined and we are free to modify $\mu(\cdot, A)$ on null sets, so linearity and monotone convergence imply that for any particular countable disjoint collection $\{A_n\} \subseteq \mathcal{S}$, we have

$$\mu\left(\omega, \bigcup_n A_n\right) = \lim_n E\left[1_{A_1}(X) + ... + 1_{A_n}(X) \,|\mathcal{G}\right] = \sum_n \mu\left(\omega, A_n\right)$$

for a set of ω having probability 1.

The issue is that this event may depend on $\{A_n\}$ and we want $\mu(\omega, \cdot)$ to satisfy the countable additivity condition for any collection $\{A_n\}$ for P-a.e. ω. If there are too many different countable collections, then the exceptional sets might pile up to give a set having positive probability, or even a non-measurable set.

To address this issue properly, we set forth a formal definition.

Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(S, \mathcal{S})$ a measurable space, $\mathcal{G} \subseteq \mathcal{F}$ a sub-σ-algebra, and $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ a random variable. A map $\mu : \Omega \times \mathcal{S} \to [0, 1]$ is said to be a *regular conditional distribution for X given $\mathcal{G}$* if

(i) For each $A \in \mathcal{S}$, $\omega \mapsto \mu(\omega, A)$ is a version of $P(X \in A \,|\mathcal{G})$,
(ii) For P-a.e. $\omega \in \Omega$, $A \mapsto \mu(\omega, A)$ is a probability measure on $(S, \mathcal{S})$.

When $(S, \mathcal{S}) = (\Omega, \mathcal{F})$ and $X(\omega) = \omega$, μ is called a *regular conditional probability.*

One reason such objects are of interest is that if μ is a regular conditional distribution for X given $\mathcal{G}$, then we can express conditional expectations of functions of X given $\mathcal{G}$ in terms of integrals against the r.c.d. (just like the ordinary change of variables formula with distribution functions).

Theorem 1.7. *Let $(\Omega, \mathcal{F}, P)$, $(S, \mathcal{S})$, $\mathcal{G}$, and X be as in the above definition. Let μ be a r.c.d. for X given $\mathcal{G}$. Then for any measurable $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$ with $E\,|f(X)| < \infty$, we have*

$$E[f(X) \,|\mathcal{G}](\omega) = \int f(x)\mu(\omega, dx) \text{ for } P\text{-a.e. } \omega$$

Proof. If $f = 1_B$ for some $B \in \mathcal{S}$, then

$$\begin{aligned} E[f(X) \,|\mathcal{G}](\omega) &= E[1_B(X) \,|\mathcal{G}](\omega) = P(X \in B \,|\mathcal{G})(\omega) \\ &= \mu(\omega, B) = \int_B \mu(\omega, dx) = \int f(x)\mu(\omega, dx) \end{aligned}$$

where the third equality (which is a.s.) is condition (i) in the definition of μ.

By linearity, the result holds for simple functions; by monotone convergence it holds for nonnegative functions; and by consideration of positive and negative parts, it holds for integrable functions.
(In each of these three steps, there are only countably many null sets that need to be discarded.) □

Of course, for all of this to be worthwhile, it is necessary that r.c.d.s actually exist. Naively, one would just take a version Y_A of $P(X \in A\,|\mathcal{G})$ for each $A \in \mathcal{S}$ and set $\mu(\omega, A) = Y_A(\omega)$. If $\mathcal{F}$ is finite this is not a problem as one can just modify each Y_A on a null set so that everything works out. But if $\mathcal{F}$ is infinite, then there are uncountably many Y_A's and the set $\{\omega : \text{there is some } \{A_n\} \text{ with } Y_{\bigcup_n A_n}(\omega) \neq \sum_n Y_{A_n}(\omega)\}$ may have positive probability or not even lie in $\mathcal{F}$ no matter how the Y_A's are chosen. One can construct examples where r.c.d.s fail to exist for essentially this reason.
However, the following theorem shows that if X takes values in a "nice" space (e.g. a complete and separable metric space), then it has a r.c.d. given any sub-σ-algebra.

Theorem 1.8. *If $(\Omega, \mathcal{F}, P)$ is a probability space, $\mathcal{G} \subseteq \mathcal{F}$ is a sub-σ-algebra, $(S, \mathcal{S})$ is a nice space, and X is a $(S, \mathcal{S})$-valued random variable on $(\Omega, \mathcal{F}, P)$, then X has a r.c.d. given $\mathcal{G}$.*

Proof. By definition, there is a bijection $\varphi : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$ with φ and φ^{-1} measurable.
Using monotonicity and throwing out a countable collection of null sets, we see that there is a set $\Omega_0 \in \mathcal{F}$ with $P(\Omega_0) = 1$ and a family of random variables $\omega \mapsto G(\omega, q)$, $q \in \mathbb{Q}$, such that $\omega \mapsto G(\omega, q)$ is a version of $P(\varphi(X) \leq q\,|\mathcal{G})$ with $q \mapsto G(\omega, q)$ is nondecreasing for all $\omega \in \Omega_0$.
As in the proof of Helly's selection theorem, if we define $F(\omega, x) = \inf\{G(\omega, q) : q > x\}$, then $F(\omega, \cdot)$ is a distribution function for all $\omega \in \Omega_0$.

* $x \mapsto F(\omega, x)$ is nondecreasing since for any $x < y$, there is a rational $x < r < y$ so, since $G(\omega, r) \leq G(\omega, s)$ for $r \leq s$,

$$F(\omega, x) = \inf\{G(\omega, q) : q > x\} \leq G(\omega, r) \leq \inf\{G(\omega, q) : q > y\} = F(\omega, y);$$

$x \mapsto F(\omega, x)$ is right continuous since for any $x \in \mathbb{R}$, $\varepsilon > 0$, there is a rational $q > x$ with $G(\omega, q) \leq F(\omega, x) + \varepsilon$ and thus for any $x < y < q$,

$$F(\omega, y) \leq G(\omega, q) \leq F(\omega, x) + \varepsilon;$$

and $x \mapsto F(\omega, x)$ satisfies the appropriate boundary conditions since it is nondecreasing with supremum 1 and infimum 0 by virtue of $q \mapsto G(\omega, q)$ satisfying these boundary conditions.
($G(\omega, \cdot)$ has the appropriate limits since $0 \leq G(\omega, q) \leq 1$ and $\int G(\omega, q) dP(\omega) = \int 1\{\varphi(X) \leq q\}\, dP$ which goes to 0 or 1 as q goes to $\mp\infty$ by monotone convergence.)

Thus each $\omega \in \Omega_0$ gives rise to a unique measure $\nu(\omega, \cdot)$ having distribution function $F(\omega, x) = \nu(\omega, (-\infty, x])$. Also, monotone convergence shows that $\omega \mapsto F(\omega, x)$ is a version of $P(\varphi(X) \leq x\,|\mathcal{G})$ for every $x \in \mathbb{R}$. As one readily checks that $\mathcal{L} = \{B \in \mathcal{B} : \nu(\omega, B) = P(\varphi(X) \in B\,|\mathcal{G})\}$ a λ-system, it follows from the π-λ theorem (with $\mathcal{P} = \{(-\infty, x] : x \in \mathbb{R}\}$) that $\nu(\omega, B)$ is a version of $P(\varphi(X) \in B\,|\mathcal{G})$.

To extract the r.c.d. in question, note that for any $A \in \mathcal{S}$, $\{X \in A\} = \{\varphi(X) \in \varphi(A)\}$, so we can take $\mu(\omega, A) := \nu(\omega, \varphi(A))$. □

2. Martingales

With conditional expectation formally defined and its most important properties established, we are ready to tackle the topic of discrete time martingales.

Recall that a filtration is an increasing sequence of sub-σ-algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$. (Really, the term "net" is more appropriate than "sequence" since often one would like to consider uncountable index sets such as $[0,T]$, but we are working in discrete time at the moment so the distinction is not important.)

We say that a collection of random variables $\{X_n\}_{n=1}^{\infty}$ is *adapted* to the filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$ if $X_n \in \mathcal{F}_n$ for all $n \in \mathbb{N}$.
Note that $\{X_n\}_{n\in\mathbb{N}}$ is always adapted to the filtration defined by $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Indeed, if $\{X_n\}_{n\in\mathbb{N}}$ is adapted to $\{\mathcal{G}_n\}_{n\in\mathbb{N}}$, then necessarily $\mathcal{G}_n \supseteq \sigma(X_1, ..., X_n)$.

Definition 2.1. We say that a sequence $\{X_n\}_{n\in\mathbb{N}}$ is a *martingale* with respect to the filtration $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ if for all $n \in \mathbb{N}$

(i) $E\,|X_n| < \infty$
(ii) $X_n \in \mathcal{F}_n$
(iii) $E\,[X_{n+1}\,|\mathcal{F}_n\,] = X_n$.

If condition (iii) is replaced with $E\,[X_{n+1}\,|\mathcal{F}_n\,] \geq X_n$, then we say that $\{X_n\}$ is a *submartingale*.

If (iii) reads $E\,[X_{n+1}\,|\mathcal{F}_n\,] \leq X_n$, then we say that $\{X_n\}$ is a *supermartingale*.

Note that, by definition, $\{X_n\}$ is a martingale if and only if it is both a submartingale and a supermartingale. Also, by linearity, if $\{X_n\}$ is a submartingale, then $\{-X_n\}$ is a supermartingale (and symmetrically).
Thus when one proves a theorem involving submartingales, say, there are immediate corollaries for martingales and supermartingales.

The standard picture of a *smartingale* (the collective term for ordinary, sub-, and super- martingales) is in terms of one's fortune after repeated bets (which need not be identical) on a game of unchanging odds.
A fair game corresponds to a martingale. If the odds are in the player's favor, then it corresponds to a submartingale. If the odds are against the player, one has a supermartingale. (If the odds are always for/against, then one still gets a sub/super-martingale even if they vary game to game.)

Example 2.1. We have already studied one martingale in some detail - simple random walk in one dimension. Specifically, let $\xi_1, \xi_2, ...$ be i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$, and let $X_n = \sum_{i=1}^{n} \xi_i$. As will be our convention henceforth, when no filtration is specified, we will take $\{\mathcal{F}_n\}$ to be the *natural filtration* $\mathcal{F}_n = \sigma(X_1, ..., X_n)$ so that condition (ii) is automatically satisfied. The first condition is also satisfied in this case since $|X_n| \leq n$. To see that this does indeed define a martingale, we compute

$$E[X_{n+1}\,|\mathcal{F}_n\,] = E[X_n + \xi_{n+1}\,|\mathcal{F}_n\,] = E[X_n\,|\mathcal{F}_n\,] + E[\xi_{n+1}\,|\mathcal{F}_n\,] = X_n + E[\xi_{n+1}] = X_n$$

where we used linearity of conditional expectation, $X_n \in \mathcal{F}_n$, and ξ_{n+1} independent of $\mathcal{F}_n$.

Example 2.2. More generally, if $X_n = \sum_{i=1}^n \xi_i$ where the ξ_n's are independent, then an identical argument shows that $\{X_n\}$ is a martingale if $E[\xi_n] = 0$ for all n; a submartingale if $E[\xi_n] \geq 0$ for all n; and a supermartingale if $E[\xi_n] \leq 0$ for all n.

Example 2.3. By linearity, if $\{X_n^{(1)}\}, ..., \{X_n^{(m)}\}$ are martingales (w.r.t. a common filtration $\{\mathcal{F}_n\}$) and $\alpha_0, \alpha_1, ..., \alpha_m \in \mathbb{R}$, then the sequence defined by $X_n = \alpha_0 + \sum_{i=1}^m \alpha_i X_n^{(i)}$ is a martingale. If $\alpha_1, ..., \alpha_m \geq 0$ and $\{X_n^{(1)}\}, ..., \{X_n^{(m)}\}$ are sub/super-martingales, then so is $\{X_n\}$.

Example 2.4. Let X be an integrable random variable on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$. Define $X_n = E[X \,|\mathcal{F}_n]$. We have

$$E\,|X_n| = E\,|E[X\,|\mathcal{F}_n]| \leq E\,[E[|X|\,|\mathcal{F}_n]] = E\,|X| < \infty$$

by Jensen's inequality and the law of total expectation;

$$X_n = E[X\,|\mathcal{F}_n] \in \mathcal{F}_n$$

by definition of conditional expectation; and

$$E[X_{n+1}\,|\mathcal{F}_n] = E\,[E[X\,|\mathcal{F}_{n+1}]\,|\mathcal{F}_n] = E[X\,|\mathcal{F}_n] = X_n$$

by the tower property.
The interpretation is that $X_1, X_2, ...$ represent increasingly better estimates of X as more information is revealed.

Example 2.5. Let $X_1, X_2, ...$ be i.i.d. with $P(X_1 = 0) = \frac{1}{2}$, $P(X_1 = 2) = \frac{1}{2}$. Define $M_n = \prod_{i=1}^n X_i$. Then $\{M_n\}$ is a martingale with respect to $\mathcal{F}_n = \sigma(X_1, ..., X_n)$ since $E[M_{n+1}\,|\mathcal{F}_n] = \frac{1}{2}(0 \cdot M_n) + \frac{1}{2}(2 \cdot M_n) = M_n$. Here we can think of M_n as representing the fortune of a gambler going "double or nothing" in a fair game with an initial bet (at time 0) of 1 unit.

The same argument shows if $X_1, X_2, ...$ are independent, bounded, and nonnegative, then $M_n = \prod_{i=1}^n X_i$ defines a martingale provided that all multiplicands have mean 1.
When they all have mean at least 1 (respectively, at most 1), M_n is a submartingale (respectively, supermartingale).

Example 2.6. The name martingale derives from a class of betting systems, the prototype of which is the following: In a fair game with an initial bet of \$1, double your bet after each loss and quit after your first win.
Mathematically, suppose that $\xi_1, \xi_2, ...$ are i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$ and define M_n recursively by $M_0 = 0$,

$$M_{n+1} = \begin{cases} 1, & M_n = 1 \\ M_n + 2^n \xi_{n+1}, & \text{else} \end{cases}.$$

Then $\{M_n\}$ is adapted to $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$ by an induction argument, and is integrable since $|M_n| < 2^n$.

It is thus a martingale since

$$\begin{aligned} E[M_{n+1}\,|\mathcal{F}_n] &= E\left[1\left\{M_n = 1\right\} + 1\left\{M_n \neq 1\right\}\left(M_n + 2^n\xi_{n+1}\right)|\mathcal{F}_n\right] \\ &= 1\left\{M_n = 1\right\} + 1\left\{M_n \neq 1\right\}\left(M_n + 2^n E[\xi_{n+1}\,|\mathcal{F}_n]\right) \\ &= 1\left\{M_n = 1\right\} + 1\left\{M_n \neq 1\right\}\left(M_n + 2^n E[\xi_{n+1}]\right) \\ &= 1\left\{M_n = 1\right\} M_n + 1\left\{M_n \neq 1\right\} M_n = M_n. \end{aligned}$$

Because you will eventually win with probability one and

$$2^n - \sum_{i=0}^{n-1} 2^i = 2^n - (2^n - 1) = 1,$$

it seems that this strategy guarantees that you will come out ahead even though the game is fair.
Moreover, if $0 < P(\xi_1 = 1) < P(\xi_1 = -1)$, then the same analysis shows that $\{M_n\}$ is a supermartingale, but you will still eventually come out ahead.

The catch is that this trick only works if you have an infinite line of credit and are allowed unlimited plays and arbitrarily large wagers. We will see later that if such stipulations are missing, then there is no system that will turn the odds of an unfair game in your favor.

*** (Sub/Super)Harmonic Functions.**

From the gambler's perspective, the definitions of sub- and super- martingales seem to be backwards (though the names are apt from the house's point of view). In fact, the nomenclature has nothing to do with choosing sides in this metaphor but rather arises from connections with potential theory.

Recall that if U is a *domain* (open connected set) in $\mathbb{R}^n$ then a continuous function $h : U \to \mathbb{R}$ is said to be *harmonic* on U if it satisfies Laplace's equation $\Delta h = 0$ on U.
An important fact about harmonic functions is

Theorem 2.1 (Mean Value Formulas). *$h \in C^2(U)$ is harmonic if and only if for every ball $\overline{B(x,r)} \subseteq U$,*

$$h(x) = \frac{1}{|B(x,r)|}\int_{B(x,r)} h(y)dy = \frac{1}{\sigma\left(\partial B(x,r)\right)}\int_{\partial B(x,r)} h(y)dS(y)$$

where $|B(x,r)| = r^n\alpha(n)$ is the volume of the ball $B(x,r)$ and $\sigma\left(\partial B(x,r)\right) = n\alpha(n)r^{n-1}$ is the surface measure of the sphere $\partial B(x,r)$; $\alpha(n) = |B(0,1)| = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}$.

Proof. Suppose that h is harmonic on $U \supseteq \overline{B(x,r)}$ and set

$$\phi(r) := \frac{1}{n\alpha(n)r^{n-1}}\int_{\partial B(x,r)} h(y)dS(y) = \frac{1}{n\alpha(n)}\int_{\partial B(0,1)} h(x+rz)dS(z)$$

Differentiating and appealing to the divergence theorem gives

$$\begin{aligned} \phi'(r) &= \frac{1}{n\alpha(n)}\int_{\partial B(0,1)} \nabla h(x+rz)\cdot z\, dS(z) = \frac{1}{n\alpha(n)r^{n-1}}\int_{\partial B(x,r)} \nabla h(y)\cdot\frac{y-x}{r}dS(y) \\ &= \frac{1}{n\alpha(n)r^{n-1}}\int_{\partial B(x,r)} \nabla h(y)\cdot\nu(y)dS(y) = \frac{1}{n\alpha(n)r^{n-1}}\int_{B(x,r)} \Delta h(y)dy = 0. \end{aligned}$$

Since $\phi(r)$ is constant, we have

$$\phi(r) = \lim_{r\to 0} \frac{1}{n\alpha(n)r^{n-1}} \int_{\partial B(x,r)} h(y)dS(y) = h(x).$$

Radial integration gives

$$\int_{B(x,r)} h(y)dy = \int_0^r \left(\int_{\partial B(x,s)} h\, dS \right) dr = \int_0^r \phi(x)n\alpha(n)s^{n-1}ds = \phi(x)\alpha(n)r^n.$$

For the converse, suppose that $\Delta h \not\equiv 0$ on U. The there is some ball $B(x,r) \subseteq U$ with, say, $\Delta h > 0$ on $B(x,r)$. Taking ϕ as above yields the contradiction

$$0 = \phi'(r) = \frac{1}{n\alpha(n)r^{n-1}} \int_{B(x,r)} \Delta h(y)dy > 0. \qquad \square$$

In other words, h is harmonic if and only if at each point in U, h is equal to its average over any ball in U centered at that point and also to its average over any sphere centered at that point.

A function $f : U \to \mathbb{R}$ is called *subharmonic* if it is *upper semicontinuous* ($\limsup_{x\to x_0} f(x) \le f(x_0)$ for all $x_0 \in U$) and for each closed ball $\overline{B(x,r)} \subseteq U$ and each harmonic function h defined on a neighborhood of $\overline{B(x,r)}$ for which $f \le h$ on $\partial B(x,r)$, one has $f \le h$ on $B(x,r)$.

* Note that the only harmonic functions in one dimension are linear functions, so subharmonic is the same as convex when $n = 1$.

The analogue of Theorem 2.1 for subharmonic functions is

Theorem 2.2. *An upper semicontinuous function f is subharmonic on a domain $U \subseteq \mathbb{R}^n$ if and only if*

$$f(x) \le \frac{1}{|B(x,r)|} \int_{B(x,r)} f(y)dy, \quad \frac{1}{\sigma\left(\partial B(x,r)\right)} \int_{\partial B(x,r)} f(y)dS(y)$$

for every ball $\overline{B(x,r)} \subseteq U$.

Proof sketch. If f is subharmonic, then it is u.s.c., so there is a sequence of continuous functions $f_n \searrow f$. For any $\overline{B(x,r)} \subseteq U$, let h_n be the harmonic function on $B(x,r)$ with $h_n(x) = f_n(x)$ on $\partial B(x,r)$. (It is known that the Dirichlet problem for the Laplace equation on a ball with continuous boundary values has a classical solution). The mean value formula for harmonic functions and $f \le f_n = h_n$ on $\partial B(x,r)$ gives

$$f(x) \le h_n(x) = \frac{1}{\sigma\left(\partial B(x,r)\right)} \int_{\partial B(x,r)} h_n(y)dS(y) = \frac{1}{\sigma\left(\partial B(x,r)\right)} \int_{\partial B(x,r)} f_n(y)dS(y)$$

for all n, and the result follows from the MCT since upper semicontinuous functions attain their maximum on compact sets and thus the f_n can be taken to be bounded from above. Radial integration takes care of the integral over $B(x,r)$.

Conversely, suppose that f satisfies the given inequality and there is $x_0 \in \overline{B(x,r)} \subseteq U$ and some harmonic function h with $f \le h$ on $\partial B(x,r)$ and $f(x_0) > h(x_0)$. Without loss, we can assume that x_0 maximizes $f - h$ on $\overline{B(x,r)}$.

By assumption, we have

$$f(x_0) - h(x_0) \leq \frac{1}{|B(x_0, s)|} \int_{B(x_0,s)} f(y) - h(y) dy$$

for sufficiently small s, so maximality implies that $f(y) - h(y) = f(x_0) - h(x_0)$ for a.e. y in some neighborhood of x_0. A connectivity argument shows that $f(y) - h(y) = f(x_0) - h(x_0) > 0$ a.e. on $\overline{B(x,r)}$, contradicting $f \leq h$ on $\partial B(x,r)$ (as $f - h$ is u.s.c.). □

A function g is *superharmonic* if and only if $-g$ is subharmonic, so one has analogous mean value inequalities for superharmonic functions.

Roughly, one thinks of martingales as corresponding to harmonic functions while submartingales (respectively, supermartingales) correspond to subharmonic (respectively, superharmonic) functions. For example, if f is subharmonic, then the value of f at x is at most the average of f over a neighborhood of x, which is sort of similar to $X_n \leq E[X_{n+1} \,|\mathcal{F}_n]$.

A concrete example of the correspondence between potential theory and martingales is that if $\{B_t\}_{t\geq 0}$ is a Brownian motion in $\mathbb{R}^n$ (which is a continuous-time martingale w.r.t. $\mathcal{F}_t = \sigma(B_s : s \leq t)$ in the sense that $B_t \in L^1(\mathcal{F}_t)$ and $E[B_t \,|\mathcal{F}_s] = B_s$ for $0 \leq s \leq t$), then under certain integrability assumptions, $Y_t = h(B_t)$ is a martingale (respectively, sub/super-martingale) if h is harmonic (respectively, sub/super-harmonic).

A more accessible connection is given by the following proposition

Proposition 2.1. *Suppose f is continuous and subharmonic on $\mathbb{R}^d$. Let $\xi_1, \xi_2, ...$ be i.i.d. uniform on $B(0,1)$, and define $S_n = x + \sum_{i=1}^n \xi_i$ for some $x \in \mathbb{R}^d$. Then $\{f(S_n)\}$ is a submartingale.*

Proof. The implicit filtration is $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n) \ni S_n$, so, since continuous functions are measurable, $f(S_n)$ is adapted to $\mathcal{F}_n$. Also, since $S_n \in B(x,n) \subseteq \overline{B(x,n)}$ by construction and continuous functions are bounded on compact sets, $E\,|f(S_n)| < \infty$. Finally, by the mean value inequality for subharmonic functions,

$$E\left[f(S_{n+1}) \,|\mathcal{F}_n\right] = E\left[f(S_n + \xi_{n+1}) \,|\mathcal{F}_n\right] = \frac{1}{|B(0,1)|} \int_{B(S_n,1)} f(y) dy \geq f(S_n).$$

Here we are using the fact that if $X \in \mathcal{G}$ and Y is independent of $\mathcal{G}$, then $E[\varphi(X,Y) \,|\mathcal{G}] = h(X)$ where $h(x) = E[\varphi(x,Y)]$. □

First Results.

An easy but extremely important fact is that the martingale property is not limited to a single time step.

Theorem 2.3. *If $\{X_n\}$ is a submartingale, then for any $n > m$, $E\left[X_n \,|\mathcal{F}_m\right] \geq X_m$.*

Proof. When $n = m+1$, the result follows from the definition of a submartingale, so we may assume for the purposes of an induction argument that $E[X_{m+k} \,|\mathcal{F}_m] \geq X_m$. Then the tower property and monotonicity give

$$E\left[X_{m+k+1} \,|\mathcal{F}_m\right] = E\left[E[X_{m+k+1} \,|\mathcal{F}_{m+k}] \,|\mathcal{F}_m\right] \geq E\left[X_{m+k} \,|\mathcal{F}_m\right] \geq X_m.$$ □

Corollary 2.1. *If $\{X_m\}$ is a supermartingale, then $E[X_n|\mathcal{F}_m] \leq X_m$ for any $n > m$, and if $\{X_m\}$ is a martingale, then $E[X_n|\mathcal{F}_m] = X_m$ for any $n > m$.*

Proof. The first claim follows from Theorem 2.3 by noting that $\{-X_m\}$ is a submartingale, and the second then follows since a martingale is both a submartingale and a supermartingale. □

The line of reasoning used in Corollary 2.1 applies to many of our subsequent results, so to avoid redundancy we will often just state a theorem for submartingales or supermartingales with the corresponding statements for the other cases taken as implicit.

Corollary 2.2. *If $\{X_m\}$ is a submartingale, then $E[X_n] \geq E[X_m]$ for any $n > m$, and analogously for supermartingales and martingales.*

Proof. Proposition 1.6, Theorem 2.3, and monotonicity imply

$$E[X_n] = E[E[X_n|\mathcal{F}_m]] \geq E[X_m].$$ □

Keeping in mind that convex functions of one variable are subharmonic, the following results provide another connection between smartingales and potential theory.

Theorem 2.4. *If $\{X_n\}$ is a martingale w.r.t. $\{\mathcal{F}_n\}$ and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function with $E|\varphi(X_n)| < \infty$ for all n, then $\{\varphi(X_n)\}$ is a submartingale w.r.t. $\{\mathcal{F}_n\}$.*

Proof. $\varphi(X_n) \in L^1(\mathcal{F}_n)$ by assumption and Jensen's inequality implies

$$E[\varphi(X_{n+1})|\mathcal{F}_n] \geq \varphi(E[X_{n+1}|\mathcal{F}_n]) = \varphi(X_n)$$

since $\{X_n\}$ is a martingale. □

Corollary 2.3. *Let $p \geq 1$ and suppose that $\{X_n\}$ is a martingale such that $E[|X_n|^p] < \infty$ for all n. Then $\{|X_n|^p\}$ is a submartingale.*

Proof. $f(x) = |x|^p$ is convex. □

Theorem 2.5. *If $\{X_n\}$ is a submartingale w.r.t. $\{\mathcal{F}_n\}$ and φ is a nondecreasing convex function with $E|\varphi(X_n)| < \infty$ for all n, then $\{\varphi(X_n)\}$ is a submartingale w.r.t. $\{\mathcal{F}_n\}$.*

Proof. By Jensen's inequality and the assumptions,

$$E[\varphi(X_{n+1})|\mathcal{F}_n] \geq \varphi(E[X_{n+1}|\mathcal{F}_n]) \geq \varphi(X_n).$$ □

Corollary 2.4. *If $\{X_n\}$ is a submartingale, then $\{(X_n - a)^+\}$ is a submartingale. If $\{X_n\}$ is a supermartingale, then $\{X_n \wedge a\}$ is a supermartingale.*

Proof. $f(x) = x \vee b$ is convex and nondecreasing, and $-[(-x) \vee (-a)] = x \wedge a$. □

In order to further develop the basic notions of martingale theory, we introduce the following definitions:

Definition. Given a filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$, we say that a sequence $\{H_n\}_{n=1}^{\infty}$ is *predictable* (or *previsible*) if $H_n \in \mathcal{F}_{n-1}$ for all $n \in \mathbb{N}$. That is, the value of H_n is determined by the information available at time $n-1$.

Definition. If $\{X_n\}_{n=0}^{\infty}$ is a smartingale and $\{H_n\}_{n=1}^{\infty}$ is predictable w.r.t. $\mathcal{F}_n = \sigma(X_0, ..., X_n)$, the process

$$(H \cdot X)_0 = 0,$$
$$(H \cdot X)_n = \sum_{m=1}^{n} H_m \left(X_m - X_{m-1}\right),\ n \geq 1$$

is called the *martingale transform* of H_n with respect to X_n.

Note that the martingale transform is defined like a Riemann sum for the "integral" of H with respect to X. With a bit of care, this analogy can be extended to develop a general theory of stochastic integration.

We typically think of $\{H_n\}$ as a gambling strategy: If X_n represents the gambler's fortune at time n if the stakes were \$1 per game and H_n represents the amount wagered on the n^{th} game (which can be based on the outcomes of the first $n-1$ games, but not on the outcome of game n), then $(H \cdot X)_n$ is the gambler's fortune at time n when they bet according $\{H_n\}$.

As an example, take $X_n = \sum_{i=0}^{n} \xi_i$ where $\xi_0 = 0$ and $\xi_1, \xi_2, ...$ are i.i.d. with $P(\xi_1 = 1) = p \in (0,1)$, $P(\xi_1 = -1) = 1-p$, and set $H_1 = 1$, $H_n = \begin{cases} 1, & \xi_{n-1} = 1 \\ 2H_{n-1}, & \xi_{n-1} = -1 \end{cases}$ for $n \geq 2$.

This corresponds to our example of the martingale betting system, except that instead of quitting after a win and walking away with one dollar, the process just starts over. Every win recovers the amount from the last losing streak and one additional dollar to boot, so the gambler has a profit of $\$k$ after the k^{th} win.
As the second Borel-Cantelli lemma ensures infinitely many wins with probability one, the gambler can eventually amass an arbitrarily large fortune no matter how much the odds are stacked against them.

However, the above reasoning is predicated on unrealistic assumptions such as infinite credit and no table limits. The ensuing results show that in practice there is no system for beating an unfavorable game.

Theorem 2.6. *Suppose that $\{X_n\}_{n=0}^{\infty}$ is a supermartingale. If $H_n \geq 0$ is predictable and each H_n is bounded, then $(H \cdot X)_n$ defines a supermartingale.*

Proof. $(H \cdot X)_n$ is integrable since H is bounded and the X_m's are integrable, and it is adapted by construction. To see that it defines a supermartingale, we compute

$$\begin{aligned} E\left[(H \cdot X)_{n+1} \,|\mathcal{F}_n\right] &= E\left[(H \cdot X)_n + H_{n+1}(X_{n+1} - X_n) \,|\mathcal{F}_n\right] \\ &= E\left[(H \cdot X)_n \,|\mathcal{F}_n\right] + E\left[H_{n+1}(X_{n+1} - X_n) \,|\mathcal{F}_n\right] \\ &= (H \cdot X)_n + H_{n+1} E\left[X_{n+1} - X_n \,|\mathcal{F}_n\right] \leq (H \cdot X)_n \end{aligned}$$

where we used $H_{n+1}, (H \cdot X)_n \in \mathcal{F}_n$, $H_n \geq 0$, and $E\left[X_{n+1} - X_n \,|\mathcal{F}_n\right] \leq 0$. □

Of course, the argument in Theorem 2.6 also applies to submartingales and to martingales (without the $H_n \geq 0$ restriction in the latter case).

An important example of a predictable sequence is given by $H_n = 1\{N \geq n\}$ where N is a stopping time.

The terminology is quite apt as the strategy is "stop gambling at time N."

It is worth observing that $(H \cdot X)$ is linear in H (and in X for that matter), so one can construct very general betting systems using stopping times.

Note that if $\{X_n\}$ is a supermartingale and $H_n = 1\{N \geq n\}$, then Theorem 2.6 shows that

$$(H \cdot X)_n = \sum_{m=1}^{n} 1\{N \geq m\}(X_m - X_{m-1}) = X_{N \wedge n} - X_0$$

is a supermartingale.

Since the constant sequence $Y_n = X_0$ is also a supermartingale and the sum of two supermartingales is a supermartingale, we have

Corollary 2.5. *If N is a stopping time and $\{X_n\}$ is a supermartingale, then $\{X_{N \wedge n}\}$ is a supermartingale, and similarly for submartingales and martingales.*

Corollary 2.5 gives us our first "optional stopping theorem," which explains why the martingale system doesn't work if we can't play forever.

Theorem 2.7. *If $\{X_n\}_{n=0}^{\infty}$ is a supermartingale and N is a stopping time with $N \leq m$ a.s. for some $m \in \mathbb{N}$, then $E[X_N] \leq E[X_0]$.*

Proof. Under the above assumptions,

$$E[X_N] = E[X_{N \wedge m}] \leq E[X_0]. \qquad \square$$

Now suppose that X_n, $n \geq 0$ is a submartingale. Fix $a, b \in \mathbb{R}$ with $a < b$ and define the sequence $N_0, N_1, N_2, \ldots$ by $N_0 = -1$ and for $k \geq 1$,

$$N_{2k-1} = \inf\{m > N_{2k-2} : X_m \leq a\}$$
$$N_{2k} = \inf\{m > N_{2k-1} : X_m \geq b\}.$$

The $N_j's$ are stopping times and $\{N_{2k-1} < m \leq N_{2k}\} = \{N_{2k-1} \leq m-1\} \cap \{N_{2k} \leq m-1\}^C \in \mathcal{F}_{m-1}$, so $H_m = 1\{N_{2k-1} < m \leq N_{2k} \text{ for some } k \geq 1\}$ defines a predictable sequence.

By construction, $X_{N_{2k-1}} \leq a$ and $X_{N_{2k}} \geq b$, so in the time interval $[N_{2k-1}, N_{2k}]$, X_m crosses from below a to above b.

$\{H_m\}_{m=1}^{\infty}$ is a gambling strategy which tries to take advantage of these "upcrossings":

It turns on after X_m dips below a and turns off once X_m rises above b. As the sum defining $(H \cdot X)_m$ telescopes in between these times, the contribution from an upcrossing starting at time N_{2k-1} and ending at time N_{2k} is $X_{N_{2k}} - X_{N_{2k-1}} \geq b - a$.

In stock market terms, this is like buying a share once its price is less than a and selling once it's greater than b.

Letting $U_n = \sup\{k : N_{2k} \leq n\}$ denote the number of upcrossings completed by time n and
$V_n = \sup\{j : N_{2j-1} \leq n\}$ the number of upcrossings begun before time n, we see that

$$(H \cdot X)_n \geq U_n(b-a) + 1\{U_n < V_n\}\left(X_n - X_{N_{2V_n-1}}\right).$$

The following lemma extends this observation to bound the expected number of upcrossings by time n.

Lemma 2.1 (Upcrossing Inequality). *If $\{X_m\}_{m=0}^{\infty}$ is a submartingale, then*

$$(b-a)E[U_n] \leq E\left[(X_n - a)^+\right] - E\left[(X_0 - a)^+\right].$$

Proof. Since $\varphi(x) = x \vee a$ is convex and nondecreasing, $Y_m = X_m \vee a$ is a submartingale.

Also, $Y_m = a$ when $X_m \leq a$ and $Y_m = X_m$ when $X_m > a$, so Y_m upcrosses $[a, b]$ the same number of times X_m does.

Moreover, $Y_m \geq a = Y_{N_{2V_n-1}}$ implies

$$(H \cdot Y)_n \geq U_n(b-a) + 1\{U_n < V_n\}\left(Y_n - Y_{N_{2V_n-1}}\right) \geq U_n(b-a).$$

Setting $K_m = 1 - H_m$, we have that $\{K_m\}$ is nonnegative, bounded, and predictable, so Corollary 2.2 and the submartingale form of Theorem 2.6 yield

$$E[(K \cdot Y)_n] \geq E\left[(K \cdot Y)_0\right] = 0.$$

Putting these facts together gives

$$\begin{aligned}(b-a)E[U_n] &\leq E\left[(H \cdot Y)_n\right] = E\left[(1 \cdot Y)_n - (K \cdot Y)_n\right] \\ &= E[Y_n - Y_0] - E[(K \cdot Y)_n] \leq E[Y_n] - E[Y_0] \\ &= E\left[(X_n - a)^+ + a\right] - E\left[(X_0 - a)^+ + a\right].\end{aligned}$$

□

The primary utility of the upcrossing inequality is to facilitate the proof of the pointwise martingale convergence theorem, which shows that smartingales behave like monotone sequences of real numbers in that they converge if appropriately bounded.

Essentially, the idea is to show that there is a set of full measure on which X_n upcrosses any interval $[a, b]$ with $a, b \in \mathbb{Q}$ only finitely many times. This allows us to conclude that X_n has an almost sure limit.

Theorem 2.8 (Martingale Convergence). *If $\{X_n\}$ is a submartingale with $\sup_n E[X_n^+] < \infty$, then there is an integrable random variable X such that $X_n \to X$ a.s. as $n \to \infty$.*

Proof.

For any $a, b \in \mathbb{Q}$ with $a < b$, let U_n be the number of upcrossings of $[a, b]$ by time n, and let $U = \lim_{n\to\infty} U_n$ be the number of upcrossings of the whole sequence. (U is a well-defined random variable because U_n is increasing.) Set $M = \sup_n E[X_n^+] < \infty$.

Since $E\left[(X_0 - a)^+\right] \geq 0$ and $E\left[(X_n - a)^+\right] \leq E\left[X_n^+\right] + |a| \leq M + |a|$, Lemma 2.1 implies that

$$E[U_n] \leq \frac{1}{b-a}\left(E\left[(X_n - a)^+\right] - E\left[(X_0 - a)^+\right]\right) \leq \frac{M + |a|}{b - a}$$

for all n, so the monotone convergence theorem gives

$$E[U] = \lim_{n\to\infty} E[U_n] \leq \frac{M + |a|}{b - a} < \infty,$$

and thus $U < \infty$ a.s.

As the above holds for any $[a, b]$ with $a, b \in \mathbb{Q}$, and the set of intervals with rational endpoints is countable, the event

$$\bigcup_{a,b\in\mathbb{Q}} \{\liminf_n X_n < a < b < \limsup_n X_n\}$$

has probability 0, hence $\limsup_n X_n = \liminf_n X_n$ a.s.

Letting X denote this common value, Fatou's lemma shows that $E[X^+] \leq \liminf_n E[X_n^+] < \infty$, so $X < \infty$ a.s. To see that $X > -\infty$ a.s., we observe that since $\{X_n\}$ is a submartingale,

$$E[X_n^-] = E[X_n^+] - E[X_n] \leq E[X_n^+] - E[X_0],$$

so another application of Fatou's lemma gives

$$E[X^-] \leq \liminf_n E[X_n^-] \leq \liminf_n E[X_n^+] - E[X_0] < \infty.$$

Thus X_n has a limit $X \in \mathbb{R}$ with $E|X| = E[X^+] + E[X^-] < \infty$. □

An immediate corollary is

Corollary 2.6. *If $X_n \geq 0$ is a supermartingale, then as $n \to \infty$, $X_n \to X$ a.s. and $E[X] \leq E[X_0]$.*

Proof. $Y_n = -X_n$ is a submartingale with $E[Y_n^+] = 0$, so it follows from Theorem 2.8 that $Y_n \to Y = -X$ a.s. The inequality follows from Fatou's lemma and the supermartingale property:

$$E[X] \leq \liminf_n E[X_n] \leq E[X_0].$$ □

Of course, by considering $X_n + K$, one may replace non-negative with bounded from below in the preceding.

It is worth observing that the assumptions in Theorem 2.8 do not guarantee convergence in L^1.

Example 2.7. Let $S_0 = 1$, and for $n \geq 1$, $S_n = S_{n-1} + \xi_n$ where $\xi_1, \xi_2, \ldots$ are i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$. That is, S_n is a simple random walk started at 1. Let $N = \inf\{n : S_n = 0\}$ be the hitting time of 0, and let $X_n = S_{n\wedge N}$ be the walk stopped at 0. Then X_n is a nonnegative martingale, so Corollary 2.6 implies that it has an almost sure limit X. Clearly, we must have $X = 0$ a.s. since if $X_n = k > 0$, then $|X_n - X_{n+1}| = 1$. However, $E[X_n] = E[X_0] = 1$ for all n, so we can't have $X_n \to X$ in L^1.

We will provide criteria for convergence in L^p before long, but before doing so we have one more general result to discuss, and then we will take some time to consider some examples in detail in order to better understand the utility of martingale convergence.

The following decomposition result is due to Doob (like essentially everything else in the classical theory of martingales) and can be useful for reducing questions about sub/super-martingales to questions about martingales.

Theorem 2.9 (Doob Decomposition)**.** *Any submartingale $\{X_n\}_{n=0}^{\infty}$ can be written in a unique way as $X_n = M_n + A_n$ where M_n is a martingale and A_n is a predictable increasing sequence with $A_0 = 0$.*

Proof. If $X_n = M_n + A_n$ with $E[M_n \,|\mathcal{F}_{n-1}] = M_{n-1}$ and $A_n \in \mathcal{F}_{n-1}$, we must have

$$\begin{aligned} E[X_n \,|\mathcal{F}_{n-1}] &= E[M_n \,|\mathcal{F}_{n-1}] + E[A_n \,|\mathcal{F}_{n-1}] \\ &= M_{n-1} + A_n = X_{n-1} - A_{n-1} + A_n. \end{aligned}$$

The recursion

$$\begin{aligned} A_0 &= 0, \\ A_n - A_{n-1} &= E[X_n \,|\mathcal{F}_{n-1}] - X_{n-1} \end{aligned}$$

uniquely defines A_n and thus $M_n = X_n - A_n$.

It remains to establish existence by checking that A_n and M_n as defined above satisfy the assumptions.

To check that A_n is indeed increasing and predictable, we note that $A_n - A_{n-1} = E[X_n \,|\mathcal{F}_{n-1}] - X_{n-1} \geq 0$ since X_n is a submartingale, and $A_n = E[X_n \,|\mathcal{F}_{n-1}] - X_{n-1} + A_{n-1} \in \mathcal{F}_{n-1}$ by induction.

Finally, rewriting the recursion for A_n as $E[X_n \,|\mathcal{F}_{n-1}] - A_n = X_{n-1} - A_{n-1}$, we see that

$$E[M_n \,|\mathcal{F}_{n-1}] = E[X_n - A_n \,|\mathcal{F}_{n-1}] = E[X_n \,|\mathcal{F}_{n-1}] - A_n = X_{n-1} - A_{n-1} = M_n,$$

so M_n is a martingale. □

The inspiration for the theorem is quite clear:

$$A_n = \sum_{k=1}^{n} \left(E[X_k \,|\mathcal{F}_{k-1}] - X_{k-1}\right)$$

is the running sum of the amount by which $E[X_n \,|\mathcal{F}_{n-1}]$ overshoots X_{n-1}, and once these "drift terms" have been subtracted off, the remainder

$$M_n = X_0 + \sum_{k=1}^{n} \left(X_k - E[X_k \,|\mathcal{F}_{k-1}]\right)$$

is a martingale.

3. Applications

Bounded Increments.

Our first application of the martingale convergence theorem involves a generalization of the second Borel-Cantelli lemma.

This will follow from a more general result concerning martingales with bounded increments:
They either converge or oscillate between $\pm\infty$.

Theorem 3.1. *Suppose $\{X_n\}$ is a martingale such that $\sup_n |X_{n+1} - X_n| \leq M$ for some $M \in \mathbb{R}$. Writing*

$$C = \Big\{ \lim_{n\to\infty} X_n \text{ exists in } \mathbb{R} \Big\},$$
$$D = \Big\{ \limsup_{n\to\infty} X_n = \infty \Big\} \cap \Big\{ \liminf_{n\to\infty} X_n = -\infty \Big\},$$

we have $P(C \cup D) = 1$.

Proof. Since $X_n - X_0$ is a martingale, we may assume without loss of generality that $X_0 = 0$.

For any $0 < K < \infty$, set $N = \inf\{n : X_n \leq -K\}$. Then $X_{n\wedge N}$ is a martingale with $X_{n\wedge N} \geq -K - M$ a.s., so applying Corollary 2.6 to $X_{n\wedge N} + K + M$ shows that X_n has a finite limit on (almost all of) $\{N = \infty\}$.

Letting $K \to \infty$ shows that $\lim_{n\to\infty} X_n$ exists a.s. on $\{\liminf_{n\to\infty} X_n > -\infty\}$.

Applying the foregoing to $-X_n$ shows that $\lim_{n\to\infty} X_n$ exists a.s. on $\{\limsup_{n\to\infty} X_n < \infty\}$.

Thus, up to a null set, $C = \Omega \setminus D$. □

Corollary 3.1. *Let $\{\mathcal{F}_n\}_{n=0}^{\infty}$ be a filtration with $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and let $A_1, A_2, \ldots$ be events with $A_n \in \mathcal{F}_n$. Then (up to a null set)*

$$\{A_n \text{ i.o.}\} = \Big\{ \sum_{n=1}^{\infty} P(A_n \,|\mathcal{F}_{n-1}) = \infty \Big\}.$$

Proof. Set $X_0 = 0$ and $X_n = \sum_{m=1}^{n} (1_{A_m} - P(A_m \,|\mathcal{F}_{m-1}))$ for $n \geq 1$. Then

$$\begin{aligned} E[X_{n+1} \,|\mathcal{F}_n] &= E\left[X_n + 1_{A_{n+1}} - P(A_{n+1} \,|\mathcal{F}_n) \,|\mathcal{F}_n\right] \\ &= X_n + E\left[1_{A_{n+1}} \,|\mathcal{F}_n\right]) - P\left(A_{n+1} \,|\mathcal{F}_n\right) = X_n, \end{aligned}$$

so X_n is a martingale with $|X_{n+1} - X_n| = \left|1_{A_{n+1}} - P\left(A_{n+1} \,|\mathcal{F}_n\right)\right| \leq 1$.

On $C = \{\lim_{n\to\infty} X_n \text{ exists and is finite}\}$, we must have

$$\sum_{n=1}^{\infty} 1_{A_m} = \infty \text{ if and only if } \sum_{n=1}^{\infty} P\left(A_n \,|\mathcal{F}_{n-1}\right) = \infty,$$

and the same is true on $D = \{\limsup_n X_n = \infty \text{ and } \liminf_n X_n = -\infty\}$.
This proves the result since $P(C \cup D) = 1$ by Theorem 3.1. □

*** Polya's Urn.**

Suppose that an urn initially contains r red balls and g green balls. At each time step we remove a ball at random, observe its color, and then replace it along with c other balls of the same color for some integer c.

Negative values of c correspond to removing balls, and in this case it is generally necessary to stop the process once further removals become impossible.

Note that $c = -1$ corresponds to sampling without replacement and $c = 0$ corresponds to sampling with replacement.

For positive values of c, each time a color is sampled increases the probability that it will be sampled in the future. Such self-reinforcing behavior is sometimes described by the phrase "the rich get richer." (Of course, the opposite holds when c is negative.)

We will assume henceforth that $c > 0$ so that the process can be continued indefinitely and so we can divide by c when convenient.

To make the foregoing rigorous, let $\xi_n = 1\{\text{green ball drawn at time } n\}$. The first observation is that the sequence $\xi_1, \xi_2, ...$, though certainly not independent, is *exchangeable* - for any $n \in \mathbb{N}$, $\sigma \in S_n$, $(\xi_1, ..., \xi_n) =_d \left(\xi_{\sigma(1)}, ..., \xi_{\sigma(n)}\right)$.

To gain intuition, we compute $P(\xi_1 = 1, \xi_2 = 0, \xi_3 = 0) = P(\xi_1 = 0, \xi_2 = 1, \xi_3 = 0)$:

Using elementary conditional probability, the left-hand side is $\left(\frac{g}{r+g}\right)\left(\frac{r}{r+g+c}\right)\left(\frac{r+c}{r+g+2c}\right)$ and the right-hand side is $\left(\frac{r}{r+g}\right)\left(\frac{g}{r+g+c}\right)\left(\frac{r+c}{r+g+2c}\right)$. The successive denominators are constant and the numerator are just permuted.

More generally,

$$P(\xi_1 = 1, ..., \xi_m = 1, \xi_{m+1} = 0, ..., \xi_n = 0)$$
$$= \left(\frac{g}{g+r}\right) \cdots \left(\frac{g+(m-1)c}{g+r+(m-1)c}\right)\left(\frac{r}{g+r+mc}\right) \cdots \left(\frac{r+(n-m-1)c}{g+r+(n-1)c}\right),$$

which is the same as $P(\xi_1 = 1_S(1), ..., \xi_n = 1_S(n))$ for any other $S \subseteq [n]$ with $|S| = m$.

Now let $X_n = \sum_{i=1}^n \xi_n$ be the number of green balls drawn at time n. The preceding discussion shows that for any $0 \le m \le n$, $P(X_n = m) = \binom{n}{m} p_{n,m}$ where

$$p_{n,m} = P(\xi_1 = 1, ..., \xi_m = 1, \xi_{m+1} = 0, ..., \xi_n = 0) = \frac{\prod_{i=0}^{m-1}(g+ic)\prod_{j=0}^{n-m-1}(r+jc)}{\prod_{k=0}^{n-1}(g+r+kc)}$$
$$= \frac{\prod_{i=0}^{m-1}\left(\frac{g}{c}+i\right)\prod_{j=0}^{n-m-1}\left(\frac{r}{c}+j\right)}{\prod_{k=0}^{n-1}\left(\frac{g+r}{c}+k\right)} = \frac{\frac{\Gamma\left(\frac{g}{c}+m\right)\Gamma\left(\frac{r}{c}+n-m\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)}}{\frac{\Gamma\left(\frac{g+r}{c}+n\right)}{\Gamma\left(\frac{g+r}{c}\right)}} = \frac{\Gamma\left(\frac{g}{c}+m\right)\Gamma\left(\frac{r}{c}+n-m\right)}{\Gamma\left(\frac{g+r}{c}+n\right)} \cdot \frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)}.$$

Consequently, we have

$$P(X_n = m) = \binom{n}{m} p_{n,m} = \frac{\Gamma(n+1)}{\Gamma(m+1)\Gamma(n-m+1)} p_{n,m}$$
$$= \frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)} \cdot \frac{\Gamma\left(m+\frac{g}{c}\right)}{\Gamma(m+1)} \cdot \frac{\Gamma\left(n-m+\frac{r}{c}\right)}{\Gamma(n-m+1)} \cdot \frac{\Gamma(n+1)}{\Gamma\left(n+\frac{g+r}{c}\right)}.$$

Stirling's approximation, $\Gamma(x+1) \approx \sqrt{2\pi x}\left(\frac{x}{e}\right)^x$, implies

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} \approx \frac{\sqrt{2\pi(x+a-1)}\left(\frac{x+a-1}{e}\right)^{x+a-1}}{\sqrt{2\pi(x+b-1)}\left(\frac{x+b-1}{e}\right)^{x+b-1}} \approx \frac{\left(1+\frac{a-1}{x}\right)^{x+a-1}}{\left(1+\frac{b-1}{x}\right)^{x+b-1}} \left(\frac{x}{e}\right)^{a-b} \approx e^{a-b}\left(\frac{x}{e}\right)^{a-b} = x^{a-b},$$

so

$$nP(X_n = m) \approx n\frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)} m^{\frac{g}{c}-1}(n-m)^{\frac{r}{c}-1} n^{1-\frac{g}{c}-\frac{r}{c}} = \frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)} \left(\frac{m}{n}\right)^{\frac{g}{c}-1} \left(1-\frac{m}{n}\right)^{\frac{r}{c}-1},$$

and thus

$$nP(X_n = m) \to \frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)} x^{\frac{g}{c}-1}(1-x)^{\frac{r}{c}-1}$$

as $m, n \to \infty$ with $\frac{m}{n} \to x$.

It follows that for any $0 < x < 1$,

$$P\left(\frac{X_n}{n} \le x\right) = P(X_n \le nx) = P(X_n = 0) + P(X_n = 1) + ... + P(X_n = \lfloor nx \rfloor)$$
$$= \int_0^{\lfloor nx \rfloor + 1} P(X_n = \lfloor y \rfloor) dy = \int_0^{\frac{\lfloor nx \rfloor + 1}{n}} nP(X_n = \lfloor nu \rfloor) du$$
$$\to \int_0^x \frac{\Gamma\left(\frac{g+r}{c}\right)}{\Gamma\left(\frac{g}{c}\right)\Gamma\left(\frac{r}{c}\right)} u^{\frac{g}{c}-1}(1-u)^{\frac{r}{c}-1} du$$

as $n \to \infty$.

This shows that $\frac{X_n}{n} \Rightarrow \text{Beta}\left(\frac{g}{c}, \frac{r}{c}\right)$, which implies that the fraction of green balls at time n, $Y_n = \frac{g + cX_n}{g + r + cn}$, satisfies

$$Y_n = \frac{X_n}{n}\left(\frac{cn}{r+g+cn}\right) + \frac{g}{g+r+cn} \Rightarrow \text{Beta}\left(\frac{g}{c}, \frac{r}{c}\right).$$

The martingale convergence theorem allows us to strengthen this conclusion to almost sure convergence:

If we can show that $Y_n \ge 0$ is a martingale, then it will follow from Corollary 2.6 that it has an almost sure limit Y, which necessarily has the Beta $\left(\frac{g}{c}, \frac{r}{c}\right)$ distribution by the previous argument.

The obvious filtration in this case is $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$. Y_n is adapted by the Doob-Dynkin Lemma and is integrable since $0 \le Y_n \le 1$, so it remains only to show that $E[Y_{n+1} | \mathcal{F}_n] = Y_n$.

To see that this is so, we observe that on $\left\{Y_n = \frac{g_n}{g_n + r_n}\right\}$,

$$E[Y_{n+1} | \mathcal{F}_n] = \frac{g_n + c}{g_n + r_n + c} \cdot \frac{g_n}{g_n + r_n} + \frac{g_n}{g_n + r_n + c} \cdot \frac{r_n}{g_n + r_n}$$
$$= \frac{g_n(g_n + c + r_n)}{(g_n + r_n + c)(g_n + r_n)} = \frac{g_n}{g_n + r_n} = Y_n.$$

Branching Processes.

Let $\{\xi_i^n\}_{i,n\in\mathbb{N}}$ be an array of i.i.d. $\mathbb{N}_0$-valued random variables. Define the sequence $Z_0, Z_1, ...$ by $Z_0 = 1$ and

$$Z_{n+1} = \begin{cases} \xi_1^{n+1} + ... + \xi_{Z_n}^{n+1}, & Z_n > 0 \\ 0, & Z_n = 0 \end{cases}.$$

Z_n is called a *Galton-Watson process*, and the interpretation is that Z_n gives the population size of the n^{th} generation where each individual gives birth to an identically distributed number of offspring and then dies. $p_k = P(\xi_i^n = k)$ is called the *offspring distribution*.

One can come up with many natural variants on this problem by changing assumptions such as i.i.d. offspring variables or one generation life-spans, but we will content ourselves with the simplest case here.

Our ultimate goal is to show that the mean of the offspring distribution determines whether or not the population is doomed to extinction. Specifically, if the population is to have a chance of surviving indefinitely, then the number of births must exceed the number of deaths on average.

In order to prove this unsurprising yet nontrivial claim, we first establish

Lemma 3.1. *Let $\mathcal{F}_n = \sigma(\xi_i^m : i \geq 1, 1 \leq m \leq n)$ and $\mu = E[\xi_i^n] \in (0, \infty)$. Then $M_n := \dfrac{Z_n}{\mu^n}$ is a martingale with respect to $\mathcal{F}_n$.*

Proof. M_n is adapted by construction and integrable by an induction argument using the ensuing computation. Linearity, monotone convergence, and the definition of $\mathcal{F}_n$ imply

$$\begin{aligned} E[Z_{n+1} \,|\mathcal{F}_n] &= E\left[\sum_{k=0}^{\infty} Z_{n+1} 1\{Z_n = k\} \,|\mathcal{F}_n\right] = \sum_{k=0}^{\infty} E[Z_{n+1} 1\{Z_n = k\} \,|\mathcal{F}_n] \\ &= \sum_{k=1}^{\infty} E\left[\left(\xi_1^{n+1} + ... + \xi_k^{n+1}\right) 1\{Z_n = k\} \,|\mathcal{F}_n\right] \\ &= \sum_{k=1}^{\infty} 1\{Z_n = k\} \, E\left[\xi_1^{n+1} + ... + \xi_k^{n+1}\right] \\ &= \sum_{k=1}^{\infty} 1\{Z_n = k\} \, k\mu = \mu Z_n, \end{aligned}$$

and the result follows upon division by μ^{n+1}. □

Since M_n is a nonnegative martingale, it has a finite almost sure limit M_∞. We begin by identifying cases where $M_\infty = 0$.

Theorem 3.2. *If $\mu < 1$, then $M_n, Z_n \to 0$ a.s.*

Proof. Since $M_n \to M_\infty \in \mathbb{R}$ a.s. and $\mu^n \to 0$, we have that $Z_n = \mu^n M_n \to 0 \cdot M_\infty = 0$ a.s.

Because Z_n is integer valued and converges to 0 a.s., we must have that $Z_n(\omega) = 0$ for all large n (depending on ω) for almost every $\omega \in \Omega$. It follows that $M_n = \frac{Z_n}{\mu^n}$ is eventually 0, hence $M_\infty = 0$. □

It makes sense that if the population has more deaths than births on average, then it will eventually die out.

The next result shows that if we exclude the trivial case where every individual has one offspring with full probability, then the same is also true if, on average, each individual replaces themselves before dying.

Theorem 3.3. *If $\mu = 1$ and $P(\xi_i^n = 1) < 1$, then $Z_n \to 0$ a.s.*

Proof. When $\mu = 1$, $Z_n = M_n$ converges almost surely to some finite Z_∞.

Because $Z_n \in \mathbb{N}_0$, we must have that $Z_n = Z_\infty$ for $n \geq N(\omega)$.

If $P(\xi_i^n = 1) < 1$, then $\mu = 1$ implies that $P(\xi_i^n = 0) > 0$, so for each $k \in \mathbb{N}$, $P(\xi_1^n = ... = \xi_k^n = 0 \text{ i.o.}) = 1$ by the second Borel-Cantelli lemma. Thus for any $k, N \in \mathbb{N}$,

$$P(Z_n = k \text{ for all } n \geq N) \leq 1 - P(\xi_1^n = ... = \xi_k^n = 0 \text{ for some } n > N) = 0,$$

so it must be the case that $Z_\infty = 0$. □

It is worth observing that since $M_n = \dfrac{Z_n}{\mu^n}$ is a martingale, $E\left[\dfrac{Z_n}{\mu^n}\right] = E\left[\dfrac{Z_0}{\mu^0}\right] = 1$, hence $E[Z_n] = \mu^n$.
Thus when $\mu < 1$, $Z_n \to 0$ (exponentially fast) in L^1. However, when $\mu = 1$, $E[Z_n] = E[Z_0] = 1$ for all n, so Z_n does not converge in L^1.

In the $\mu \leq 1$ cases, we were able to conclude not only that $Z_n \to 0$ a.s., but also that $M_n \to 0$ a.s.

When $\mu > 1$, we will show that there is positive probability that the population never goes extinct, but we point out that this does not enable us to conclude that M_∞ is nonzero with positive probability. Necessary and sufficient conditions for this to occur are stated in Durrett.

Theorem 3.4. *If $\mu > 1$, then $P(Z_n > 0 \text{ for all } n) > 0$.*

Proof. For $s \in [0, 1]$, let

$$G(s) = E\left[s^{\xi_i^n}\right] = \sum_{k=0}^{\infty} p_k s^k$$

be the *probability generating function* for the offspring distribution.

As $\sum_{k=0}^{\infty} p_k s^k$ is a power series whose interval of convergence contains $[-1, 1]$, we may differentiate termwise to obtain

$$G'(s) = \sum_{k=1}^{\infty} k p_k s^{k-1} \geq 0$$

$$G''(s) = \sum_{k=2}^{\infty} k(k-1) p_k s^{k-2} \geq 0$$

for $s \in [0, 1)$.

Since the DCT implies that $\lim_{s \to 1^-} G'(s) = \sum_{k=1}^{\infty} k p_k = \mu > 1$, there is a $k \geq 2$ such that $p_k > 0$, hence $G'(s), G''(s) > 0$ for $s \in (0, 1)$.

The reason we care about the p.g.f. is that the proof relies on showing that the probability of extinction is given by the unique fixed point of G in $[0, 1)$.

Claim (a). If $\theta_m = P(Z_m = 0)$, then $\theta_m = \sum_{k=0}^{\infty} p_k \theta_{m-1}^k = G(\theta_{m-1})$.

Proof. If $Z_1 = k$, which happens with probability p_k, then $Z_m = 0$ if and only if all k lineages die out in the next $m-1$ time steps. By independence, this happens with probability θ_{m-1}^k. Summing over the different possibilities for k establishes the claim. □

Our next step is to establish the existence and uniqueness of the purported fixed point:

Claim (b). There is a unique $\rho < 1$ such that $G(\rho) = \rho$.

Proof. Since $\lim_{s \to 1^-} G'(s) = \mu > 1$, and G' is continuous on $[0,1)$, there is an $\varepsilon > 0$ such that $G'(s) > 1$ on $(1-\varepsilon, 1)$. As G is continuous on $[1-\varepsilon, 1]$ and $G(1) = 1$, the mean value theorem implies that

$$1 - G(s) = G(1) - G(s) = G'(c)(1-s) > 1 - s,$$

hence $G(s) < s$, on $(1-\varepsilon, 1)$.

Also, $G(0) = p_0 \geq 0$. If $G(0) = 0$, take $\rho = 0$. Otherwise, letting $F(x) = G(x) - x$, we have $F(0) > 0$ and $F(1 - \frac{\varepsilon}{2}) < 0$, so the existence of ρ follows from the intermediate value theorem.

To see that this is the only fixed point less than 1, observe that $G'' > 0$ on $(0,1)$, so G is strictly convex and thus for any $x \in (\rho, 1)$,

$$G(x) = G\left(\lambda\rho + (1-\lambda)\cdot 1\right) < \lambda G(\rho) + (1-\lambda)G(1) = \lambda\rho + (1-\lambda)\cdot 1 = x.$$

There can be no fixed point in $(0, \rho)$ either as the above computation (with x and ρ interchanged) would then imply that $G(\rho) < \rho$. □

The final step is to show that ρ is indeed the extinction probability:

Claim (c). $\theta_m \nearrow \rho$ as $m \nearrow \infty$.

Proof. $\theta_n = G(\theta_{n-1})$ is an increasing sequence because $\theta_0 = 0 \leq G(0) = \theta_1$ and $G'(x) \geq 0$ for $x \geq 0$.

Similarly, $\sup_n \theta_n \leq \rho$ since $\theta_0 = 0 \leq \rho$ and $\theta_m \leq \rho$ implies $\theta_{m+1} = G(\theta_m) \leq G(\rho) = \rho$.

As θ_n is a bounded increasing sequence, it converges to some $\theta_\infty \leq \rho$, so, since G is continuous, we have

$$\theta_\infty = \lim_{n\to\infty} \theta_n = \lim_{n\to\infty} G(\theta_{n-1}) = G(\theta_\infty).$$

Because $\theta_\infty \leq \rho$ is a fixed point of G, the previous claim implies $\theta_\infty = \rho$. □

Finally, $\{Z_m = 0\} \subseteq \{Z_{m+1} = 0\}$ for all m, so continuity from below implies

$$\begin{aligned} P(Z_n = 0 \text{ for some } n) &= P\left(\bigcup_{m=1}^{\infty} \{Z_m = 0\}\right) = \lim_{m\to\infty} P(Z_m = 0) \\ &= \lim_{m\to\infty} \theta_m = \rho < 1 \end{aligned}$$

and the proof is complete. □

4. L^p Convergence

In order to obtain an L^p version of Theorem 2.8 we begin with a generalization of Theorem 2.7.

Theorem 4.1. *If X_n is a submartingale, M and N are stopping times with $M \leq N$, and there is a $k \in \mathbb{N}$ with $P(N \leq k) = 1$, then $E[X_M] \leq E[X_N]$.*

Proof. Since $\{M < n \leq N\} = \{M \leq n-1\} \cap \{N \leq n-1\}^C \in \mathcal{F}_{n-1}$, $K_n = 1\{M < n \leq N\} \geq 0$ is predictable, hence

$$(K \cdot X)_n = \sum_{m=1}^{n} 1\{M < m \leq N\}(X_m - X_{m-1}) = \sum_{m=M\wedge n+1}^{N\wedge n} (X_m - X_{m-1}) = X_{N\wedge n} - X_{M\wedge n}$$

is a submartingale.

Consequently,

$$E[X_N] - E[X_M] = E[X_{N\wedge k}] - E[X_{M\wedge k}] = E[(K \cdot X)_k] \geq E[(K \cdot X)_0] = 0. \qquad \square$$

In particular, we have

Corollary 4.1. *If X_n is a submartingale and N is a stopping time with $P(N \leq k) = 1$ for some $k \in \mathbb{N}$, then $E[X_0] \leq E[X_N] \leq E[X_k]$.*

Proof. For the first inequality, take $M = 0$ in Theorem 4.1.

For the second, take $M = N$, $N = k$. $\square$

Our next step in proving the L^p martingale convergence theorem is

Theorem 4.2 (Doob's Inequality)**.** *Let X_n be a submartingale, $\widetilde{X}_n = \max_{0\leq m\leq n} X_m^+$, $\lambda > 0$, and $A = \{\widetilde{X}_n \geq \lambda\}$. Then*

$$\lambda P(A) \leq E[X_n 1_A] \leq E[X_n^+ 1_A] \leq E[X_n^+].$$

Proof. Let $N = \inf\{m : X_m \geq \lambda\} \wedge n$. If $\omega \in A$, then there is a smallest $m \leq n$ such that

$$\lambda \leq X_m(\omega) = X_{N(\omega)}(\omega).$$

Also, since $X_N = X_n$ on A^C, Corollary 4.1 implies

$$E[X_N 1_A] + E[X_n 1_{A^C}] = E[X_N 1_A] + E[X_N 1_{A^C}] = E[X_N] \leq E[X_n] = E[X_n 1_A] + E[X_n 1_{A^C}].$$

Thus, as in the proof of Chebychev's inequality, we have

$$\lambda P(A) = E[\lambda 1_A] \leq E[X_N 1_A] \leq E[X_n 1_A].$$

The other inequalities are trivial since $X_n 1_A \leq X_n^+ 1_A \leq X_n^+$. $\square$

A tangential application of Doob's inequality is

Theorem 4.3 (Kolmogorov's Maximal Inequality)**.** *If $\xi_1, \xi_2, \ldots$ are independent with $E[\xi_i] = 0$ and $\mathrm{Var}(\xi_i) \in (0, \infty)$, then $S_n = \sum_{i=1}^n \xi_i$ satisfies*

$$P\left(\max_{1\le m\le n} |S_n| \ge x\right) \le x^{-2}\mathrm{Var}(S_n) \text{ for all } x > 0.$$

Proof. S_n^2 is a submartingale (by convexity), so taking $\lambda = x^2$ in Theorem 4.2 gives

$$x^2 P\left(\max_{1\le m\le n} |S_n| \ge x\right) = x^2 P\left(\max_{1\le m\le n} S_n^2 \ge x^2\right) \le E\left[S_n^2\right] = \mathrm{Var}(S_n). \qquad \square$$

More important for the task at hand is

Theorem 4.4 (L^p Maximum Inequality)**.** *If X_n is a submartingale, then for all $1 < p < \infty$,*

$$E\left[\widetilde{X}_n^p\right] \le \left(\frac{p}{p-1}\right)^p E\left[(X_n^+)^p\right].$$

Proof. For any fixed $M > 0$, we have

$$\left\{\widetilde{X}_n \wedge M \ge \lambda\right\} = \begin{cases} \left\{\widetilde{X}_n \ge \lambda\right\}, & M \ge \lambda \\ \emptyset, & M < \lambda \end{cases},$$

so the layer cake representation, Doob's inequality, Tonelli's theorem, and Hölder's inequality give

$$\begin{aligned}
E\left[\left(\widetilde{X}_n \wedge M\right)^p\right] &= \int_0^\infty p\lambda^{p-1} P\left(\widetilde{X}_n \wedge M \ge \lambda\right) d\lambda \\
&= \int_0^M p\lambda^{p-1} P\left(\widetilde{X}_n \ge \lambda\right) d\lambda \\
&\le \int_0^M p\lambda^{p-1} \left(\frac{1}{\lambda}\int X_n^+ 1\left\{\widetilde{X}_n \ge \lambda\right\} dP\right) d\lambda \\
&= \int X_n^+ \int_0^{\widetilde{X}_n \wedge M} p\lambda^{p-2} d\lambda \, dP \\
&= \frac{p}{p-1}\int X_n^+ \left(\widetilde{X}_n \wedge M\right)^{p-1} dP \\
&= \frac{p}{p-1} E\left|X_n^+ \left(\widetilde{X}_n \wedge M\right)^{p-1}\right| \\
&\le \frac{p}{p-1} E\left[(X_n^+)^p\right]^{\frac{1}{p}} E\left[\left|\widetilde{X}_n \wedge M\right|^p\right]^{\frac{p-1}{p}}.
\end{aligned}$$

Dividing through by $E\left[\left|\widetilde{X}_n \wedge M\right|^p\right]^{\frac{p-1}{p}} > 0$ shows that

$$E\left[\left(\widetilde{X}_n \wedge M\right)^p\right]^{\frac{1}{p}} \le \frac{p}{p-1} E\left[(X_n^+)^p\right]^{\frac{1}{p}},$$

and thus

$$E\left[\left(\widetilde{X}_n \wedge M\right)^p\right] \leq \left(\frac{p}{p-1}\right)^p E\left[(X_n^+)^p\right].$$

Sending $M \to \infty$ and invoking monotone convergence completes the proof. □

Since $|Y_n|$ is a submartingale whenever Y_n is a martingale, we have

Corollary 4.2. *If Y_n is a martingale and $Y_n^* = \max_{1\leq m\leq n} |Y_n|$, then*

$$E\big[\,(Y_n^*)^p\,\big] \leq \left(\frac{p}{p-1}\right)^p E\big[\,|Y_n|^p\,\big].$$

With the maximum inequality behind us, it is a small step to show

Theorem 4.5 (L^p Convergence)**.** *If $1 < p < \infty$ and X_n is a martingale with $\sup_n E\big[\,|X_n|^p\,\big] < \infty$, then $X_n \to X$ a.s. and in L^p.*

Proof. $E\big[X_n^+\big]^p \leq E\big[\,|X_n|\,\big]^p \leq E\big[\,|X_n|^p\,\big] < \infty$, so the martingale convergence theorem implies $X_n \to X$ a.s.

Also,

$$E\left[\left(\max_{1\leq m\leq n} |X_n|\right)^p\right] \leq \left(\frac{p}{p-1}\right)^p E\big[\,|X_n|^p\,\big]$$

by Corollary 4.2, so letting $n \to \infty$ and using monotone convergence shows that $\sup_n |X_n| \in L^p$.

Because $|X_n - X|^p \leq (2\sup_n |X_n|)^p$, dominated convergence gives $E\big[\,|X_n - X|^p\,\big] \to 0$. □

Uniform Integrability.

It remains to consider the $p = 1$ case. Essentially, this is just the Vitali convergence theorem, but we will go ahead and do everything carefully.

Definition. A collection of random variables $\{X_i\}_{i\in I}$ is said to be *uniformly integrable* if

$$\lim_{M\to\infty}\left(\sup_{i\in I} E\big[\,|X_i|\,1\,\{|X_i| > M\}\,\big]\right) = 0.$$

Uniform integrability implies uniform L^1 bounds since we can take M large enough that the supremum in the definition is less than 1, say, and thus conclude that

$$\begin{aligned}\sup_{i\in I} E\left[|X_i|\right] &= \sup_{i\in I}\left(E\big[\,|X_i|\,1\,\{|X_i| \leq M\}\,\big] + E\big[\,|X_i|\,1\,\{|X_i| > M\}\,\big]\right)\\ &\leq \sup_{i\in I} E\big[\,|X_i|\,1\,\{|X_i| \leq M\}\,\big] + \sup_{i\in I} E\big[\,|X_i|\,1\,\{|X_i| > M\}\,\big] \leq M + 1 < \infty.\end{aligned}$$

However, one should observe that uniform integrability is a less restrictive assumption than domination by an integrable function since if $X \geq 0$ is integrable and $|X_i| \leq X$ for all $i \in I$, then
$|X_i| 1\{|X_i| > M\} \leq |X| 1\{|X| > M\}$ for all $i \in I$, so monotonicity and dominated convergence give

$$\lim_{M\to\infty} \left(\sup_{i\in I} E\big[\, |X_i| 1\{|X_i| > M\}\,\big] \right) \leq \lim_{M\to\infty} E\big[\, |X| 1\{|X| > M\}\,\big] = 0.$$

As an example of a u.i. family which is not dominated by an integrable random variable, consider Lebesgue measure on $(0,1)$ and $X_n(\omega) = \omega^{-1} 1_{(\frac{1}{n+1}, \frac{1}{n})}(\omega)$.

Any finite collection of integrable random variables is clearly uniformly integrable, but countable collections need not be.
For example, if $X_n = n1_{(0,\frac{1}{n})}$ on $[0,1]$ with Lebesgue measure, then $E\big[\, |X_n| 1\{|X_n| > M\}\,\big] = 1$ for all $n > M$. (This also shows that uniform integrability is a stronger assumption than uniform L^1 bounds.)

Our next theorem shows that one can have very large u.i. families.

Theorem 4.6. *Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X \in L^1$, the collection $\{E[X|\mathcal{G}] : \mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}\}$ is uniformly integrable.*

Proof. We first note that since $X \in L^1$, if A_n is a sequence of events with $P(A_n) \to 0$, then $E\big[\, |X| 1_{A_n}\big] \to 0$ by the DCT for convergence in probability.
It follows that for every $\varepsilon > 0$, there is a $\delta > 0$ such that $P(A) < \delta$ implies $E\big[\, |X| 1_A\big] < \varepsilon$ - if not, there exists an $\varepsilon > 0$ and a sequence $A_1, A_2, \ldots$ with $P(A_n) < \frac{1}{n}$ and $E\big[\, |X| 1_{A_n}\big] \geq \varepsilon$, a contradiction.
Now let $\varepsilon > 0$ be given and choose δ as above. Taking $M > \frac{E|X|}{\delta}$, it follows from Jensen's inequality that for any $\mathcal{G} \subseteq \mathcal{F}$,

$$\begin{aligned} E\big[\, |E[X|\mathcal{G}]| 1\{|E[X|\mathcal{G}]| > M\}\,\big] &\leq E\big[E\big[|X|\,\big|\,\mathcal{G}\big] 1\big\{\big|E\big[X\big|\mathcal{G}\big]\big| > M\big\}\,\big] \\ &\leq E\big[E\big[|X|\,\big|\,\mathcal{G}\big] 1\big\{E\big[|X|\,\big|\,\mathcal{G}\big] > M\big\}\,\big] \\ &= E\big[\, |X| 1\big\{E\big[|X|\,\big|\,\mathcal{G}\big] > M\big\}\,\big] \end{aligned}$$

where the final equality follows from the definition of conditional expectation by observing that $\big\{E\big[|X|\,\big|\mathcal{G}\big] > M\big\} \in \mathcal{G}$.

Using Chebychev's inequality, we have $P\left(E\big[|X|\,\big|\,\mathcal{G}\big] > M\right) \leq \frac{E\left[E\left[|X|\,\middle|\,\mathcal{G}\right]\right]}{M} = \frac{E|X|}{M} < \delta$.
Thus,

$$E\big[\, \big|E\big[X\big|\mathcal{G}\big]\big| 1\big\{\big|E\big[X\big|\mathcal{G}\big]\big| > M\big\}\,\big] \leq E\big[\, |X| 1\big\{E\big[|X|\,\big|\,\mathcal{G}\big] > M\big\}\,\big] < \varepsilon$$

for every $\mathcal{G} \subseteq \mathcal{F}$, and the result follows since ε was arbitrary. □

Uniform integrability is the condition needed to upgrade convergence in probability to convergence in L^1.

Theorem 4.7. *If $X_n \to_p X$, then the following are equivalent:*

(i): *$\{X_n\}_{n=0}^{\infty}$ is uniformly integrable.*
(ii): *$X_n \to X$ in L^1.*
(iii): *$E|X_n| \to E|X| < \infty$.*

Proof. Assume $\{X_n\}$ is u.i. and let $\varphi_M(x) = \begin{cases} M, & x > M \\ x, & |x| \leq M \\ -M, & x < -M \end{cases}$.

Then

$$\begin{aligned} |X_n - X| &\leq |X_n - \varphi_M(X_n)| + |\varphi_M(X_n) - \varphi_M(X)| + |\varphi_M(X) - X| \\ &= (|X_n| - M)^+ + |\varphi_M(X_n) - \varphi_M(X)| + (|X| - M)^+ \\ &\leq |X_n| \, 1\{|X_n| > M\} + |\varphi_M(X_n) - \varphi_M(X)| + |X| \, 1\{|X| > M\}, \end{aligned}$$

and thus

$$E|X_n - X| \leq E[|X_n| \, 1\{|X_n| > M\}] + E|\varphi_M(X_n) - \varphi_M(X)| + E[|X| \, 1\{|X| > M\}].$$

Since $X_n \to_p X$ and φ_M is bounded and continuous, $E|\varphi_M(X_n) - \varphi_M(X)| \to 0$. (Convergence in probability is preserved by continuous functions and the convergence in probability generalization of bounded convergence then applies.)

Uniform integrability ensures that the first term can be made less than any given $\varepsilon > 0$ by choosing M sufficiently large.

It also ensures that $\sup_n E|X_n| < \infty$, so the convergence in probability version of Fatou's lemma implies that $E|X| < \infty$, and thus M can be taken large enough that the third term is less than ε as well.

Therefore, given $\varepsilon > 0$, we can choose M so that

$$E|X_n - X| < \varepsilon + E|\varphi_M(X_n) - \varphi_M(X)| + \varepsilon \to 2\varepsilon \text{ as } n \to \infty,$$

hence (i) implies (ii).

That (ii) implies (iii) is standard:

$$\big|E|X_n| - E|X|\big| \leq E\big||X_n| - |X|\big| \leq E|X_n - X| \to 0.$$

Finally, suppose that $E|X_n| \to E|X| < \infty$ and define $\psi_M(x) = \begin{cases} x, & 0 \leq x \leq M-1 \\ (M-1)(M-x), & M-1 < x < M \\ 0, & \text{otherwise} \end{cases}$.

Given $\varepsilon > 0$, one can choose M large enough that $E|X| - E[\psi_M(|X|)] < \varepsilon$ by the DCT.

Also, as in the first part of the proof, $X_n \to_p X$ and $\psi_M(|\cdot|) \in C_b$ implies $E[\psi_M(|X_n|)] \to E[\psi_M(|X|)]$.

Thus there is an $N \in \mathbb{N}$ such that for all $n \geq N$, $\big|E|X_n| - E|X|\big|, |E[\psi_M(|X_n|)] - E[\psi_M(|X|)]| < \varepsilon$, hence

$$\begin{aligned} E[|X_n| \, 1\{|X_n| > M\}] &\leq E|X_n| - E[\psi(|X_n|)] \\ &\leq \big|E|X_n| - E|X|\big| + (E|X| - E[\psi(|X|)]) + |E[\psi(|X|)] - E[\psi(|X_n|)]| < 3\varepsilon. \end{aligned}$$

By increasing M if need be, we can ensure that $E[|X_n| \, 1\{|X_n| > M\}] < 3\varepsilon$ for the (finitely many) $n < N$, and uniform integrability is established. □

We are finally able to provide necessary and sufficient conditions for L^1 convergence.

Theorem 4.8. *For a submartingale, the following are equivalent.*

(i): *It is uniformly integrable.*
(ii): *It converges a.s. and in L^1.*
(iii): *It converges in L^1.*

Proof. Uniform integrability implies that $\sup_n E|X_n| < \infty$ so the martingale convergence theorem implies $X_n \to X$ a.s., and Theorem 4.7 implies $X_n \to X$ in L^1, hence (i) implies (ii).

As (ii) tautologically implies (iii), it remains only to show that L^1 convergence implies uniform integrability.

But this is also a simple consequence of Theorem 4.7 since $X_n \to X$ in L^1 implies $X_n \to_p X$. □

When X_n is a martingale, we can actually say a little bit more.

Theorem 4.9. *For a martingale, the following are equivalent.*

(i): *It is uniformly integrable.*
(ii): *It converges a.s. and in L^1.*
(iii): *It converges in L^1.*
(iv): *There is an $X \in L^1$ with $X_n = E[X|\mathcal{F}_n]$.*

Proof. Since martingales are submartingales, $(i) \Rightarrow (ii) \Rightarrow (iii)$ by Theorem 4.8.

Now assume that $X_n \to X$ in L^1. Then for all $A \in \mathcal{F}$,

$$|E[X_n 1_A] - E[X 1_A]| = |E[(X_n - X) 1_A]| \leq E|(X_n - X) 1_A| \leq E|X_n - X| \to 0,$$

hence $E[X_n 1_A] \to E[X 1_A]$.

Also, since X_n is a martingale, we have $E[X_n|\mathcal{F}_m] = X_m$ for all $m < n$. Thus if $A \in \mathcal{F}_m$, then

$$E[X_m 1_A] = E\big[E[X_n|\mathcal{F}_m] 1_A\big] = E\big[E[X_n 1_A|\mathcal{F}_m]\big] = E[X_n 1_A].$$

Putting these facts together shows that $E[X_m 1_A] = E[X 1_A]$ for all $A \in \mathcal{F}_m$, which, by definition of conditional expectation, implies that $X_m = E[X|\mathcal{F}_m]$.

Since Theorem 4.6 shows that $(iv) \Rightarrow (i)$, the chain is complete. □

The nontrivial part of Theorem 4.9 (conditional on preceding results) was the conclusion that if a martingale X_n converges to a random variable X in L^1, then $X_n = E[X|\mathcal{F}_n]$. Our next result can be seen as a variation on this theme.

Theorem 4.10 (Lévy's Forward Theorem)**.** *Suppose that $\mathcal{F}_n \nearrow \mathcal{F}_\infty$ – i.e. $\{\mathcal{F}_n\}$ is a filtration and $\mathcal{F}_\infty = \sigma\left(\bigcup_n \mathcal{F}_n\right)$. Then for any integrable X,*

$$E[X|\mathcal{F}_n] \to E[X|\mathcal{F}_\infty] \text{ a.s. and in } L^1.$$

Proof. We showed in Example 2.4 that $X_n = E[X\,|\mathcal{F}_n]$ is a martingale and Theorem 4.6 implies it is uniformly integrable.

It follows, therefore, from Theorem 4.9 that $E[X\,|\mathcal{F}_n]$ converges to a limit X_∞ a.s. and in L^1, and that $E[X\,|\mathcal{F}_n] = X_n = E[X_\infty\,|\mathcal{F}_n]$.

By definition of conditional expectation, this means that for all $A \in \mathcal{F}_n$,

$$\int_A X dP = \int_A X_n dP = \int_A X_\infty dP.$$

Our claim that $X_\infty = E[X\,|\mathcal{F}_\infty]$ will follow if we can show that the above holds for all $A \in \mathcal{F}_\infty$.

But this is a consequence of the $\pi - \lambda$ theorem since $\mathcal{P} = \bigcup_n \mathcal{F}_n$ is a π-system which generates $\mathcal{F}_\infty$ and is contained in the λ-system $\mathcal{L} = \left\{A : \int_A X dP = \int_A X_\infty dP\right\}$.

($\mathcal{P}$ is a π-system since $\{\mathcal{F}_n\}$ is a filtration and $\mathcal{L}$ is a λ-system since X, X_∞ are integrable.) □

An immediate consequence of Theorem 4.10 is

Theorem 4.11 (Lévy's $0-1$ Law)**.** *If $\mathcal{F}_n \nearrow \mathcal{F}_\infty$ and $A \in \mathcal{F}_\infty$, then $P(A\,|\mathcal{F}_n) \to 1_A$ a.s.*

Though Theorem 4.11 may seem trivial, it should be noted that it implies Kolmogorov's $0-1$ law:

If $X_1, X_2, ...$ are independent and A belongs to the tail field $\mathcal{T} = \bigcap_n \sigma(X_n, X_{n+1}, ...)$, then 1_A is independent of each $\mathcal{F}_n = \sigma(X_1, ..., X_n)$, so $P(A\,|\mathcal{F}_n) = P(A)$. As $\mathcal{T} \subseteq \sigma(X_1, X_2, \ldots) = \mathcal{F}_\infty$, Theorem 4.11 implies that this converges to 1_A a.s., so we must have $P(A) = 1_A$ a.s., hence $P(A) \in \{0, 1\}$.

5. Optional Stopping

We round out our discussion of smartingales with a look at optional sampling theorems, of which we have already seen one example:

If X_n is a submartingale and M, N are stopping times with $M \le N \le k$ a.s., then

$$E[X_0] \le E[X_M] \le E[X_N] \le E[X_k].$$

We now look into what kind of conditions on X_n allow us to reach similar conclusions for unbounded stopping times.

Lemma 5.1. *If X_n is a uniformly integrable submartingale and N is any stopping time, then $X_{N\wedge n}$ is uniformly integrable.*

Proof. X_m^+ is a submartingale and $N \wedge n \le n$ is a stopping time, so $E[X_{N\wedge n}^+] \le E[X_n^+]$ by Corollary 4.1. Since X_m^+ is u.i., we have

$$\sup_n E[X_{N\wedge n}^+] \le \sup_n E[X_n^+] < \infty,$$

so the martingale convergence theorem implies $X_{N\wedge n} \to X_N$ a.s. (where $X_\infty = \lim_{n\to\infty} X_n$) and $E|X_N| < \infty$. Because $E|X_N| < \infty$ and $\{X_n\}$ is u.i., for any $\varepsilon > 0$, we can choose K so that

$$E\big[|X_N|\,;|X_N| > K\big], \sup_n E\big[|X_n|\,;|X_n| > K\big] < \frac{\varepsilon}{2},$$

hence

$$\begin{aligned} E\big[|X_{N\wedge n}|\,;|X_{N\wedge n}| > K\big] &= E\big[|X_n|\,;|X_n| > K,\, N > n\big] + E\big[|X_N|\,;|X_N| > K,\, N \le n\big] \\ &\le \sup_n E\big[|X_n|\,;|X_n| > K\big] + E\big[|X_N|\,;|X_N| > K\big] < \varepsilon \end{aligned}$$

for all n, showing that $\{X_{N\wedge n}\}$ is u.i. □

It is worth observing that the final sentence in the preceding proof establishes

Corollary 5.1. *If $E|X_N| < \infty$ and $X_n 1\{N > n\}$ is u.i., then $X_{N\wedge n}$ is u.i.*

The utility of Lemma 5.1 is that it enables us to prove

Theorem 5.1. *If X_n is a uniformly integrable submartingale, then for any stopping time $N \le \infty$, $E[X_0] \le E[X_N] \le E[X_\infty]$ where $X_\infty = \lim_{n\to\infty} X_n$.*

Proof. Corollary 4.1 gives $E[X_0] \le E[X_{N\wedge n}] \le E[X_n]$ for all n. Since X_n and $X_{N\wedge n}$ are u.i., the L^1 martingale convergence theorem implies $E[X_{N\wedge n}] \to E[X_N]$ and $E[X_n] \to E[X_\infty]$, so the result follows by taking $n \to \infty$ in the above inequality. □

Observe that if X_n is our "double after losing" martingale and $N = \inf\{n : X_n = 1\}$, then $X_N = 1 \ne 0 = X_0$. Thus the conclusion of Theorem 5.1 need not hold if X_n is not u.i.

More generally, Theorem 5.1 shows that any such system will fail if we only have finite credit, because then the corresponding martingale would be bounded and thus uniformly integrable.

A useful consequence of Theorem 5.1 is

Theorem 5.2. *If $L \le M$ are stopping times and $Y_{M\wedge n}$ is a u.i. submartingale, then*

$$E[Y_0] \le E[Y_L] \le E[Y_M] \text{ and } Y_L \le E[Y_M\,|\mathcal{F}_L\,].$$

Proof. The first claim follows from Theorem 5.1 by setting $X_n = Y_{M\wedge n}$, $N = L$, so that

$$E[Y_0] = E[X_0], \quad E[Y_L] = E[Y_{M\wedge L}] = E[X_N], \quad E[Y_M] = \lim_{n\to\infty} E[Y_{M\wedge n}] = E[X_\infty].$$

Now for $A \in \mathcal{F}_L$, let $N = \begin{cases} L & \text{on } A \\ M & \text{on } A^C \end{cases}$. This is a stopping time since $L \le M$ implies $A \in \mathcal{F}_L \subseteq \mathcal{F}_M$, hence $\{N = n\} = (A \cap \{L = n\}) \cup \big(A^C \cap \{M = n\}\big) \in \mathcal{F}_n$.

As $N \le M$ by construction, the preceding shows that

$$E[Y_L; A] + E[Y_M; A^C] = E[Y_N] \le E[Y_M] = E[Y_M; A] + E[Y_M; A^C],$$

hence

$$E[Y_L; A] \le E[Y_M; A] = E\big[E[Y_M 1_A\,|\mathcal{F}_L\,]\big] = E\big[E[Y_M\,|\mathcal{F}_L\,]; A\big].$$

Taking $A = A_\varepsilon := \{Y_L - E[Y_M\,|\mathcal{F}_L\,] \ge \varepsilon\}$ shows that $P(A_\varepsilon) = 0$ for all $\varepsilon > 0$ and the desired result follows. □

We have shown that optional stopping holds for bounded stopping times and for bounded smartingales (as bounded implies uniformly integrable), and noted that these results show that you need infinite time and credit, respectively, to ensure victory in an unfavorable game. Our last optional stopping theorem lies somewhere in between these two and shows that another way for casinos to guard against length of play strategies is to place caps on bets.

Theorem 5.3. *Suppose X_n is a submartingale with $E\big[\,|X_{n+1} - X_n|\,|\mathcal{F}_n\,\big] \le B$ a.s. If N is a stopping time with $E[N] < \infty$, then $X_{N\wedge n}$ is u.i. and thus $E[X_N] \ge E[X_0]$.*

Proof. Since

$$|X_{N\wedge n}| = \left|X_0 + \sum_{m=0}^{n-1} (X_{m+1} - X_m)\,1\{N > m\}\right| \le |X_0| + \sum_{m=0}^{\infty} |X_{m+1} - X_m|\,1\{N > m\},$$

it will follow that $X_{N\wedge n}$ is u.i. once we show that the right-hand side is integrable.

To see this, observe that $\{N > m\} = \{N \le m\}^C \in \mathcal{F}_m$, so it follows from our assumptions that

$$\begin{aligned} E\big[\,|X_{m+1} - X_m|\,1\{N > m\}\,\big] &= E\big[E[|X_{m+1} - X_m|\,|\mathcal{F}_m]1\{N > m\}\,\big] \\ &\le E\,[B1\{N > m\}] = BP\,(N > m)\,, \end{aligned}$$

and thus

$$E\Big[\,|X_0| + \sum_{m=0}^{\infty} |X_{m+1} - X_m|\,1\{N > m\}\Big] \le E\,|X_0| + \sum_{m=0}^{\infty} BP\,(N > m) = E\,|X_0| + BE\,[N] < \infty.$$

□

Our final optional stopping theorem applies to arbitrary stopping times and requires that the smartingale be a.s. bounded in the appropriate direction.

Theorem 5.4. *If X_n is a nonnegative supermartingale and N is a stopping time, then $E[X_0] \geq E[X_N]$.*

Proof. Corollary 2.6 shows that $X_\infty = \lim_{n\to\infty} X_n$ exists, so X_N is well-defined.

Also, $E[X_0] \geq E[X_{N\wedge n}]$ for all $n \in \mathbb{N}$ by Theorem 2.7.

Since monotone convergence implies

$$E[X_N; N < \infty] = \lim_{n\to\infty} E[X_N; N \leq n]$$

and Fatou's lemma implies

$$E[X_N; N = \infty] \leq \liminf_{n\to\infty} E[X_n; N > n],$$

we have

$$E[X_0] \geq \liminf_{n\to\infty} E[X_{N\wedge n}] = \liminf_{n\to\infty} (E[X_n; N > n] + E[X_N; N \leq n]) \geq E[X_N]. \qquad \square$$

Example 5.1 (Gambler's Ruin)**.** Suppose that in successive flips of an unfair coin, we win one dollar if the coin comes up heads and lose one dollar if it comes up tails. If we start with nothing, then our fortune at time n behaves like asymmetric simple random walk, $S_n = \sum_{i=1}^n \xi_i$ where $\xi_1, \xi_2, ...$ are i.i.d. with $P(\xi_i = 1) = 1 - P(\xi_i = -1) = p$.

Clearly S_n is a submartingale w.r.t. $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$ if $p > \frac{1}{2}$ and a supermartingale if $p < \frac{1}{2}$.

Since we don't want to have to consider the cases separately, we look at $X_n = \left(\frac{1-p}{p}\right)^{S_n}$, which we claim is a martingale: X_n is bounded and $\mathcal{F}_n$-measurable, and

$$\begin{aligned} E[X_{n+1}|\mathcal{F}_n] &= E\left[\left(\frac{1-p}{p}\right)^{S_n}\left(\frac{1-p}{p}\right)^{\xi_{n+1}}\middle|\mathcal{F}_n\right] \\ &= \left(\frac{1-p}{p}\right)^{S_n} E\left[\left(\frac{1-p}{p}\right)^{\xi_{n+1}}\right] \\ &= X_n\left[p\left(\frac{1-p}{p}\right) + (1-p)\left(\frac{p}{1-p}\right)\right] = X_n. \end{aligned}$$

Now let's suppose we decide beforehand that we will quit once we lose a dollars or win b dollars, whichever happens first. That is, we walk away at time $N = T_{-a} \wedge T_b$ where $T_x = \inf\{m : S_m = x\}$.

A natural question is how likely is it the game ends in ruin, $r = P(T_{-a} < T_b)$.

We first note that $S_{N\wedge n}$ is uniformly bounded and thus is uniformly integrable. It follows from the L^1 martingale convergence theorem that $S_{N\wedge n}$ has a limit almost surely and in L^1. As convergence to a point in $(-a, b)$ is impossible, it must be the case that $N < \infty$ a.s.

That $S_{N\wedge n}$ is bounded also implies $X_{N\wedge n} = \left(\frac{1-p}{p}\right)^{S_{N\wedge n}}$ is u.i., so we can apply the martingale form of Theorem 5.1 to conclude that

$$\begin{aligned} 1 = E[X_0] = E[X_N] &= \left(\frac{1-p}{p}\right)^{-a} r + \left(\frac{1-p}{p}\right)^{b}(1-r) \\ &= r\left[\left(\frac{1-p}{p}\right)^{-a} - \left(\frac{1-p}{p}\right)^{b}\right] + \left(\frac{1-p}{p}\right)^{b}. \end{aligned}$$

The probability of ruin is thus

$$r = \frac{1 - \left(\frac{1-p}{p}\right)^b}{\left(\frac{p}{1-p}\right)^a - \left(\frac{1-p}{p}\right)^b}.$$

Example 5.2 (Patterns in Coin Tossing)**.** We are interested in the expected waiting time for the first occurrence of HTH in a sequence of independent tosses of a fair coin. More formally, suppose that $X_1, X_2, ...$ are i.i.d. with $P(X_i = H) = P(X_i = T) = \frac{1}{2}$, and set $\tau_{HTH} = \inf\{t \geq 3 : X_{t-2} = H, X_{t-1} = T, X_t = H\}$. We want to compute $E[\tau_{HTH}]$.

Somewhat surprisingly, our analysis is simplified by incorporating a casino and an army of gamblers into the model. The setup is as follows. The casino offers even odds on successive tosses of a fair coin $(X_1, X_2, ...)$. Gamblers arrive one at a time with the k^{th} gambler joining the game just before X_k is observed and placing a \$1 bet on heads. If $X_k = T$, he loses his dollar and quits playing. If $X_k = H$, the casino pays him \$2, which he then wagers on $X_{k+1} = T$. If he loses, then he walks away with nothing and is out his initial \$1 stake. Otherwise, he bets his \$4 fortune on $X_{k+2} = H$. Regardless of the outcome, he quits after this round. Since the game is fair, the casino's net profit from the k^{th} round, $\xi_k \in \sigma(X_1, ..., X_k)$, has $E[\xi_k] = 0$. It follows that the casino's profit from the first n rounds, $M_n = \sum_{k=1}^n \xi_k$, is a martingale w.r.t. $\sigma(X_1, ..., X_n)$. Also, $M_{\tau_{HTH}} = \tau_{HTH} - 10$ because each of the τ_{HTH} gamblers payed a \$1 entrance fee, and all gamblers except the $(\tau_{HTH} - 2)nd$ (who has \$8) and the $\tau_{HTH}th$ (who has \$2) walked away with nothing.

If we can show that optional stopping applies, then it will follow that $0 = E[M_0] = E[M_{\tau_{HTH}}] = E[\tau_{HTH}] - 10$. To see that this is so, we first note that τ_{HTH} is stochastically dominated by $3Y$ where $Y \sim \text{Geom}(\frac{1}{8})$, thus $E[\tau_{HTH}] < \infty$, hence $E|M_{\tau_{HTH}}| \leq E[\tau_{HTH}] + 10 < \infty$.

Also, since $|M_n| \leq 7n$ (as none of the n gamblers have a net loss or gain exceeding 7), we see that

$$\begin{aligned}\int |M_n 1\{\tau_{HTH} > n\}| \, dP &\leq 7n \int 1\{\tau_{HTH} > n\} \, dP = 7nP(\tau_{HTH} > n) \\ &\leq 7nP\left(Y > \frac{n}{3}\right) = 7n\left(\frac{7}{8}\right)^{\lfloor \frac{n}{3} \rfloor} \to 0 \text{ as } n \to \infty.\end{aligned}$$

It follows that $M_n 1\{\tau_{HTH} > n\}$ is u.i., so Corollary 5.1 shows that $M_{\tau_{HTH} \wedge n}$ is u.i., hence Theorem 5.2 implies $E[M_{\tau_{HTH}}] = E[M_0]$ and we conclude that $E[\tau_{HTH}] = 10$.

The exact same logic applies for words of different lengths and alphabets of various sizes.

It is interesting to note that if you carry out this analysis for the sequence HHH, then you find that $E[\tau_{HHH}] = 8 + 4 + 2 = 14$, so coin toss patterns of the same length can have different expected times to first occurrence.

6. Markov Chains

In words, a Markov chain is a random process in which the future depends on the past only through the present. More formally,

Definition. Let $(S, \mathcal{S})$ be a measurable space. A sequence of S-valued random variables $X_0, X_1, X_2, ...$ on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ is said to be a *Markov chain* with respect to $\mathcal{F}_n$ if $X_n \in \mathcal{F}_n$ and for all $B \in \mathcal{S}$,

$$P(X_{n+1} \in B \,|\mathcal{F}_n) = P(X_{n+1} \in B \,|X_n).$$

S is called the *state space* of the chain, and the law of X_0 is called the *initial distribution*.

* We take it as implicit in the name "Markov chain" (as opposed to "Markov process") that we are working in discrete time.

As the above definition is fairly difficult to work with directly, we introduce the following useful construct.

Definition. A function $p : S \times \mathcal{S} \to \mathbb{R}$ is called a *transition probability* if

(1) For each $x \in S$, $A \mapsto p(x, A)$ is a probability measure on $(S, \mathcal{S})$,
(2) For each $A \in \mathcal{S}$, $x \mapsto p(x, A)$ is a measurable function.

We say that X_n is a Markov chain (w.r.t. $\mathcal{F}_n$) with transition probabilities p_n if

$$P(X_{n+1} \in B \,|\mathcal{F}_n) = p_n(X_n, B).$$

If $(S, \mathcal{S})$ is nice or S is countable, there is no loss of generality in supposing the existence of transition probabilities as we are then assured of the existence of a r.c.d. for X_{n+1} given $\sigma(X_n)$:

$$\mu_n(\omega, B) = P(X_{n+1} \in B \,|X_n) = P(X_{n+1} \in B \,|\mathcal{F}_n).$$

The last problem on the first homework shows that we can take $\mu_n(\omega, B) = p_n(X_n(\omega), B)$ for some transition probability p_n.

Conversely, if we are given an initial distribution μ on $(S, \mathcal{S})$ and a sequence of transition probabilities $p_0, p_1, ...$, we can define a consistent sequence of finite dimensional distributions by

$$\nu_n(X_0 \in B_0, X_1 \in B_1, ..., X_n \in B_n) = \int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n).$$

If $(S, \mathcal{S})$ is nice, Kolmogorov's extension theorem guarantees the existence of a measure P_μ on the sequence space $(S^{\mathbb{N}_0}, \mathcal{S}^{\mathbb{N}_0})$ such that the coordinate maps $X_n(\omega) = \omega_n$ have the desired distributions.

* Note that the extension theorem also applies when S is countable since we can then identify S with a subset of $\mathbb{Z} \subseteq \mathbb{R}$.

Through a slight abuse of notation, when $\mu = \delta_x$ is the point mass at x, we will write P_x for P_{δ_x}.

It is worth observing that the family of measures $\{P_x\}_{x \in S}$ is fundamental in the sense that $P_\mu(A) = \int P_x(A)\mu(dx)$ for any initial distribution μ on $(S, \mathcal{S})$.

To see that the construction from Kolmogorov's theorem defines a Markov chain with respect to $\mathcal{F}_n = \sigma(X_0, X_1, ..., X_n)$ having transition probabilities $\{p_n\}$, we need to prove that

$$\int_A 1_B(X_{n+1})\, dP_\mu = \int_A p_n(X_n, B)\, dP_\mu$$

for all $n \in \mathbb{N}_0$, $A \in \mathcal{F}_n$, $B \in \mathcal{S}$.

Since the collection of $(n+1)$-cylinders is a π-system which generates $\mathcal{F}_n$, a $\pi - \lambda$ argument shows that it suffices to consider $A = \{X_0 \in B_0, X_1 \in B_1, ..., X_n \in B_n\}$ with $B_i \in \mathcal{S}$.

We compute

$$\begin{aligned}\int_A 1_B(X_{n+1})\, dP_\mu &= P_\mu(A, X_{n+1} \in B) = P_\mu(X_0 \in B_0, X_1 \in B_1, ..., X_n \in B_n, X_{n+1} \in B) \\ &= \int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n) \int_B p_n(x_n, dx_{n+1}) \\ &= \int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n) p_n(x_n, B).\end{aligned}$$

To finish up, we reconstruct the integral:
If $f(x_n) = 1_C(x_n)$, then

$$\begin{aligned}\int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n) 1_C(x_n) &= P_\mu(X_0 \in B_0, ..., X_n \in B_n \cap C) \\ &= P_\mu(A, X_n \in C) = \int_A 1_C(X_n)\, dP_\mu.\end{aligned}$$

Linearity shows the result holds for $f(x_n)$ simple, and the bounded convergence theorem extends it to bounded measurable f, such as $f(x_n) = p_n(x_n, B)$.

In summary, we have shown that

Theorem 6.1. *If $(S, \mathcal{S})$ is nice (or S is countable), then for any distribution μ and any sequence of transition probabilities $p_0, p_1, ...$, there exists an S-valued Markov chain $\{X_n\}$ such that $X_0 \sim \mu$ and $P(X_{n+1} \in B \,|X_0, X_1, ..., X_n\,) = p_n(X_n, B)$.*

In order to verify that the preceding Kolmogorov construction is indeed the right one, we prove

Theorem 6.2. *Any Markov chain X_n on $(S, \mathcal{S})$ having initial distribution μ and transition probabilities p_n has finite dimensional distributions satisfying*

$$P(X_0 \in B_0, X_1 \in B_1, ..., X_n \in B_n) = \int_{B_0} \mu(dx_0) \int_{B_1} p_0(x_0, dx_1) \cdots \int_{B_n} p_{n-1}(x_{n-1}, dx_n)$$

for all $n \in \mathbb{N}_0$, $B_0, ..., B_n \in \mathcal{S}$.

To this end, we first record the following useful consequence of the $\pi - \lambda$ theorem.

Theorem 6.3 (Functional Monotone Class Theorem)**.** *Let $\mathcal{A}$ be a π-system containing Ω and let $\mathcal{H}$ be a collection of functions $f : \Omega \to \mathbb{R}$ which satisfies*

(i) *If $A \in \mathcal{A}$, then $1_A \in \mathcal{H}$.*
(ii) *If $f, g \in \mathcal{H}$ and $c \in \mathbb{R}$, then $f + g$ and cf are in $\mathcal{H}$.*
(iii) *If $f_1, f_2, ... \in \mathcal{H}$ are nonnegative with $f_n \nearrow f$, then $f \in \mathcal{H}$.*

Then $\mathcal{H}$ contains all functions which are measurable with respect to $\sigma(\mathcal{A})$.

Proof. $\mathcal{L} = \{A : 1_A \in \mathcal{H}\}$ is a λ-system since (i) implies $1_\Omega \in \mathcal{H}$; (ii) implies $1_{B\setminus C} = 1_B - 1_C \in \mathcal{H}$ if $B, C \in \mathcal{L}$ with $C \subseteq B$; and (iii) implies $1_A = \lim_n 1_{A_n} \in \mathcal{H}$ if $A_n \in \mathcal{L}$ with $A_n \nearrow A$.

As (i) shows that the π-system $\mathcal{A}$ is contained in $\mathcal{L}$, it follows from the $\pi - \lambda$ theorem that $\sigma(\mathcal{A}) \subseteq \mathcal{L}$, thus $\mathcal{H}$ contains all indicators of events in $\sigma(\mathcal{A})$.

It then follows from (ii) that $\mathcal{H}$ contains all simple functions and from (iii) that it contains all nonnegative measurable functions. Taking positive and negative parts and using (ii) gives the result. □

(Often, condition (iii) only supposes that $f \in \mathcal{H}$ when $f_n \nearrow f$ with f bounded, and the conclusion is that $\mathcal{H}$ contains all bounded $\sigma(\mathcal{A})$-measurable functions. The argument is the same, and that version can be more convenient when $\mathcal{H}$ is defined in terms of expectations.)

Proof of Theorem 6.2. The proof of Theorem 6.1 shows that

$$\mathcal{H} = \left\{ f : S \to \mathbb{R} \text{ such that } f \text{ is bounded and } E[f(X_{n+1}) \,|\mathcal{F}_n] = \int p_n(X_n, dy) f(y) \text{ for all } n \in \mathbb{N}_0 \right\}$$

satisfies the conditions of Theorem 6.3 with $\mathcal{A} = \mathcal{S}$, hence $E[f(X_{n+1}) \,|\mathcal{F}_n] = \int p_n(X_n, dy) f(y)$ for all bounded $f \in \mathcal{S}$.

Accordingly, for any bounded measurable $f_0, ..., f_n$,

$$\begin{aligned} E\left[\prod_{m=0}^{n} f_m(X_m)\right] &= E\left[E\left[\prod_{m=0}^{n} f_m(X_m)\middle|\mathcal{F}_{n-1}\right]\right] \\ &= E\left[\prod_{m=0}^{n-1} f_m(X_m) \cdot E\left[f_n(X_n)\,|\mathcal{F}_{n-1}\right]\right] \\ &= E\left[\prod_{m=0}^{n-1} f_m(X_m) \cdot \int p_{n-1}(X_{n-1}, dy) f_n(y)\right]. \end{aligned}$$

Since $\int p_{n-1}(X_{n-1}, dy) f_n(y)$ is a bounded measurable function of X_{n-1}, it follows by induction that if $\mu = \mathscr{L}(X_0)$, then

$$E\left[\prod_{m=0}^{n} f_m(X_m)\right] = \int \mu(dx_0) f_0(x_0) \int p_0(x_0, dx_1) f_1(x_1) \cdots \int p_{n-1}(x_{n-1}, dx_n) f_n(x_n),$$

establishing Theorem 6.2. □

The preceding results show that we can describe a Markov chain X_n by specifying the transition probabilities p_n. In practice, the transition probabilities are the fundamental objects for analyzing Markov chains.

Given transition probabilities, we can assume that the X_n's are the coordinate maps on the sequence space $\left(S^{\mathbb{N}_0}, \mathcal{S}^{\mathbb{N}_0}\right)$.

This construction gives us a measure P_μ for each initial distribution μ, which makes X_n a Markov chain with transition probabilities p_n.

It also enables us to work with the shift operators $(\theta^n \omega)_i = \omega_{i+n}$.

To keep things simple, we will restrict our attention henceforth to *temporally homogeneous* Markov chains, in which there is a single transition probability $p = p_0 = p_1 = ...$

Example 6.1 (Random Walk)**.** Let $\xi_1, \xi_2, ... \in \mathbb{R}^d$ be i.i.d. random vectors with distribution μ, and define $X_n = x_0 + \xi_1 + ... + \xi_n$. Then X_n defines a Markov chain with initial distribution δ_{x_0} and transition probability $p(x, B) = \mu(B - x)$.

If the ξ_k's are independent but not identically distributed, we still get a Markov chain, but it's no longer time homogeneous.

All other examples of Markov chains that we will consider will have countable state space $S = \{s_1, s_2, ...\}$ equipped with the σ-algebra $\mathcal{S} = 2^S$.

In this case, the transition probabilities are specified by functions of the form $p : S \times S \to [0, 1]$ with $\sum_{t \in S} p(s, t) = 1$ for all $s \in S$. (The corresponding transition probability $\widetilde{p} : S \times \mathcal{S} \to \mathbb{R}$ is given by $\widetilde{p}(x, B) = \sum_{y \in B} p(x, y)$.) The interpretation is $p(s, t) = P(X_{n+1} = t \,|X_n = s\,)$.

Example 6.2 (Branching Processes)**.** Let $\xi_1, \xi_2, ... \in \mathbb{N}_0$ be i.i.d. The Galton-Watson process can be viewed as a Markov chain on $\mathbb{N}_0$ with transition function $p(i, j) = P\left(\sum_{k=1}^{i} \xi_k = j\right)$. (If the current population size is i, and individual k has ξ_k offspring, the next generation will have population size $\sum_{k=1}^{i} \xi_k$.)

Example 6.3 (Birth and Death Chains)**.** Birth and death chains are defined by the condition $X_n \in \mathbb{N}_0$ with $|X_n - X_{n+1}| \leq 1$. In terms of transition probabilities, this means that $p_0 + r_0 = 1$ and $p_k + r_k + q_k = 1$ for $k \geq 1$ where $p_n = p(n, n + 1)$, $r_n = p(n, n)$, and $q_n = p(n, n - 1)$. One can think of the associated chains as giving population sizes in successive generations in which at most one birth or death can occur per generation.

Example 6.4 (M/G/1 Queue)**.** The $M/G/1$ queue is a model of line lengths at a service station. M stands for Markovian (or memoryless) and means that customers arrive according to a rate λ Poisson process; G indicates that the service times follow a general distribution F; and 1 is because there is a single server. We assume that the line can be arbitrarily long and that the priority is "first come, first serve."

The time steps correspond to new customers being served, so that X_n is the length of the queue when customer n begins their service. $X_0 = x$ is the number of people in line when service opens with customer 0. Let $a_k = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^k}{k!} dF(t)$ be the probability that k customers arrive during a service time, and let ξ_n denote the net number of customers to enter the queue during the service of customer n, keeping in mind that customer n completed their service in this time period. Our assumptions imply that that $\xi_0, \xi_1, ...$ are i.i.d with $P(\xi_n = k - 1) = a_k$, and we have $X_{n+1} = (X_n + \xi_n)^+$. We take positive parts because if there is no one waiting when the n^{th} customer begins their service ($X_n = 0$) and no customers arrive during this service time ($\xi_n = -1$), then the next queue length is 0 since we don't start counting until the next customer arrives and begins service.

It is not difficult to see that X_n defines a Markov chain with transition probabilities

$$p(0, 0) = a_0 + a_1,$$
$$p(j, j + k - 1) = a_k \text{ if } j \geq 1 \text{ or } k > 1.$$

Example 6.5 (Random Walks on Graphs)**.** Let $G = (V, E)$ be a simple undirected graph with vertex set V and edge set E. For $u, v \in V$, write $u \sim v$ if $\{u, v\} \in E$. Assume that $\sup_{v \in V} \deg(v) < \infty$ where $\deg(v) = \sum_{u \in V} 1\{u \sim v\}$ is the degree of v. A simple random walk on G proceeds by moving from the present vertex to a neighbor chosen uniformly at random - that is, $p(u, v) = \dfrac{1}{\deg(u)} 1\{v \sim u\}$.

More generally, suppose that $G = (V, \overrightarrow{E})$ is a directed graph (possibly containing self-loops) and let $w : \overrightarrow{E} \to [0, \infty)$. One can define a random walk on V by $p(u, v) = \dfrac{w(\{u, v\})}{\sum_{x:\{u,x\} \in \overrightarrow{E}} w(\{u, x\})}$.

In fact, every Markov chain on a countable state space S can be interpreted as a random walk on the directed graph having vertices indexed by S, edge set $\{\{u, v\} : p(u, v) > 0\}$, and edge weights $w(\{u, v\}) = p(u, v)$.

Example 6.6 (Random Walks on Groups/Card Shuffling)**.** Any probability μ on a countable group G induces a random walk via $p(g, h) = \mu(hg^{-1})$. The Markov chain is defined by $X_{n+1} = g_{n+1} X_n$ where $g_1, g_2, \ldots$ are chosen independently from μ.

For example, let $G = (\mathbb{Z}/2\mathbb{Z})^d$, and let $\mu(x) = \frac{1}{d}$ if x has exactly one coordinate equal to one, $\mu(x) = 0$ otherwise. The associated Markov chain, X_n, is equivalent to simple random walk on the hypercube.

If we define $\|x\| = |\{i \in [d] : x_i = 1\}|$, then one can verify that $Y_n = \|X_n\|$ is a Markov chain. In fact, Y_n is equivalent to the Ehrenfest chain mentioned in Durrett (Example 6.2.5).

As another example, let $G = S_n$, and let $\mu(\tau) = \binom{n}{2}^{-1} 1\{\tau = (ij)\}$ be the uniform distribution on transpositions. We can think of permutations as representing arrangements of a deck of cards: $\sigma \in S_n$ corresponds to the ordering in which $\sigma(k)$ is the label of the kth card from the top. (Equivalently, the card labeled l is in position $\sigma^{-1}(l)$.)

Left-multiplying σ by $\tau = (ij)$ corresponds to interchanging card i and card j in the deck:

$$\tau \circ \sigma(k) = \begin{cases} \sigma(k), & \sigma(k) \notin \{i, j\} \\ i, & \sigma(k) = j \\ j, & \sigma(k) = i \end{cases}.$$

Thus we can think of the random walk in terms of repeatedly shuffling the deck by randomly transposing pairs of cards.

In card shuffling applications, it is often more convenient to multiply on the right – so $X_{n+1} = X_n \sigma$, $p(\sigma, \pi) = \mu(\sigma^{-1}\pi)$ – as we typically want shuffles to act on positions rather than labels.

For the random transposition case, right-multiplying σ by $\tau = (ij)$ corresponds to interchanging the card in position i with the card in position j:

$$\sigma \circ \tau(k) = \begin{cases} \sigma(k), & k \notin \{i, j\} \\ \sigma(j), & k = i \\ \sigma(i), & k = j \end{cases}.$$

In this example, the two conventions are essentially equivalent, but typically there is a distinction.

Consider shuffling by removing the top card and inserting it in a random position. Here we need to multiply on the right by permutations distributed according to $\mu(\sigma) = \frac{1}{n} 1\{\sigma = (1 \cdots k) \text{ for some } k \in [n]\}$.
Left multiplication by $(1 \cdots k)$ would correspond to replacing the card labeled k with the card labeled 1 and the card labeled j with that labeled $j + 1$ for $j < k$. This requires looking at the cards.

The inverse of this "top-to-random shuffle" – namely, placing a randomly chosen card on the top of the deck – corresponds to right-multiplication by a cycle of the form $(k \cdots 1) = (1 \cdots k)^{-1}$ with k chosen uniformly from $[n]$.

Note that the left-invariant walk (having kernel $p(x,y) = \mu(x^{-1}y)$) transforms into the right-invariant walk driven by $\check{\mu}(g) = \mu(g^{-1})$ under the anti-automorphism $x \mapsto x^{-1}$, so it suffices to stick with one convention for developing theory and then translate the results when a particular model is better suited to the other choice.

When $S = \{s_1, ..., s_n\}$ is finite (as in some of the latter examples), the transition probabilities can be encoded in a *transition matrix* $K \in M_n(\mathbb{R})$ defined by $K_{i,j} = p(s_i, s_j)$. One nice thing about finite state space Markov chains is that one can often prove powerful results by using ideas from linear algebra.

Some texts have different conventions regarding the indexing of transition matrices. For us, the x,y-entry represents the probability of moving from x to y in a single step. It follows that probability distributions are represented by row vectors and functions by column vectors.

That is, if K is the transition matrix for X_n, μ is a probability measure on $(S, \mathcal{S})$, and $f : S \to \mathbb{R}$, then

$$(\mu K)(y) = \sum_{x \in S} \mu(x) K(x,y) = P(X_{n+1} = y \,|X_n \sim \mu)$$

and

$$(Kf)(x) = \sum_{y \in S} K(x,y) f(y) = E[f(X_{n+1}) \,|X_n = x].$$

For countably infinite state spaces, the "transition matrix" is infinite, so not all linear algebraic results carry over directly. However, the operator perspective is still convenient.

For example, Theorem 6.2 implies that

$$P_\mu(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \mu(x_0) \prod_{m=1}^{n} p(x_{m-1}, x_m).$$

When $n = 1$, we have

$$P_\mu(X_1 = y) = \sum_x P_\mu(X_0 = x, X_1 = y) = \sum_x \mu(x) p(x,y) = (\mu p)(y).$$

When $n = 2$, $\mu = \delta_x$,

$$P_x(X_2 = z) = \sum_y P_x(X_1 = y, X_2 = z) = \sum_y p(x,y) p(y,z) = p^2(x,z).$$

It follows by induction that

$$P_x(X_n = z) = \sum_y P_x(X_{n-1} = y, X_n = z) = \sum_y p^{n-1}(x,y) p(y,z) = p^n(x,z),$$

and thus

$$P_\mu(X_n = z) = \sum_x \mu(x) P_x(X_n = z) = \sum_x \mu(x) p^n(x,z) = (\mu p^n)(z).$$

7. Extensions of the Markov Property

The Markov property reads $E[1_B(X_{n+1})|\mathcal{F}_n] = E[1_B(X_{n+1})|X_n]$ for all $B \in \mathcal{S}$. By the usual argument of approximating by simple functions, we see that this is equivalent to $E[h(X_{n+1})|\mathcal{F}_n] = E[h(X_{n+1})|X_n]$ for all bounded measurable h.

In this section, we show that for discrete time Markov chains, the Markov property extends to the whole future and to random times.

Specifically, we have that for all bounded measurable h and all $n \in \mathbb{N}_0$,

$$E[h(X_n, X_{n+1}, \ldots)|\mathcal{F}_n] = E[h(X_n, X_{n+1}, \ldots)|X_n].$$

Moreover, if N is a stopping time, then the above holds with N in place of n when we restrict to the event $\{N < \infty\}$.

We will assume throughout that the X_n are coordinate maps on the sequence space $(\Omega, \mathcal{F}) = (S^{\mathbb{N}_0}, \mathcal{S}^{\mathbb{N}_0})$ and that $\mathcal{F}_n = \sigma(X_0, X_1, ..., X_n)$. For each probability measure μ on $(S, \mathcal{S})$, we write P_μ for the measure on $(\Omega, \mathcal{F})$ that makes X_n a Markov chain with initial distribution μ and transition probability p.

Also, we recall that the maps $\theta^n : \Omega \to \Omega$ act by shifting coordinates: $(\theta^n\omega)_i = \omega_{i+n}$.

We begin by showing that the Markov property is not limited to a single time step.

Theorem 7.1. *If $Y : \Omega \to \mathbb{R}$ is bounded and measurable, then*

$$E_\mu[Y \circ \theta^n | \mathcal{F}_n] = E_{X_n}[Y]$$

where the subscript on the left indicates that the conditional expectation is taken with respect to P_μ, and the expression on the right is $\varphi(x) = E_x[Y]$ evaluated at $x = X_n$.

Proof. Let $A = \{\omega : \omega_0 \in A_0, ..., \omega_n \in A_n\}$ for some $A_0, ..., A_n \in \mathcal{S}$ and let $g_0, ..., g_m : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{B})$ be bounded and measurable.

Applying the formula

$$E_\mu\Big[\prod_{i=0}^{N} f_i(X_i)\Big] = \int \mu(dx_0) f_0(x_0) \int p(x_0, dx_1) f_1(x_1) \cdots \int p(x_{N-1}, dx_N) f_N(x_N)$$

with $f_k = 1_{A_k}$ for $k < n$, $f_n = 1_{A_n} g_0$, and $f_{n+j} = g_j$ for $1 \le j \le m$, we have

$$\begin{aligned} E_\mu\Big[\prod_{k=0}^{m} g_k(X_{n+k}); A\Big] &= \int_{A_0} \mu(dx_0) \int_{A_1} p(x_0, dx_1) \cdots \int_{A_n} p(x_{n-1}, dx_n) \\ &\qquad \cdot g_0(x_n) \int p(x_n, dx_{n+1}) g_1(x_{n+1}) \cdots \int p(x_{m+n-1}, dx_{m+n}) g_m(x_{m+n}) \\ &= E_\mu\Big[E_{X_n}\Big[\prod_{k=0}^{m} g_k(X_k)\Big]; A\Big]. \end{aligned}$$

The collection of sets A for which this holds is a λ-system, and the collection of sets for which it has been proved is a π-system that generates $\mathcal{F}_n$, so the $\pi - \lambda$ theorem shows that it is true for all $A \in \mathcal{F}_n$.

Thus if $Y(\omega) = \prod_{k=0}^{m} g_k(\omega_k)$ for $g_0, ..., g_m$ bounded and measurable, then

$$\int_A Y \circ \theta^n \, dP_\mu = \int_A \prod_{k=0}^{m} g_k(X_{n+k}) \, dP_\mu = E_\mu\Big[\prod_{k=0}^{m} g_k(X_{n+k}); A\Big]$$
$$= E_\mu\Big[E_{X_n}\Big[\prod_{k=0}^{m} g_k(X_k)\Big]; A\Big] = \int_A E_{X_n}\Big[\prod_{k=0}^{m} g_k(X_k)\Big] dP_\mu = \int_A E_{X_n}[Y] \, dP_\mu$$

for all $A \in \mathcal{F}_n$, hence $E_\mu[Y \circ \theta^n \,|\mathcal{F}_n] = E_{X_n}[Y]$ for all such Y.

To complete the proof, we observe that the collection $\mathcal{A}$ of events of the form $\{\omega_0 \in A_0, \ldots, \omega_k \in A_k\}$ is a π-system that generates $\mathcal{S}^{\mathbb{N}_0}$. The set of functions $\mathcal{H} = \{Y : E_\mu[Y \circ \theta^n \,|\mathcal{F}_n] = E_{X_n}[Y]\}$ contains the indicators of events in $\mathcal{A}$ (take $g_k = 1_{A_k}$ in the preceding), and is certainly closed under sums, scalar multiples, and increasing limits, so Theorem 6.3 implies that $\mathcal{H}$ contains all bounded measurable Y. □

Thus, conditional on $\mathcal{F}_n$, $X_n, X_{n+1}, ...$ has the same distribution as a copy of the chain started at X_n. Markov chains are forgetful; they start fresh at every step.

It should be noted that Theorem 7.1 depends on the assumption of time homogeneity.
In general, writing $Y(\omega) = h(\omega_0, \omega_1, \ldots)$, so that $Y \circ \theta^n = h(X_n, X_{n+1}, \ldots)$, the proof of Theorem 7.1 gives

$$E[h(X_n, X_{n+1}, ...) \,|\mathcal{F}_n] = E[h(X_n, X_{n+1}, ...) \,|X_n].$$

Another interpretation of this statement of the Markov property is that the past and the future are conditionally independent given the present:

Corollary 7.1. *If $A \in \sigma(X_0, ..., X_n)$ and $B \in \sigma(X_n, X_{n+1}, ...)$, then for any initial distribution μ,*

$$P_\mu(A \cap B \,|X_n) = P_\mu(A \,|X_n) \, P_\mu(B \,|X_n).$$

Proof. By Theorem 7.1 and basic properties of conditional expectation,

$$P_\mu(A \cap B \,|X_n) = E_\mu[1_A 1_B \,|X_n] = E_\mu[E_\mu[1_A 1_B \,|\mathcal{F}_n] \,|X_n]$$
$$= E_\mu[1_A E_\mu[1_B \,|\mathcal{F}_n] \,|X_n] = E_\mu[1_A E_\mu[1_B \,|X_n] \,|X_n]$$
$$= E_\mu[1_A \,|X_n] \, E_\mu[1_B \,|X_n] = P_\mu(A \,|X_n) \, P_\mu(B \,|X_n).$$ □

A useful application of the Markov property is the following intuitive decomposition result for expressing multi-step transition probabilities in terms of convolution (matrix multiplication):

Proposition 7.1 (Chapman-Kolmogorov)**.** *If X_n is time homogeneous with discrete state space, then*

$$P_x(X_{m+n} = z) = \sum_y P_x(X_m = y) \, P_y(X_n = z).$$

Proof. Since $1_{\{z\}}(X_{m+n}(\omega)) = \left(1_{\{z\}} \circ X_n \circ \theta^m\right)(\omega)$, Theorem 7.1 shows that

$$P_x(X_{m+n} = z) = E_x[1_{\{z\}}(X_{m+n})] = E_x\left[E_x\left[\left(1_{\{z\}} \circ X_n\right) \circ \theta^m \,|\mathcal{F}_m\right]\right]$$
$$= E_x\left[E_{X_m}\left[1_{\{z\}} \circ X_n\right]\right] = E_x[P_{X_m}(X_n = z)] = \sum_y P_x(X_m = y) \, P_y(X_n = z).$$ □

Our second extension is known as the Strong Markov Property which generalizes the original definition by replacing deterministic times with stopping times.

Recall that if N is a stopping time with respect to a filtration $\mathcal{F}_n$, then the stopped σ-algebra $\mathcal{F}_N$ consists of all events $A \in \mathcal{F}$ such that $A \cap \{N = n\} \in \mathcal{F}_n$ for all n.

Also, remember that we defined random shifts by

$$\theta^N \omega = \begin{cases} \theta^n \omega, & \omega \in \{N = n\} \\ \Delta, & \omega \in \{N = \infty\} \end{cases}$$

where Δ is an extra point we add to Ω for convenience. In what follows, we will restrict our attention to $\{N < \infty\}$, so this extra point need not concern us.

Theorem 7.2 (Strong Markov Property)**.** *Let N be a stopping time and suppose that $Y : \Omega \to \mathbb{R}$ is bounded and measurable. Then*

$$E_\mu \left[Y \circ \theta^N \, | \mathcal{F}_N \right] = E_{X_N}[Y] \text{ on } \{N < \infty\}.$$

Proof. For any $A \in \mathcal{F}_N$,

$$\begin{aligned} E_\mu \left[Y \circ \theta^N ; A \cap \{N < \infty\} \right] &= \sum_{n=0}^{\infty} E_\mu \left[Y \circ \theta^n ; A \cap \{N = n\} \right] \\ &= \sum_{n=0}^{\infty} E_\mu \left[E_{X_n}[Y] ; A \cap \{N = n\} \right] \\ &= E_\mu \left[E_{X_N}[Y] ; A \cap \{N < \infty\} \right], \end{aligned}$$

and the result follows by definition of conditional expectation. □

The above proof is representative of many results for discrete stopping times - one sums over possible values of N and then applies existing results to the summands. This trick doesn't work in continuous time and the corresponding theorems can be much less trivial.

While every discrete time Markov process has the strong Markov property, the two notions do not necessarily coincide in continuous time. (For example, $B_t 1\{B_0 \neq 0\}$ is Markov but not strong Markov.)

8. Classifying States

We will restrict our attention henceforth to chains having a countable state space.

Define $T_y^0 = 0$ and

$$T_y^k = \inf\left\{n > T_y^{k-1} : X_n = y\right\}$$

(where $\inf \emptyset = \infty$), so that T_y^k is the time of the k^{th} visit to y at positive times.

Set $T_y = T_y^1$ and let $\rho_{xy} = P_x\left(T_y < \infty\right)$ be the probability that the chain started at x visits y in finitely many steps.

Theorem 8.1. $P_x\left(T_y^k < \infty\right) = \rho_{xy}\rho_{yy}^{k-1}$.

Proof. The result follows from the definition of ρ_{xy} when $k = 1$, so can assume that $k \geq 2$.

Now let $Y(\omega) = 1\left\{\omega_n = y \text{ for some } n \in \mathbb{N}\right\} = 1\left\{T_y < \infty\right\}$.

Setting $N = T_y^{k-1}$, we have $Y \circ \theta^N = 1\left\{\omega_n = y \text{ for some } n > T_y^{k-1}\right\} = 1\left\{T_y^k < \infty\right\}$ on $\{N < \infty\}$.

Also,

$$E_x[Y \circ \theta^N \,|\mathcal{F}_N\,] = E_{X_N}[Y] \text{ on } \{N < \infty\}$$

by the strong Markov property.

As $X_N = y$ on $\{N < \infty\}$, the right-hand side is $E_{X_N}[Y] = E_y[Y] = P_y\left(T_y < \infty\right) = \rho_{yy}$.

Thus, since $1\{N < \infty\} \in \mathcal{F}_N$, we have

$$\begin{aligned} P_x\left(T_y^k < \infty\right) &= E_x\left[1\left\{T_y^k < \infty\right\}\right] = E_x\left[Y \circ \theta^N; N < \infty\right] \\ &= E_x\left[E_x[Y \circ \theta^N \,|\mathcal{F}_N\,]; N < \infty\right] \\ &= E_x\left[\rho_{yy}; N < \infty\right] = \rho_{yy} P_x\left(T_y^{k-1} < \infty\right), \end{aligned}$$

and the result follows by induction. □

Definition. We say that $y \in S$ is a *recurrent state* if $\rho_{yy} = 1$ and a *transient state* if $\rho_{yy} < 1$.

If every $x \in S$ is a recurrent state, we say that the chain is *recurrent*.

If y is recurrent, then Theorem 8.1 shows that $P_y\left(T_y^k < \infty\right) = \rho_{yy}^k = 1$ for all $k \in \mathbb{N}$, hence $P_y\left(X_n = y \text{ i.o.}\right) = 1$.

If y is transient and we let $N(y) = \sum_{k=1}^{\infty} 1\left\{X_k = y\right\}$ be the number of visits to y at positive times, then

$$\begin{aligned} E_x[N(y)] &= \sum_{k=1}^{\infty} P_x\left(N(y) \geq k\right) = \sum_{k=1}^{\infty} P_x\left(T_y^k < \infty\right) \\ &= \sum_{k=1}^{\infty} \rho_{xy}\rho_{yy}^{k-1} = \rho_{xy} \sum_{k=0}^{\infty} \rho_{yy}^k = \frac{\rho_{xy}}{1 - \rho_{yy}} < \infty. \end{aligned}$$

Combining these observations gives

Theorem 8.2. *y is recurrent if and only if $E_y\left[N(y)\right] = \infty$.*

Definition. We say that y is *accessible* from x if $\rho_{xy} > 0$. If $x = y$ or x is accessible from y and y is accessible from x, we say that x and y *communicate*.

Our next theorem says that if x is recurrent, then so are all states accessible from x. (Recurrence is contagious.)

Theorem 8.3. *If x is recurrent and $\rho_{xy} > 0$, then y is recurrent and $\rho_{yx} = 1$.*

Proof. We first prove that $\rho_{yx} = 1$ by showing that $\rho_{xy} > 0$ and $\rho_{yx} < 1$ implies $\rho_{xx} < 1$.
Let $K = \inf\left\{k : p^k(x,y) > 0\right\}$. ($K < \infty$ since $\rho_{xy} > 0$.) Then there is a sequence $y_1, ..., y_{K-1}$ such that $p(x,y_1)p(y_1,y_2)\cdots p(y_{K-1},y) > 0$. Moreover, since K is minimal, $y_k \neq x$ for $k = 1, ..., K-1$.
If $\rho_{yx} < 1$, then we have

$$1 - \rho_{xx} = P_x\left(T_x = \infty\right) \geq p(x,y_1)p(y_1,y_2)\cdots p(y_{K-1},y)(1-\rho_{yx}) > 0,$$

a contradiction.

To prove that y is recurrent, we note that $\rho_{yx} > 0$ implies that there is an L with $p^L(y,x) > 0$. Because

$$p^{L+n+K}(y,y) \geq p^L(y,x)p^n(x,x)p^K(x,y),$$

we see that

$$\begin{aligned} E_y\left[N(y)\right] &= \sum_{k=1}^{\infty} E_y\left[1\left\{X_k = y\right\}\right] = \sum_{k=1}^{\infty} P_y\left(X_k = y\right) = \sum_{k=1}^{\infty} p^k(y,y) \geq \sum_{k=L+K+1}^{\infty} p^k(y,y) \\ &= \sum_{n=1}^{\infty} p^{L+n+K}(y,y) \geq p^L(y,x)p^K(x,y)\sum_{n=1}^{\infty} p^n(x,x) = p^L(y,x)p^K(x,y)E_x\left[N(x)\right] = \infty, \end{aligned}$$

so y is recurrent by Theorem 8.2. □

Theorem 8.3 allows us to conclude that states accessible from x are recurrent provided that we already know x is recurrent. It is useful at this point to introduce the following definitions.

Definition. A set $D \subseteq S$ is called a *communicating class* if all states in D communicate: $x, y \in D$ implies $x = y$ or $\rho_{xy} > 0$. A communicating class in which every state is accessible from itself ($\rho_{xx} > 0$ for all $x \in D$) is called *irreducible*. (In general, a set $B \subseteq S$ is irreducible if $\rho_{xy} > 0$ for all $x, y \in B$.)
If S itself is irreducible, we say that the chain is *irreducible*.

Proposition 8.1. *The communicating classes partition the state space.*

Proof. "Communicates with" is reflexive and symmetric by definition, thus we need only establish transitivity. This is trivial if x, y, z are not all distinct, so (because of symmetry) it suffices to show that $\rho_{xy}, \rho_{yz} > 0$ implies $\rho_{xz} > 0$.
But this is a simple consequence of the strong Markov property since $P_x\left(T_z \circ \theta^{T_y} < \infty \left| \mathcal{F}_{T_y}\right.\right) = P_y\left(T_z < \infty\right)$ on $\{T_y < \infty\}$, thus

$$\begin{aligned} \rho_{xz} = P_x\left(T_z < \infty\right) &\geq P_x\left(T_z \circ \theta^{T_y} < \infty; T_y < \infty\right) \\ &= E_x\left[P_x\left(T_z \circ \theta^{T_y} < \infty \left| \mathcal{F}_{T_y}\right.\right); T_y < \infty\right] \\ &= E_x\left[P_y\left(T_z < \infty\right); T_y < \infty\right] \\ &= P_y\left(T_z < \infty\right) E_x\left[T_y < \infty\right] = \rho_{yz}\rho_{xy} > 0. \end{aligned}$$

(Or you could just use $p^{K+L}(x,z) \geq p^K(x,y)p^L(y,z)$...) □

Note that communicating classes consisting of a single element, x, are not necessarily irreducible as it may be the case that $\rho_{xx} = 0$.

However, the proof of transitivity shows that every class containing more than one state is irreducible.

Also, any communicating class containing a recurrent state is irreducible as it either contains multiple elements or consists of a single recurrent state.

More importantly, Theorem 8.3 shows that recurrence is a class property - either every element in a communicating class is recurrent or every element is transitive.

Thus if we have identified a communicating class, we can check recurrence for all members simultaneously by testing a single element.

In many cases, this is simplified by the next result.

Definition. A set $C \subseteq S$ is said to be *closed* if it contains all points accessible from any of its elements: $x \in C$ and $\rho_{xy} > 0$ implies $y \in C$. The reason for the name is that if C is closed and $x \in C$, then $P_x(X_n \in C) = 1$ - there is no escaping C.
If a singleton $\{z\}$ is closed, we say that the state z is *absorbing*.

Theorem 8.4. *If C is a finite closed set, then it contains a recurrent state. If, in addition, C is a communicating class, then all states in C are recurrent.*

Proof. It suffices to prove the first statement as the second then follows from Theorem 8.3.

To this end, suppose that C is a finite closed set with $\rho_{yy} < 1$ for all $y \in C$. Then $E_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty$ for all $x, y \in C$, so, since C is finite, we have the contradiction that for any $x \in C$,

$$\infty > \sum_{y \in C} E_x[N(y)] = \sum_{y \in C} \sum_{n=1}^{\infty} p^n(x,y) = \sum_{n=1}^{\infty} \sum_{y \in C} p^n(x,y) = \sum_{n=1}^{\infty} 1 = \infty$$

where the penultimate equality follows from the fact that C is closed. □

Corollary 8.1. *Every irreducible Markov chain on a finite state space is recurrent.*

Proof. S is closed. □

Theorem 8.3 provides a simple test of transience:
If there exists a $y \in S$ such that $\rho_{xy} > 0$ and $\rho_{yx} < 1$, then $[x]$ is transient.

Theorem 8.4 gives a similar recurrence test for states in a finite communicating class:
If $|[x]| < \infty$ and $\rho_{xy} > 0$ implies $\rho_{yx} > 0$, then $[x]$ is recurrent.

(The assumptions imply that $[x]$ is closed since $\rho_{xw} > 0$ implies $\rho_{wx} > 0$, hence $w \in [x]$; and $y \in [x] \setminus \{x\}$ and $\rho_{yz} > 0$ implies $\rho_{xz} \geq \rho_{xy}\rho_{yz} > 0$, so $\rho_{zx} > 0$ as well, hence $z \in [x]$.)

To recap, we can partition the state space into communicating classes, each of which is either recurrent or transient.

All classes containing at least two states (as well as certain single-state classes, such as those consisting of an absorbing state) are irreducible.

Communicating classes are not necessarily closed, but Theorem 8.3 shows that any communicating class containing a recurrent element is both closed and irreducible.

It follows from Proposition 8.1 that the set of recurrent states $R = \{x \in S : \rho_{xx} = 1\}$ can be expressed as a disjoint union of closed and irreducible communicating classes.

We conclude this discussion by considering recurrence/transience behavior in some concrete examples.

Example 8.1 (Random Walks on Finite Groups)**.** If G is a finite group and X_n is a Markov chain on G with transition function $p(g,h) = \mu(hg^{-1})$ for some probability μ on G, then X_n is irreducible if and only if the support of μ, $\Sigma = \{g \in G : \mu(g) > 0\}$, generates G. If so, then for any $r, s \in G$, there exist $g_{i_1}, ..., g_{i_k} \in \Sigma$ such that $g_{i_k} \cdots g_{i_1} = sr^{-1}$, hence $\rho_{rs} \geq p^k(r,s) \geq \mu(g_{i_1}) \cdots \mu(g_{i_k}) > 0$. If not, there is some $g \in G$ which cannot be expressed as a finite product of terms in Σ, so $\rho_{eg} = 0$ where e is the identity in G.

Whether or not the chain is irreducible, all states are recurrent. If X_n is not irreducible, then Σ generates a proper subgroup $H < G$. The communicating classes are precisely the right cosets of H, and they are all closed.

Example 8.2 (Branching Processes)**.**

If the offspring distribution has mass $p_0 > 0$ at zero (i.e. there is positive probability that an individual has no children), then for every $k \geq 1$, $\rho_{k0} \geq p_0^k > 0$. Since $\rho_{0k} = 0$ for all such k, we see that every state $k \geq 1$ is transient. 0 is recurrent because $p(0,0) = 1$.

Example 8.3 (Birth and Death Chains on $\mathbb{N}_0$)**.**

Denote

$$p(i, i+1) = p_i, \quad p(i, i-1) = q_i, \quad p(i,i) = r_i$$

where $q_0 = 0$ and $p_{k-1}, q_k > 0$ for $k \geq 1$. (The latter condition ensures that the chain is irreducible.)

Let $N = \inf\{n \geq 0 : X_n = 0\}$. We will define a function $\varphi : \mathbb{N}_0 \to \mathbb{R}$ so that $\varphi(X_{N \wedge n})$ is a martingale.

We begin by imposing the conditions $\varphi(0) = 0$ and $\varphi(1) = 1$. For the martingale property to hold when $X_n = k \geq 1$, we must have

$$\varphi(k) = p_k \varphi(k+1) + r_k \varphi(k) + q_k \varphi(k-1),$$

or

$$(p_k + q_k)\varphi(k) = (1 - r_k)\varphi(k) = p_k \varphi(k+1) + q_k \varphi(k-1).$$

Dividing by p_k and rearranging gives

$$\varphi(k+1) - \varphi(k) = \frac{q_k}{p_k}\big(\varphi(k) - \varphi(k-1)\big).$$

Since $\varphi(1) - \varphi(0) = 1$, we have

$$\varphi(m+1) - \varphi(m) = \prod_{j=1}^{m} \frac{q_j}{p_j} \text{ for } m \geq 1,$$

hence

$$\begin{aligned}\varphi(n) &= \varphi(1) + \sum_{m=1}^{n-1} \big(\varphi(m+1) - \varphi(m)\big) \\ &= 1 + \sum_{m=1}^{n-1} \prod_{j=1}^{m} \frac{q_j}{p_j} = \sum_{m=0}^{n-1} \prod_{j=1}^{m} \frac{q_j}{p_j} \text{ for } n \geq 1\end{aligned}$$

where we adopt the convention that the empty product equals 1.

Now for any $a < x < b$, let $T = T_a \wedge T_b$ where $T_z = \inf\{n \geq 1 : X_n = z\}$.

Since $\varphi\left(X_{T \wedge n}\right)$ is a bounded martingale and $X_T \in \{a, b\}$ P_x-a.s., optional stopping gives

$$\varphi(x) = E_x[\varphi(X_T)] = \varphi(b) P_x\left(T_a > T_b\right) + \varphi(a)\left[1 - P_x\left(T_a > T_b\right)\right],$$

hence

$$P_x(T_a > T_b) = \frac{\varphi(x) - \varphi(a)}{\varphi(b) - \varphi(a)}.$$

Taking $a = 0$, $b = M$ gives

$$P_x\left(T_0 > T_M\right) = \frac{\varphi(x)}{\varphi(M)},$$

so letting $M \to \infty$ and observing that $T_M \geq M - x$, we see that 0 is recurrent if and only if

$$\sum_{m=0}^{M} \prod_{j=1}^{m} \frac{q_j}{p_j} = \varphi(M) \to \infty \text{ as } M \to \infty.$$

(φ is increasing and thus has a limit $\varphi(\infty) \in [1, \infty]$. If $\varphi(\infty) = \infty$, then $P_1(T_0 = \infty) = 0$, hence 0 is recurrent as the chain started at 0 is either 0 or 1 after one time step. If $\varphi(\infty) < \infty$, then $P_1(T_0 = \infty) = \frac{1}{\varphi(\infty)} > 0$, so as long as $p_0 > 0$, there is positive probability that the chain started at 0 never returns.)

9. Stationary Measures

Definition. If X_n is a Markov chain with state space $(S, \mathcal{S})$ and transition function p, we say that a measure $\mu \not\equiv 0$ on $(S, \mathcal{S})$ is a *stationary measure* for X_n if it satisfies the *equilibrium equation*

$$\mu(y) = \sum_{x \in S} \mu(x) p(x, y)$$

for every $y \in S$. If μ is a probability measure, it is called a *stationary distribution*.

Of course, if a stationary measure μ is finite, then $\pi(y) = \dfrac{\mu(y)}{\sum_x \mu(x)}$ is a stationary distribution.

If π is a stationary distribution for X_n, then the equilibrium equation reads $P_\pi(X_1 = y) = \pi(y)$.

Using the Markov property and induction, it follows that $P_\pi(X_n = y) = \pi(y)$ for all n, hence the name stationary.

From the transition operator perspective, the equilibrium equation reads $\mu p = \mu$, so a stationary measure is a left eigenfunction with eigenvalue 1.

Example 9.1 (Random Walk on $\mathbb{Z}^d$)**.** Here $p(x, y) = f(y - x)$ where $f \geq 0$ and $\sum_z f(z) = 1$. In this case, $\mu \equiv 1$ is a stationary measure since

$$\mu(y) = 1 = \sum_z f(z) = \sum_x 1 \cdot f(y - x) = \sum_x \mu(x) p(x, y).$$

Of course $\nu \equiv k$ is also a stationary measure, and in general, any positive multiple of a stationary measure is stationary.

Note that $\nu \equiv k$ is not a stationary distribution since $\sum_x \nu(x) = \infty$.

Example 9.2 (Asymmetric Simple Random Walk)**.** Here $S = \mathbb{Z}$ and $p(x, x+1) = p$, $p(x, x-1) = q = 1 - p$. The preceding example shows that $\mu \equiv 1$ is a stationary measure. If $p \neq q$, another stationary measure is given by $\nu(x) = \left(\frac{p}{q}\right)^x$ since

$$\begin{aligned}
\sum_x \nu(x) p(x, y) &= \nu(y - 1) p(y - 1, y) + \nu(y + 1) p(y + 1, y) \\
&= \left(\frac{p}{q}\right)^{y-1} p + \left(\frac{p}{q}\right)^{y+1} q = \frac{p^y q}{q^y} + \frac{p^{y+1}}{q^y} \\
&= \frac{p^y (q + p)}{q^y} = \frac{p^y}{q^y} = \nu(y).
\end{aligned}$$

Example 9.3 (Random Walks on a Finite Group)**.** Suppose that G is a finite group and X_n is a Markov chain on G with transition probabilities $p(x, y) = f(yx^{-1})$ for some probability f on G. Then $\pi \equiv \frac{1}{|G|}$ is a stationary distribution since

$$\pi(y) = \frac{1}{|G|} = \frac{1}{|G|} \sum_{g \in G} f(g) = \frac{1}{|G|} \sum_{x \in G} f(yx^{-1}) = \sum_{x \in G} \pi(x) p(x, y).$$

Example 9.4 (Birth and Death Chains)**.** Here $S = \mathbb{N}_0$, $p(i, i+1) = p_i$, $p(i,i) = r_i$, and $p(i, i-1) = q_i$, with $q_0 = 0$ and $p(i,j) = 0$ for $|i-j| > 1$.

The measure $\mu(k) = \prod_{j=1}^{k} \frac{p_{j-1}}{q_j}$ has

$$\mu(i)p(i,i+1) = p_i \prod_{j=1}^{i} \frac{p_{j-1}}{q_j} = q_{i+1} \prod_{j=1}^{i+1} \frac{p_{j-1}}{q_j} = \mu(i+1)p(i+1,1).$$

Since $p(x,y) = 0$ if $|x-y| > 1$, this implies that μ satisfies the *detailed balance equations*:

$$\mu(x)p(x,y) = \mu(y)p(y,x) \text{ for all } x, y \in S.$$

Summing over x, we have

$$\sum_x \mu(x)p(x,y) = \mu(y) \sum_x p(y,x) = \mu(y),$$

so any such measure is stationary.

If X_n has a stationary distribution which satisfies the detailed balance equations, we say that X_n is reversible.

To see the reason for the nomenclature, observe that if X_n is reversible with respect to π, then

$$\begin{aligned} P_\pi(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) &= \pi(x_0)p(x_0,x_1)p(x_1,x_2)\cdots p(x_{n-1},x_{n-1}) \\ &= p(x_1,x_0)\pi(x_1)p(x_1,x_2)\cdots p(x_{n-1},x_n) \\ &= p(x_2,x_1)p(x_1,x_0)\pi(x_2)\cdots p(x_{n-1},x_n) \\ &= ... = \pi(x_n)p(x_n,x_{n-1})\cdots p(x_1,x_0) \\ &= P_\pi(X_0 = x_n, X_1 = x_{n-1}, ..., X_n = x_0), \end{aligned}$$

thus in stationarity $(X_0, X_1, ..., X_n) =_d (X_n, ..., X_1, X_0)$.

Random walks on countable groups (like the first three examples) are reversible with respect to counting measure precisely when $f(g) = f(g^{-1})$ for all $g \in G$.

Example 9.5 (Simple Random Walk on a Finite Graph)**.** If $G = (V, E)$ is a simple, undirected graph with $|V| < \infty$ and X_n is a Markov chain on V with transition probabilities $p(x,y) = \frac{1}{\deg(x)} 1\{x \sim y\}$, then $\pi(x) = \frac{\deg(x)}{2|E|}$ is a stationary distribution for X_n since

$$\pi(x)p(x,y) = \frac{1}{2|E|} 1\{x \sim y\} = \pi(y)p(y,x).$$

π is a probability measure on V by the "handshaking lemma," $\sum_{x \in V} \deg(x) = 2|E|$, which follows by observing that if we orient the edges in any way, then $|E| = \sum_x \deg^+(x) = \sum_x \deg^-(x)$ (where $\deg^+$ and $\deg^-$ denote the outdegree and indegree, respectively), hence $2|E| = \sum_x \left(\deg^+(x) + \deg^-(x)\right) = \sum_x \deg(x)$.

Reversibility is a very convenient feature for Markov chains to have, but as the term "detailed balance" suggests, it is much less generic than one might infer from the preceding examples.

For instance, the M/G/1 queue has no reversible measures since if $x > y+1$, then $p(x,y) = 0$, but $p(y,x) > 0$.

We will see shortly that there is a more complicated potential obstruction to reversibility in the case of irreducible Markov chains, but first we present a lemma which is useful in its own right.

Lemma 9.1. *If p is irreducible and μ is a stationary measure for p, then $\mu(x) > 0$ for all states x.*

Proof. Since $\mu(x) \not\equiv 0$, there must be some x_0 with $\mu(x_0) > 0$. Assume that $N = \{y : \mu(y) = 0\} \neq \emptyset$, and let $x_0, x_1, ..., x_k$ be a sequence of minimal length such that $p(x_{i-1}, x_i) > 0$ for each $i = 1, ..., k$ and $x_k \in N$. Such a sequence exists by irreducibility. Since $x_k \in N$, we have that

$$0 = \mu(x_k) = \sum_w \mu(w)p(w, x_k) \geq \mu(x_{k-1})p(x_{k-1}, x_k).$$

It follows that $\mu(x_{k-1}) = 0$, contradicting minimality.

(Alternatively, for any $y \in S$, there is an n with $p^n(x_0, y) > 0$, so $\mu(y) = (\mu p^n)(y) \geq \mu(x_0)p^n(x_0, y) > 0$.) □

Theorem 9.1. *Suppose that p is irreducible. A necessary and sufficient condition for the existence of a reversible measure is*

(i) $p(x, y) > 0$ *implies* $p(y, x) > 0$,

(ii) *For any loop* $x_0, x_1, \ldots, x_n = x_0$ *with* $\prod_{i=1}^n p(x_{i-1}, x_i) > 0$,

$$\prod_{i=1}^n \frac{p(x_{i-1}, x_i)}{p(x_i, x_{i-1})} = 1.$$

Proof. To prove necessity, we note that any stationary measure has $\mu(x) > 0$ for all x (by Lemma 9.1), so the detailed balance equations imply that (i) holds. To check the cycle condition, (ii), we observe that $p(x_{i-1}, x_i) = \frac{\mu(x_i)}{\mu(x_{i-1})} p(x_i, x_{i-1})$, so, since $x_0 = x_n$, we have

$$\prod_{i=1}^n \frac{p(x_{i-1}, x_i)}{p(x_i, x_{i-1})} = \prod_{i=1}^n \frac{\mu(x_i)}{\mu(x_{i-1})} = \frac{\mu(x_n)}{\mu(x_0)} \prod_{i=1}^{n-1} \frac{\mu(x_i)}{\mu(x_i)} = 1.$$

To show that these conditions are sufficient as well, fix $s \in S$ and set $\mu(s) = 1$. By irreducibility, for any $x \in S \setminus \{s\}$, there is a sequence $s = x_0, x_1, \ldots, x_n = x$ with $\prod_{i=1}^n p(x_{i-1}, x_i) > 0$, and we set

$$\mu(x) = \prod_{i=1}^n \frac{p(x_{i-1}, x_i)}{p(x_i, x_{i-1})}.$$

If $s = y_0, y_1, \ldots, y_m = x$ is another such sequence, then (i) implies that $\prod_{j=1}^m p(y_j, y_{j-1}) > 0$, and $s = x_0, x_1, \ldots, x_n = y_m, y_{m-1}, \ldots, y_0 = s$ is a loop, so (ii) implies that

$$\prod_{i=1}^n \frac{p(x_{i-1}, x_i)}{p(x_i, x_{i-1})} \cdot \prod_{j=1}^m \frac{p(y_j, y_{j-1})}{p(y_{j-1}, y_j)} = 1.$$

It follows that μ does not depend on the particular path chosen.

Finally, detailed balance is satisfied since if $p(x, y) > 0$, then consideration of the path $s = x_0, x_1, ..., x, y$ shows that

$$\mu(y) = \mu(x)\frac{p(x, y)}{p(y, x)}.$$ □

Though the existence of a reversible measure implies the existence of a stationary measure, it is not a necessary condition. The following theorem shows that any chain having a recurrent state has a stationary measure.

Theorem 9.2. *If x is a recurrent state and $T_x = \inf\{n \geq 1 : X_n = x\}$, then*

$$\mu_x(y) = E_x\left[\sum_{n=0}^{T_x-1} 1\{X_n = y\}\right] = \sum_{n=0}^{\infty} P_x(X_n = y, T_x > n)$$

defines a stationary measure.

The intuition is that $\mu_x(y)$ is the expected number of visits to y in $\{0, 1, ..., T_x - 1\}$ for the chain started at x and $(\mu_x p)(y)$ is the expected number of visits to y in $\{1, ..., T_x\}$, which is equal since $X_0 = X_{T_x} = x$.

Proof. Tonelli's theorem shows that

$$\begin{aligned}
\sum_y \mu_x(y)p(y,z) &= \sum_y p(y,z) \sum_{n=0}^{\infty} P_x(X_n = y, T_x > n) \\
&= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, T_x > n)p(y,z) \\
&= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, X_{n+1} = z, T_x > n)
\end{aligned}$$

for all $z \in S$.

When $z \neq x$, we have

$$\begin{aligned}
\sum_y \mu_x(y)p(y,z) &= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, X_{n+1} = z, T_x > n) \\
&= \sum_{n=0}^{\infty} P_x(X_{n+1} = z, T_x > n+1) \\
&= \sum_{m=1}^{\infty} P_x(X_m = z, T_x > m) \\
&= \sum_{n=0}^{\infty} P_x(X_n = z, T_x > n) = \mu_x(z)
\end{aligned}$$

since $P_x(X_0 = z) = 0$.

For the $z = x$ case, we note that

$$\mu_x(x) = \sum_{n=0}^{\infty} P_x(X_n = x, T_x > n) = P_x(X_0 = x, T_x > 0) = 1,$$

so

$$\begin{aligned}
\sum_y \mu_x(y)p(y,x) &= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, X_{n+1} = x, T_x > n) \\
&= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, T_x = n+1) \\
&= \sum_{n=0}^{\infty} P_x(T_x = n+1) \\
&= \sum_{m=1}^{\infty} P_x(T_x = m) = 1 = \mu_x(x)
\end{aligned}$$

since x recurrent implies $T_x \in [1, \infty)$ a.s.

Finally, we observe that $\mu_x(y) < \infty$ for all y. To see that this is so, note that since x is recurrent, it follows from Theorem 8.3 that if $\rho_{xy} > 0$, then $\rho_{yx} = 1$, hence $p^n(y,x) > 0$ for some $n \in \mathbb{N}$.
As $\mu_x = \mu_x p$ implies that $\mu_x = \mu_x p^n$, we have

$$1 = \mu_x(x) = (\mu_x p^n)(x) = \sum_w \mu_x(w) p^n(w,x) \geq \mu_x(y) p^n(y,x),$$

thus $\mu_x(y) \leq \frac{1}{p^n(y,x)} < \infty$.
On the other hand, if $\rho_{xy} = 0$, then the definition of μ_x implies that $\mu_x(y) = 0 < \infty$. □

To complement the previous existence result, we have

Theorem 9.3. *If p is irreducible and recurrent, then the stationary measure is unique up to multiplication by a positive constant.*

Proof. Fix $s \in S$ and define T_s, μ_s as in Theorem 9.2. For any stationary measure ν and any $z \neq s$, we have

$$\nu(z) = \sum_y \nu(y) p(y,z) = \nu(s) p(s,z) + \sum_{y \neq s} \nu(y) p(y,z).$$

Repeating this decomposition gives

$$\begin{aligned}
\nu(z) &= \nu(s)p(s,z) + \sum_{y \neq s} \left(\nu(s)p(s,y) + \sum_{x \neq s} \nu(x)p(x,y) \right) p(y,z) \\
&= \nu(s)p(s,z) + \sum_{y \neq s} \nu(s)p(s,y)p(y,z) + \sum_{y \neq s} \sum_{x \neq s} \nu(x)p(x,y)p(y,z) \\
&= \nu(s) P_s(X_1 = z) + \nu(s) P_s(X_1 \neq s, X_2 = z) + P_\nu(X_0 \neq s, X_1 \neq s, X_2 = z).
\end{aligned}$$

Continuing in this fashion yields

$$\begin{aligned}
\nu(z) &= \nu(s) \sum_{m=1}^{n} P_s(X_1, ..., X_{m-1} \neq s, X_m = z) + P_\nu(X_0, ..., X_{n-1} \neq s, X_n = z) \\
&\geq \nu(s) \sum_{m=1}^{n} P_s(X_1, ..., X_{m-1} \neq s, X_m = z) = \nu(s) \sum_{m=0}^{n} P_s(X_m = z, T_s > m).
\end{aligned}$$

Letting $n \to \infty$ shows that $\nu(z) \geq \nu(s)\mu_s(z)$ for all $z \neq s$.

Since $\mu_s(s) = 1$, we have $\nu(s) \geq \nu(s)\mu_s(s)$ as well, hence $\nu(x) \geq \nu(s)\mu_s(x)$ for all $x \in S$.

Now ν and μ_s are stationary and $\mu_s(s) = 1$, so

$$\sum_x \nu(x) p^n(x,s) = \nu(s) = \nu(s)\mu_s(s) = \nu(s) \sum_x \mu_s(x) p^n(x,s),$$

and thus

$$\sum_x \left(\nu(x) - \nu(s)\mu_s(x) \right) p^n(x,s) = 0$$

for all $n \in \mathbb{N}$.

As $\nu(x) \geq \nu(s)\mu_s(x)$, this implies that $\nu(x) = \nu(s)\mu_s(x)$ for all x with $p^n(x,s) > 0$. Because p is irreducible and n is arbitrary, we conclude that $\nu(x) = \nu(s)\mu_s(x)$ for all $x \in S$. □

The foregoing guarantees existence and uniqueness (up to scaling) of stationary measures under certain relatively mild assumptions.

However, it may be the case that some (and thus every) stationary measure is infinite, so that no stationary distribution exists.

Theorem 9.4. *If there is a stationary distribution π, then all states $y \in S$ with $\pi(y) > 0$ are recurrent.*

Proof. Suppose that $\pi(y) > 0$ and recall that $N(y) = \sum_{k=1}^{\infty} 1\{X_k = y\}$ satisfies $E_x[N(y)] = \rho_{xy} \sum_{k=0}^{\infty} \rho_{yy}^k$. Thus if we start the chain in stationarity, we have

$$E_\pi[N(y)] = \sum_x \pi(x) E_x[N(y)] = \sum_x \pi(x) \rho_{xy} \sum_{k=0}^{\infty} \rho_{yy}^k \leq \sum_x \pi(x) \sum_{k=0}^{\infty} \rho_{yy}^k = \sum_{k=0}^{\infty} \rho_{yy}^k.$$

On the other hand, since $\pi p^n = \pi$, we see that

$$E_\pi[N(y)] = \sum_{n=1}^{\infty} P_\pi(X_n = y) = \sum_{n=1}^{\infty} \sum_x \pi(x) p^n(x, y) = \sum_{n=1}^{\infty} \pi(y) = \infty.$$

Combining these equations shows that

$$\sum_{k=0}^{\infty} \rho_{yy}^k \geq E_x[N(y)] = \infty,$$

hence $\rho_{yy} = 1$. □

Now Lemma 9.1 says that if p is irreducible, then any stationary measure μ has $\mu(x) > 0$ for all x, thus Theorem 9.4 shows that if an irreducible Markov chain has a transient state, then it cannot have a stationary distribution.

We will see shortly that recurrence is not quite sufficient for an irreducible Markov chain to have a stationary distribution, but first we show that if one exists (which is necessarily unique by Theorem 9.3), then it must be given by the following theorem.

Theorem 9.5. *If p is irreducible and has a stationary distribution π, then $\pi(x) = \dfrac{1}{E_x[T_x]}$.*

Proof. Irreducibility implies that $\pi(x) > 0$ for all $x \in S$ by Lemma 9.1, so Theorem 9.4 shows that all states are recurrent.

It follows from Theorem 9.2 that

$$\mu_x(y) = \sum_{n=0}^{\infty} P_x(X_n = y, T_x > n)$$

defines a stationary measure with $\mu_x(x) = 1$.

By Tonelli's theorem and the layer cake representation, we have

$$\begin{aligned} \sum_y \mu_x(y) &= \sum_y \sum_{n=0}^{\infty} P_x(X_n = y, T_x > n) \\ &= \sum_{n=0}^{\infty} \sum_y P_x(X_n = y, T_x > n) = \sum_{n=0}^{\infty} P_x(T_x > n) = E_x[T_x]. \end{aligned}$$

By Theorem 9.3, the stationary measure is unique up to positive scaling, so the unique stationary distribution is given by

$$\pi(x) = \frac{\mu_x(x)}{E_x[T_x]} = \frac{1}{E_x[T_x]}.$$ □

Observe that Theorem 9.5 gives the interesting identity

$$\sum_x \frac{p(x,y)}{E_x[T_x]} = \sum_x \pi(x)p(x,y) = \pi(y) = \frac{1}{E_y[T_y]}.$$

A notable example of an irreducible Markov chain which does not have a stationary distribution is simple random walk on $\mathbb{Z}$: We have seen that the expected first return time is infinite, so Theorem 9.5 shows that there cannot be a stationary distribution.

Definition. If a state x has $E_x[T_x] < \infty$, we say that x is *positive recurrent*. If x is a recurrent state with $E_x[T_x] = \infty$, then x is called *null recurrent*.

Theorem 9.6. *If p is irreducible, then the following are equivalent.*

- **(i):** *Some x is positive recurrent*
- **(ii):** *There is a (unique) stationary distribution*
- **(iii):** *All states are positive recurrent.*

Proof. If x is positive recurrent, then the preceding proof shows that

$$\pi(y) = \frac{\mu_x(y)}{E_x[T_x]} = \frac{1}{E_x[T_x]} \sum_{n=0}^{\infty} P_x(X_n = y, T_x > n)$$

defines a stationary distribution (which is unique by Theorem 9.3), thus (i) implies (ii).
If π is a stationary distribution, then Theorem 9.5 shows that $\pi(y) = \frac{1}{E_y[T_y]}$ for all y. Since Lemma 9.1 implies that $\pi(y) > 0$ for all y, we must have $E_y[T_y] < \infty$, hence (ii) implies (iii).
As (iii) trivially implies (ii), the proof is complete. □

We observe that Theorem 9.2 and the proof of Theorem 9.5 show that if any state x is positive recurrent, then $\pi_x(y) = \frac{\mu_x(y)}{E_x[T_x]}$ defines a stationary distribution, regardless of irreducibility. Also, since x is recurrent, the communicating class $[x]$ is closed and irreducible. Applying Theorem 9.6 to the chain restricted to $[x]$ shows that positive recurrence is a class property.

Thus we have an extra layer in our classification of states: Each communicating class is either recurrent or transient, and each recurrent class is either positive recurrent or null recurrent. Moreover, for each positive recurrent class, there is a unique stationary distribution which is positive on all states in the class and 0 for all other states.

If a state y is transient, then for all states x, we have

$$\sum_{n=1}^{\infty} p^n(x,y) = \sum_{n=1}^{\infty} P_x(X_n = y) = E_x[N(y)] = \frac{\rho_{xy}}{1-\rho_{yy}} < \infty,$$

hence $p^n(x,y) \to 0$ as $n \to \infty$.
As termwise convergence to 0 is not sufficient for summability, the converse is not necessarily true.

Let $N_n(y) = \sum_{m=1}^{n} 1\{X_m = y\}$ be the number of visits to y by time n.

Theorem 10.1. *Suppose that y is recurrent. Then for any $x \in S$,*

$$\lim_{n\to\infty} \frac{1}{n} N_n(y) = \frac{1}{E_y[T_y]} 1\{T_y < \infty\} \; P_x\text{-a.s.}$$

Proof. We begin by considering the chain started at y. Let $R(k) = \inf\{n \geq 1 : N_n(y) = k\}$ be the time of the k^{th} return to y.

Set $t_1 = R(1) = T_y$ and $t_k = R(k) - R(k-1)$ for $k \geq 2$. Since we have assumed that $X_0 = y$, the strong Markov property implies that $t_1, t_2, ...$ are i.i.d. Thus it follows from the strong law of large numbers that

$$\frac{R(n)}{n} = \frac{1}{n} \sum_{k=1}^{n} t_k \to E_y[T_y] \; P_y\text{-a.s.}$$

Observing that $R(N_n(y))$ is the time of the last visit to y by time n and $R(N_n(y)+1)$ is the time of the first visit to y after time n, we see that $R(N_n(y)) \leq n < R(N_n(y)+1)$, thus

$$\frac{R(N_n(y))}{N_n(y)} \leq \frac{n}{N_n(y)} < \frac{R(N_n(y)+1)}{N_n(y)+1} \cdot \frac{N_n(y)+1}{N_n(y)}.$$

Letting $n \to \infty$ and noting that $N_n(y) \to \infty$ a.s. (because y is recurrent) yields

$$\frac{n}{N_n(y)} \to E_y[T_y] \; P_y\text{-a.s.}$$

When the initial state is $x \neq y$, we first note that if $T_y = \infty$, then $N_n(y) = 0$ for all n, hence

$$\frac{N_n(y)}{n} \to 0 \text{ on } \{T_y = \infty\}.$$

The strong Markov property shows that, conditional on $\{T_y < \infty\}$, $t_2, t_3, ...$ are i.i.d. with $P_x(t_k = n) = P_y(T_y = n)$. It follows that

$$\frac{R(k)}{k} = \frac{t_1}{k} + \frac{t_2 + ... + t_k}{k-1} \cdot \frac{k-1}{k} \to 0 + E_y[T_y] \; P_x\text{-a.s.}$$

Thus, arguing as before, we see that for all $x \in S$,

$$\frac{N_n(y)}{n} \to \frac{1}{E_y[T_y]} \; P_x\text{-a.s.}$$

on $\{T_y < \infty\}$. Adding our observation about the $\{T_y = \infty\}$ case completes the proof □

Theorem 10.1 helps to explain the terminology for recurrent states: y is positive recurrent if, when we start the chain at y, the asymptotic fraction of time spent at y is positive. y is null-recurrent if this fraction is 0.

To connect this result with our opening remarks about the $n \to \infty$ behavior of $p^n(x,y)$, we note that $\frac{N_n(y)}{n} \leq 1$, so the bounded convergence theorem gives

$$\frac{1}{n}\sum_{m=1}^{n} p^m(x,y) = \frac{E_x[N_n(y)]}{n} \to E_x\left[\frac{1}{E_y[T_y]}1\{T_y < \infty\}\right] = \frac{P_x(T_y < \infty)}{E_y[T_y]} = \frac{\rho_{xy}}{E_y[T_y]}.$$

(This holds for y transient as well since $E_y[T_y] = \infty$ in that case.)

In particular, if y is positive recurrent and accessible from x, then $\frac{\rho_{xy}}{E_y[T_y]} > 0$, so it cannot be the case that $p^n(x,y) \to 0$. In other words, if $\rho_{xy} > 0$ and $p^n(x,y) \to 0$, then y is transient or null recurrent.

More precisely, we have

Corollary 10.1. *A recurrent class $[w]$ is null recurrent if and only if $\frac{1}{n}\sum_{m=1}^{n} p^m(x,y) \to 0$ for any/all $x, y \in [w]$.*

We are now in a position to upgrade Corollary 8.1 to

Theorem 10.2. *Every irreducible Markov chain on a finite state space is positive recurrent and thus has a unique stationary distribution.*

Proof. We know that irreducible finite state space chains are recurrent. Suppose that such a chain was null recurrent, then for any $x, y \in S$ $\frac{1}{n}\sum_{m=1}^{n} p^m(x,y) \to 0$. But since S is finite, this implies that

$$1 = \frac{1}{n}\sum_{m=1}^{n}\sum_{y} p^n(x,y) = \sum_{y}\frac{1}{n}\sum_{m=1}^{n} p^n(x,y) \to 0,$$

a contradiction. □

The preceding analysis shows that the Cesàro mean $\frac{1}{n}\sum_{m=1}^{n} p^m(x,y)$ always converges.

If p is irreducible and positive recurrent, the limit is $\frac{\rho_{xy}}{E_y[T_y]} = \frac{1}{E_y[T_y]} = \pi(y)$.

However, the following simple examples show that the sequence $p^n(x,y)$ may not converge in the ordinary sense:

Example 10.1. Consider the chain on $\mathbb{Z}/m\mathbb{Z}$ with transition probabilities $p(x, x+1) = 1$, where addition is taken modulo m. This chain is irreducible and positive recurrent (with uniform stationary distribution), but $p^k(x,y) = 1$ if $k \equiv y - x \pmod{m}$ and $p^k(x,y) = 0$ otherwise, hence $p^k(x,y)$ is divergent for all x, y.

Example 10.2. Consider the random transposition shuffle which proceeds by choosing two distinct cards at random and interchanging them. This is the random walk on S_n driven by the uniform measure on transpositions. By construction, $p^k(\sigma,\eta) = 0$ if k is even and $\mathrm{sgn}(\sigma)\mathrm{sgn}(\eta) = -1$ or k is odd and $\mathrm{sgn}(\sigma)\mathrm{sgn}(\eta) = 1$. However, one can show that for any $k \geq n$, $p^k(x,y) \geq \frac{1}{\binom{n}{2}^n}$ if the parity of $\sigma\eta$ and k agree.

In both cases, the periodicity problem can be sidestepped by adding "laziness" - that is, expanding the support of the measure driving the walk to include the identity - so that the chain has some positive probability of staying put at each step.

Inspired by the obstructions to convergence in the preceding examples, we are led to make the following definition,

Definition. For any $x \in S$, set $I_x = \{n \geq 1 : p^n(x,x) > 0\}$. $d_x = \gcd(I_x)$ is called the *period* of x where $\gcd(\emptyset) = \infty$.

Note that $I_x \subseteq \mathbb{N}$ is nonempty precisely when $\rho_{xx} > 0$. In this case, $d_x \leq \min(I_x)$. In particular, $d_x = 1$ if $p(x,x) > 0$.

Lemma 10.1. *If $x \sim y$, then $d_y = d_x$.*

Proof. We may assume that $x \neq y$ so that $\rho_{xy}, \rho_{yx} > 0$. Let K and L be such that $p^K(x,y), p^L(y,x) > 0$. Then

$$p^{K+L}(y,y) \geq p^L(y,x)p^K(x,y) > 0,$$

so $d_y \,|(K+L)$.

If n is such that $p^n(x,x) > 0$, then

$$p^{K+n+L}(y,y) \geq p^L(y,x)p^n(x,x)p^K(x,y) > 0,$$

so we also have that $d_y \,|(K+L+n)$, and thus $d_y \,|n$. As $n \in I_x$ is arbitrary, it follows that $d_y \,|d_x$.

Interchanging the roles of x and y shows that $d_x \,|d_y$ as well, and we conclude that $d_y = d_x$. □

Definition. We say that a state x is *aperiodic* if $d_x = 1$. If all states in a recurrent Markov chain are aperiodic, we say that the chain is aperiodic.

Lemma 10.2. *If $d_x = 1$, then there is an $m_x \in \mathbb{N}$ such that $p^m(x,x) > 0$ for all $m > m_x$.*

Proof. We first observe that there is a finite $F_x \subset I_x$ such that $\gcd(F_x) = \gcd(I_x) = 1$. To see that this is so, note that $d(n) = \gcd(I_x \cap [1,n])$ is a nonincreasing $\mathbb{N}$-valued function and thus can only decrease a finite number of times. Let $N = \max\{n \in \mathbb{N} : d(n) < d(n-1)\}$ and set $F_x = I_x \cap [1,N] = \{b_1, ..., b_n\}$.

Now the Euclidean Algorithm shows that there are integers $a_1, ..., a_n$ with $\sum_{i=1}^n a_i b_i = 1$.

Set $a = \max_i |a_i|$, $b = \sum_i b_i$. For any $m \in \mathbb{N}$, there exist $q, r \in \mathbb{N}_0$ with $r < b$ such that

$$m = qb + r = q\sum_i b_i + r\sum_i a_i b_i = \sum_i (q + ra_i)b_i.$$

If $q \geq ab$, then $q + ra_i \geq 0$ for all i. In other words, every integer greater than ab^2 can be written as a sum of elements in $F_x \subset I_x$. Since I_x is closed under addition, this means that I_x contains every integer greater than $m_x = ab^2$. □

If K is the transition matrix for an irreducible and aperiodic Markov chain with finite state space S, then it follows from Lemma 10.2 that there is an $N \in \mathbb{N}$ such that $K^n(x,y) > 0$ for all $x, y \in S$ whenever $n \geq N$:

For each x, y, there is an $n(x,y) \in \mathbb{N}$ such that $p^{n(x,y)}(x,y) > 0$ by irreducibility. Since S is finite, $M = \max_{(x,y)} n(x,y)$ exists in $\mathbb{N}$.

Letting $L = \max_x m_x$ with m_x as in Lemma 10.2, we see that for any $n \geq L + M$ and any $x, y \in S$, $p^n(x,y) \geq p^{n(x,y)}(x,y)p^{n-n(x,y)}(y,y) > 0$ since $n(x,y) \leq M$, hence $n - n(x,y) \geq L \geq m_y$.

Now since K is stochastic, it has spectral radius 1, and since K^N is a positive matrix, the Perron-Frobenius theorem ensures that 1 is a simple eigenvalue.

Perron-Frobenius also shows that K^N has a left eigenvector π with $\pi K^N = \pi$, $\pi(x) > 0$ for all x, and $\sum_x \pi(x) = 1$. It follows that K has unique and strictly positive stationary distribution π.

Moreover, letting $w \equiv 1$ denote the appropriately normalized right eigenvector with eigenvalue 1, Perron-Frobenius implies $\lim_{m\to\infty} \left(K^N\right)^m = w\pi$, the matrix with all rows equal to π.

Finally, writing K in Jordan normal form and recalling that the eigenvalues satisfy $\lambda_0 = 1 > |\lambda_1| \geq |\lambda_2| \geq ...$, we see that K^n converges as well (with exponential rate given by $|\lambda_1|$), hence $K^n(x,y) \to \pi(y)$ for all $x, y \in S$.

When the state space is countably infinite, we no longer have these linear algebra tools and the above argument does not apply. However, if we tack on the assumption of positive recurrence (so that a stationary distribution exists), we can still arrive at the same conclusion.

Theorem 10.3. *Suppose that p is irreducible, aperiodic, and positive recurrent with countable state space S. Then there is a probability measure π on S which satisfies*

$$\lim_{n\to\infty} p^n(x,y) = \pi(y)$$

for all states x, y.

Proof. Define a transition probability $\widetilde{p}$ on $S \times S$ by

$$\widetilde{p}\left((x_1,y_1),(x_2,y_2)\right) = p(x_1,x_2)p(y_1,y_2).$$

For any $x, y \in S$, aperiodicity gives $p^m(x,x) > 0$ whenever $m > m_x$, and irreducibility shows that there exists $n(x,y) \in \mathbb{N}$ with $p^{n(x,y)}(x,y) > 0$.

It follows that for any $(x_1,y_1),(x_2,y_2) \in S \times S$,

$$\begin{aligned}\widetilde{p}^{n(x_1,x_2)+n(y_1,y_2)+m}\left((x_1,y_1),(x_2,y_2)\right) &= p^{n(x_1,x_2)+n(y_1,y_2)+m}(x_1,x_2)p^{n(x_1,x_2)+n(y_1,y_2)+m}(y_1,y_2) \\ &\geq p^{n(x_1,x_2)}(x_1,x_2)p^{n(y_1,y_2)+m}(x_2,x_2)p^{n(y_1,y_2)}(y_1,y_2)p^{n(x_1,x_2)+m}(y_2,y_2) > 0\end{aligned}$$

whenever $m \geq m_{x_2} \vee m_{y_2}$, hence $\widetilde{p}$ is irreducible.

Also, since p is irreducible and positive recurrent, it has a unique stationary distribution, π.

Define $\widetilde{\pi}((x,y)) = \pi(x)\pi(y)$. Then $\widetilde{\pi}$ is a stationary distribution for $\widetilde{p}$ since

$$\sum_{(w,y)\in S\times S} \widetilde{\pi}((w,y))\,\widetilde{p}((w,y),(x,z)) = \sum_{w\in S}\pi(w)p(w,x)\sum_{y\in s}\pi(y)p(y,z) = \pi(x)\pi(z) = \widetilde{\pi}((x,z)).$$

Since $\widetilde{p}$ is irreducible and has a stationary distribution, it follows from Theorem 9.6 that all states in $S\times S$ are positive recurrent.

Now let (X_n, Y_n) be a Markov chain on $S\times S$ with transition probability $\widetilde{p}$ and let $T = \inf\{n : X_n = Y_n\}$ be the hitting time of the diagonal $D = \{(y,y) : y\in S\}$. Since $\widetilde{p}$ is irreducible and recurrent, the hitting time of any point in D is a.s. finite (by Theorem 8.3), thus $T<\infty$ a.s.

By considering the time and location of the first intersection and appealing to the Markov property, we see that

$$\begin{aligned}
P(X_n = y, T\le n) &= \sum_{m=1}^{n}\sum_{x\in S} P(T=m, X_m = x, X_n = y)\\
&= \sum_{m=1}^{n}\sum_{x\in S} P(T=m, X_m = x)P(X_n = y\,|X_m = x\,)\\
&= \sum_{m=1}^{n}\sum_{x\in S} P(T=m, Y_m = x)P(Y_n = y\,|Y_m = x\,)\\
&= P(Y_n = y, T\le n).
\end{aligned}$$

It follows that

$$\begin{aligned}
P(X_n = y) &= P(X_n = y, T\le n) + P(X_n = y, T > n)\\
&= P(Y_n = y, T\le n) + P(X_n = y, T > n)
\end{aligned}$$

Since

$$P(Y_n = y) = P(Y_n = y, T\le n) + P(Y_n = y, T > n),$$

we have

$$\begin{aligned}
|P(X_n = y) - P(Y_n = y)| &= |P(X_n = y, T > n) - P(Y_n = y, T > n)|\\
&\le P(X_n = y, T > n) + P(Y_n = y, T > n),
\end{aligned}$$

and thus

$$\sum_y |P(X_n = y) - P(Y_n = y)| \le 2P(T > n).$$

If we take $X_0 = x$, $Y_0 \sim \pi$, this says that

$$\sum_y |p^n(x,y) - \pi(y)| \le 2P(T>n)\to 0$$

and the proof is complete. □

11. Coupling

Let p denote the transition function for a time homogeneous Markov chain $X_0, X_1, ...$ on a countable state space S, so that $p(x,y) = P(X_{t+1} = y \,|X_t = x)$ for all $t \in \mathbb{N}_0$, $x, y \in S$.

The k-step transitions are given recursively by

$$p^k(x,z) = P(X_{t+k} = z \,|X_t = x) = \sum_{y \in S} p^{k-1}(x,y)p(y,z).$$

In the $|S| = N < \infty$ case, we think of p as an $N \times N$ matrix and the above is just the formula for matrix exponentiation.

In general, p is an operator which acts on functions ("column vectors") by

$$(pf)(x) = \sum_y p(x,y)f(y) = E[f(X_{t+1}) \,|X_t = x]$$

and acts on probabilities ("row vectors") by

$$(\mu p)(y) = \sum_x \mu(x)p(x,y) = P(X_{t+1} = y \,|X_t \sim \mu).$$

Thus if the chain has initial distribution $\mu_0 = \mathscr{L}(X_0)$, then the distribution of X_t is $\mu_0 p^t$.

We know that if p is irreducible, aperiodic, and positive recurrent, then it has a unique stationary distribution π and $\lim_{n\to\infty} p^n(x,y) = \pi(y)$ for all $x, y \in S$.

(If S is finite, the assumption of positive recurrence is redundant.)

In applications such as MCMC, we want to know how fast the chain converges - that is, how big does t need to be for $\mu_0 p^t$ to be close to π.

To make this question more precise, we need a metric on probabilities.

Definition. The *total variation distance* between probabilities μ and ν on a countable set S is defined as

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{s \in S} |\mu(s) - \nu(s)|.$$

It is routine to verify that total variation defines a metric, and that we have the equivalent characterizations

$$\begin{aligned} \|\mu - \nu\|_{TV} &= \max_{A \subseteq S} (\mu(A) - \nu(A)) \\ &= \frac{1}{2} \max_{\|f\|_\infty \leq 1} (E_\mu[f] - E_\nu[f]). \end{aligned}$$

(The maxima are attained by $A = \{s : \mu(s) > \nu(s)\}$ and $f = 1_A - 1_{A^C}$.)

Also, it is clear from the original definition that $\|\mu_n - \nu\|_{TV} \to 0$ implies $\mu_n \to \nu$ pointwise. The converse implication does not necessarily hold when the state space is infinite.

Note that the proof of Theorem 10.3 actually established that $p_x^n \to \pi$ in total variation.

When speaking of Markov chain mixing, we often want a measure of distance which is independent of the initial distribution, so we define

$$d(t) = \sup_{\mu_0} \left\| \mu_0 p^t - \pi \right\|_{TV}.$$

In the case where the initial distribution is a point mass, we write $p_x^t = \delta_x p^t$.

It is left as an exercise to show that an equivalent definition is

$$d(t) = \sup_{x \in S} \left\| p_x^t - \pi \right\|_{TV}.$$

Also, it is often useful to consider the distance between two copies of the chain started at different states, and we define

$$\overline{d}(t) = \sup_{x,y \in S} \left\| p_x^t - p_y^t \right\|_{TV}.$$

Another good exercise is to establish the inequality

$$d(t) \leq \overline{d}(t) \leq 2d(t).$$

One of the main goals in the modern theory of Markov chains is to estimate the ε-mixing time

$$t_{mix}(\varepsilon) = \min \left\{ t \in \mathbb{N}_0 : d(t) \leq \varepsilon \right\}.$$

Remarkably, in many cases of interest the mixing time is asymptotically independent of ε.

Here we are thinking of a sequence of chains $p_{(1)}, p_{(2)}, ...$ where $p_{(n)}$ has state space $S_{(n)}$, stationary distribution $\pi_{(n)}$, and ε-mixing time $t_{mix}^{(n)}(\varepsilon)$.

For example, $p_{(n)}$ may represent a particular method of shuffling n cards, simple random walk on the n-dimensional hypercube, Glauber dynamics for the Ising model on the torus $(\mathbb{Z}/n\mathbb{Z})^d$, and so forth.

We say that the sequence $p_{(n)}$ exhibits the *cutoff phenomenon* if

$$\lim_{n \to \infty} \frac{t_{mix}^{(n)}(\varepsilon)}{t_{mix}^{(n)}(1-\varepsilon)} = 1$$

for all $\varepsilon \in (0,1)$.

This means that $t_{mix}(\varepsilon_1)$ and $t_{mix}(\varepsilon_2)$ differ only in lower order terms.

Writing

$$d_{(n)}(t) = \sup_{x \in S_{(n)}} \left\| \delta_x p_{(n)}^t - \pi_{(n)} \right\|_{TV},$$

this is equivalent to the requirement that

$$\lim_{n \to \infty} d_{(n)} \left(c t_{mix}^{(n)} \right) = \begin{cases} 1, & c < 1 \\ 0, & c > 1 \end{cases}$$

where $t_{mix}^{(n)} = t_{mix}^{(n)} \left(\frac{1}{4} \right)$. (The choice of $\frac{1}{4}$ is arbitrary but standard.)

That is, when time is scaled by $t_{mix}^{(n)}$, $d_{(n)}(t)$ approaches a step function. Essentially, the distance to equilibrium stays near 1 for a while and then abruptly drops and tends rapidly to 0.

A classical method of bounding mixing times is based on the notion of coupling. (In fact, this was the technique used to prove Theorem 10.3.)

Definition. If μ and ν are probabilities on $(S, \mathcal{S})$, we say that (X, Y) is a *coupling* of μ and ν if X and Y are S-valued random variables on a common probability space $(\Omega, \mathcal{F}, P)$ such that $P(X \in A) = \mu(A)$ and $P(Y \in A) = \nu(A)$ for all $A \in \mathcal{S}$.

Lemma 11.1. *If (X, Y) is a coupling of μ and ν, then $\|\mu - \nu\|_{TV} \leq P(X \neq Y)$.*

Proof. For any $A \in \mathcal{S}$,

$$\begin{aligned}
\mu(A) - \nu(A) &= P(X \in A) - P(Y \in A) \\
&= P(X \in A, X \neq Y) + P(X \in A, X = Y) \\
&\quad - P(Y \in A, X \neq Y) - P(Y \in A, X = Y) \\
&= P(X \in A, X \neq Y) - P(Y \in A, X \neq Y) \\
&\leq P(X \in A, X \neq Y) \leq P(X \neq Y).
\end{aligned}$$

□

When S is countable, one can show that there always exist couplings for which the inequality is an equality, hence we have the additional definition of total variation:

$$\|\mu - \nu\|_{TV} = \min\{P(X \neq Y) : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

* Briefly, one defines w on $S \times S$ by

$$\begin{aligned}
w(z, z) &= \min\{\mu(z), \nu(z)\}, \\
w(x, y) &= \frac{(\mu(x) - w(x, x))\,(\nu(y) - w(y, y))}{1 - \sum_z w(z, z)}
\end{aligned}$$

and checks that $(X, Y) \sim w$ is a coupling with $P(X \neq Y) = \|\mu - \nu\|_{TV}$. We leave the verification of this claim as an exercise.

Definition. A *coupling of a transition probability p* on a countable state space S is an $S \times S$-valued process $\{(X_t, Y_t)\}_{t \in \mathbb{N}_0}$ defined on some probability space $(\Omega, \mathcal{F}, P)$ such that, marginally, $\{X_t\}$ and $\{Y_t\}$ are each Markov chains with transition probability p.

Note that we do not require $\{X_t\}$ and $\{Y_t\}$ to proceed independently or that $\{(X_t, Y_t)\}$ is a Markov chain.

Theorem 11.1. *Suppose that $\{(X_t, Y_t)\}_{t \in \mathbb{N}_0}$ is a coupling of p with $X_0 \sim \mu$, $Y_0 \sim \nu$. If T is a random time such that $X_t = Y_t$ on $\{T \leq t\}$, then*

$$\left\|\mu p^t - \nu p^t\right\|_{TV} \leq P\left(T > t\right).$$

Proof. By construction, (X_t, Y_t) is a coupling of μp^t and νp^t. Since $\{X_t \neq Y_t\} \subseteq \{T > t\}$, the coupling lemma implies

$$\left\|\mu p^t - \nu p^t\right\|_{TV} \leq P\left(X_t \neq Y_t\right) \leq P\left(T > t\right).$$

□

When $\nu = \pi$, Theorem 11.1 gives the bound $\|\mu p^t - \pi\|_{TV} \leq P(T > t)$.

It can also be convenient to take $\mu = \delta_x$ and $\nu = \delta_y$ to get $\left\|p_x^t - p_y^t\right\|_{TV} \leq P(T_{x,y} > t)$.

One can then bound the distance to stationarity using

$$d(t) \leq \overline{d}(t) \leq \sup_{x,y} P(T_{x,y} > t).$$

Typically, one considers couplings $\{(X_t, Y_t)\}_{t \in \mathbb{N}_0}$ with the property that $X_t = Y_t$ implies $X_{t+1} = Y_{t+1}$.

In this case, we can take $T = \inf\{t \geq 0 : X_t = Y_t\}$.

Definition. A *faithful coupling* of a Markov chain with transition matrix p and state space S is a Markov chain on $S \times S$ whose transition probability, q, satisfies

(1) $\sum_{y' \in S} q\left((x,y),(x',y')\right) = p(x,x')$ for all $x, y, x' \in S$.
(2) $\sum_{x' \in S} q\left((x,y),(x',y')\right) = p(y,y')$ for all $x, y, y' \in S$.

Faithful couplings can be modified so that the two trajectories stay together after colliding.

To wit, suppose that $\{(X_t, Y_t)\}_{t=0}^{\infty}$ is a faithful coupling of p and let $T = \inf\{t \geq 0 : X_t = Y_t\}$.

Define $Z_t = \begin{cases} Y_t, & t \leq T \\ X_t, & t > T \end{cases}$. Then $Z_0 \sim \mathscr{L}(Y_0)$, $X_{t+1} = Z_{t+1}$ whenever $X_t = Z_t$, and the strong Markov property implies that (X_t, Z_t) is a coupling of p. (This trick may not work for unfaithful couplings.)

The general idea is that we start one copy of the chain in a specified distribution, let another copy begin in stationarity, and then let them evolve according to the same transition mechanism until they meet and proceed simultaneously forever after. As the second chain was stationary to begin with, it remains so for all time, thus the first chain must have equilibrated by the time they couple.

Though the preceding argument captures the intuition, it is not strictly correct as it overlooks a subtle point: $Y_t \sim \pi$ for all t does not guarantee that $Y_T \sim \pi$ for a stopping time T.

For example, consider the chain with state space $\{x, y\}$ and transition probabilities $p(x,y) = 1$, $p(y,x) = p(y,y) = \frac{1}{2}$. It is easy to see that $\pi(x) = \frac{1}{3}$, $\pi(y) = \frac{2}{3}$ is stationary for p. If we let $\{X_t\}$ be a copy of the chain started at y, let $\{Y_t\}$ be another copy of the chain with initial distribution π, and let T be the coupling time of $\{X_t\}$ and $\{Y_t\}$, then we necessarily have that $Y_T = y$ since $W_t = x$ implies that $W_{t-1} = y$ for any chain $\{W_t\}$ having transition probability p.

Nonetheless, the coupling bound on variation distance still holds and can be quite useful.

Example 11.1 (Lazy Random Walk on the Hypercube)**.** Here $S = (\mathbb{Z}/2\mathbb{Z})^d$, $p(x,x) = \frac{1}{2}$, $p(x,y) = \frac{1}{2d}$ if x and y differ in exactly one coordinate, and $p(x,z) = 0$ otherwise.

That is, at each step we flip a fair coin. If it comes up heads, we stay put. If it comes up tails, we choose one of our d neighbors uniformly at random and move there.

As an irreducible random walk on a finite group (or a simple random walk on a finite regular graph), the stationary distribution is uniform. The $\frac{1}{2}$ holding probabilities ensure aperiodicity, so the convergence theorem implies that distribution of the position at time t will approach the uniform distribution as $t \to \infty$. We want to know how fast it converges.

To this end, we let $X_0 = x$ and let Y_0 have the uniform distribution on S. Let $U_1, U_2, ...$ be i.i.d. uniform on $\{1, ..., d\}$ and let $V_1, V_2, ...$ be i.i.d. uniform on $\{0, 1\}$. All random variables are taken to be independent. At time t, the U_t^{th} coordinate of each chain is set to V_t and the others remain as they were.

In other words, at every time step we pick a coordinate at random and set its value (in both chains) to 0 or 1 according to a toss of a fair coin.

It is easy to see that $\{X_t\}$ and $\{Y_t\}$ are each evolving according to p. Moreover, the two chains agree in coordinate U_t from time t onward, so they have coupled by time $T = \inf\{t : \{U_1, ..., U_t\} = \{1, ..., d\}\}$. Since T does not depend on the initial state, Theorem 11.1 shows that

$$d(t) = \sup_x \left\| p_x^t - \pi \right\|_{TV} \leq P(T > t).$$

Thus the variation bound reduces to a coupon collector problem:

If we let $A_k^t = \{k \notin \{U_1, ..., U_t\}\}$ for $k = 1, ..., d$, then

$$P(T > t) = P\left(\bigcup_{k=1}^{d} A_k^t\right) \leq dP\left(A_1^t\right) = d\left(1 - \frac{1}{d}\right)^t \leq de^{-\frac{t}{d}}.$$

Therefore, if $t = d\log(d) + cd$ where $c > 0$ is chosen so that $t \in \mathbb{N}$, then $d(t) \leq e^{-c}$, hence the mixing time is $O\left(d\log(d)\right)$.

* Using a more sophisticated coupling or other techniques such as Fourier analysis, along with lower bound arguments like Wilson's method, one can show that the correct rate is $\frac{1}{2}d\log(d)$.

We conclude our discussion with a look at *grand couplings*.

The idea here is to build copies of the chain started at each possible initial state using a common source of randomness.

In other words, we wish to construct a collection of random variables $\{X_t^x : x \in S, t \in \mathbb{N}_0\}$ on a common probability space $(\Omega, \mathcal{F}, P)$ such that for each $x \in S$, $\{X_t^x\}_{t=0}^{\infty}$ is a Markov chain with transition probability p and initial state $X_0^x = x$. To apply the coupling bound, we also need the various trajectories to stay together after their first meeting.

One way to achieve such a setup is through a *random mapping representation* of p, which is a pair (f, Z) such that Z is a random variable on $(\Omega, \mathcal{F}, P)$ and $f : S \times \Omega \to S$ is a (deterministic) function satisfying $P\left(f(x, Z) = y\right) = p(x, y)$ for all $x, y \in S$.

It is left as an exercise to show that every transition probability on a countable state space S admits a random mapping representation.

(Hint: Let $Z \sim U(0, 1)$ and consider the array $F_{i,j} = \sum_{k=1}^{j} p(s_i, s_k)$ where p is the transition function and $\{s_1, s_2, ...\}$ is an enumeration of the state space.)

If (f, Z) is a random mapping representation for p and $Z, Z_0, Z_1, ...$ is an i.i.d. sequence on $(\Omega, \mathcal{F}, P)$, then a grand coupling is given by taking $X_0^x = x$ and $X_{t+1}^x = f(X_t^x, Z_t)$ for each $x \in S$, $t \in \mathbb{N}_0$.

Random mapping representations can be represented more compactly as iterates of random functions by writing $f_t(\cdot) = f(\cdot, Z_t)$.

If we define

$$F_i^j = f_{j-1} \circ f_{j-2} \circ \cdots \circ f_{i+1} \circ f_i \text{ for } i < j,$$

then $X_t = F_0^t(X_0)$ is a *forward simulation* of the Markov chain with transition probability p and initial distribution $\mathscr{L}(X_0)$.

The *coalescence time* is defined by

$$T_c = \inf\{t : F_0^t \text{ is a constant function}\}.$$

It follows from the coupling lemma and the fact that $d(t) \leq \overline{d}(t)$ that we have the variation bound

$$d(t) \leq P(T_c > t).$$

* Note that the unique element in the range of $F_0^{T_c}$ is not necessarily distributed according to π as evidenced by the two-state example given previously. However, if

$$R_c = \inf\{t : F_{-t}^0 \text{ is a constant function}\},$$

then $F_{-R_c}^0(x) \sim \pi$.
This observation lies at the heart of the perfect sampling scheme known as *coupling from the past.*

(Roughly, by time homogeneity and the convergence theorem,

$$\lim_{t\to\infty} P\left(F_{-t}^0(x) = y\right) = \lim_{t\to\infty} P\left(F_0^t(x) = y\right) = \pi(y),$$

so $F_{-\infty}^0, F_0^\infty \sim \pi$.

If $r > R_c$, then

$$F_{-r}^0 = f_{-1} \circ \cdots \circ f_{-R_c} \circ f_{-R_c-1} \circ \cdots \circ f_{-r} = F_{-R_c}^0 \circ f_{-R_c-1} \circ \cdots \circ f_{-r} = F_{-R_c}^0,$$

hence $F_{-R_c}^0 = F_{-\infty}^0 \sim \pi$.

This trick does not work in the forward direction, because if $t > T_c$, then

$$F_0^t = f_{t-1} \circ \cdots \circ f_{T_c} \circ f_{T_c-1} \circ \cdots \circ f_0 = f_{t-1} \circ \cdots \circ f_{T_c} \circ F_0^{T_c}$$

is not equal to $F_0^{T_c}$ in general.)

Example 11.2 (Metropolis chain for proper q-colorings)**.** Let $G = (V, E)$ be a graph. A proper q-coloring of G is an element $x \in \{1, 2, ..., q\}^V$ such that $x(u) \neq x(v)$ whenever $\{u, v\} \in E$ (henceforth $u \sim v$).

We can generate an approximation to the uniform distribution on the set Λ of proper q-colorings of G using the Metropolis algorithm:

From any coloring x, select a vertex v uniformly from V and a color j uniformly from $[q]$. If $j \neq x(u)$ for any $u \sim v$, assign v the color j. Otherwise do nothing.

By applying these dynamics on the larger space $\widetilde{\Lambda} = [q]^V$, we can use a grand coupling to prove

Theorem 11.2. *Let G be a graph with n vertices and maximal degree Δ. For the Metropolis chain on proper q-colorings of G with $q > 3\Delta$, we have*

$$t_{mix}(\varepsilon) \leq \left(1 - \frac{3\Delta}{q}\right)^{-1} n \log\left(\frac{n}{\varepsilon}\right).$$

Proof. Let $(v_0, k_0), (v_1, k_1), ...$ be i.i.d. uniform on $V \times [q]$, and for each $x \in \widetilde{\Lambda}$, define $\{X_t^x\}_{t \geq 0}$ by $X_0^x = x$ and

$$X_{t+1}^x(v) = \begin{cases} k_t, & v = v_t, X_t^x(u) \neq k_t \text{ for all } u \sim v \\ X_t^x(v), & \text{else} \end{cases}.$$

Define the metric ρ on $\widetilde{\Lambda}$ by

$$\rho(x, y) = \sum_{v \in V} 1\{x(v) \neq y(v)\}.$$

Also, for each $x \in \widetilde{\Lambda}$, $v \in V$, denote the set of colors of neighbors of v in x by

$$\mathcal{N}(x, v) = \{j \in [q] : x(u) = j \text{ for some } u \sim v\}.$$

Now suppose that $x, y \in \widetilde{\Lambda}$ are such that $\rho(x, y) = 1$. Then x and y differ at only one vertex, say w. Let's see what happens to the distance after an update.

Since $\mathcal{N}(x, w) = \mathcal{N}(y, w)$ by assumption, we have

$$P\left(\rho\left(X_1^x, X_1^y\right) = 0\right) = P\left(v_0 = w, k_0 \notin \mathcal{N}(x, w)\right) = \frac{1}{n} \cdot \frac{q - |\mathcal{N}(x, w)|}{q} \geq \frac{q - \Delta}{nq}.$$

Similarly,

$$\begin{aligned} P\left(\rho\left(X_1^x, X_1^y\right) = 2\right) &\leq P\left(v_0 \sim w, k_0 = y(w)\right) + P\left(v_0 \sim w, k_0 = x(w)\right) \\ &= \frac{|\{u \in V : u \sim w\}|}{n} \cdot \frac{1}{q} + \frac{|\{u \in V : u \sim w\}|}{n} \cdot \frac{1}{q} \leq \frac{2\Delta}{nq}. \end{aligned}$$

The only other possible value for $\rho(X_1^x, X_1^y)$ is 1, so we have

$$\begin{aligned} E\left[\rho\left(X_1^x, X_1^y\right) - 1\right] &= -1 \cdot P\left(\rho(X_1^x, X_1^y) = 0\right) + 0 \cdot P\left(\rho(X_1^x, X_1^y) = 1\right) + 1 \cdot P\left(\rho(X_1^x, X_1^y) = 2\right) \\ &\leq \frac{2\Delta}{nq} - \frac{q - \Delta}{nq} = \frac{3\Delta - q}{nq}, \end{aligned}$$

or

$$E\left[\rho\left(X_1^x, X_1^y\right)\right] \leq 1 - \frac{q - 3\Delta}{nq}.$$

If $x, z \in \widetilde{\Lambda}$ are such that $\rho(x, z) = r$, then there exist $x_0 = x, x_1, ..., x_{r-1}, x_r = z$ such that $\rho(x_{i-1}, x_i) = 1$ for $i = 1, ..., r$.

It follows from the triangle inequality and the preceding estimate that

$$E\left[\rho\left(X_1^x, X_1^z\right)\right] \leq \sum_{k=1}^{r} E\left[\rho\left(X_1^{x_{k-1}}, X_1^{x_k}\right)\right] \leq r\left(1 - \frac{q - 3\Delta}{nq}\right) = \rho(x, z)\left(1 - \frac{q - 3\Delta}{nq}\right).$$

Since the chain is time homogeneous, we have

$$E\left[\rho\left(X_t^x, X_t^z\right) \middle| X_{t-1}^x = x_{t-1}, X_{t-1}^z = z_{t-1}\right] = E\left[\rho\left(X_1^{x_{t-1}}, X_1^{z_{t-1}}\right)\right] \le \rho(x_{t-1}, z_{t-1})\left(1 - \frac{q-3\Delta}{nq}\right),$$

so taking expectation with respect to $\left(X_{t-1}^x, X_{t-1}^z\right)$ gives

$$E\left[\rho\left(X_t^x, X_t^z\right)\right] \le E\left[\rho(X_{t-1}^x, X_{t-1}^z)\right]\left(1 - \frac{q-3\Delta}{nq}\right),$$

and thus

$$E\left[\rho\left(X_t^x, X_t^z\right)\right] \le \rho(x,z)\left(1 - \frac{q-3\Delta}{nq}\right)^t$$

by induction.

Now Chebychev's inequality shows that

$$\begin{aligned} P\left(X_t^x \ne X_t^z\right) &= P\left(\rho\left(X_t^x, X_t^z\right) \ge 1\right) \le E\left[\rho\left(X_t^x, X_t^z\right)\right] \\ &\le \rho(x,z)\left(1 - \frac{q-3\Delta}{nq}\right)^t \le n\left(1 - \frac{q-3\Delta}{nq}\right)^t \end{aligned}$$

for all $x, z \in \widetilde{\Lambda}$, hence

$$\begin{aligned} d(t) \le \overline{d}(t) &\le \max_{x,z\in\Lambda} P\left(X_t^x \ne X_t^z\right) \le \max_{x,z\in\widetilde{\Lambda}} P\left(X_t^x \ne X_t^z\right) \\ &\le n\left(1 - \frac{q-3\Delta}{nq}\right)^t \le n\exp\left(-\frac{q-3\Delta}{nq}t\right). \end{aligned}$$

Since we assumed that $q > 3\Delta$, taking $t \ge \left(1 - \frac{3\Delta}{q}\right)^{-1} n\log\left(\frac{n}{\varepsilon}\right)$ gives

$$d(t) \le n\exp\left(-\frac{q-3\Delta}{nq}t\right) \le ne^{-\log\left(\frac{n}{\varepsilon}\right)} = \varepsilon. \qquad \square$$

12. Probability Preserving Dynamical Systems

Much of our investigation of probability theory has revolved around the long term behavior of sequences of random variables. We continue this theme with a cursory look at ergodic theory. Roughly speaking, ergodic theorems assert that under certain stability and irreducibility conditions time averages converge to space averages. As usual, we begin with some definitions.

Definition. A sequence $X_0, X_1, ...$ is said to be *stationary* if $(X_0, X_1, ...) =_d (X_k, X_{k+1}, ...)$ for all $k \in \mathbb{N}$.
Equivalently, $X_0, X_1, ...$ is stationary if for every $n, k \in \mathbb{N}_0$, we have $(X_0, ..., X_n) =_d (X_k, ..., X_{n+k})$.

We have already seen several examples of stationary sequences. For instance, i.i.d. sequences are stationary, and more generally, so are exchangeable sequences.
Another example of a stationary sequence is a Markov chain $X_0, X_1, ...$ started in equilibrium.

To treat the general case, we introduce the following construct.

Definition. Given a probability space $(\Omega, \mathcal{F}, P)$, a measurable map $T : \Omega \to \Omega$ is said to be *probability preserving* if $P\left(T^{-1}A\right) = P(A)$ for all $A \in \mathcal{F}$, where $T^{-1}A = \{\omega \in \Omega : T\omega \in A\}$ denotes the preimage of A under T.
We say that the tuple $(\Omega, \mathcal{F}, P, T)$ is a *probability preserving dynamical system.*
Iterates of T and T^{-1} are defined inductively by $T^0\omega = \omega$ and, for $n \geq 1$,

$$T^n = T \circ T^{n-1},$$
$$T^{-n} = T^{-1} \circ T^{-(n-1)} = (T^n)^{-1}.$$

* We use the inverse image in our definitions because $A \in \mathcal{F}$ does not necessarily imply that $TA \in \mathcal{F}$.
Also, beware that some authors say that "P is an invariant measure for T" rather than "T preserves P."
Finally, observe that since the push-forward measure $T_*P = P \circ T^{-1}$ is equal to P, the change of variables formula shows that

$$\int_\Omega f \circ T dP = \int_\Omega f dT_*P = \int_\Omega f dP$$

for all f for which the latter integral is defined.

If X is a random variable on $(\Omega, \mathcal{F}, P)$ and $T : \Omega \to \Omega$ is probability preserving, then $X_n(\omega) = X(T^n\omega)$ defines a stationary sequence since for any $n, k \in \mathbb{N}$ and any Borel set $B \in \mathcal{B}^{n+1}$, if $A = \{\omega : (X_0(\omega), ..., X_n(\omega)) \in B\}$, then

$$P\left((X_k, ..., X_{n+k}) \in B\right) = P\left(T^{-k}A\right) = P\left(A\right) = P\left((X_0, ..., X_n) \in B\right).$$

In fact, every stationary sequence taking values in a nice space can be expressed in this form:

If $Y_0, Y_1, ...$ is a stationary sequence of random variables taking values in a nice space $(S, \mathcal{S})$, then the Kolmogorov extension theorem gives a measure P on $\left(S^{\mathbb{N}_0}, \mathcal{S}^{\mathbb{N}_0}\right)$ such that the coordinate projections $X_n(\omega) = \omega_n$ satisfy $(X_0, X_1, ...) =_d (Y_0, Y_1, ...)$. If we let $X = X_0$ and $T = \theta$ (the shift map), then T is probability preserving and $X_n(\omega) = \omega_n = (\theta^n\omega)_0 = X(T^n\omega)$.

In light of the preceding observations, we will assume henceforth that we are working with stationary sequences of the form $X_n(\omega) = X(T^n\omega)$ for some $(S, \mathcal{S})$-valued random variable X defined on a probability space $(\Omega, \mathcal{F}, P)$ and some probability preserving map $T : \Omega \to \Omega$.

For subsequent results, we will need a few more definitions.

Definition. Let $(\Omega, \mathcal{F}, P, T)$ be a probability preserving dynamical system. We say that an event $A \in \mathcal{F}$ is *invariant* if $T^{-1}A = A$ up to null sets - that is, $P\left(\left(T^{-1}A\right) \triangle A\right) = 0$ where $\triangle$ denotes the symmetric difference, $E \triangle F = (E \setminus F) \cup (F \setminus E)$.
A random variable X is called *invariant* if $X \circ T = X$ a.s.

It is left as an exercise to show

Proposition 12.1. *$\mathcal{I} = \{A \in \mathcal{F} : A$ is invariant$\}$ is a sub-σ-field of $\mathcal{F}$, and $X \in \mathcal{I}$ if and only if X is invariant.*

Definition. We say that T is *ergodic* if for every invariant event $A \in \mathcal{I}$, we have $P(A) \in \{0, 1\}$.

Ergodicity is a kind of irreducibility requirement: T is ergodic if Ω cannot be decomposed as $\Omega = A \sqcup B$ with $A, B \in \mathcal{I}$ and $P(A), P(B) > 0$.
Ergodic maps are good at mixing things up in the sense that they don't fix nontrivial subsets.

A useful test for ergodicity is given by

Proposition 12.2. *T is ergodic if and only if every invariant $X : \Omega \to \mathbb{R}$ is a.s. constant.*

Proof. Suppose that T is ergodic and $X \circ T = X$ a.s. For any $a \in \mathbb{R}$, the set $E_a = \{\omega \in \Omega : X(\omega) < a\}$ is clearly invariant since $T^{-1}E_a = \{\omega : X(T\omega) < a\} = E_a$ a.s. Ergodicity implies that $P(E_a) \in \{0, 1\}$ for all $a \in \mathbb{R}$, hence X is a.s. constant.
Conversely, suppose that the only invariant random variables are a.s. constant, and let $A \in \mathcal{I}$. Then 1_A is an invariant random variable, and thus is a.s. constant, and we conclude that $P(A) \in \{0, 1\}$. □

Note that the above proof also holds if we restrict our attention to classes of random variables containing the indicator functions, such as $X \in L^p(\Omega, \mathcal{F}, P)$.

* For the sake of clarity, we will sometimes use the notation of functions rather than random variables, but ultimately the two are the same.

Example 12.1 (Rotation of the Circle)**.** Let $(\Omega, \mathcal{F}, P)$ be $[0, 1)$ with the Borel sets and Lebesgue measure. For $\alpha \in (0, 1)$, define $T_\alpha : \Omega \to \Omega$ by $T_\alpha x = x + \alpha \pmod 1$. T_α is clearly measurable and probability preserving.
* We consider T_α a rotation since $[0, 1)$ may be identified with $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ via the map $x \mapsto e^{2\pi i x}$.
If $\alpha \in \mathbb{Q}$, then $\alpha = \frac{m}{n}$ for some $m, n \in \mathbb{N}$ with $(m, n) = 1$, so for any measurable $B \in \left[0, \frac{1}{2n}\right]$ with $P(B) > 0$, the set $\mathcal{O}(B) = \bigcup_{k=0}^{n-1} \left(B + \frac{k}{n} \pmod 1\right)$ is invariant with $P(\mathcal{O}(B)) \in (0, 1)$, thus T_α is not ergodic.
(Alternatively, the function $f(x) = e^{2\pi i n x}$ is nonconstant and invariant.)

However, T_α is ergodic whenever $\alpha \notin \mathbb{Q}$. To see that this is so, we note that any square-integrable $f : [0,1) \to \mathbb{R}$ has Fourier expansion $\sum_{k\in\mathbb{Z}} c_k e^{2\pi ikx}$ where $c_k = \int_0^1 f(x)e^{-2\pi ikx}dx$ and

$$\sum_{k=-n}^{n} c_k e^{2\pi ikx} \to f(x) \text{ in } L^2([0,1)).$$

If f is invariant, then we have

$$f(x) = f(T_\alpha x) = \sum_{k\in\mathbb{Z}} c_k e^{2\pi ik(x+\alpha(\text{mod } 1))} = \sum_{k\in\mathbb{Z}} c_k e^{2\pi ik\alpha} e^{2\pi ikx}$$

(where equality is in the L^2 sense). Uniqueness of Fourier coefficients implies that $c_k = c_k e^{2\pi ik\alpha}$ for all $k \in \mathbb{Z}$. Since α is irrational, $e^{2\pi ik\alpha} \neq 1$ for $k \in \mathbb{Z} \setminus \{0\}$, so it must be the case that $c_k = 0$ for $k \neq 0$.
It follows that any invariant $f \in L^2$ is constant, which shows that T_α is ergodic.

Example 12.2 (Affine Expanding Maps)**.** Again take $(\Omega, \mathcal{F}, P)$ to be $[0,1)$ with Lebesgue measure.
For any integer $d \geq 2$, define $T_d x = dx \,(\text{mod } 1)$. Note that $T_d^{-1}x = \left\{\frac{x}{d} + \frac{j}{d} : j = 0, 1, ..., d-1\right\}$. (Throughout this example, all operations are taken modulo 1 and we suppress the mod notation for convenience.)
For any interval $I = [a, b)$ with $0 \leq a < b < 1$, $T_d^{-1}I$ can be expressed as the disjoint union $T_d^{-1}I = \bigsqcup_{j=0}^{d-1} I_j$ with $I_j = \left[\frac{a+j}{d}, \frac{b+j}{d}\right)$, so

$$P\left(T_d^{-1}I\right) = \sum_{j=0}^{d-1}\left(\frac{b+j}{d} - \frac{a+j}{d}\right) = d\left(\frac{b-a}{d}\right) = P(I).$$

Thus, by a $\pi - \lambda$ argument, T_d is probability preserving.
To establish ergodicity, we suppose that $f \in L^2$ is invariant with Fourier series $\sum_k c_k e^{2\pi ikx}$.
Then

$$f\left(T_d x\right) = \sum_{k\in\mathbb{Z}} c_k e^{2\pi ik(dx(\text{mod}1))} = \sum_{k\in\mathbb{Z}} c_k e^{2\pi ikdx},$$

so comparing coefficients with $f(x) = \sum_k c_k e^{2\pi ikx}$ shows that $c_k = c_{dk}$. Iterating yields $c_k = c_{dk} = c_{d^2k} = ...$
Since Parseval's identity gives

$$\sum_{k\in\mathbb{Z}} |c_k|^2 = \int_0^1 |f(x)|^2 \, dx < \infty,$$

it must be the case that $c_k = 0$ for all $k \neq 0$, hence f is constant.

Before moving on to the ergodic theorems, we present the following simple yet incredible result due to Henri Poincaré.

Theorem 12.1 (Poincaré Recurrence Theorem)**.** *Let $(\Omega, \mathcal{F}, P, T)$ be a probability preserving dynamical system, and let $U \in \mathcal{F}$. Then for almost every $\omega \in U$, $T^n\omega \in U$ infinitely often.*
If T is ergodic and $P(U) > 0$, then $P\left(T^n\omega \in U \text{ i.o.}\right) = 1$.

Proof. Let $U_n = \bigcup_{j=n}^{\infty} T^{-j}U$ be the set of points $\omega \in \Omega$ that enter U at least once at or after time n.
Then $U_{n+1} = T^{-1}U_n$, so

$$P(U_{n+1}) = P\left(T^{-1}U_n\right) = P(U_n),$$

hence $P(U_n) = P(U_0)$ for all n by induction.

Setting $W = \bigcap_{n=0}^{\infty} U_n = \{\omega \in \Omega : T^n\omega \in U \text{ i.o.}\}$ and noting that $U_0 \supseteq U_1 \supseteq U_2 \ldots$, it follows from continuity from above that

$$P(W) = \lim_{n\to\infty} P(U_n) = P(U_0).$$

The set of points in question is $V = U \cap W$. Since $U, W \subseteq U_0$ and $P(U_0) = P(W)$, we conclude that $P(U) = P(V)$.

For the second claim, note that W is invariant, so ergodicity implies that $P(W) \in \{0, 1\}$.
Since $P(W) \geq P(V) = P(U) > 0$, we must have $P(W) = 1$. □

To put Theorem 12.1 in perspective, suppose that Ω is a separable metric space and $\mathcal{F}$ contains the Borel sets for the metric topology. Then we can cover Ω with countably many open balls of arbitrarily small radius, and almost every point in each ball returns to its neighborhood of origin infinitely often. It follows that $P\left(\liminf_{n\to\infty} d\left(T^n x, x\right) = 0\right) = 1$.

This interpretation of the recurrence theorem lends itself to all sorts of amusing inferences about thermodynamics, cosmology, and so forth, but we will resist the urge to indulge in such speculations here.

A natural question to ask after seeing Theorem 12.1 is "How long does it take to return?" The following result (due to Mark Kac) says that if T is ergodic, the expected first return time for a point in A is $\dfrac{1}{P(A)}$.

Theorem 12.2. *Let $(\Omega, \mathcal{F}, P, T)$ be a probability preserving dynamical system with T ergodic, and suppose that $A \in \mathcal{F}$ has $P(A) > 0$. Define $\tau_A(\omega) = \inf\{n \geq 1 : T^n\omega \in A\}$. Then*

$$\int_A \tau_A dP = 1.$$

Proof.

* In what follows, equality and inclusion are taken to mean "up to a null set." Because the relevant definitions are stated in these terms and all constructions are countable, there is no harm in being loose on this point and the argument is much clearer without constant qualification.

For each $n \in \mathbb{N}$, set $A_n = \{\omega \in A : \tau_A(\omega) = n\}$. The A_n's are disjoint and the recurrence theorem implies that their union is A.
Consequently, we have

$$\sum_{n=1}^{\infty} P(A_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = P(A).$$

Similarly, for $n \geq 1$, define $B_n = \{\omega \in \Omega : \tau_A(\omega) = n\}$, and let $B = \cup_{n=1}^{\infty} B_n$.
Since

$$T^{-1}B_n = \{\omega \in \Omega : T\omega \in B_n\} = \{\omega \in \Omega : T^{n+1}\omega \in A, T^k\omega \notin A \text{ for } 2 \leq k \leq n\} = B_{n+1} \cup T^{-1}A_n,$$

we see that

$$\begin{aligned} T^{-1}B &= \bigcup_{n=1}^{\infty} \left(B_{n+1} \cup T^{-1}A_n\right) = \left(\bigcup_{m=2}^{\infty} B_m\right) \cup T^{-1} \bigcup_{n=1}^{\infty} A_n \\ &= \left(\bigcup_{m=2}^{\infty} B_m\right) \cup T^{-1}A = \bigcup_{n=1}^{\infty} B_n = B, \end{aligned}$$

hence B is an invariant set.

Moreover, $A_n \subseteq B_n$ by definition, so $A \subseteq B$, and thus $P(B) \geq P(A) > 0$. Therefore, it follows from ergodicity that $P(B) = 1$.

Since the B_n's are disjoint, we have

$$\sum_{n=1}^{\infty} P(B_n) = P(B) = 1.$$

Now we can write

$$\int_A \tau_A dP = \sum_{n=1}^{\infty} nP(A_n) = \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(A_k),$$

so if we can show that $\sum_{k=n}^{\infty} P(A_k) = P(B_n)$, then we will obtain

$$\int_A \tau_A dP = \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(A_k) = \sum_{n=1}^{\infty} P(B_n) = 1$$

as desired.

When $m = 1$, we have $B_1 = T^{-1}A$ by definition, so $P(B_1) = P(A) = \sum_{k=1}^{\infty} P(A_k)$.

Now suppose that $\sum_{k=m}^{\infty} P(A_k) = P(B_m)$. Since $T^{-1}B_n = B_{n+1} \cup T^{-1}A_n$ with B_{n+1} and $T^{-1}A_n$ disjoint (because $\omega \in T^{-1}A_n$ implies $T\omega \in A_n \subseteq A$ and thus $\omega \notin B_{n+1}$), the fact that T preserves P implies

$$P(B_n) = P\left(T^{-1}B_n\right) = P(B_{n+1}) + P(T^{-1}A_n) = P(B_{n+1}) + P(A_n),$$

so the inductive hypothesis gives

$$P(B_{m+1}) = P(B_m) - P(A_m) = \sum_{k=m}^{\infty} P(A_k) - P(A_m) = \sum_{k=m+1}^{\infty} P(A_k).$$

This completes the inductive step and the proof. □

13. Ergodic Theorems

Suppose that $(\Omega, \mathcal{F}, P, T)$ is a probability preserving dynamical system. If f is a real-valued measurable function on Ω (i.e. a random variable), we define the n^{th} *Birkhoff sum* as

$$f^n(x) = \sum_{k=0}^{n-1} f\left(T^k x\right),$$

so that $\frac{1}{n} f^n$ gives the *time average* over orbit segments of length n.

Ergodic theorems are statements of the form $\frac{1}{n} f^n \to \overline{f}$ where $\overline{f}$ is invariant with $\int_\Omega \overline{f} dP = \int_\Omega f dP$.

If $g = f \cdot 1_A$ for $A \in \mathcal{I}$, then

$$\frac{1}{n} g^n = \frac{1}{n} \sum_{k=0}^{n-1} (f \cdot 1_A) \circ T^k = \frac{1}{n} \sum_{k=0}^{n-1} (f \circ T_k) 1_{T^{-k} A} = \frac{1}{n} \sum_{k=0}^{n-1} (f \circ T_k) 1_A = \frac{1}{n} f^n \cdot 1_A,$$

so

$$\int_A \overline{f} dP = \int_\Omega \overline{g} dP = \int_\Omega g dP = \int_A f dP.$$

(In the cases we consider, g will satisfy the assumptions of an ergodic theorem if f does.)

Thus, in probabilistic terms, $\overline{f} = E[f \,|\mathcal{I}]$.

When T is ergodic, $\overline{f}$ must be a.s. constant, hence $\overline{f} = E[f]$ - the *space average*.

The first proof of an ergodic theorem was by John von Neumann in 1931, establishing convergence in L^2 (and thus in L^1 by an approximation argument). This result is known as the *mean ergodic theorem*.

von Neumann's paper was not published until 1932, by which time George Birkhoff (after hearing about the former's work) had already published a proof of the *pointwise ergodic theorem* establishing a.s. convergence.

We begin with a cute proof of the mean ergodic theorem in L^2 due to Frigyes Riesz.

Theorem 13.1. *Let $S = \{g \in L^2 : g \circ T = g\}$ and let P_S be the projection from L^2 onto S. For all $f \in L^2$, one has $\frac{1}{n} f^n \to P_S f$ in L^2.*

Proof. Define the *Koopman operator* $U : L^2 \to L^2$ by $Uf = f \circ T$. Then

$$\begin{aligned} \langle Uf, Ug \rangle &= \int (Uf)(x)(\overline{Ug})(x) dP(x) \\ &= \int (f\overline{g}) \circ T(x) dP(x) \\ &= \int (f\overline{g})(x) dP(x) = \langle f, g \rangle \end{aligned}$$

where the penultimate equality used the fact that T preserves P.

* This shows that $\langle f, g \rangle = \langle Uf, Ug \rangle = \langle f, U^* U g \rangle$ for all $f, g \in L^2$, so $U^* U$ is the identity. A bounded linear operator with this property is said to to be an *isometry*. If U is surjective as well, then UU^* is also the identity and we say that U is *unitary*.

Now set $W = \{Uh - h : h \in L^2\}$. We will show that $W^\perp = S$.

To see that $S \subseteq W^\perp$, let $f \in S$, $g \in W$ so that $f = Uf$ and $g = Uh - h$ for some $h \in L^2$. Then

$$\langle f, g\rangle = \langle f, Uh - h\rangle = \langle f, Uh\rangle - \langle f, h\rangle = \langle Uf, Uh\rangle - \langle f, h\rangle = 0,$$

so, since $g \in W$ was arbitrary, we have that $f \in W^\perp$.

For the reverse inclusion, let $f \in W^\perp$ so that for every $h \in L^2$, we have $\langle f, Uh - h\rangle = 0$, hence

$$\langle f, h\rangle = \langle f, Uh\rangle = \langle U^* f, h\rangle$$

Since h was arbitrary, this implies that $U^* f = f$.
Accordingly, we have

$$\begin{aligned}\|Uf - f\|_2^2 &= \langle Uf - f, Uf - f\rangle \\ &= \langle Uf, Uf\rangle - \langle Uf, f\rangle - \langle f, Uf\rangle + \langle f, f\rangle \\ &= \|f\|_2^2 - \langle f, U^* f\rangle - \langle U^* f, f\rangle + \|f\|_2^2 \\ &= 2\|f\|_2^2 - 2\langle f, f\rangle = 0,\end{aligned}$$

and we conclude that $f \in S$.

Since W is clearly a subspace of L^2 (though not necessarily closed), we have $L^2 = \overline{W} \oplus W^\perp = \overline{W} \oplus S$.

(If V is a closed subspace of a Hilbert space H, then $H = V \oplus V^\perp$.
Since $W \subseteq \overline{W}$, $\overline{W}^\perp \subseteq W^\perp$, so $H = \overline{W} \oplus \overline{W}^\perp \subseteq \overline{W} + W^\perp \subseteq H$.
To see that the latter sum is direct, note that $x \in \overline{W} \cap W^\perp$ implies $\langle x, x\rangle = \langle x, \lim_n x_n\rangle = \lim_n \langle x, x_n\rangle = 0$.)

Now if $f \in S$, then $U^k f = f$ for all $k \geq 0$, so

$$\frac{1}{n} f^n = \frac{1}{n}\sum_{k=0}^{n-1} U^k f = \frac{1}{n} \cdot nf = f,$$

hence $\frac{1}{n} f^n \to f = P_S f$.

If $g = Uh - h \in W$, then

$$g^n = \sum_{k=0}^{n-1} U^k g = \sum_{k=0}^{n-1} \left(U^{k+1} h - U^k h\right) = U^n h - h,$$

thus $\frac{1}{n} g^n = \frac{1}{n}\left(U^n h - h\right) \to 0 = P_S g$. (Recall that convergence is in the L^2 sense.)

If $g \in \overline{W}$, there exists a sequence $g_i \in W$ with $g_i \to g$, so for any $\varepsilon > 0$, taking i so that $\|g - g_i\|_2 < \varepsilon$ shows that we can take n large enough that

$$\left\|\frac{1}{n}\sum_{k=0}^{n-1} U^k g\right\|_2 \leq \frac{1}{n}\sum_{k=0}^{n-1} \left\|U^k (g - g_i)\right\|_2 + \left\|\frac{1}{n}\sum_{k=0}^{n-1} U^k g_i\right\|_2 < 2\varepsilon,$$

since U is an isometry and $\frac{1}{n}\sum_{k=0}^{n-1} U^k g_i \to 0$, so $\frac{1}{n} g^n \to 0$ for $g \in \overline{W}$.

Finally, if $F \in L^2$, then $F = f + g$ with $f \in S$, $g \in \overline{W}$, so $\frac{1}{n} F^n = \frac{1}{n} f^n + \frac{1}{n} g^n \to f = P_S F$ in L^2. □

* Note that the limit function $\overline{f} = P_S f$ is the projection of f onto the space of T-invariant functions in L^2. Theorem 1.6 shows that this is $E[f\,|\mathcal{I}]$.

Also, it is worth observing that the above proof shows more generally that if U is a linear isometry on a Hilbert space H, then $\left\|\frac{1}{n}\sum_{k=0}^{n-1} U^k x - P_{\ker(U-I)} x\right\| \to 0$ for all $x \in H$.

Analogous to the development of conditional expectation in terms of projection, we have

Corollary 13.1. *If $f \in L^1$, then $\frac{1}{n}f^n \to \overline{f}$ in L^1 where $\overline{f}$ is invariant with $\int_\Omega \overline{f}dP = \int_\Omega fdP$.*

Proof. Since bounded L^1 functions are dense in L^1 and are square integrable, for any $\varepsilon > 0$, there is a $g \in L^1_b = \{h \in L^1 : \|h\|_\infty < \infty\}$ such that $\|f - g\|_1 < \varepsilon$. Since $g \in L^1_b \subseteq L^2$, Theorem 13.1 shows that $\frac{1}{n}g^n \to \overline{g}$ in L^2 where $\overline{g} = P_S g$.

As $\|\cdot\|_1 \leq \|\cdot\|_2$ (by Hölder's inequality), we have that $\frac{1}{n}g^n \to \overline{g}$ in L^1.

Now choose n big enough that $\left\|\frac{1}{n}g^n - \overline{g}\right\|_1 < \varepsilon$. Since

$$\left\|\frac{1}{n}f^n - \frac{1}{n}g^n\right\|_1 \leq \frac{1}{n}\sum_{k=0}^{n-1}\left\|(f-g)\circ T^k\right\|_1 \leq \|f-g\|_1 < \varepsilon,$$

we have that $\left\|\frac{1}{n}f^n - \overline{g}\right\|_1 \leq \left\|\frac{1}{n}f^n - \frac{1}{n}g^n\right\|_1 + \left\|\frac{1}{n}g^n - \overline{g}\right\|_1 < 2\varepsilon$ for n sufficiently large.

Eliminating $\overline{g}$ shows that for all large m, n, we have

$$\left\|\frac{1}{n}f^n - \frac{1}{m}f^m\right\|_1 \leq \left\|\frac{1}{n}f^n - \overline{g}\right\|_1 + \left\|\overline{g} - \frac{1}{m}f^m\right\|_1 < 4\varepsilon,$$

so $\frac{1}{n}f^n$ has a limit $\overline{f}$ in L^1 (by completeness).

Invariance follows from uniqueness of L^1 limits since

$$\frac{1}{n}f^n \circ T = \frac{1}{n}\sum_{k=1}^{n} f\circ T^k = \frac{n+1}{n}\cdot\frac{1}{n+1}\sum_{k=0}^{n} f\circ T^k - \frac{1}{n}f \to_{L^1} \overline{f},$$

and

$$\lim_{n\to\infty}\int\left|\frac{1}{n}f^n\circ T - \overline{f}\circ T\right|dP = \lim_{n\to\infty}\int\left|\left(\frac{1}{n}f^n - \overline{f}\right)\circ T\right|dP = \lim_{n\to\infty}\int\left|\frac{1}{n}f^n - \overline{f}\right|dP = 0.$$

Finally, L^1 convergence, Fubini's theorem, and the fact that T preserves P imply

$$\int_\Omega \overline{f}dP = \lim_{n\to\infty}\int_\Omega \frac{1}{n}f^n dP = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\int_\Omega f\circ T^k dP = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\int_\Omega fdP = \int_\Omega fdP.$$

□

The key to proving the pointwise ergodic theorem is

Theorem 13.2 (Maximal Ergodic Theorem)**.** *Suppose that $f \in L^1$ and define $M(f) = \{x \in \Omega : \sup_{n\geq 1} f^n(x) > 0\}$. Then $\int_{M(f)} fdP \geq 0$.*

Proof. Write $F_N(x) = \max_{0\leq k\leq N} f^k(x)$ where we adopt the convention that $f^0 \equiv 0$.

Then $F_1 \leq F_2 \leq \ldots$ pointwise, so the sets $M_N(f) = \{x \in \Omega : F_N(x) > 0\}$ form a nested increasing sequence with $M(f) = \bigcup_{N=1}^\infty M_N(f)$.

Now for every $k = 0, 1, ..., N$,

$$\begin{aligned} F_N(Tx) + f(x) &= f(x) + \max_{0\leq j\leq N} f^j(Tx) \geq f(x) + f^k(Tx) \\ &= f(x) + \sum_{j=0}^{k-1} f(T^{j+1}x) = \sum_{j=0}^{k} f(T^j x) = f^{k+1}(x), \end{aligned}$$

hence $f(x) \geq f^k(x) - F_N(Tx)$ for $k = 1, ..., N+1$.

Because $F_N(x) = \max_{1\le k\le N} f^k(x)$ for $x \in M_N(f)$, this shows that $f(x) \ge F_N(x) - F_N(Tx)$ for $x \in M_N(f)$. As F_N is nonnegative and equal to 0 on $\Omega \setminus M_N(f)$, we have

$$\begin{aligned}\int_{M_N(f)} f dP &\ge \int_{M_N(f)} F_N dP - \int_{M_N(f)} F_N \circ T dP \\ &= \int_\Omega F_N dP - \int_{M_N(f)} F_N \circ T dP \\ &\ge \int_\Omega F_N dP - \int_\Omega F_N \circ T dP = 0.\end{aligned}$$

Finally, since $M_N(f) \nearrow M(f)$, the dominated convergence theorem shows that

$$\int_{M(f)} f dP = \lim_{N\to\infty} \int_{M_N(f)} f dP \ge 0.$$

□

Corollary 13.2. *If $M_\alpha(f) = \{x : \sup_{n\ge1} \frac{1}{n} f^n(x) > \alpha\}$, then $\int_{M_\alpha(f)} f dP \ge \alpha P(M_\alpha(f))$.*

Proof. Let $g = f - \alpha$. Then

$$g^n = \sum_{k=0}^{n-1} (f-\alpha)\circ T^k = \sum_{k=0}^{n-1} \left(f\circ T^k - \alpha\right) = f_n - n\alpha,$$

so $\frac{1}{n} g^n = \frac{1}{n} f^n - \alpha$, and thus

$$\begin{aligned}M_\alpha(f) &= \left\{x : \sup_{n\ge1} \frac{1}{n} f^n(x) > \alpha\right\} = \left\{x : \sup_{n\ge1} \frac{1}{n} f^n(x) - \alpha > 0\right\} \\ &= \left\{x : \sup_{n\ge1} \frac{1}{n} g^n(x) > 0\right\} = \left\{x : \sup_{n\ge1} g^n(x) > 0\right\} = M(g).\end{aligned}$$

Therefore, the maximal ergodic theorem implies

$$0 \le \int_{M(g)} g dP = \int_{M_\alpha(f)} (f-\alpha)\, dP = \int_{M_\alpha(f)} f dP - \alpha P\left(M_\alpha(f)\right).$$

□

Corollary 13.3. *If $A \subseteq M(f)$ is T-invariant, then $\int_A f dP \ge 0$.*

Proof. Since $T^{-1}A = A$ up to a null set, $1_A \circ T = 1_A$ a.s., so if $g = f \cdot 1_A$, then

$$g^n = \sum_{k=0}^{n-1} (f\cdot 1_A)\circ T^k = \sum_{k=0}^{n-1} \left(f\circ T^k\right)\cdot 1_A = f^n \cdot 1_A.$$

It follows that

$$M(g) = \left\{x \in \Omega : \sup_{n\ge1} g^n(x) > 0\right\} = \left\{x \in A : \sup_{n\ge1} f^n(x) > 0\right\} = A \cap M(f) = A,$$

and thus

$$\int_A f dP = \int_{M(g)} g dP \ge 0.$$

□

We are now in a position to prove

Theorem 13.3 (Pointwise Ergodic Theorem)**.** *For any $f \in L^1$,*

$$\lim_{n\to\infty} \frac{1}{n} f^n = \overline{f} \text{ a.s.}$$

where $\overline{f} \in L^1$ is invariant with $\int_\Omega f dP = \int_\Omega \overline{f} dP$.

Proof. Set

$$f^+(x) = \limsup_{n\to\infty} \frac{1}{n} f^n(x),$$
$$f^-(x) = \liminf_{n\to\infty} \frac{1}{n} f^n(x).$$

Clearly f^+ and f^- are invariant with $f^-(x) \le f^+(x)$ for all x. We wish to show that $f^+ = f^-$ a.s., so we need to show that $M = \{x : f^-(x) < f^+(x)\}$ has $P(M) = 0$.
For $\alpha, \beta \in \mathbb{Q}$, let $M_{\alpha,\beta} = \{x : f^-(x) < \alpha, f^+(x) > \beta\}$. Then $M = \bigcup_{\alpha,\beta\in\mathbb{Q}} M_{\alpha,\beta}$, so it suffices to show that $P(M_{\alpha,\beta}) = 0$ for all $\alpha, \beta \in \mathbb{Q}$ with $\alpha < \beta$.

We note at the outset that the invariance of f^+ and f^- imply that $M_{\alpha,\beta}$ is an invariant set.
Now let $M_\beta^+ = \{x : f^+(x) > \beta\}$. If $x \in M_\beta^+$, then there is some $n \in \mathbb{N}$ such that $\frac{1}{n} f^n(x) > \beta$, so $(f-\beta)^n(x) = f^n(x) - n\beta > 0$, hence $x \in M(f-\beta)$.
Since $M_{\alpha,\beta} \subseteq M_\beta^+ \subseteq M(f-\beta)$ is invariant, Corollary 13.3 shows that $\int_{M_{\alpha,\beta}} (f-\beta) dP \ge 0$, hence $\int_{M_{\alpha,\beta}} f dP \ge \beta P(M_{\alpha,\beta})$.

Similarly, if $x \in M_\alpha^- := \{x : f^-(x) < \alpha\}$, then there is an m with $\frac{1}{m} f^m < \alpha$, so $(\alpha - f)^m > 0$.
It follows that $M_{\alpha,\beta} \subseteq M_\alpha^- \subseteq M(\alpha - f)$, so that $\int_{M_{\alpha,\beta}} f dP \le \alpha P(M_{\alpha,\beta})$.

Thus we have shown that

$$\beta P(M_{\alpha,\beta}) \le \int_{M_{\alpha,\beta}} f dP \le \alpha P(M_{\alpha,\beta}),$$

so since $\alpha < \beta$, we conclude that $P(M_{\alpha,\beta}) = 0$ as desired. This shows that $\frac{1}{n} f^n$ has an almost sure limit f^*.

By Corollary 13.1, we also have that $\frac{1}{n} f^n \to \overline{f}$ in L^1, and Fatou's lemma gives

$$\int |f^* - \overline{f}| dP = \int \liminf_n \left| \frac{1}{n} f^n - \overline{f} \right| dP \le \liminf_n \int \left| \frac{1}{n} f^n - \overline{f} \right| dP = 0,$$

so $f^* = \overline{f}$ a.s.

In light of previous observations, this shows that $\frac{1}{n} f^n$ converges a.s. to $\overline{f} = E[f|\mathcal{I}]$. □

It is left as a homework exercise to use Theorem 13.3 to extend the mean ergodic theorem to

Theorem 13.4. *If $f \in L^p$, $1 \le p < \infty$, then $\frac{1}{n} f^n \to E[f|\mathcal{I}]$ in L^p.*

We conclude our discussion of ergodic theorems with some examples.

Example 13.1 (Strong Law of Large Numbers)**.** Suppose that $X_1, X_2, ...$ is an i.i.d. sequence with $X_1 \in L^1$. We can think of the X_i's as coordinate projections on the probability space $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P)$ from Kolmogorov's theorem with P arising from finite dimensional distributions given by product measure.

If θ is the shift map, then $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}}, P, \theta)$ is a probability preserving dynamical system.
An invariant set A has $\{\omega \in A\} = \{\theta\omega \in A\} \in \sigma(X_2, X_3, ...)$. Iterating shows that
$A \in \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, ...) = \mathcal{T}$, the tail field.
Since Kolmogorov's $0-1$ law shows that $\mathcal{T}$ is trivial, this shows that $\mathcal{I}$ is trivial, hence θ is ergodic.

The pointwise ergodic theorem gives

$$\frac{1}{n}\sum_{k=1}^{n} X_k = \frac{1}{n}\sum_{k=0}^{n-1} X_1 \circ \theta^k \to E[X_1 \,|\mathcal{I}] = E[X_1],$$

the strong law of large numbers!
(Corollary 13.1 shows that we also have convergence in L^1.)

Example 13.2 (Markov Ergodic Theorem)**.** Suppose that $\{X_n\}$ is a Markov chain with countable state space S and stationary distribution π with $\pi(x) > 0$ for all $x \in S$. If $X_0 \sim \pi$, then $X_0, X_1, ...$ is stationary and the shift map is probability preserving.

If R is a recurrent communicating class, then $X_0 \in R$ implies $X_n \in R$ for all n, so $\{\omega : X_0(\omega) \in R\} \in \mathcal{I}$. Thus θ is not ergodic if the chain is not irreducible.

If the chain is irreducible and A is invariant, then, taking $\mathcal{F}_n = \sigma(X_0, ..., X_n)$ as usual, it follows from the Markov property and the invariance of A that

$$E_\pi[1_A \,|\mathcal{F}_n] = E_\pi[1_A \circ \theta^n \,|\mathcal{F}_n] = h(X_n)$$

where $h(x) = E_x[1_A]$.

Levy's $0-1$ law implies that $E_\pi[1_A \,|\mathcal{F}_n] \to 1_A$ a.s., and since the chain is irreducible and recurrent (by irreducibility and the existence of a stationary distribution), for every $x \in S$, $P_\pi(X_n = x \text{ i.o.}) = 1$. This means that $h(X_n) = h(x)$ infinitely often for any $x \in S$, so it must be the case that $h \equiv 0$ or $h \equiv 1$ a.s. In other words, $P_\pi(A) \in \{0, 1\}$, so the shift map is ergodic.

For any $f \in L^1(\pi)$, applying the ergodic theorem to $f \circ X_0$ yields

$$\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \to \sum_x f(x)\pi(x) \text{ a.s. and in } L^1.$$

Example 13.3 (Equidistribution Theorem)**.** As a final example, recall that irrational rotation is ergodic. Thus Theorem 13.3 shows that for any $\alpha \in (0,1) \setminus \mathbb{Q}$, the map $T\omega = \omega + \alpha \,(\text{mod } 1)$ satisfies

$$\frac{1}{n}\sum_{k=0}^{n-1} 1\{T^k\omega \in A\} \to \lambda(A) \text{ a.s.}$$

for all $A \in \mathcal{B}_{[0,1)}$, where λ is Lebesgue measure.

In fact, we can show

Claim. If $A = [a, b)$, then the exceptional set is $\emptyset$.

Proof. Write $A_k = \left[a + \frac{1}{k}, b - \frac{1}{k}\right)$. If $b - a > \frac{2}{k}$, then the ergodic theorem shows that

$$\frac{1}{n}\sum_{j=0}^{n-1} 1\{T^j\omega \in A_k\} \to b - a - \frac{2}{k}$$

for all $\omega \in \Omega_k$ where $\lambda(\Omega_k) = 1$.

Let $G = \bigcap_{k > \frac{2}{b-a}} \Omega_k$. Then $\lambda(G) = 1$, so G is dense in $[0,1)$, thus for any $x \in [0,1)$, $k \in \mathbb{N} \cap \left(\frac{2}{b-a}, \infty\right)$, we can find $\omega_k \in G$ so that $|\omega_k - x| < \frac{1}{k}$.
Since $T^j\omega_k \in A_k$ implies that $T^j x \in A$, we have

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} 1\left\{T^j x \in A\right\} \geq \liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} 1\left\{T^j\omega_k \in A_k\right\} = b - a - \frac{2}{k}.$$

As k was arbitrary, we conclude that

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} 1\left\{T^j x \in A\right\} \geq b - a.$$

A similar argument with A^C shows that

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} 1\left\{T^j x \in A\right\} \leq b - a,$$

and the claim follows. □

A neat application of this result concerns the distribution of leading digits in the decimal expansion of powers of 2. The first few terms of the sequence of initial digits are

$$1, 2, 4, 8, 1, 3, 6, 1, 2, 5, 1, 2, 4, 8, 1, 3, 5, 1, ...$$

If 2^j has leading digit k, then $k \cdot 10^m \leq 2^j < (k+1) \cdot 10^m$ for some $m \in \mathbb{N}_0$, thus

$$m + \log_{10}(k) \leq j\log_{10}(2) < m + \log_{10}(k+1).$$

If T is rotation by $\alpha = \log_{10}(2) \in (0,1) \setminus \mathbb{Q}$, then the above inequality shows that

$$\log_{10}(k) \leq T^j 0 < \log_{10}(k+1).$$

Thus if

$$N_k(n) = \left\{j \in [0, n-1] \cap \mathbb{Z} : k \text{ is the first digit in the decimal expansion of } 2^j\right\},$$

then we have

$$\frac{|N_k(n)|}{n} = \frac{1}{n}\sum_{j=0}^{n-1} 1\left\{\log_{10}(k) \leq T^j 0 < \log_{10}(k+1)\right\} \to \log_{10}(k+1) - \log_{10}(k) = \log_{10}\left(\frac{k+1}{k}\right).$$

(Of course, one can play the same game with powers and bases other than 2 and 10 and can consider the leading r digits as well by the same basic reasoning.)

The distribution $p(k) = \log_{10}\left(\frac{k+1}{k}\right)$ on $\{1, ..., 9\}$ is known as *Benford's law* and it has been shown to model many real world data sets surprisingly well. It is sometimes used on financial and scientific data as a method of fraud detection.

The general idea is that if you have a collection of numbers that range over several orders of magnitude and are approximately uniformly distributed on a logarithmic scale, then the distribution of leading digits should be close to Benford's law:

x has leading digit d if there is some m with $d \cdot 10^m \leq x < (d+1) \cdot 10^m$, or

$$\log_{10}(d) + m \leq \log_{10}(x) < \log_{10}(d+1) + m,$$

and the width of the interval $[\log_{10}(d) + m, \log_{10}(d+1) + m)$ is $\log_{10}\left(\frac{d+1}{d}\right)$.

The reason that one wants the data to span several orders of magnitude is that a sample point has leading digit d if it falls in any of the intervals $[\log_{10}(d) + m, \log_{10}(d+1) + m)$, $[\log_{10}(d) + m + 1, \log_{10}(d+1) + m + 1)$, $[\log_{10}(d) + m + 2, \log_{10}(d+1) + m + 2)$, etc..., and averaging over more intervals blurs out local deviations from log uniformity.

Note that by a standard change of variables argument, a random variable whose logarithm is uniform has density $f(x) \propto x^{-1}$. Thus Benford's law describes the leading digits of random variables which obey power laws with exponent 1.

One way that this leading digit law might come up is if you were to examine the size of an exponential growth process at a random time. Basically, this is because exponential growth translates to linear growth on a logarithmic scale.

As with the power law description, what we are ultimately picking up on is some kind of self-similarity or scale invariance. This is one reason that Benford's law is so often observed empirically: In many financial applications, you expect the same basic picture whether you're working with dollars or yen. Similarly, if you are looking at scientific data, then it generally shouldn't matter too much whether length is measured in centimeters or furlongs.

A final example in which the law might arise is a process which is subject to successive multiplicative perturbations, so that its logarithm undergoes additive perturbations (i.e. random walk). The CLT shows that after many time steps, the law of the log should be approximately normal with huge variance, and this looks uniform in the central regime.

14. Brownian Motion

Historically, Brownian motion refers to the random movement of particles suspended in a fluid famously described by Robert Brown in 1827.

In 1905, Albert Einstein explained this phenomenon in terms of the motion of water molecules. Einstein's analysis and the experimental verification of his predictions were a major force behind the widespread acceptance of the atomic hypothesis.

Five years prior to the publication of Einstein's paper, Louis Bachelier gave a mathematical treatment of Brownian motion in the context of evaluating stock options. This work (Bachelier's PhD thesis) helped usher in the modern era of mathematical finance.

The first fully rigorous mathematical derivation of Brownian motion was due to Norbert Wiener in a series of papers in the early 1920s.

It is this *Wiener process* - the mathematical construct rather than its various physical manifestations - with which we will concern ourselves here, though we retain the colloquial term "Brownian motion."

Definition. A real-valued stochastic process $\{B(t) : t \geq 0\}$ is called a (linear) *Brownian motion* if

(1) For all times $0 \leq t_1 < \cdots < t_n$, $B(t_1), B(t_2) - B(t_1), \ldots, B(t_n) - B(t_{n-1})$ are independent random variables,
(2) For all $t \geq 0$, $h > 0$, the increment $B(t+h) - B(t)$ is normal with mean 0 and variance h,
(3) The function $t \mapsto B(t)$ is almost surely continuous.

If $B(0) = x$ a.s., we say that $\{B(t) : t \geq 0\}$ is a Brownian motion started at x.
The case $x = 0$ is known as *standard Brownian motion*.

We will sometimes talk about Brownian motion on $[0, T]$ for some $T \geq 0$. This just means that properties 1-3 hold when we restrict our attention to times in $[0, T]$.

We will only concern ourselves with the one-dimensional case here, but we observe that if $B_1(t), ..., B_d(t)$ are independent linear Brownian motions started at $x_1, ..., x_d$, respectively, then $B(t) = (B_1(t), ..., B_d(t))^T$ is a d-dimensional Brownian motion started at $(x_1, ..., x_d)^T$.

On the one hand, we can regard $\{B(t) : t \geq 0\}$ as a collection of random variables $\omega \mapsto B(t, \omega)$ defined on some underlying probability space $(\Omega, \mathcal{F}, P)$ and indexed by $t \in [0, \infty)$.

Alternatively, Brownian motion can be thought of as a random function. Specifically, Property 3 allows us to interpret Brownian motion as a random variable taking values in the space of continuous functions $C([0, \infty), \mathbb{R})$. The target σ-algebra is the Borel sets for the topology of uniform convergence on compact sets (which coincides with the natural product σ-algebra generated by the coordinate maps $\pi_t : f \mapsto f(t)$).

The reason we don't work in the space of measurable functions from $[0, \infty)$ to $\mathbb{R}$ with the product σ-algebra is that every measurable set is then determined by the values of the functions at a countable number of points, hence the set of continuous functions is not even measurable!

For a fixed $\omega \in \Omega$, the map $t \mapsto B(t, \omega)$ is called a *sample path* or *trajectory*.

* We will switch freely between the equivalent notations $B(t), B_t, B(t, \omega), B_t(\omega)$ depending on readability and whether we want to emphasize the role of ω.

Before establishing that the process we have defined actually exists, we make a few simple observations in order to familiarize ourselves with our object of study.

Proposition 14.1 (Translation Invariance)**.** *If $\{B_t : t \geq 0\}$ is a Brownian motion started at x, then for any $y \in \mathbb{R}$, $\{B_t + y : t \geq 0\}$ is a Brownian motion started at $x + y$.*

Proof. Properties 1 and 2 hold since the y terms cancel out and Property 3 holds since adding a constant preserves continuity. The starting state is $B_0 + y = x + y$. □

The exact same argument shows that if $\{B_t\}_{t\geq 0}$ is a standard Brownian motion and Y is independent of $\{B_t\}_{t\geq 0}$, then the process $\{X_t\}_{t\geq 0}$ defined by $X_t = B_t + Y$ is a Brownian motion with $X_0 = Y$ a.s. Thus there is no real loss of generality in restricting our attention to standard Brownian motion.

In a similar vein, we have the following Markov property for time shifts.

Proposition 14.2 (Time-shift Invariance)**.** *If $\{B(t) : t \geq 0\}$ is a Brownian motion and $s \geq 0$, then $\{B(s+t) - B(s) : t \geq 0\}$ is a standard Brownian motion and is independent of the process $\{B(t) : 0 \leq t \leq s\}$.*

Proof. Clearly the process starts at $B(s) - B(s) = 0$. Properties 1 and 2 follow from cancellation of the $B(s)$ terms and Property 3 holds since the composition of a.s. continuous functions is a.s. continuous.

Now, stochastic processes $\{X(s) : s \in S\}$, $\{Y(t) : t \in T\}$ are independent if for every $m, n \in \mathbb{N}$, $s_1, \ldots, s_m \in S$, $t_1, ..., t_n \in T$, the random vectors $(X(s_1), \ldots, X(s_m))$ and $(Y(t_1), \ldots, Y(t_n))$ are independent.

Since Brownian motion has independent increments, for any $t_1, \ldots, t_n \geq 0$, $0 \leq s_1, \ldots, s_m \leq s$, the vectors $(B(s + t_1) - B(s), \ldots, B(s + t_n) - B(s))$ and $(B(s_1), \ldots, B(s_m))$ are independent. (The former is built from disjoint increments to the right of s and the latter from disjoint increments to the left.) □

Another simple but extremely useful invariance property is

Proposition 14.3 (Diffusive Scaling)**.** *If $\{B(t) : t \geq 0\}$ is a standard Brownian motion and $a \neq 0$, then the process $\{X(t) : t \geq 0\}$ defined by $X(t) = \frac{1}{a}B(a^2t)$ is also a standard Brownian motion.*

Proof. Again, we just have to verify the defining properties. Clearly $X(0) = \dfrac{1}{a}B(0) = 0$.

If $0 \leq t_1 < \cdots < t_n$, then $0 \leq a^2t_1 < \cdots < a^2t_n$, so $B(a^2t_1), B(a^2t_2) - B(a^2t_1), \ldots, B(a^2t_n) - B(a^2t_{n-1})$ are independent, hence

$$X(t_1), X(t_2) - X(t_1), ..., X(t_n) - X(t_{n-1}) = \frac{B(at_1^2)}{a}, \frac{B(a^2t_2) - B(a^2t_1)}{a}, \ldots, \frac{B(a^2t_n) - B(a^2t_{n-1})}{a}$$

are independent.

Similarly, for $t \geq 0$, $h > 0$, $B(a^2t + a^2h) - B(a^2t)$ is normally distributed with mean 0 and variance a^2h, so

$$X(t + h) - X(t) = \frac{B(a^2t + a^2h) - B(a^2t)}{a}$$

is normal with mean 0 and variance h.

Finally, $X(t) = \dfrac{1}{a}B(a^2t)$ is a composition of a.s. continuous functions and thus is a.s. continuous. □

Observe that taking $a = -1$ shows that standard Brownian motion is symmetric about 0.

It is sometimes more convenient to express the scaling relation as $\{B_{at}\}_{t\geq 0} =_d \{\sqrt{a}B_t\}_{t\geq 0}$ for $a > 0$.

At this point, it is useful to give the following alternative characterization of Brownian motion.

Theorem 14.1. *A real-valued process $\{B_t\}_{t\geq 0}$ with $B_0 = 0$ is a standard Brownian motion if and only if*

a) B_t *is a Gaussian process - i.e. all of its finite dimensional distributions are multivariate normal,*

b) $E[B_t] = 0$ *and* $E[B_s B_t] = s \wedge t$,

c) With probability one, $t \mapsto B_t$ *is continuous.*

Proof. To see that Brownian motion is a Gaussian process, fix times $0 < t_1 < ... < t_n$ and define

$$M = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & -1 & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} \frac{1}{\sqrt{t_1}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{t_2 - t_1}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{t_n - t_{n-1}}} \end{bmatrix}.$$

Properties 1 and 2 show that the random vector

$$X = DM \left(B(t_1), ..., B(t_n)\right)^T = \left(\frac{B(t_1)}{\sqrt{t_1}}, \frac{B(t_2) - B(t_1)}{\sqrt{t_2 - t_1}}, \ldots, \frac{B(t_n) - B(t_{n-1})}{\sqrt{t_n - t_{n-1}}}\right)^T$$

has independent standard normal entries.
Since M and D are nonsingular, the matrix $A = M^{-1}D^{-1}$ is well-defined, so we have
$\left(B(t_1), \ldots, B(t_n)\right)^T = AX$, which is multivariate normal (with mean 0 and covariance AA^T) by definition.
Now suppose that $s \leq t$. Properties 1 and 2 show that $B_s \sim \mathcal{N}(0, s)$ and $B_t - B_s \sim \mathcal{N}(0, t - s)$ are independent, so we have

$$E[B_s] = 0, \qquad E[B_s B_t] = E[B_s(B_t - B_s)] + E[B_s^2] = s.$$

For the other direction, note that multivariate normals have independent entries if and only if all covariances are 0. If $\{B(t)\}_{t\geq 0}$ satisfies a and b, then for any $r \leq s \leq t \leq u$,

$$\begin{aligned} E\left[\left(B(s) - B(r)\right)\left(B(u) - B(t)\right)\right] &= E\left[B(s)B(u)\right] - E\left[B(s)B(t)\right] - E\left[B(r)B(u)\right] + E\left[B(r)B(t)\right] \\ &= s - s - r + r = 0, \end{aligned}$$

so Property 1 holds.
Similarly, for any $t \geq 0$, $h > 0$, $B(t+h) - B(h)$ is a difference of mean zero normals and thus is normal with mean zero and variance

$$\begin{aligned} \mathrm{Var}\left(B(t+h) - B(t)\right) &= \mathrm{Var}\left(B(t+h)\right) + \mathrm{Var}\left(B(t)\right) - 2\mathrm{Cov}\left(B(t+h), B(t)\right) \\ &= (t+h) + t - 2E\left[B(t+h)B(t)\right] = 2t + h - 2t = h. \end{aligned}$$

□

With this alternative definition, we can prove two more simple results.

Proposition 14.4 (Time Inversion)**.** *Suppose that $\{B(t) : t \geq 0\}$ is a standard Brownian motion. Then the process $\{X(t) : t \geq 0\}$ defined by*

$$X(t) = \begin{cases} 0, & t = 0 \\ tB\left(\frac{1}{t}\right), & t > 0 \end{cases}$$

is also a standard Brownian motion.

Proof. Since for any $t_1, ..., t_n > 0$, $\left(B(t_1^{-1}), ..., B(t_n^{-1})\right)^T$ is a Gaussian vector, so is

$$(X(t_1), ..., X(t_n))^T = \begin{bmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & t_n \end{bmatrix} \left(B\left(\frac{1}{t_1}\right), ..., B\left(\frac{1}{t_n}\right)\right)^T.$$

Also, $E[X(0)] = 0$ and $E[X(t)] = tE\left[B\left(\frac{1}{t}\right)\right] = 0$ for $t > 0$, and for $0 < s \leq t$, $E[X(0)X(t)] = 0$ and

$$E[X(s)X(t)] = stE\left[B\left(\frac{1}{s}\right)B\left(\frac{1}{t}\right)\right] = st \cdot \frac{1}{t} = s.$$

Finally, the paths $t \mapsto X(t)$ are a.s. continuous for $t > 0$ since compositions of continuous functions are continuous, so it remains only to demonstrate that $\lim_{t\to 0^+} X(t) = 0$.
Because $\mathbb{Q}^+ = \mathbb{Q} \cap (0, \infty)$ is countable, the foregoing shows that $\{X(t) : t \in \mathbb{Q}^+\}$ has the same distribution as $\{B(t) : t \in \mathbb{Q}^+\}$, so we must have

$$P\left(\lim_{\substack{t\to 0 \\ t\in\mathbb{Q}^+}} X(t) = 0\right) = P\left(\lim_{\substack{t\to 0 \\ t\in\mathbb{Q}^+}} B(t) = 0\right) = 1.$$

Since $\mathbb{Q}^+$ is dense in $(0, \infty)$, we conclude that $\lim_{t\to 0^+} X(t) = 0$ a.s. and the proof is complete. □

Time inversion gives an easy proof of the following fact about the long term behavior of Brownian motion.

Proposition 14.5 (Brownian SLLN)**.** *If $\{B(t) : t \geq 0\}$ is a standard Brownian motion, then*

$$\lim_{t\to\infty} \frac{B(t)}{t} = 0 \text{ a.s.}$$

Proof. Let $\{X(t) : t \geq 0\}$ be as in Proposition 14.4. Then

$$\lim_{t\to\infty} \frac{B(t)}{t} = \lim_{t\to\infty} X\left(\frac{1}{t}\right) = \lim_{s\to 0^+} X(s) = 0 \text{ a.s.}$$ □

Wiener's Theorem.

Now that we're a little more comfortable working with Brownian motion, it's time to prove that it actually exists. The main complication lies in the continuity requirement.

To illustrate this issue, observe that if $\{B(t) : t \geq 0\}$ is a Brownian motion and U is an independent random variable which is uniformly distributed on $[0, 1]$, then the process $\left\{\widetilde{B}(t) : t \geq 0\right\}$ defined by

$$\widetilde{B}(t) = \begin{cases} B(t), & t \neq U \\ 0, & t = U \end{cases}$$

has the same finite dimensional distributions as Brownian motion since $P(U \in S) = 0$ for any finite $S \subseteq [0, 1]$. However, $\widetilde{B}(t)$ is discontinuous whenever $B(U) \neq 0$ – that is, with probability one.

There are several ways to go about proving existence. We will pursue a fairly straightforward approach due to Paul Lévy. The basic idea is to construct standard Brownian motion on $[0, 1]$ as a uniform limit of continuous functions having the right finite dimensional distributions on sets of dyadic rationals and then patch together independent copies to obtain Brownian motion on $[0, \infty)$.

Let $\mathcal{D}_n = \left\{ \frac{k}{2^n} : k = 0, 1, ..., 2^n \right\}$ for $n = 0, 1, ...$, and set $\mathcal{D} = \bigcup_{n=0}^{\infty} \mathcal{D}_n$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space on which a collection $\{Z_d : d \in \mathcal{D}\}$ of independent standard normals can be defined. (Kolmogorov's extension theorem gives one such space.)

For each $n \in \mathbb{N}_0$, we define random variables $B(d)$, $d \in \mathcal{D}_n$ so that

(1) For all $q < r \leq s < t$ in $\mathcal{D}_n$, $B(t) - B(s) \sim \mathcal{N}(0, t-s)$ and $B(r) - B(q) \sim \mathcal{N}(0, r-q)$ are independent.

(2) The collections $\{B(d) : d \in \mathcal{D}_n\}$ and $\{Z_t : t \in \mathcal{D} \setminus \mathcal{D}_n\}$ are independent.

We begin by taking $B(0) = 0$ and $B(1) = Z_1$. Now suppose for the sake of induction that $\{B(d) : d \in \mathcal{D}_{n-1}\}$ satisfies (1) and (2).

We then define $B(d)$ for $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$ by

$$B(d) = \frac{B\left(d - 2^{-n}\right) + B\left(d + 2^{-n}\right)}{2} + \frac{Z_d}{2^{\frac{n+1}{2}}}.$$

Since the first term is the average of B at points in $\mathcal{D}_{n-1}$ and the $Z_d's$ are independent, the inductive hypothesis shows that $B(d)$ is the sum of two random variables which are independent of $\{Z_d : t \in \mathcal{D} \setminus \mathcal{D}_n\}$, so Property (2) is satisfied.

The inductive hypothesis also shows that $\frac{1}{2}\left[B\left(d + 2^{-n}\right) - B\left(d - 2^{-n}\right)\right]$ and $2^{-\frac{n+1}{2}} Z_d$ are independent normals having mean 0 and variance $2^{-(n+1)}$, so their sum

$$\frac{1}{2}\left[B\left(d + 2^{-n}\right) - B\left(d - 2^{-n}\right)\right] + 2^{-\frac{n+1}{2}} Z_d$$
$$= \frac{B\left(d + 2^{-n}\right) + B\left(d - 2^{-n}\right)}{2} - B\left(d - 2^{-n}\right) + \frac{Z_d}{2^{\frac{n+1}{2}}} = B(d) - B\left(d - 2^{-n}\right)$$

and their difference

$$\frac{1}{2}\left[B\left(d + 2^{-n}\right) - B\left(d - 2^{-n}\right)\right] - 2^{-\frac{n+1}{2}} Z_d$$
$$= B\left(d + 2^{-n}\right) - \frac{B\left(d + 2^{-n}\right) + B\left(d - 2^{-n}\right)}{2} - 2^{-\frac{n+1}{2}} Z_d = B\left(d + 2^{-n}\right) - B(d)$$

are independent and normally distributed with mean 0 and variance 2^{-n} (as proved in the homework).

Since the vector of increments $\left(B(d) - B\left(d - 2^{-n}\right) : d \in \mathcal{D}_n \setminus \{0\}\right)$ is multinormal (as its coordinates are linear combinations of independent normals), it has independent entries if they are pairwise independent.

We have already seen that $B\left(d + 2^{-n}\right) - B(d)$ and $B(d) - B\left(d - 2^{-n}\right)$ are independent for $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$.

It remains to consider the case where the two intervals lie on opposite sides of some $d \in \mathcal{D}_{n-1}$. For such pairs, choose $d \in \mathcal{D}_j$ with j minimal so that the intervals lie in $[d - 2^{-j}, d]$ and $[d, d + 2^{-j}]$, respectively.

The inductive hypothesis ensures that $B(d) - B\left(d - 2^{-j}\right)$ and $B\left(d + 2^{-j}\right) - B(d)$ are independent.

Since the increments in question are constructed from $B(d) - B\left(d - 2^{-j}\right)$ and $B\left(d + 2^{-j}\right) - B(d)$, respectively, using disjoint subsets of the $\{Z_t : t \in \mathcal{D}_n\}$, the desired independence follows.

As the increments $\left\{B(d) - B\left(d - 2^{-n}\right) : d \in \mathcal{D}_n \setminus \{0\}\right\}$ are i.i.d. $\mathcal{N}(0, 2^{-n})$, Property (1) is verified and the inductive step is complete.

At this point, we have defined $B(t)$ on the set $\mathcal{D}$ of dyadic rationals so that the finite dimensional distributions are as desired.

The next step is to extend the definition to all of $[0,1]$ by interpolation.

Let

$$F_0(t) = \begin{cases} Z_1, & t = 1 \\ 0, & t = 0 \\ \text{linear}, & \text{in between} \end{cases},$$

and, for $n \geq 1$,

$$F_n(t) = \begin{cases} 2^{-\frac{n+1}{2}} Z_t, & t \in \mathcal{D}_n \setminus \mathcal{D}_{n-1} \\ 0, & t = \mathcal{D}_{n-1} \\ \text{linear}, & \text{between consecutive points in } \mathcal{D}_n \end{cases}.$$

The F_i's are continuous, and we claim that for all $n \in \mathbb{N}_0$, $d \in \mathcal{D}_n$,

$$B(d) = \sum_{i=0}^{n} F_i(d) = \sum_{i=0}^{\infty} F_i(d).$$

(We only need to justify the first equality since $d \in \mathcal{D}_n$ implies $F_i(d) = 0$ for $i > n$.)

The proof is a simple induction argument:

It holds for $n = 0$ by construction. Assume that it holds for $n-1$.

We need to verify that $B(d) = \sum_{i=0}^{n} F_i(d)$ for $d \in \mathcal{D}_n \setminus \mathcal{D}_{n-1}$.

Since F_i is linear on $[d - 2^{-n}, d + 2^{-n}]$ for $0 \leq i \leq n$, we have

$$\sum_{i=0}^{n-1} F_i(d) = \sum_{i=0}^{n-1} \frac{F_i\left(d - 2^{-n}\right) + F_i\left(d + 2^{-n}\right)}{2} = \frac{B\left(d - 2^{-n}\right) + B\left(d + 2^{-n}\right)}{2}.$$

The claim follows from the definition of $B(d)$ since $F_n(d) = 2^{-\frac{n+1}{2}} Z_t$.

We now need to show that the sum $\sum_{i=0}^{\infty} F_i(t)$ is uniformly convergent.

To this end, we observe that the Z_d's are standard normal, so

$$P\left(|Z_d| \geq u\right) = \frac{2}{\sqrt{2\pi}} \int_u^{\infty} e^{-\frac{x^2}{2}} dx \leq \frac{2}{\sqrt{2\pi}} \int_u^{\infty} \frac{x}{u} e^{-\frac{x^2}{2}} dx = \frac{2}{u\sqrt{2\pi}} \int_{\frac{1}{2}u^2}^{\infty} e^{-y} dy = \frac{2}{u\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$

for all $u > 0$.

In particular,

$$P\left(|Z_d| \geq \sqrt{4n\log(2)}\right) \leq \frac{2}{\sqrt{8\pi n \log(2)}} e^{-2n\log(2)} \leq 2^{-2n}$$

for all $n \in \mathbb{N}$.

Consequently,

$$\sum_{n=0}^{\infty} P\left(|Z_d| \geq \sqrt{4n\log(2)} \text{ for some } d \in \mathcal{D}_n\right) \leq \sum_{n=0}^{\infty} \sum_{d \in \mathcal{D}_n} P\left(|Z_d| \geq \sqrt{4n\log(2)}\right) \leq \sum_{n=0}^{\infty} \frac{2^n + 1}{2^{2n}} < \infty.$$

Therefore, the first Borel-Cantelli lemma shows that there is a (random, but a.s. finite) N such that $n \geq N$ and $d \in \mathcal{D}_n$ implies$|Z_d| < \sqrt{4n\log(2)}$, hence

$$\|F_n\|_\infty \leq 2^{-\frac{n+1}{2}}\sqrt{4n\log(2)}$$

whenever $n \geq N$.

It follows that, almost surely, $B(t) = \sum_{i=0}^{\infty} F_i(t)$ is a uniform limit of continuous functions and thus is continuous.

To see that $\{B(t) : t \in [0,1]\}$ is a standard Brownian motion on $[0,1]$, we just have to verify that it has the right finite dimensional distributions.

But this is an easy consequence of continuity since we have already established the claim on the dense set $\mathcal{D} \subseteq [0,1]$.

Indeed, let $0 \leq t_1 < ... < t_n \leq 1$. Then there exist $t_{1,k} \leq ... \leq t_{n,k}$ in $\mathcal{D}$ with $t_{i,k} \to t_i$ as $k \to \infty$, so continuity gives

$$\lim_{k\to\infty} B(t_{i+1k}) - B(t_{i,k}) = B(t_{i+1}) - B(t_i) \text{ a.s.}$$

for $i = 1, ..., n-1$.

As

$$\lim_{k\to\infty} E\left[B(t_{i+1,k}) - B(t_{i,k})\right] = 0$$

and

$$\begin{aligned}\lim_{k\to\infty} &\operatorname{Cov}\left(B(t_{i+1,k}) - B(t_{i,k}), B(t_{j+1,k}) - B(t_{j,k})\right) \\ &= \lim_{k\to\infty} 1\{i=j\}\left(t_{i+1,k} - t_{i,k}\right) = 1\{i=j\}\left(t_{i+1} - t_i\right),\end{aligned}$$

we see that $(B(t_2) - B(t_1), ..., B(t_n) - B(t_{n-1}))$ is multivariate normal with mean $(0, ..., 0)^T$ and covariance $\Sigma = \operatorname{Diag}(t_2 - t_1, \ldots, t_n - t_{n-1})$.

(An easy characteristic functions argument shows that the limit is indeed Gaussian.)

Thus we have constructed a continuous process $B : [0,1] \to \mathbb{R}$ with the same finite dimensional distributions as standard Brownian motion on $[0,1]$.

To extend this to all positive times, take a sequence $B_0, B_1, ...$ of independent $C([0,1], \mathbb{R})$-valued random variables having the distribution of this process and glue them together end to end.

Specifically, define $\{B(t) : t \geq 0\}$ by

$$B(t) = B_{\lfloor t \rfloor}(t - \lfloor t \rfloor) + \sum_{i=0}^{\lfloor t \rfloor - 1} B_i(1).$$

One readily checks that B is an a.s. continuous function from $[0,\infty)$ to $\mathbb{R}$ which has the right finite dimensional distributions.

15. Sample Path Properties

Having shown that Brownian motion exists and established some simple invariance principles, we now turn our attention to a few basic properties of Brownian paths.

We begin with an easy result which demonstrates some of the peculiarities of Brownian motion.

Theorem 15.1. *Almost surely, Brownian motion is not monotone on any interval $[a,b]$, $0 \leq a < b < \infty$.*

Proof. Fix a nondegenerate interval $[a,b]$. We can partition $[a,b]$ into n subintervals $[a_{i-1}, a_i]$ by picking $a = a_0 < a_1 < ... < a_n = b$.

If B_t is monotone on $[a,b]$, then each of the increments $B(a_i) - B(a_{i-1})$ has the same sign.

Since the increments are independent mean zero normals, the probability of this occurring is $2 \cdot 2^{-n}$.

Thus for any $0 \leq a < b < \infty$, $n \in \mathbb{N}$, $P(B_t \text{ is monotone on } [a,b]) \leq 2^{-(n-1)}$.

Sending $n \to \infty$ shows that $\{B_t\}_{t \geq 0}$ is not monotone on any particular interval almost surely.

Since there are countably many intervals with rational endpoints and every nondegenerate interval contains some such interval, the result follows. □

Our next goal is to show that Brownian motion is a.s. locally γ-Hölder continuous for $\gamma < \frac{1}{2}$.
The key to doing so is the following result which deduces local Hölder continuity from moment bounds.

Theorem 15.2 (Kolmogorov's Continuity Theorem)**.** *Let $X(t)$ be a stochastic process with a.s. continuous sample paths such that*

$$E\left[|X(t) - X(s)|^{\beta}\right] \leq C\,|t-s|^{1+\alpha}$$

for some constants $\alpha, \beta, C > 0$ and all times $s, t \geq 0$.

Then for each $0 \leq \gamma < \frac{\alpha}{\beta}$, $T > 0$, and almost every ω, there exists a constant $K = K(\omega, \gamma, T)$ such that

$$|X(t,\omega) - X(s,\omega)| \leq K\,|t-s|^{\gamma} \text{ for all } 0 \leq s, t \leq T.$$

That is, the sample paths $t \mapsto X(t,\omega)$ are a.s. γ-Hölder continuous on $[0,T]$.

Proof. By time scaling, it suffices to consider the case $T = 1$.

Let $\mathcal{D}$ denote the dyadic rationals in $[0,1]$ as in the proof of Wiener's theorem, and let $\gamma \in \left[0, \frac{\alpha}{\beta}\right)$.

Chebychev's inequality and our assumption give

$$\begin{aligned}
P\left(\max_{1 \leq k \leq 2^n} \left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right| \geq 2^{-\gamma n}\right) &= P\left(\bigcup_{k=1}^{2^n} \left\{\left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right| \geq 2^{-\gamma n}\right\}\right) \\
&\leq \sum_{k=1}^{2^n} P\left(\left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right| \geq 2^{-\gamma n}\right) \\
&\leq \sum_{k=1}^{2^n} \frac{E\left[\left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right|^{\beta}\right]}{2^{-\beta\gamma n}} \\
&\leq 2^{\beta\gamma n} C \sum_{k=1}^{2^n} \left(\frac{1}{2^n}\right)^{\alpha+1} = 2^{(\beta\gamma - \alpha)n} C.
\end{aligned}$$

Since $\beta\gamma < \alpha$, we see that

$$\sum_{n=1}^{\infty} P\left(\max_{1\le k\le 2^n}\left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right| \ge 2^{-\gamma n}\right) \le C\sum_{n=1}^{\infty} 2^{(\beta\gamma-\alpha)n} < \infty,$$

so the first Borel-Cantelli lemma implies the existence of a set Ω^* with $P(\Omega^*) = 1$ such that for every $\omega \in \Omega^*$, there is an $N(\omega) \in \mathbb{N}$ with

$$\max_{1\le k\le 2^n}\left|X\left(\frac{k}{2^n},\omega\right) - X\left(\frac{k-1}{2^n},\omega\right)\right| < 2^{-\gamma n}$$

for all $n \ge N(\omega)$.

Choosing $K_0 = K_0(\omega,\gamma)$ large enough gives

$$\max_{1\le k\le 2^n}\left|X\left(\frac{k}{2^n}\right) - X\left(\frac{k-1}{2^n}\right)\right| < 2^{-\gamma n}K_0$$

for all $n \in \mathbb{N}$ on Ω_*.

Now suppose that $s,t \in \mathcal{D}$ with $s < t$, and let $n \in \mathbb{N}$ be such that $2^{-n} \le t - s < 2^{-(n-1)}$.

We can write

$$s = \frac{i}{2^n} - \frac{1}{2^{p_1}} - ... - \frac{1}{2^{p_k}}, \quad t = \frac{j}{2^n} + \frac{1}{2^{q_1}} + ... + \frac{1}{2^{q_\ell}}$$

with $n < p_1 < ... < p_k$, $n < q_1 < ... < q_\ell$ and $i \le j$.

Moreover, since $\frac{j}{2^n} - \frac{i}{2^n} \le t - s < 2^{-(n-1)}$, we must have $|i-j| \le 1$, hence

$$\left|X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right)\right| < 2^{-\gamma n}K_0.$$

Also,

$$\left|X\left(\frac{i}{2^n} - \frac{1}{2^{p_1}} - ... - \frac{1}{2^{p_r}}\right) - X\left(\frac{i}{2^n} - \frac{1}{2^{p_1}} - ... - \frac{1}{2^{p_{r-1}}}\right)\right| \le 2^{-\gamma p_r}K_0$$

for $r = 1, ..., k$, so, since $n < p_1 < ... < p_k$,

$$\begin{aligned}\left|X(s) - X\left(\frac{i}{2^n}\right)\right| &\le K_0\sum_{r=1}^{k} 2^{-\gamma p_r} \le K_0\sum_{r=1}^{k} 2^{-\gamma(n+r)} \\ &\le 2^{-\gamma n}K_0\sum_{r=1}^{\infty} 2^{-\gamma r} \le 2^{-\gamma n}K_1\end{aligned}$$

where $K_1 = (1 \vee \sum_{r=1}^{\infty} 2^{-\gamma r})\, K_0$.

The exact same reasoning shows that $\left|X\left(\frac{j}{2^n}\right) - X(t)\right| \le 2^{-\gamma n}K_1$.

It follows that

$$\begin{aligned}|X(t) - X(s)| &\le \left|X(t) - X\left(\frac{j}{2^n}\right)\right| + \left|X\left(\frac{j}{2^n}\right) - X\left(\frac{i}{2^n}\right)\right| + \left|X\left(\frac{i}{2^n}\right) - X(s)\right| \\ &\le 2^{-\gamma n}K \le K|t-s|^{\gamma}, \quad K = 3K_1,\end{aligned}$$

hence the paths are a.s. γ-Hölder continuous on $\mathcal{D}$.

Because $\mathcal{D}$ is dense in $[0,1]$ and $t \mapsto X(t)$ is a.s. continuous, the Hölder condition holds a.s. for all $s,t \in [0,1]$. □

Applying the above result to Brownian motion yields

Theorem 15.3. *Let $\{B_t\}_{t\geq 0}$ be a standard Brownian motion, and let $T > 0$. Then with full probability, $t \mapsto B_t$ is γ-Hölder continuous on $[0,T]$ for all $\gamma < \frac{1}{2}$.*

Proof. The scaling relation (Proposition 14.3) with $a = t - s$ shows that

$$E\left[|B_t - B_s|^\beta\right] = E\left[|B_{t-s}|^\beta\right] = C_\beta |t-s|^{\frac{\beta}{2}}$$

where

$$C_\beta = E\left[|B_1|^\beta\right] = \frac{1}{\sqrt{2\pi}} \int |x|^\beta e^{-\frac{x^2}{2}} dx.$$

Thus when $\beta > 2$, Theorem 15.2 applies with $\alpha = \frac{\beta}{2} - 1$, so $t \mapsto B_t$ is γ-Hölder continuous for $0 \leq \gamma < \frac{\alpha}{\beta} = \frac{1}{2} - \frac{1}{\beta}$. The result follows by sending β to infinity. □

In fact, the preceding result is optimal in the sense that

Theorem 15.4. *$\{B_t\}_{t\in[0,1]}$ is not Hölder continuous for any exponent $\gamma \geq \frac{1}{2}$.*

Proof. (Homework) We wish to show that

$$\sup_{s,t\in[0,1]} \frac{|B_t - B_s|}{|t-s|^\gamma} = \infty \text{ a.s.}$$

whenever $\gamma \geq \frac{1}{2}$.

Since $x^{\gamma_1} \geq x^{\gamma_2}$ for $0 < \gamma_1 \leq \gamma_2$, $x \in [0,1]$, it suffices to establish the result for $\gamma = \frac{1}{2}$.

Define

$$K(\omega) := \sup_{0\leq s<t\leq 1} \frac{|B_t(\omega) - B_s(\omega)|}{\sqrt{t-s}},$$
$$K_{i,n}(\omega) := \sup_{\frac{i-1}{n}\leq s<t\leq \frac{i}{n}} \frac{|B_t(\omega) - B_s(\omega)|}{\sqrt{t-s}} \text{ for } n \in \mathbb{N}, i \in [n].$$

The independent increments property of Brownian motion shows that $K_{1,n}, ..., K_{n,n}$ are independent.

Moreover, the scaling relation shows that $\{X_t\}_{t\in[0,1]} =_d \{B_t\}_{t\in[0,1]}$ where $X_t := \sqrt{n}\left(B_{\frac{i+t}{n}} - B_{\frac{i}{n}}\right)$, hence

$$\begin{aligned} K_{i,n} &= \sup_{\frac{i-1}{n}\leq s<t\leq \frac{i}{n}} \frac{|B_t - B_s|}{\sqrt{t-s}} = \sup_{0\leq s<t\leq 1} \frac{\left|B_{\frac{t}{n}+\frac{i-1}{n}} - B_{\frac{s}{n}+\frac{i-1}{n}}\right|}{\sqrt{\left(\frac{t}{n}+\frac{i-1}{n}\right) - \left(\frac{s}{n}+\frac{i-1}{n}\right)}} \\ &= \sup_{0\leq s<t\leq 1} \frac{\sqrt{n}\left|B_{\frac{i+t}{n}-\frac{1}{n}} - B_{\frac{i+s}{n}-\frac{1}{n}}\right|}{\sqrt{t-s}} =_d \sup_{0\leq s<t\leq 1} \frac{\sqrt{n}\left|B_{\frac{i+t}{n}} - B_{\frac{i+s}{n}}\right|}{\sqrt{t-s}} \\ &= \sup_{0\leq s<t\leq 1} \frac{|X_t - X_s|}{\sqrt{t-s}} =_d \sup_{0\leq s<t\leq 1} \frac{|B_t - B_s|}{\sqrt{t-s}} = K. \end{aligned}$$

Because $K \geq \max_{1\leq i\leq n} K_{i,n}$ for all $n \in \mathbb{N}$ by construction, it follows that for any $M > 0$, we have

$$P(K \leq M) \leq P(K_{1,n} \leq M, ..., K_{n,n} \leq M) \leq P(K \leq M)^n.$$

Since $\frac{B_t - B_s}{\sqrt{t-s}} \sim \mathcal{N}(0,1)$ for all $0 \leq s < t \leq 1$, we have that $P(K > M) > 0$ for all $M > 0$.

Therefore, the preceding inequality implies that $P(K \leq M) = 0$ for all $M > 0$, hence $K = \infty$ a.s. □

One can actually show that Brownian motion is almost surely not γ-Hölder continuous at any point t (which is a stronger statement than not being γ-Hölder continuous over an interval) whenever $\gamma > \frac{1}{2}$.

This can be verified by appropriately modifying the proof of the following theorem.

Theorem 15.5. *With probability one, Brownian paths are not Lipschitz continuous at any point.*

Proof. Fix a constant $C \in (0, \infty)$ and let

$$A_n = \left\{ \omega : \text{there is an } s \in [0,1] \text{ such that } |B_t(\omega) - B_s(\omega)| \leq C|t-s| \text{ whenever } |t-s| \leq \frac{3}{n} \right\}.$$

For $1 \leq k \leq n-2$, let

$$Y_{n,k} = \max\left\{ \left| B\left(\frac{k}{n}\right) - B\left(\frac{k-1}{n}\right) \right|, \left| B\left(\frac{k+1}{n}\right) - B\left(\frac{k}{n}\right) \right|, \left| B\left(\frac{k+2}{n}\right) - B\left(\frac{k+1}{n}\right) \right| \right\},$$

and set

$$B_n = \left\{ \min_{1 \leq k \leq n-2} Y_{n,k} \leq \frac{6C}{n} \right\}.$$

We will show that $A_n \subseteq B_n$. To this end, suppose that $\omega \in A_n$. Then there is an $s \in [0,1]$ with $|B_t(\omega) - B_s(\omega)| \leq C|t-s|$ whenever $|t-s| \leq \frac{3}{n}$.

Let $k = \max\{j \in [n-2] : \frac{j-1}{n} \leq s\}$. Then $\left|\frac{j}{n} - s\right| \leq \frac{3}{n}$ for $j = k-1, k, k+1, k+2$, so

$$\begin{aligned} \left| B\left(\frac{\ell}{n}, \omega\right) - B\left(\frac{\ell-1}{n}, \omega\right) \right| &\leq \left| B\left(\frac{\ell}{n}, \omega\right) - B(s, \omega) \right| + \left| B(s, \omega) - B\left(\frac{\ell-1}{n}, \omega\right) \right| \\ &\leq C\left(\left| \frac{\ell}{n} - s \right| + \left| s - \frac{\ell-1}{n} \right| \right) \leq C\left(\frac{3}{n} + \frac{3}{n} \right) = \frac{6C}{n} \end{aligned}$$

for $\ell = k, k+1, k+2$, hence $Y_{n,k}(\omega) \leq \frac{6C}{n}$.

Now for any $k = 1, ..., n-2$,

$$P\left(Y_{n,k} \leq \frac{6C}{n}\right) = \prod_{j=k}^{k+2} P\left(\left| B\left(\frac{j}{n}\right) - B\left(\frac{j-1}{n}\right) \right| \leq \frac{6C}{n} \right) = P\left(\left| B\left(\frac{1}{n}\right) \right| \leq \frac{6C}{n} \right)^3$$

since Brownian motion has stationary independent increments.

Applying the scaling relation with $a = \frac{1}{n}$ and then using $B(1) \sim \mathcal{N}(0,1)$ yields

$$\begin{aligned} P\left(Y_{n,k} \leq \frac{6C}{n}\right) &= P\left(\left| B\left(\frac{1}{n}\right) \right| \leq \frac{6C}{n} \right)^3 \\ &= P\left(\left| \frac{1}{\sqrt{n}} B(1) \right| \leq \frac{6C}{n} \right)^3 = P\left(|B(1)| \leq \frac{6C}{\sqrt{n}} \right)^3 \\ &= \left(\frac{2}{\sqrt{2\pi}} \int_0^{\frac{6C}{\sqrt{n}}} e^{-\frac{x^2}{2}} dx \right)^3 \leq \left(\frac{12C}{\sqrt{2\pi n}} \right)^3 \end{aligned}$$

where the final inequality is because $e^{-\frac{x^2}{2}} \leq 1$ for $x \geq 0$.

Combining our observations shows that

$$P(A_n) \leq P(B_n) = P\left(\bigcup_{k=1}^{n-2} \left\{ Y_{n,k} \leq \frac{6C}{n} \right\} \right) \leq (n-2)\left(\frac{12C}{\sqrt{2\pi n}} \right)^3 \to 0$$

as $n \to \infty$.

Since $A_1 \subseteq A_2 \subseteq \ldots$, $P(A_n)$ is increasing in n, so we conclude that $P(A_n) = 0$ for all n.
As C was arbitrary, we have shown that $t \mapsto B_t$ is not Lipschitz at any point in $[0,1]$ with full probability.
The exact same argument shows that

$$A_n^m = \left\{ \text{there exists an } s \in [m, m+1] \text{ such that } |B_t - B_s| \leq C|t-s| \text{ whenever } |t-s| \leq \frac{3}{n} \right\}$$

has probability zero for all $m, n \in \mathbb{N}$, and the result follows from countable subadditivity. □

Note that if f is differentiable at t, then there is a $\delta > 0$ such that $|f(t) - f(s)| \leq (|f'(t)| + 1)|t-s|$ whenever $|t-s| < \delta$.
In other words, a function is Lipschitz at every point at which it is differentiable. Taking the contrapositive and appealing to Theorem 15.5 yields

Theorem 15.6 (Paley, Wiener, Zygmund)**.** *Brownian motion is a.s. nowhere differentiable.*

In fact, if we define the *upper and lower right derivatives* of a function by

$$D^* f(t) = \limsup_{h \searrow 0} \frac{f(t+h) - f(t)}{h}, \qquad D_* f(t) = \liminf_{h \searrow 0} \frac{f(t+h) - f(t)}{h},$$

then one can show that

$$P\left(\bigcap_{t \geq 0} \left(\{D_* B(t) = -\infty\} \cup \{D^* B(t) = \infty\} \right) \right) = 1.$$

Indeed, if there is some $t_0 \in [0,1]$ such that $-\infty < D_* B(t_0) \leq D^* B(t_0) < \infty$, then

$$\limsup_{h \searrow 0} \frac{|B(t_0+h) - B(t_0)|}{h} < \infty.$$

Since $B(t)$ is a.s. continuous (and thus uniformly bounded on compact sets), this means that

$$\sup_{h \in [0,1]} \frac{|B(t_0+h) - B(t_0)|}{h} \leq M$$

for some a.s. finite M, so that $B(t)$ is "Lipschitz from the right" at t_0.
An argument similar to the proof of Theorem 15.5 shows that for any $C \in (0, \infty)$, the event

$$\{\omega : \text{there is some } s \in [0,1] \text{ such that } |B(s+h, \omega) - B(s, \omega)| \leq Ch \text{ for all } h \in [0,1]\}$$

has probability zero.

The Hölder continuity results discussed here are not the most precise answer to the question "How continuous is Brownian motion?"

In 1937, Paul Lévy proved the following theorem regarding Brownian motion's modulus of continuity:

Theorem 15.7. *Almost surely,*

$$\limsup_{h\to 0} \sup_{0\le t\le 1-h} \frac{|B(t+h)-B(t)|}{\sqrt{2h\log\left(\frac{1}{h}\right)}} = 1.$$

Though the proof of Theorem 15.7 is not difficult, we leave it to independent pursuit so that we may explore more topics.

Before moving on, however, we note that the arguments in Theorems 15.2 and 15.3 provide an alternative path to proving the existence of Brownian motion.

Namely, one uses the extension theorem to get a process on the dyadic rationals with appropriate finite dimensional distributions.
One then argues as before to show that this process satisfies a Hölder condition.
As $\mathcal{D}$ is dense in $[0,1]$, there is a unique continuous extension to $[0,1]$.
By an easy limiting argument, this continuous version has the correct finite dimensional distributions.

Though I prefer the approach we have taken since it is more concrete, the above argument has the advantage that it also works for other continuous processes specified by their finite dimensional distributions.

Another benefit is that the starting point is a part of the finite dimensional distributions, so as with Markov chains, we have one process $B_t(\omega) = \omega(t)$ and a family of probability measures $\{P_x\}_{x\in\mathbb{R}}$ such that under P_x, B_t is a Brownian motion with $P_x(B_0 = x) = 1$.

Measures corresponding to general initial distributions then arise as mixtures: $P_\mu(A) = \int P_x(A)d\mu(x)$.

16. Markov Properties

Our next object is to establish that Brownian motion is a (strong) Markov process and explore some simple consequences of this fact.

In order to discuss Markov properties, it is necessary to have a referent filtration.
The definitions in continuous time are as one would expect from the discrete case:

Given a probability space $(\Omega, \mathcal{F}, P)$, a filtration is a family of sub-σ-algebras $\{\mathcal{F}(t) : t \geq 0\}$ with $\mathcal{F}(s) \subseteq \mathcal{F}(t) \subseteq \mathcal{F}$ for all $0 \leq s \leq t$.
A process $\{X(t) : t \geq 0\}$ on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}(t)\}_{t\geq 0}, P)$ is adapted if $X(t)$ is $\mathcal{F}(t)$-measurable for all $t \geq 0$.

If $\{B(t) : t \geq 0\}$ is a standard Brownian motion, the obvious candidate is $\mathcal{F}^0(t) = \sigma(B(s) : 0 \leq s \leq t)$.
$B(t)$ is adapted to this filtration by construction, and Proposition 14.2 shows that $\{B(t+s) - B(s) : t \geq 0\}$ is independent of $\mathcal{F}^0(s)$.

It turns out that a more convenient filtration is given by $\mathcal{F}^+(s) = \bigcap_{t>s} \mathcal{F}^0(t)$.
$\{\mathcal{F}^+(t) : t \geq 0\}$ is clearly a filtration with $\mathcal{F}^0(t) \subseteq \mathcal{F}^+(t)$.
Intuitively, $\mathcal{F}^0(t)$ contains all of the information up to time t, while $\mathcal{F}^+(t)$ contains, in addition, an infinitesimal peak into the future: $A \in \mathcal{F}^+(t)$ if and only if $A \in \mathcal{F}^0(t+\varepsilon)$ for all $\varepsilon > 0$.
One advantage of working with the latter is that it is *right-continuous*:

$$\bigcap_{t>s} \mathcal{F}^+(t) = \bigcap_{t>s} \left(\bigcap_{u>t} \mathcal{F}^0(u) \right) = \bigcap_{u>s} \mathcal{F}^0(u) = \mathcal{F}^+(s).$$

It is with respect to this right-continuous filtration that we state

Theorem 16.1 (Markov Property)**.** *For every $s \geq 0$, $\{B_{t+s} - B_s : t \geq 0\}$ is independent of $\mathcal{F}^+(s)$.*

Proof. Let s_n be a strictly decreasing sequence with $\lim_{n\to\infty} s_n = s$.
Let $A \in \mathcal{F}^+(s)$, let $t_1, ..., t_m \geq 0$, and let $F : \mathbb{R}^m \to \mathbb{R}$ be bounded and continuous.
Since $A \in \mathcal{F}^+(s)$ implies $A \in \mathcal{F}^0(s_n)$ for all n, Proposition 14.2 shows that

$$E\left[F\left(B_{t_1+s_n} - B_{s_n}, ..., B_{t_m+s_n} - B_{s_n}\right) 1_A\right] = E\left[F\left(B_{t_1+s_n} - B_{s_n}, ..., B_{t_m+s_n} - B_{s_n}\right)\right] P(A)$$

for all n.
Sample path continuity implies

$$\lim_{n\to\infty} B_{t+s_n} - B_{s_n} = B_{t+s} - B_s \text{ a.s.}$$

for all $t \geq 0$, so two applications of the dominated convergence theorem give

$$\begin{aligned} E\left[F\left(B_{t_1+s} - B_s, ..., B_{t_m+s} - B_s\right) 1_A\right] &= \lim_{n\to\infty} E\left[F\left(B_{t_1+s_n} - B_{s_n}, ..., B_{t_m+s_n} - B_{s_n}\right) 1_A\right] \\ &= \lim_{n\to\infty} E\left[F\left(B_{t_1+s_n} - B_{s_n}, ..., B_{t_m+s_n} - B_{s_n}\right)\right] P(A) \\ &= E\left[F\left(B_{t_1+s} - B_s, ..., B_{t_m+s} - B_s\right)\right] P(A). \end{aligned}$$

As $A \in \mathcal{F}^+(s)$, $t_1, ..., t_m \geq 0$, and $F \in C_b(\mathbb{R}^m, \mathbb{R})$ were arbitrary, the claim follows. □

In other words, conditional on $\mathcal{F}^+(s)$, the process $\{B(t+s) : t \geq 0\}$ is a Brownian motion started at $B(s)$.

When specialized to the *germ σ-algebra* $\mathcal{F}^+(0)$, Theorem 16.1 yields

Theorem 16.2 (Blumenthal's $0-1$ Law)**.** *The σ-algebra $\mathcal{F}^+(0)$ is trivial.*

Proof. Let $A \in \mathcal{F}^+(0)$. Then $A \in \sigma(B_t : t \geq 0)$, so Theorem 16.1 implies that A is independent of $\mathcal{F}^+(0)$. In particular, A is independent of itself, so $P(A) = P(A \cap A) = P(A)P(A)$, hence $P(A) \in \{0, 1\}$. □

An interesting consequence of Theorem 16.2 is that standard Brownian motion has positive values, negative values, and zeros in any time interval $(0, \varepsilon)$ almost surely.

Theorem 16.3. *Suppose that $\{B(t) : t \geq 0\}$ is a standard Brownian motion.*
Define $\tau = \inf\{t > 0 : B(t) > 0\}$ and $T = \inf\{t > 0 : B(t) = 0\}$. Then $P(\tau = 0) = P(T = 0) = 1$.

Proof. Let $A_k = \left\{\tau < \frac{1}{k}\right\} = \left\{B_t > 0 \text{ for some } 0 < t < \frac{1}{k}\right\}$. Then

$$\{\tau = 0\} = \bigcap_{k=n}^{\infty} A_k \in \mathcal{F}^0\left(\frac{1}{n}\right)$$

for all $n \in \mathbb{N}$, so $\{\tau = 0\} \in \mathcal{F}^+(0)$, hence $P(\tau = 0) \in \{0, 1\}$.
To see that $P(\tau = 0) > 0$, observe that for all $n \in \mathbb{N}$,

$$P(A_n) \geq P\left(B_{\frac{1}{2n}} > 0\right) = \frac{1}{2}$$

since $B_{\frac{1}{2n}} \sim \mathcal{N}\left(0, \frac{1}{2n}\right)$, so

$$P(\tau = 0) = P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n) \geq \frac{1}{2}.$$

The exact same argument (or the fact that $\{B_t\}_{t\geq 0} =_d \{-B_t\}_{t\geq 0}$) shows that $\sigma = \inf\{t > 0 : B(t) < 0\} = 0$ a.s. Since B is a.s. continuous, the intermediate value theorem implies $P(T = 0) = 1$. □

Because of time inversion, results concerning the $t \to 0$ behavior of Brownian motion can be used to understand the behavior as $t \to \infty$.

Define $\mathcal{G}(t) = \sigma(B(s) : s \geq t)$ and let $\mathcal{T} = \bigcap_{t\geq 0} \mathcal{G}(t)$ be the σ-algebra of tail events.
Since $\mathcal{T}$ is mapped onto the germ σ-algebra under time inversion, we have

Theorem 16.4. *The tail σ-algebra for standard Brownian motion is trivial.*

A typical application of Theorem 16.4 is given by

Theorem 16.5. *If B_t is a standard Brownian motion, then with probability one,*

$$\limsup_{t\to\infty} \frac{B_t}{\sqrt{t}} = \infty, \quad \liminf_{t\to\infty} \frac{B_t}{\sqrt{t}} = -\infty.$$

Proof. It suffices to prove the first statement as the second then follows from symmetry.

Let $K \in (0,\infty)$, and recall the inequality

$$P\left(A_n \text{ i.o.}\right) = P\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_n\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_n\right)$$
$$\geq \lim_{n\to\infty}\left(\sup_{m\geq n} P\left(A_n\right)\right) = \limsup_{n\to\infty} P(A_n).$$

It follows from Brownian scaling and the fact that $B_1 \sim \mathcal{N}(0,1)$ that

$$P\left(\frac{B_n}{\sqrt{n}} \geq K \text{ i.o}\right) \geq \limsup_{n\to\infty} P\left(\frac{B_n}{\sqrt{n}} \geq K\right) = P(B_1 \geq K) > 0.$$

Since $\left\{\frac{B_n}{\sqrt{n}} \geq K \text{ i.o}\right\} \in \mathcal{T}$, Theorem 16.4 implies $P\left(\frac{B_n}{\sqrt{n}} \geq K \text{ i.o}\right) = 1$ and the result follows since K was arbitrary. □

Writing P_x for the probability measure which makes $\{B(t) : t \geq 0\}$ a Brownian motion started at x, the previous theorem, along with a.s. continuity and translation invariance shows that

Theorem 16.6. *Let B_t be a Brownian motion and let $A = \bigcap_{n\in\mathbb{N}}\{B_t = 0 \text{ for some } t \geq n\}$. Then $P_x(A) = 1$ for all x.*

Our final application of the Markov property concerns the local and global maxima of Brownian motion.

Theorem 16.7. *For a Brownian motion $\{B(t) : 0 \leq t \leq 1\}$, almost surely*

(1) *Every local maximum is a strict local maximum.*
(2) *The set of times where local maxima are attained is countable and dense.*
(3) *The global maximum is attained at a unique time.*

Proof. We begin by showing that for any closed time intervals with disjoint interiors, the maxima of Brownian motion over each are almost surely distinct:

Let $0 \leq a_1 < b_1 \leq a_2 < b_2 \leq 1$, and let $m_1(\omega) = \max_{t\in[a_1,b_1]} B(t,\omega)$, $m_2(\omega) = \max_{t\in[a_2,b_2]} B(t,\omega)$. (Since $B(t)$ is almost surely continuous and the intervals in question are compact, m_1 and m_2 are well-defined with full probability.)

Theorem 16.3 and the Markov property show that for any $\varepsilon > 0$, there almost surely exists some $a_2 < t < a_2 + \varepsilon$ with $B(t) - B(a_2) > 0$, hence $B(a_2) < m_2$ a.s.

Thus for almost every ω, there is a smallest $n = n(\omega) \in \mathbb{N}$ such that $m_2(\omega)$ is the maximum over $\left[a_2 + \frac{1}{n}, b_2\right]$. By considering each of these intervals, it suffices to assume in the proof that $b_1 < a_2$.

Applying the Markov property at time b_1, we have that $B(a_2) - B(b_1)$ is independent of $m_1 - B(b_1)$.

Applying the Markov property at time a_2 shows that $m_2 - B(a_2)$ is independent of each of these increments.

Now we can write the event $m_1 = m_2$ as

$$B(a_2) - B(b_1) = \left(m_1 - B(b_1)\right) - \left(m_2 - B(a_2)\right).$$

Conditioning on the values of $m_1 - B(b_1)$ and $m_2 - B(a_2)$, we see that the right hand side is constant and the left is a continuous random variable, so this event has probability 0.

(1) We just proved that all non-overlapping pairs of nondegenerate closed intervals with rational endpoints have distinct maxima. If Brownian motion had a non-strict local maximum, there would be two such intervals having the same maximum.

(2) The maximum over any nondegenerate closed interval with rational endpoints is a.s. not attained at the endpoints, so there is a local maximum between any two rational numbers, hence the set of local maxima is dense. Since every local maximum is a strict local maximum, there are no more maxima than there are intervals with rational endpoints, which is countable.

(3) For any $q \in \mathbb{Q} \cap [0,1]$, the maxima over $[0,q]$ and $[q,1]$ are distinct. If the global maximum was obtained at two points $t_1 < t_2$, there would be a rational $q \in (t_1, t_2)$ such that the maxima over $[0,q]$ and $[q,1]$ agree.

□

Strong Markov Property.

To state the strong Markov property, we need to understand stopping times in the continuous setting. Though there are many similarities, beware that not all results for discrete time carry over.

Definition. Given a filtered probability space $\Big(\Omega, \mathcal{F}, \{\mathcal{F}(t)\}_{t\geq 0}, P\Big)$, we say that a $[0,\infty]$-valued random variable T is a *stopping time* if $\{T \leq t\} \in \mathcal{F}(t)$ for all $t \geq 0$.

One of the reasons that it is so convenient to work with right-continuous filtrations is

Theorem 16.8. *If $\{\mathcal{F}(t)\}_{t\geq 0}$ is right-continuous, then a $[0,\infty]$-valued random variable T is a stopping time if $\{T < t\} \in \mathcal{F}(t)$ for all $t \geq 0$.*

Proof. For such a T, we have

$$\{T \leq t\} = \bigcap_{k=1}^{\infty} \Big\{T < t + \frac{1}{k}\Big\} \in \bigcap_{n=1}^{\infty} \mathcal{F}\Big(t + \frac{1}{n}\Big) = \mathcal{F}(t). \qquad \square$$

Thus for right-continuous filtrations, we recover our definition from the discrete setting:

Corollary 16.1. *If $\{\mathcal{F}(t)\}_{t\geq 0}$ is right-continuous, then a $[0,\infty]$-valued random variable T is a stopping time if $\{T = t\} \in \mathcal{F}(t)$ for all $t \geq 0$.*

Proof.

$$\{T = t\} = \{T \leq t\} \setminus \{T < t\} \in \mathcal{F}(t). \qquad \square$$

To illustrate the utility of Theorem 16.8 (and get some practice working with stopping times), observe that

- If T_n is an increasing sequence of stopping times with respect to $\{\mathcal{F}(t)\}_{t\geq 0}$ and $T_n \nearrow T$, then T is also a stopping time with respect to $\{\mathcal{F}(t)\}_{t\geq 0}$:
$$\{T \leq t\} = \bigcap_{n=1}^{\infty} \{T_n \leq t\} \in \mathcal{F}(t).$$

- If T_n is a decreasing sequence of stopping times with $T_n \searrow T$, then so is T provided that the filtration is right-continuous:
$$\{T < t\} = \bigcup_{n=1}^{\infty} \{T_n < t\} \in \mathcal{F}(t).$$

- If K is a closed set, then $T_K = \inf\{t \geq 0 : B_t \in K\}$ is a stopping time with respect to $\{\mathcal{F}^0(t)\}$ (and thus $\mathcal{F}^+(t)$):
Let D be a countable dense subset of K. Then
$$\{T_K \leq t\} = \bigcap_{n=1}^{\infty} \bigcup_{s\in\mathbb{Q}\cap(0,t)} \bigcup_{x\in D} \left\{ |B(s) - x| < \frac{1}{n} \right\} \in \mathcal{F}^0(t).$$

- If U is an open set, then $T_U = \inf\{t \geq 0 : B_t \in U\}$ is a stopping time with respect to $\{\mathcal{F}^+(t)\}_{t\geq 0}$:

  By continuity,
  $$\{T_U \leq t\} = \bigcap_{s>t} \{T_U < s\} = \bigcap_{s>t} \bigcup_{r\in\mathbb{Q}\cap(0,s)} \{B(r) \in U\} \in \mathcal{F}^+(t).$$
  However, T_U is not necessarily a stopping time with respect to $\{\mathcal{F}^0(t)\}_{t\geq 0}$. Indeed, suppose that U is bounded and $\overline{U}$ does not contain the starting point. Then we can find a path $\gamma : [0,t] \to \mathbb{R}$ with $\gamma(t) \in \partial U$ and $\{\gamma(s) : 0 \leq s < t\} \cap \overline{U} = \emptyset$. Clearly $\mathcal{F}^0(t)$ contains no nontrivial subset of $\{B(s) = \gamma(s) \text{ for all } 0 \leq s \leq t\}$. But if $\{T_U \leq t\} \in \mathcal{F}^0(t)$, then $\{B(s) = \gamma(s) \text{ for all } 0 \leq s \leq t, T = t\}$ would be a nontrivial subset, a contradiction.
  (The first statement is because $\mathcal{F}^0(t)$ only contains information about the paths up to time t, and the second is because B could either enter U immediately after hitting its boundary or could remain in the complement for a while.)

Definition. If T is a stopping time with respect to $\{\mathcal{F}^+(t) : t \geq 0\}$, the stopped σ-algebra is given by
$$\mathcal{F}^+(T) = \left\{A \in \mathcal{F} : A \cap \{T \leq t\} \in \mathcal{F}^+(t) \text{ for all } t \geq 0\right\}.$$

Theorem 16.9 (Strong Markov Property)**.** *If $\{B_t : t \geq 0\}$ is a Brownian motion and T is an a.s. finite stopping time with respect to $\{\mathcal{F}^+(t) : t \geq 0\}$, then the process*
$$\{B_{T+t} - B_T : t \geq 0\}$$
is a standard Brownian motion independent of $\mathcal{F}^+(T)$.

Proof. For $n \in \mathbb{N}$, define T_n by $T_n = \dfrac{m+1}{2^n}$ if $\dfrac{m}{2^n} \le T < \dfrac{m+1}{2^n}$.

Define the processes

$$B_{n,k}(t) = B\Big(t + \frac{k}{2^n}\Big) - B\Big(\frac{k}{2^n}\Big), \quad B_n^*(t) = B(T_n + t) - B(T_n).$$

We will show that $B_n^*(t)$ is a standard Brownian motion independent of $\mathcal{F}^+(T_n)$.
To this end, let $E \in \mathcal{F}^+(T_n)$. For every event $\{B_n^* \in A\}$, we have

$$\begin{aligned} P\left(\{B_n^* \in A\} \cap E\right) &= \sum_{k=1}^{\infty} P\Big(\big\{B_{n,k} \in A\big\} \cap E \cap \big\{T_n = k/2^n\big\}\Big) \\ &= \sum_{k=1}^{\infty} P\Big(B_{n,k} \in A\Big) P\Big(E \cap \big\{T_n = k/2^n\big\}\Big) \end{aligned}$$

since the Markov property implies $\big\{B_{n,k} \in A\big\}$ is independent of $E \cap \big\{T_n = k/2^n\big\} \in \mathcal{F}^+\big(k/2^n\big)$.

Because $B_{n,k}$ is a standard Brownian motion, $P\big(B_{n,k} \in A\big) = P\big(\widetilde{B} \in A\big)$ does not depend on n or k, so

$$\begin{aligned} P\left(\{B_n^* \in A\} \cap E\right) &= \sum_{k=1}^{\infty} P\Big(B_{n,k} \in A\Big) P\Big(E \cap \big\{T_n = k/2^n\big\}\Big) \\ &= \sum_{k=1}^{\infty} P\Big(\widetilde{B} \in A\Big) P\Big(E \cap \big\{T_n = k/2^n\big\}\Big) = P\big(\widetilde{B} \in A\big) P(E). \end{aligned}$$

Taking $E = \Omega$ shows that B_n^* is a standard Brownian motion, hence

$$P\left(\{B_n^* \in A\} \cap E\right) = P\left(\{B_n^* \in A\} \cap E\right) = P\big(B_n^* \in A\big) P(E)$$

for all A, E, establishing the independence of B_n^* and $\mathcal{F}^+(T_n)$.

Now $T_n \searrow T$, so it follows from continuity that

$$\begin{aligned} &\left(B\left(T_n + t_2\right) - B\left(T_n + t_1\right), \ldots, B\left(T_n + t_m\right) - B\left(T_n + t_{m-1}\right)\right) \\ &\qquad \to \left(B\left(T + t_2\right) - B\left(T + t_1\right), \ldots, B\left(T + t_m\right) - B\left(T + t_{m-1}\right)\right). \end{aligned}$$

Since the left hand side is multivariate normal with mean zero and covariance $\operatorname{diag}\left(t_2 - t_1, ..., t_m - t_{m-1}\right)$ for all n, the right hand side is as well.
As $\{B\left(T + t\right) - B\left(T\right) : t \ge 0\}$ is clearly a.s. continuous, we see that it is a Brownian motion.

To finish up, we need to show that $\{B\left(T + t\right) - B\left(T\right) : t \ge 0\}$ is independent of $\mathcal{F}^+(T)$.

This is a simple consequence of the fact that $T \le T_n$ implies $\mathcal{F}^+(T) \subseteq \mathcal{F}^+(T_n)$

(because $A \cap \{T_n \le t\} = (A \cap \{T \le t\}) \cap \{T_n \le t\} \in \mathcal{F}^+(t)$ if $A \in \mathcal{F}^+(T)$).

Specifically, since $\{B(T_n + t) - B(T_n) : t \ge 0\}$ is independent of $\mathcal{F}^+(T_n) \supseteq \mathcal{F}^+(T)$, if $A \in \mathcal{F}^+(T)$, $t_1, ..., t_m \ge 0$, and $F \in C_b\left(\mathbb{R}^m, \mathbb{R}\right)$, then continuity and the dominated convergence theorem give

$$\begin{aligned} &E\left[F\left(B(T + t_1) - B(T), ..., B(T + t_m) - B(T)\right) 1_A\right] \\ &\qquad = \lim_{n \to \infty} E\left[F\left(B(T_n + t_1) - B(T_n), ..., B(T_n + t_m) - B(T_n)\right) 1_A\right] \\ &\qquad\qquad = \lim_{n \to \infty} E\left[F\left(B(T_n + t_1) - B(T_n), ..., B(T_n + t_m) - B(T_n)\right)\right] P(A) \\ &\qquad\qquad\qquad = E\left[F\left(B(T + t_1) - B(T), ..., B(T + t_m) - B(T)\right)\right] P(A). \end{aligned}$$

□

It is easy to get bogged down in the details of these kind of arguments, but observe that the general strategy is quite simple. Namely, we approximate T discretely from above, sum over the possible values of T_n, and apply the ordinary Markov property, just as we did for Markov chains. The rest is just tying up loose ends.

An alternative form of Theorem 16.9 that is sometimes useful is

For any $x \in \mathbb{R}$ and any bounded measurable $f : C([0,\infty), \mathbb{R}) \to \mathbb{R}$, we have

$$E_x\left[f\left(\{B(T+t) : t \geq 0\}\right) \middle| \mathcal{F}^+(T)\right] = E_{B(T)}\left[f\left(\left\{\widetilde{B}(t) : t \geq 0\right\}\right)\right]$$

where expectation on the right is with respect to a Brownian motion $\left\{\widetilde{B}(t) : t \geq 0\right\}$ started at $B(T)$.

An intriguing consequence of the strong Markov property is

Theorem 16.10 (Reflection Principle)**.** *If T is an a.s. finite stopping time and $\{B(t) : t \geq 0\}$ is a standard Brownian motion, then the process $\{B^*(t) : t \geq 0\}$ defined by*

$$B^*(t) = B(t)1\{T \leq t\} + (2B(T) - B(t))\,1\{T > t\}$$

is also a standard Brownian motion. (B^* is called Brownian motion reflected at T*.)*

Proof. The strong Markov property implies that $B^{(T)} = \{B_{T+t} - B_T : t \geq 0\}$ is a standard Brownian motion which is independent of $B = \{B_t : 0 \leq t \leq T\}$, hence $-B^{(T)} = \{B_T - B_{T+t} : t \geq 0\}$ is as well.

The concatenation map which glues a continuous path $\{g(t) : t \geq 0\}$ to a finite continuous path $\{f(t) : 0 \leq t \leq T\}$ to form the continuous path $\Psi_T(f,g)(t) = f(t)1\{0 \leq t \leq T\} + (f(T) - g(0) + g(t-T))\,1\{t > T\}$ is evidently measurable.

Applying Ψ_T to B and $B^{(T)}$ gives the original process B, and applying Ψ_T to B and $-B^{(T)}$ gives B^*. The result follows since $\left(B, B^{(T)}\right) =_d \left(B, -B^{(T)}\right)$. □

It is an interesting fact that Brownian motion reflected at a random time is still a Brownian motion, but the real beauty of this result lies in its consequences. For example, with the aid of Theorem 16.10, we can deduce the marginal distributions of the running maximum of Brownian motion,

$$M(t) = \max_{0 \leq s \leq t} B(s).$$

Theorem 16.11. *If $a > 0$, then*

$$P_0\left(M(t) > a\right) = 2P_0\left(B(t) > a\right) = P_0\left(|B(t)| > a\right).$$

Proof. The second equality follows from the symmetry of standard Brownian motion.

For the first, let $T = \inf\{t \geq 0 : B(t) = a\}$, and let $B^*(t)$ be Brownian motion reflected at T.

Then

$$\{M(t) > a\} = \{B(t) > a\} \bigsqcup \{M(t) > a, B(t) \leq a\}.$$

The right-hand side is a disjoint union, the second term of which coincides with $\{B^*(t) \geq a\}$, so Theorem 16.10 gives

$$P_0\left(M(t) > a\right) = P_0\left(B(t) > a\right) + P_0\left(B^*(t) > a\right) = 2P_0\left(B(t) > a\right).$$ □

Our next application of the strong Markov property concerns the zero set of Brownian motion.

First, we take a brief detour to recall a few facts from undergraduate real analysis.

Definition. A point x in a topological space $(X, \mathcal{T})$ is called an *isolated point* of $S \subseteq X$ if there exists some $U \in \mathcal{T}$ with $x \in U$ and $U \cap S = \{x\}$.

A point $x \in \mathbb{R}$ is *isolated from the right* (with respect to S) if there is some $\varepsilon > 0$ such that $(x, x+\varepsilon) \cap S = \emptyset$, and is *isolated from the left* if there is some $\varepsilon > 0$ such that $(x - \varepsilon, x) \cap S = \emptyset$.

$x \in \mathbb{R}$ is isolated if and only if it is isolated from the left and from the right.

Definition. A *perfect subset* of a topological space $(X, \mathcal{T})$ is a closed set with no isolated points.

Proposition. *A perfect subset of $\mathbb{R}$ (with the usual metric topology) is uncountable.*

Proof. We first note that if $x \in S$ and U is an open set containing x, then $U \cap S$ is infinite - otherwise, the open ball centered at x with radius $\frac{1}{2} \min_{y \in U \cap S} |x - y|$ would intersect S only in x.

Now suppose that S can be enumerated as $S = \{x_1, x_2, ...\}$. Let U_1 be a bounded open set containing x_1. Then U_1 contains infinitely many points in S. Let $k_1 = 1$ and $k_2 = \min\{k \geq k_1 : x_k \in U_1\}$. Let U_2 be an open set containing x_{k_2} such that $\overline{U_2} \subseteq U_1 \setminus \{x_{k_1}\}$. Then set $k_3 = \min\{k \geq k_2 : x_k \in U_2\}$, let U_3 be an open set containing x_{k_3} with $\overline{U_3} \subseteq U_2 \setminus \{x_{k_2}\}$, and so on...

Define $V = \bigcap_{n=1}^{\infty} \left(\overline{U_n} \cap S\right)$. Since S and $\overline{U_n}$ are closed subsets of $\mathbb{R}$ and $\overline{U_n}$ is bounded, V is the intersection of a nested decreasing sequence of nonempty compact sets and thus is nonempty. But, by construction, $V \subseteq S$ does not contain any x_k, contradicting the assumption that S can be enumerated. □

Returning to Brownian motion, we have

Theorem 16.12. *Let $\{B(t) : t \geq 0\}$ be a Brownian motion and define*

$$\text{Zeros} = \{t \geq 0 : B(t) = 0\}.$$

Then, almost surely, Zeros *is a perfect set and thus is uncountable.*

Proof. Since Brownian motion is a.s. continuous, Zeros is closed with probability one. In light of the preceding proposition, it remains only to show that Zeros almost surely has no isolated point.

To this end, we define for each nonnegative $q \in \mathbb{Q}$, $\tau_q = \inf\{t \geq q : B(t) = 0\}$. Then τ_q is a stopping time which is a.s. finite by Theorem 16.6. Since Zeros is a.s. closed, $P(\tau_q \in \text{Zeros}) = 1$. Theorem 16.3 and the strong Markov property applied to τ_q show that, almost surely, τ_q is not isolated from the right for all $q \in \mathbb{Q} \cap [0, \infty)$.

We will be done if we can show that all other points in Zeros are not isolated from the left. For such t, take $q_n \nearrow t$ with $q_n \in \mathbb{Q}$ and define $t_n = \tau_{q_n}$. Then $q_n \leq t_n < t$, so $t_n \nearrow t$, hence t is not isolated from the left. □

Markov Processes.

At this point, we need to define precisely what we mean by a continuous time Markov process.

Definition. A function $p : [0,\infty) \times \mathbb{R} \times \mathcal{B} \to [0,1]$ is a *Markov transition kernel* on $\mathbb{R}$ if

(1) For all $A \in \mathcal{B}$, $p(\cdot,\cdot,A) : [0,\infty) \times \mathbb{R} \to [0,1]$ is a measurable function of (t,x).

(2) For all $t \in [0,\infty)$, $x \in \mathbb{R}$, $p(t,x,\cdot)$ is a probability measure on $(\mathbb{R},\mathcal{B})$.

(3) For all $A \in \mathcal{B}$, $x \in \mathbb{R}$, $s,t > 0$, $p(s+t,x,A) = \int p(s,x,dy)p(t,y,A)$.

Given a filtration $\{\mathcal{F}(t) : t \geq 0\}$, an adapted process $\{X(t) : t \geq 0\}$ is a *Markov process with transition kernel* p if for all $t > s \geq 0$ and all $A \in \mathcal{B}$,

$$P\left(X(t) \in A \,|\mathcal{F}(s)\right) = p\left(t-s, X(s), A\right) \text{ almost surely.}$$

In words, $p(t,x,A)$ is the probability that the process started at x will be in A at time t.

Theorem 16.1 shows that Brownian motion is a Markov process with respect to $\{\mathcal{F}^+(t) : t \geq 0\}$.

The transition kernel is

$$p(t,x,A) = \frac{1}{\sqrt{2\pi t}} \int_A e^{-\frac{(y-x)^2}{2t}} dy.$$

That is, $p(t,x,\cdot)$ is the normal distribution with mean x and variance t.
(The Chapman-Kolmogorov condition here is the statement that the sum of independent normals is normal.)

Likewise, if $\{B(t) : t \geq 0\}$ is a standard Brownian motion, the *reflected Brownian motion* $\{X(t) : t \geq 0\}$ given by $X(t) = |B(t)|$ is a Markov process (w.r.t. $\{\mathcal{F}^+(t) : t \geq 0\}$) with transition kernel $p(t,x,\cdot)$ the *modulus normal distribution* - i.e. the law of $|W|$ with $W \sim \mathcal{N}(x,t)$.

The following theorem shows that the difference of the maximum process $M(t) = \sup_{0 \leq s \leq t} B(s)$ and the underlying Brownian motion $B(t)$ is a reflected Brownian motion.

Theorem 16.13. *Let $\{B(t) : t \geq 0\}$ be a Brownian motion and let $\{M(t) : t \geq 0\}$ be the running maximum process. Then the process $\{Y(t) : t \geq 0\}$ given by $Y(t) = M(t) - B(t)$ is a reflected Brownian motion.*

Proof. Since $M(0) = B(0)$, we can assume without loss that $B(t)$ is a standard Brownian motion.
We proceed by fixing $s \geq 0$ and defining the processes $\left\{\widehat{B}(t) : t \geq 0\right\}$ and $\left\{\widehat{M}(t) : t \geq 0\right\}$ by

$$\widehat{B}(t) = B(s+t) - B(s), \quad \widehat{M}(t) = \max_{0 \leq u \leq t} \widehat{B}(u).$$

Since $Y(t)$ is clearly $\mathcal{F}^+(t)$-measurable, we need only show that

$$P\left(Y(s+t) \in A \,\middle|\mathcal{F}^+(s)\right) = p(t, Y(s), A)$$

for all $t \geq 0$, $A \in \mathcal{B}$, where $p(t,y,\cdot) = \mathscr{L}\left(|W|\right)$, $W \sim \mathcal{N}(y,t)$.

As $\widehat{B}(t) \sim \mathcal{N}(0,t)$ is independent of $\mathcal{F}^+(s)$, this is equivalent to showing that, conditional on $\mathcal{F}^+(s)$, $Y(s+t)$ has the same distribution as $\left|\widehat{B}(t) + Y(s)\right|$.

Writing $M(s+t) = M(s) \vee \left(B(s) + \widehat{M}(t)\right)$ - the max over $[0, s+t]$ is the maximum of the max over $[0,s]$ and the max over $[s, s+t]$ - shows that

$$\begin{aligned} Y(s+t) &= M(s+t) - B(s+t) \\ &= \left[M(s) \vee \left(B(s) + \widehat{M}(t)\right)\right] - \left(\widehat{B}(t) + B(s)\right) \\ &= \left[M(s) - \left(\widehat{B}(t) + B(s)\right)\right] \vee \left[\left(B(s) + \widehat{M}(t)\right) - \left(\widehat{B}(t) + B(s)\right)\right] \\ &= \left(Y(s) - \widehat{B}(t)\right) \vee \left(\widehat{M}(t) - \widehat{B}(t)\right) \\ &= \left(Y(s) \vee \widehat{M}(t)\right) - \widehat{B}(t). \end{aligned}$$

We will be done if we prove that for every $y \geq 0$, $\left(y \vee \widehat{M}(t)\right) - \widehat{B}(t) =_d \left|\widehat{B}(t) + y\right|$.

To see this, observe that for $a \geq 0$,

$$\begin{aligned} P\left(\left(y \vee \widehat{M}(t)\right) - \widehat{B}(t) > a\right) &= P\left(y - \widehat{B}(t) > a\right) + P\left(y - \widehat{B}(t) \leq a, \widehat{M}(t) - \widehat{B}(t) > a\right) \\ &= P\left(y + \widehat{B}(t) > a\right) + P\left(y - \widehat{B}(t) \leq a, \widehat{M}(t) - \widehat{B}(t) > a\right) \end{aligned}$$

since $\widehat{B}(t) =_d -\widehat{B}(t)$.

Thus equality in distribution will follow upon showing that

$$P\left(y - \widehat{B}(t) \leq a, \widehat{M}(t) - \widehat{B}(t) > a\right) = P\left(y + \widehat{B}(t) < -a\right).$$

To this end, let $\{W(u) : 0 \leq u \leq t\}$ be the time reversed Brownian motion $W(u) = \widehat{B}(t-u) - \widehat{B}(t)$ (which was shown in the homework to be a standard Brownian motion on $[0,t]$), and set $M_W(t) = \max_{0 \leq u \leq t} W(u)$.

We have

$$M_W(t) = \max_{0 \leq u \leq t}\left(\widehat{B}(t-u) - \widehat{B}(t)\right) = \max_{0 \leq u \leq t} \widehat{B}(u) - \widehat{B}(t) = \widehat{M}(t) - \widehat{B}(t)$$

and $W(t) = -\widehat{B}(t)$, so

$$P\left(y - \widehat{B}(t) \leq a, \widehat{M}(t) - \widehat{B}(t) > a\right) = P\left(y + W(t) \leq a, M_W(t) > a\right).$$

Finally, let $\{W^*(u) : 0 \leq u \leq t\}$ be the process formed by reflecting W at $\tau_a = \inf\{u : W(u) = a\}$.

The reflection principle shows that $\{W^*(u) : 0 \leq u \leq t\}$ is a standard Brownian motion on $[0,t]$, hence $W^*(t) =_d -\widehat{B}(t)$.

Moreover, it is clear that $\{y + W(t) \leq a, M_W(t) > a\} = \{W^*(t) \geq a + y\}$, so we end up with

$$\begin{aligned} P\left(y - \widehat{B}(t) \leq a, \widehat{M}(t) - \widehat{B}(t) > a\right) &= P\left(y + W(t) \leq a, M_W(t) > a\right) \\ &= P\left(W^*(t) \geq a + y\right) = P\left(-\widehat{B}(t) \geq a + y\right) \\ &= P\left(\widehat{B}(t) \leq -a - y\right) = P\left(\widehat{B}(t) + y < -a\right), \end{aligned}$$

(as $\widehat{B}(t)$ is a continuous random variable) and the proof is complete. □

While the foregoing establishes that $\{M(t) - B(t) : t \geq 0\}$ is a Markov process, $\{M(t) : t \geq 0\}$ clearly is not. However, the next result shows that the process which records the times when new maxima are achieved is Markovian.

Theorem 16.14. *For $a \geq 0$, define the stopping time*

$$T_a = \inf\{t \geq 0 : B(t) = a\}.$$

Then $\{T_a : a \geq 0\}$ is a Markov process with transition kernel given by the densities

$$p(a,t,s) = \frac{a}{\sqrt{2\pi(s-t)^3}} \exp\left(-\frac{a^2}{2(s-t)}\right) 1\{s > t\} \text{ for } a > 0.$$

Proof. Fix $a \geq b \geq 0$ and observe that for all $t \geq 0$,

$$\{T_a - T_b = t\} = \{B(T_b + s) - B(T_b) < a - b \text{ for all } s < t\} \cap \{B(T_b + t) - B(T_b) = b - a\}.$$

The strong Markov property shows that this event is independent of $\mathcal{F}^+(T_b)$, and thus of $\{T_c : c \leq b\}$. Therefore, $\{T_a : a \geq 0\}$ is Markovian with respect to its natural filtration.

To compute the transition density, observe that the strong Markov property implies that $T_{a-b} =_d T_a - T_b$ so Theorem 16.11 gives

$$\begin{aligned} P(T_a - T_b \leq t) = P(T_{a-b} \leq t) &= P\Big(\max_{0 \leq s \leq t} B(s) \geq a - b\Big) \\ &= P\big(|B(t)| \geq a - b\big) = \frac{2}{\sqrt{2\pi t}} \int_{a-b}^{\infty} e^{-\frac{x^2}{2t}} dx = \int_0^t \frac{a-b}{\sqrt{2\pi s^3}} e^{-\frac{(a-b)^2}{2s}} ds \end{aligned}$$

where the final equality used the substitution $x = (a-b)\sqrt{\frac{t}{s}}$. □

The process in Theorem 16.14 is called a *stable subordinator of index* $\frac{1}{2}$.

A subordinator is a real-valued nondecreasing *Lévy process* - that is, a process with stationary independent increments which is continuous in probability: for all $t \geq 0$, $\varepsilon > 0$, $\lim_{h \to 0} P(|X_{t+h} - X_t| > \varepsilon) = 0$.

It is stable with index α if $X(0) = 0$ and the scaling relation $t^{-\frac{1}{\alpha}} X(t) =_d X(1)$ holds for all $t > 0$.

Thus, for example, standard Brownian motion is stable with index 2, but is not a subordinator.

Conversely, the Poisson process is a subordinator, but is not stable.

The process $\{T_a : a \geq 0\}$ is nondecreasing and continuous in probability by continuity of Brownian paths. It has stationary independent increments since the transition densities are shift invariant in the sense that $p(a,t,s) = p(a,0,s-t)$ for all $a, s, t \geq 0$.

The self-similarity is a consequence of Brownian scaling:

$$\begin{aligned} T_a &= \inf\{t \geq 0 : B_t = a\} = a^2 \inf\{t \geq 0 : B_{a^2 t} = a\} \\ &= a^2 \inf\{t \geq 0 : aB_t = a\} = a^2 T_1. \end{aligned}$$

17. Martingale Properties

In the previous section, we were able to infer several interesting facts about Brownian motion and related processes by appealing to the Markov property.

We will now explore Brownian motion as a (continuous time) martingale and use the opportunity to introduce some standard concepts in the study of stochastic processes.

Definition. A real-valued process $\{X(t) : t \geq 0\}$ is said to be a *martingale* with respect to a filtration $\{\mathcal{F}(t) : t \geq 0\}$ if it is adapted, integrable, and satisfies $E[X(t)|\mathcal{F}(s)] = X(s)$ for all $0 \leq s \leq t$.

It is a sub/super-martingale if the equality is replaced with an appropriate inequality.

Example 17.1. Brownian motion is clearly a martingale since for all $0 \leq s \leq t$,

$$\begin{aligned} E\left[B(t)\left|\mathcal{F}^+(s)\right.\right] &= E\left[(B(t) - B(s)) + B(s)\left|\mathcal{F}^+(s)\right.\right] \\ &= E[B(t) - B(s)] + B(s) = B(s). \end{aligned}$$

Before elaborating on the martingale property of Brownian motion, we will take a moment to extend some familiar results to continuous time.

The basic strategy is to apply discrete time analogues to approximating sequences.

In order to present these results in greater generality, we need a few more definitions.

We say that a filtered probability space $\left(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P\right)$ satisfies the *usual conditions* if

(1) $\mathcal{F}$ is complete with respect to P – that is, if $A \in \mathcal{F}$ has $P(A) = 0$ and $B \subseteq A$, then $B \in \mathcal{F}$.
(2) $\mathcal{F}_0$ contains all P-null sets.
(3) $\{\mathcal{F}_t\}_{t\geq 0}$ is right-continuous.

Given a probability space $(\Omega, \mathcal{F}^o, P^o)$, its completion is the space $(\Omega, \mathcal{F}, P)$ where $\mathcal{F} = \sigma(\mathcal{F}^o \cup \mathcal{N})$ with $\mathcal{N} = \{E \subseteq \Omega : E \subseteq F$ for some $F \in \mathcal{F}^o$ with $P^o(F) = 0\}$, and $P(E) = \inf\{P^o(F) : E \subseteq F \in \mathcal{F}\}$.

The *usual augmentation* $\left(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P\right)$ of the filtered space $\left(\Omega, \mathcal{F}^o, \{\mathcal{F}^o_t\}_{t\geq 0}, P^o\right)$ is formed by taking $(\Omega, \mathcal{F}, P)$ to be the completion of $(\Omega, \mathcal{F}^o, P^o)$ and setting $\mathcal{F}_t = \bigcap_{s>t} \sigma(\mathcal{F}^o_s \cup \mathcal{N})$.

All of our results in the previous section carry over without change to the augmented filtration for Brownian motion since adding null sets doesn't make any difference in the arguments.

The reason that we are now concerned with completeness is that many processes of interest do not enjoy the continuity properties of Brownian motion, but we still want to be able to argue by discrete approximation.

This approach generally works for *càdlàg* (right-continuous with left limits) processes, and we will see that under the usual conditions, martingales have càdlàg versions.

(We say that $\{Y_t : t \geq 0\}$ is a *version* of $\{X_t : t \geq 0\}$ if for all $t \geq 0$, $P(X_t \neq Y_t) = 0$. This is a strictly weaker notion than being *indistinguishable*: $P(X_t \neq Y_t$ for some $t \geq 0) = 0$, though the two coincide when the processes are right-continuous.)

Theorem 17.1 (Doob's inequalities). *If X_t is a martingale or nonnegative submartingale with càdlàg sample paths, then*

(1)
$$P\left(\sup_{s\le t}|X_s|\ge\lambda\right)\le\frac{E|X_t|}{\lambda}\ \text{for all}\ \lambda>0.$$

(2)
$$E\left[\sup_{s\le t}|X_t|^p\right]\le\left(\frac{p}{p-1}\right)^p E\left[|X_t|^p\right]\ \text{for all}\ 1<p<\infty.$$

Proof. Since the absolue value of a martingale is a nonnegative submartingale, it suffices to consider the case where $X_t\ge 0$ is a submartingale.

Let $\mathcal{D}_n(t)=\left\{\frac{kt}{2^n}:0\le k\le 2^n\right\}$, and set $Y_k^{(n)}=X_{\frac{kt}{2^n}}$, $\mathcal{G}_k^{(n)}=\mathcal{F}_{\frac{kt}{2^n}}$.

Then $Y_k^{(n)}$ is clearly a discrete time submartingale with respect to $\mathcal{G}_k^{(n)}$.

Setting $A_n(t)=\left\{\sup_{s\in\mathcal{D}_n(t)}|X_s|>\lambda\right\}$, it follows from Theorem 4.2 that

$$P\left(A_n(t)\right)=P\left(\max_{0\le k\le 2^n}Y_k^{(n)}>\lambda\right)\le\frac{E\left|Y_{2^n}^{(n)}\right|}{\lambda}=\frac{E|X_t|}{\lambda}.$$

Since $A_1(t)\subseteq A_2(t)\subseteq\cdots$ and X_t is right-continuous, we have $\bigcup_n A_n(t)=\left\{\sup_{s\le t}|X_s|>\lambda\right\}$ and thus

$$P\left(\sup_{s\le t}|X_s|>\lambda\right)=\lim_{n\to\infty}P\left(A_n(t)\right)\le\frac{E|X_t|}{\lambda}.$$

Plugging in $\lambda-\varepsilon$ for λ and taking $\varepsilon\searrow 0$ gives (1).

Similarly, Theorem 4.4 gives

$$E\left[\sup_{0\le k\le 2^n}\left|Y_k^{(n)}\right|^p\right]\le\left(\frac{p}{p-1}\right)^p E\left[\left|Y_{2^n}^{(n)}\right|^p\right]=\left(\frac{p}{p-1}\right)^p E\left[|X_t|^p\right].$$

Right-continuity implies that $\sup_{0\le k\le 2^n}\left|Y_k^{(n)}\right|^p\nearrow\sup_{s\le t}|X_s|^p$, and (2) follows from Fatou's lemma. □

In a similar vein, we can prove continuous time optional stopping theorems, such as

Theorem 17.2. *Suppose that $\{X(t):t\ge 0\}$ is a martingale with càdlàg sample paths and $S\le T$ are stopping times. If there exists some integrable random variable Y such that $|X(t\wedge T)|\le Y$ a.s. for all $t\ge 0$, then*

$$E\left[X(T)\,|\mathcal{F}(S)\right]=X(S)\ \text{a.s.}$$

Proof. Fix $N\in\mathbb{N}$ and define a discrete time martingale $Y_n=X\left(T\wedge\frac{n}{2^N}\right)$ and stopping times $S'=\lfloor 2^N S\rfloor+1$, $T'=\lfloor 2^N T\rfloor+1$. The referent filtration is taken to be $\mathcal{G}_n=\mathcal{F}\left(\frac{n}{2^N}\right)$.

$\{Y_n:n\in\mathbb{N}\}$ is clearly uniformly integrable, so, writing $S_N=2^{-N}\left(\lfloor 2^N S\rfloor+1\right)$, Theorem 5.2 gives

$$E\left[X(T)\,|\mathcal{F}\left(S_N\right)\right]=E\left[Y_{T'}\,|\mathcal{G}_{S'}\right]=Y_{S'}=X\left(T\wedge S_N\right).$$

Since $S_N\searrow S$ as $N\nearrow\infty$, dominated convergence and right-continuity show that for any $A\in\mathcal{F}(S)$,

$$\begin{aligned}\int_A X(T)dP&=\lim_{N\to\infty}\int_A E\left[X(T)\,|\mathcal{F}(S_N)\right]dP\\&=\lim_{N\to\infty}\int_A X\left(T\wedge S_N\right)dP=\int_A\lim_{N\to\infty}X\left(T\wedge S_N\right)dP=\int_A X(S)dP.\end{aligned}$$

□

To wrap up our general discussion of martingales, we prove a continuous time martingale convergence theorem and show that there is no loss in assuming càdlàg paths as long as the usual conditions are satisfied.

We will use the notation $\mathcal{D}_n = \left\{\frac{k}{2^n} : k \in \mathbb{N}_0\right\}$, $\mathcal{D} = \bigcup_{n\in\mathbb{N}_0} \mathcal{D}_n$.

Theorem 17.3. *Suppose that $\{X_t : t \geq 0\}$ is a submartingale with respect to some filtration $\{\mathcal{F}_t : t \geq 0\}$ and that $\sup_{t\geq 0} E|X_t| < \infty$. Then*

(1) $\lim_{t\to\infty} X_t$ *exists in $\mathbb{R}$ almost surely.*
(2) *With probability one, X_t has finite left and right limits along $\mathcal{D}$.*

Proof. Theorem 17.1 ensures that for all $\lambda > 0$,

$$P\left(\sup_{t\in\mathcal{D}_n\cap[0,n]} X_t \geq \lambda\right) \leq \frac{E|X_n|}{\lambda},$$

so the monotone convergence theorem gives

$$P\left(\sup_{t\in\mathcal{D}} X_t \geq \lambda\right) \leq \sup_{t\geq 0} \frac{E|X_t|}{\lambda}.$$

Since $\lambda > 0$ is arbitrary, we see that $\{|X_t| : t \in \mathcal{D}\}$ is a.s. a bounded set.

Accordingly, the only way for (1) or (2) to fail is if there is some pair of rationals $a < b$ such that the number of upcrossings of $[a, b]$ by $\{X_t : t \in \mathcal{D}\}$ is infinite.

But Lemma 2.1 shows that if U_n is the number of upcrossings by $\{X_t : t \in \mathcal{D}_n \cap [0, n]\}$, then

$$E[U_n] \leq \frac{E|X_n|}{b-a}.$$

Taking $n \to \infty$, Fatou's lemma shows that the number of upcrossings by $\{X_t : t \in \mathcal{D}\}$ has finite expectation and thus is finite with full probability. The result follows since there are countably many pairs of rationals. □

The reason for taking the seemingly circuitous route through the upcrossing inequality was to establish the second claim in Theorem 17.3. The utility of this fact lies in

Theorem 17.4. *Let $\{\mathcal{F}_t\}$ be a filtration satisfying the usual conditions. If X_t is a martingale with respect to $\{\mathcal{F}_t\}$, then there is a version of $\{X_t : t \geq 0\}$ which is a martingale and has càdlàg paths.*

Proof. Jensen's inequality implies that $|X_t|$ is a submartingale, so $E|X_t| \leq E|X_N| < \infty$ for all $t \leq N$.

Accordingly, Theorem 17.3 shows that for any $N \in \mathbb{N}$, $X_{t\wedge N}$ almost surely has left and right limits along $\mathcal{D}$.

Since N is arbitrary, we must have that X_t has left and right limits along $\mathcal{D}$ a.s.

Now define

$$Y_t = \lim_{u\in\mathcal{D}, u\to t^+} X_u.$$

Then Y_t has càdlàg paths by construction.
Also, since $\{\mathcal{F}_t\}$ is right-continuous and $\{X_t\}$ is adapted, Y_t is $\mathcal{F}_t$-measurable for all $t \geq 0$.

We now observe that for fixed $N \in \mathbb{N}$, $\{X_t : 0 \leq t \leq N\}$ is uniformly integrable:

Let $\varepsilon > 0$. Since X_N is integrable, there is a $\delta > 0$ such that $P(A) < \delta$ implies $E[X_N; A] < \varepsilon$. (See the proof of Theorem 4.6 for details.)

When $M > \frac{E|X_N|}{\delta}$, we have

$$P(|X_t| \geq M) \leq \frac{E|X_t|}{M} \leq \frac{E|X_N|}{M} < \delta$$

for all $0 \leq t \leq N$ because $|X_t|$ is a submartingale.

Since $\{|X_t| \geq M\} \in \mathcal{F}_t$, the submartingale property implies

$$E[|X_t|; |X_t| \geq M] \leq E[|X_N|; |X_t| \geq M] < \varepsilon$$

for all $t \in [0, N]$, hence $\{X_t : 0 \leq t \leq N\}$ is u.i.

Now let $t < N$. If $E \in \mathcal{F}_t$, then the Vitali convergence theorem gives

$$E[Y_t; E] = E\left[\lim_{u \in \mathcal{D}, u \to t^+} X_u; E\right] = \lim_{u \in \mathcal{D}, u \to t^+} E[X_u; E] = E[X_t; E].$$

Since $Y_t \in \mathcal{F}_t$, this proves that $X_t = Y_t$ a.s.

Because N is arbitrary, we conclude that $\{Y_t : t \geq 0\}$ is a version of $\{X_t : t \geq 0\}$, and we are done since this implies that Y_t is a martingale: For $s \leq t$, $E[Y_t | \mathcal{F}_s] = E[X_t | \mathcal{F}_s] = X_s = Y_s$ almost surely. □

Now that we are comfortable with martingales in continuous time and have the foregoing results at our disposal, we can return to Brownian motion. We begin with

Theorem 17.5 (Wald's lemma)**.** *Let $\{B(t) : t \geq 0\}$ be a standard Brownian motion and let T be a stopping time with $E[T] < \infty$. Then $E[B(T)] = 0$.*

Proof. Define

$$M_k = \max_{0 \leq t \leq 1} |B(t+k) - B(k)|, \quad M = \sum_{k=0}^{\lfloor T \rfloor} M_k.$$

Then independent increments and the fact that M_k is independent of $\{T \geq k\} = \{T < k\}^C \in \mathcal{F}^+(k)$ yield

$$\begin{aligned} E[M] &= E\left[\sum_{k=0}^{\lfloor T \rfloor} M_k\right] = \sum_{k=0}^{\infty} E[M_k 1\{T \geq k\}] = \sum_{k=0}^{\infty} E[M_k] P(T \geq k) \\ &= E[M_0]\left(1 + \sum_{k=1}^{\infty} P(T \geq k)\right) \leq E[M_0]\left(1 + \int_0^{\infty} P(T \geq t)\, dt\right) = E[M_0](1 + E[T]). \end{aligned}$$

The reflection principle shows that

$$\begin{aligned} E[M_0(t)] &= \int_0^{\infty} P\left(\max_{0 \leq t \leq 1} |B(t)| > x\right) dx \leq 1 + \int_1^{\infty} 2P\left(\max_{0 \leq t \leq 1} B(t) > x\right) dx \\ &= 1 + 2\int_1^{\infty} P(|B(1)| > x)\, dx = 1 + 2\int_1^{\infty} \left(2\int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{s^2}{2}} ds\right) dx \\ &\leq 1 + \frac{4}{\sqrt{2\pi}} \int_1^{\infty} \int_x^{\infty} \frac{s}{x} e^{-\frac{s^2}{2}} ds dx = 1 + \frac{4}{\sqrt{2\pi}} \int_1^{\infty} x^{-1} e^{-\frac{x^2}{2}} dx < \infty, \end{aligned}$$

so $E[M] < \infty$.

Since $\sup_{t \geq 0} |B(t \wedge T)| \leq M \in L^1$, Theorem 17.2 gives

$$E[B(T)] = E\left[E\left[B(T) | \mathcal{F}^+(0)\right]\right] = E[B(0)] = 0.$$ □

In order to compute the second moment of $B(T)$, we need the following results, the second of which gives another concrete example of a continuous martingale.

Lemma 17.1. *Let $S \leq T$ be stopping times with $E[T] < \infty$. Then*

$$E\left[B(T)^2\right] = E\left[B(S)^2\right] + E\left[\left(B(T) - B(S)\right)^2\right].$$

Proof.

$$\begin{aligned} E\left[B(T)^2\right] &= E\left[E\left[\left(B(T) - B(S)\right)^2 + 2B(T)B(S) - B(S)^2 \,\middle|\, \mathcal{F}^+(S)\right]\right] \\ &= E\left[E\left[\left(B(T) - B(S)\right)^2 + 2B(S)\left(B(T) - B(S)\right) + B(S)^2 \,\middle|\, \mathcal{F}^+(S)\right]\right] \\ &= E\left[\left(B(T) - B(S)\right)^2\right] + 2E\left[B(S)E\left[B(T) - B(S) \,\middle|\, \mathcal{F}^+(S)\right]\right] + E\left[B(S)^2\right]. \end{aligned}$$

Since $E[T] < \infty$ implies $E\left[T - S \,|\mathcal{F}(S)\right] < \infty$ a.s., the strong Markov property at time S in conjunction with Wald's lemma shows that $E\left[B(T) - B(S) \,|\mathcal{F}^+(S)\right] = 0$ a.s. □

Proposition 17.1. *Let $\{B(t) : t \geq 0\}$ be a Brownian motion. Then the process $\left\{B(t)^2 - t : t \geq 0\right\}$ is a martingale.*

Proof. $B(t)^2 - t$ is clearly integrable and $\mathcal{F}^+(t)$-measurable for all $t \geq 0$, and for any $0 \leq s \leq t$,

$$\begin{aligned} E\left[B(t)^2 - t \,\middle|\, \mathcal{F}^+(s)\right] &= E\left[\left(B(t) - B(s)\right)^2 + 2B(t)B(s) - B(s)^2 - t \,\middle|\, \mathcal{F}^+(s)\right] \\ &= E\left[\left(B(t) - B(s)\right)^2\right] + 2B(s)\left[B(t) \,\middle|\, \mathcal{F}^+(s)\right] - B(s)^2 - t \\ &= (t - s) + 2B(s)^2 - B(s)^2 - t = B(s)^2 - s. \end{aligned}$$

□

We can now prove Wald's second lemma.

Theorem 17.6. *If T is a stopping time for standard Brownian motion and $E[T] < \infty$, then $E\left[B(T)^2\right] = E[T]$.*

Proof. Consider the martingale $\left\{B(t)^2 - t : t \geq 0\right\}$ and define stopping times $T_n = \inf\{t \geq 0 : |B(t)| = n\}$. Then the process $\left\{B\left(t \wedge T \wedge T_n\right)^2 - \left(t \wedge T \wedge T_n\right) : t \geq 0\right\}$ is uniformly dominated by the integrable random variable $n^2 + T$, so

Theorem 17.2 with $S = 0$ implies $E\left[B\left(T \wedge T_n\right)^2 - \left(T \wedge T_n\right)\right] = 0$, or $E\left[B\left(T \wedge T_n\right)^2\right] = E\left[T \wedge T_n\right]$.

Applying Lemma 17.1 to $T \wedge T_n \leq T$ shows that $E\left[B(T)^2\right] \geq E\left[B\left(T \wedge T_n\right)^2\right]$, hence

$$E\left[B(T)^2\right] \geq \lim_{n\to\infty} E\left[B\left(T \wedge T_n\right)^2\right] = \lim_{n\to\infty} E\left[T \wedge T_n\right] = E[T]$$

by monotone convergence.

On the other hand, Fatou's lemma gives

$$E\left[B(T)^2\right] = E\left[\liminf_{n\to\infty} B\left(T \wedge T_n\right)^2\right] \leq \liminf_{n\to\infty} E\left[B\left(T \wedge T_n\right)^2\right] = \liminf_{n\to\infty} E\left[T \wedge T_n\right] = E[T],$$

and the proof is complete. □

A typical application of Wald's lemmas is computing exit times and exit probabilities for Brownian motion. For example, we have the following "gambler's ruin" type result.

Theorem 17.7. *Let $\{B(t) : t \geq 0\}$ be a standard Brownian motion, and let $T_A = \inf\{t \geq 0 : B(t) \in A\}$ denote the hitting time of a Borel set A. If $a < 0 < b$, then $T = T_{\{a,b\}}$ satisfies*

(1)
$$P(B(T) = a) = \frac{b}{|a| + b} \text{ and } P(B(T) = b) = \frac{|a|}{|a| + b}.$$

(2)
$$E[T] = |ab|.$$

Proof. $|B(T \wedge t)| \leq |a| \vee b$, so it follows from Theorem 17.2 that

$$\begin{aligned} 0 = E[B(0)] = E[B(T)] &= aP(B(T) = a) + bP(B(T) = b) \\ &= aP(B(T) = a) + b[1 - P(B(T) = a)] \\ &= b + (a - b)P(B(T) = a) \\ &= b - (b + |a|)P(B(T) = a), \end{aligned}$$

which gives (1).

To see that T has finite mean, observe that

$$\begin{aligned} E[T] = \int_0^\infty P(T > t)dt &\leq \int_0^\infty P(T > \lfloor t \rfloor)dt = \sum_{k=0}^\infty P(T > k) \\ &\leq 1 + \sum_{k=1}^\infty P\left(\bigcap_{j=1}^k \{B(t) \in (a, b) \text{ for all } j - 1 \leq t < j\}\right) \\ &\leq 1 + \sum_{k=1}^\infty P\left(\max_{0 \leq t \leq 1} |B(t)| \leq b - a\right)^k < \infty. \end{aligned}$$

Therefore, Wald's second lemma implies

$$\begin{aligned} E[T] = E\left[B(T)^2\right] &= a^2 P(T = a) + b^2 P(T = b) \\ &= \frac{|a|^2 b}{|a| + b} + \frac{|a| b^2}{|a| + b} = \frac{|a| b(|a| + b)}{|a| + b} = |ab|. \end{aligned}$$

□

Proposition 17.1 showed that the process $\left\{B(t)^2 - t : t \geq 0\right\}$ is a martingale with respect to $\mathcal{F}^+(t)$. Another important example is given by geometric Brownian motion.

Proposition 17.2. *Let $\{B(t) : t \geq 0\}$ be a standard Brownian motion.*

Then the process $\left\{\exp\left(\sigma B(t) - \frac{\sigma^2 t}{2}\right) : t \geq 0\right\}$ is a martingale for all $\sigma > 0$.

Proof. For $0 \le s \le t$,

$$E\left[\exp\left(\sigma B(t) - \frac{\sigma^2 t}{2}\right)\middle|\, \mathcal{F}^+(s)\right] = \exp\left(\sigma B(s)\right) E\left[\exp\left(\sigma\left(B(t) - B(s)\right)\right)|\, \mathcal{F}^+(s)\right] \exp\left(-\frac{\sigma^2 t}{2}\right).$$

Since $\sigma\left(B(t) - B(s)\right) \sim \mathcal{N}\left(0, \sigma^2(t-s)\right)$ is independent of $\mathcal{F}^+(s)$, the middle term is

$$\begin{aligned}\int_{-\infty}^{\infty} e^x \frac{1}{\sqrt{2\pi\sigma^2(t-s)}} e^{-\frac{x^2}{2\sigma^2(t-s)}}\, dx &= \frac{1}{\sqrt{2\pi\sigma^2(t-s)}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\sigma^2(t-s)x}{2\sigma^2(t-s)}\right) dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2(t-s)}} \int_{-\infty}^{\infty} \exp\left(-\frac{\left(x - \sigma^2(t-s)\right)^2 - \sigma^4(t-s)^2}{2\sigma^2(t-s)}\right) dx = \exp\left(\frac{\sigma^2(t-s)}{2}\right),\end{aligned}$$

thus

$$E\left[\exp\left(\sigma B(t) - \frac{\sigma^2 t}{2}\right)\middle|\, \mathcal{F}^+(s)\right] = \exp\left(\sigma B(s) + \frac{\sigma^2(t-s)}{2} - \frac{\sigma^2 t}{2}\right) = \exp\left(\sigma B(s) - \frac{\sigma^2 s}{2}\right). \qquad \square$$

The derived martingales in Propositions 17.1 and 17.2 are of the form $f_k(B(t), t)$ with $f_1(x,t) = x^2 - t$ and $f_2(x,t) = \exp\left(\sigma x - \frac{\sigma^2 t}{2}\right)$. Note that both f_1 and f_2 solve the *backward equation* $\dfrac{\partial u}{\partial t} = -\dfrac{1}{2}\dfrac{\partial^2 u}{\partial x^2}$.

The following formal derivation show that under relatively mild conditions, if f solves the backward equation, then $\{f\left(B(t), t\right) : t \ge 0\}$ is a martingale.

First observe that the transition density $p_t(x,y) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)$ satisfies the *forward equation* $\dfrac{\partial u}{\partial t} = \dfrac{1}{2}\dfrac{\partial^2 u}{\partial y^2}$. (This is the heat equation with diffusivity $\frac{1}{2}$.)

Indeed,

$$\frac{\partial}{\partial t}\left[\frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)\right] = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)\left(-\frac{1}{2t} + \frac{(y-x)^2}{2t^2}\right)$$

and

$$\begin{aligned}\frac{\partial^2}{\partial y^2}\left[\frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)\right] &= \frac{1}{\sqrt{2\pi t}} \frac{\partial}{\partial y}\left[-\frac{y-x}{t} \exp\left(-\frac{(y-x)^2}{2t}\right)\right] \\ &= \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t}\right)\left(-\frac{1}{t} + \frac{(y-x)^2}{t^2}\right).\end{aligned}$$

If one can justify differentiating under the integral, then

$$\begin{aligned}\frac{\partial}{\partial t} E_x\left[f\left(B(t), t\right)\right] &= \int \frac{\partial}{\partial t}\left[f(y,t) p_t(x,y)\right] dy \\ &= \int \left(\frac{\partial}{\partial t} f(y,t)\right) p_t(x,y) dy + \int f(y,t)\left(\frac{\partial}{\partial t} p_t(x,y)\right) dy.\end{aligned}$$

If the following integration by parts steps are valid, then the latter integral is

$$\begin{aligned}\int f(y,t)\left(\frac{\partial}{\partial t} p_t(x,y)\right) dy &= \frac{1}{2}\int f(y,t)\left(\frac{\partial^2}{\partial y^2} p_t(x,y)\right) dy \\ &= -\frac{1}{2}\int \left(\frac{\partial}{\partial y} f(y,t)\right)\left(\frac{\partial}{\partial y} p_t(x,y)\right) dy \\ &= \frac{1}{2}\int \left(\frac{\partial^2}{\partial y^2} f(y,t)\right) p_t(x,y) dy.\end{aligned}$$

Putting these equations together gives

$$\begin{aligned}\frac{\partial}{\partial t}E_x\left[f\left(B(t),t\right)\right] &= \int\left(\frac{\partial}{\partial t}f(y,t)\right)p_t(x,y)dy + \int f(y,t)\left(\frac{\partial}{\partial t}p_t(x,y)\right)dy \\ &= \int\left(\frac{\partial}{\partial t}f(y,t)\right)p_t(x,y)dy + \frac{1}{2}\int\left(\frac{\partial^2}{\partial y^2}f(y,t)\right)p_t(x,y)dy \\ &= \int\left(\frac{\partial}{\partial t}f(y,t) + \frac{1}{2}\frac{\partial^2}{\partial y^2}f(y,t)\right)p_t(x,y)dy = 0\end{aligned}$$

since f satisfies the backward equation.

Finally, the Markov property shows that

$$E\left[f\left(B(t),t\right)\middle|\mathcal{F}^+(s)\right] = E_{B(s)}\left[f\left(B(t),t\right) - f\left(B(s),s\right)\right] + f\left(B(s),s\right) = f\left(B(s),s\right)$$

because $E_x\left[f\left(B(t),t\right)\right]$ is constant in t.

An important example of the above heuristic is $f(x,t) = g(x)$ with $\Delta g = 0$.

That is, under mild integrability assumptions, harmonic functions of Brownian motion are martingales.

18. Donsker's Theorem

Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = 0$ and $\text{Var}(X_1) = 1$, and set $S_n = \sum_{i=1}^n X_i$.

We can extend the random walk S_n from $\mathbb{N}_0$ to $[0,\infty)$ by linear interpolation: $S(t) = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\left(S_{\lfloor t+1 \rfloor} - S_{\lfloor t \rfloor}\right)$.

Scaling diffusively gives a sequence of continuous functions $S^n(t) = S(nt)/\sqrt{n}$.

The CLT implies that the law of $S^n(1)$ converges weakly to $\mathcal{N}(0,1) = \mathscr{L}(B(1))$ where $\{B(t)\}_{t\in[0,1]}$ is a standard Brownian motion.

We will prove the stronger statement that the process $\{S^n(t)\}_{t\in[0,1]}$ converges in distribution to $\{B(t)\}_{t\in[0,1]}$. Thus we can think of Brownian motion as the scaling limit of random walk.

Here we are viewing the processes as random variables taking values in $C[0,1]$ (the space of continuous functions from $[0,1]$ to $\mathbb{R}$ endowed with the uniform metric), so weak convergence means that for every bounded and continuous functional $\Lambda : C[0,1] \to \mathbb{R}$, $\lim_{n\to\infty} E[\Lambda(S^n)] = E[\Lambda(B)]$.

One approach to proving this functional CLT begins by defining $\left\{\widetilde{S}_n(t) : t \geq 0\right\}$ by $\widetilde{S}_n(t) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nt \rfloor} X_k$.

Since $\text{Cov}\left(\widetilde{S}_n(s), \widetilde{S}_n(t)\right) = \frac{\lfloor ns \rfloor \wedge \lfloor nt \rfloor}{n} \to s \wedge t$ for all $s, t \geq 0$, the multivariate CLT shows that for any $t_1, \ldots, t_m \geq 0$, $\left(\widetilde{S}_n(t_1), \ldots, \widetilde{S}_n(t_m)\right) \Rightarrow (B(t_1), \ldots, B(t_m))$ as $n \to \infty$.

Linear interpolation yields the continuous process $S^n(t) = \widetilde{S}_n(t) + \left(\frac{nt - \lfloor nt \rfloor}{\sqrt{n}}\right) X_{\lfloor nt \rfloor + 1}$, and we have that

$$\begin{aligned} P\left(\left|S^n(t) - \widetilde{S}_n(t)\right| > \varepsilon\right) &= P\left(\left|\left(\frac{nt - \lfloor nt \rfloor}{\sqrt{n}}\right) X_{\lfloor nt \rfloor + 1}\right|^2 > \varepsilon^2\right) \\ &\leq \frac{(nt - \lfloor nt \rfloor)^2}{n\varepsilon^2} E\left[X^2_{\lfloor nt \rfloor + 1}\right] \leq \frac{1}{n\varepsilon^2} \to 0, \end{aligned}$$

so Slutsky's theorem shows that the finite dimensional distributions of $\{S^n(t)\}_{t\geq 0}$ converge weakly to those of Brownian motion.

One can then deduce weak convergence of the processes by demonstrating tightness.

Making this argument rigorous involves more analysis than probability, so we'll pursue an alternative approach based on Skorokhod embedding.

Theorem 18.1 (Skorokhod's embedding theorem)**.** *If X is a random variable with mean zero and finite variance, then there is a stopping time T for $\{\mathcal{F}^+(t)\}_{t\geq 0}$ such that $B(T) =_d X$ and $E[T] = E\left[X^2\right]$.*

Note that if X is supported on $\{a, b\}$ with $a < 0 < b$, then the mean zero condition implies that

$$P(X = a) = \frac{b}{b-a}, \; P(X = b) = \frac{-a}{b-a}.$$

Letting $T = T_{a,b} = \inf\{t : B(t) \notin (a,b)\}$, Theorem 17.7 shows that $B(T) =_d X$ and $E[T] = -ab = E\left[X^2\right]$.

We can use this observation to prove Theorem 18.1 by considering an appropriate martingale.

Definition. A martingale $\{X_n\}_{n=0}^\infty$ is *binary splitting* if whenever $x_0, \ldots, x_n \in \mathbb{R}$ is such that the event $A(x_0, \ldots, x_n) = \{X_0 = x_0, \ldots, X_n = x_n\}$ has positive probability, then conditional on $A(x_0, \ldots, x_n)$, X_{n+1} is supported on at most two values.

Lemma 18.1. *If X is a random variable on $(\Omega, \mathcal{F}, P)$ with $E\left[X^2\right] < \infty$, then there is a binary splitting martingale $\{X_n\}$ with $X_n \to X$ a.s. and in L^2.*

Proof. Set $\mathcal{G}_0 = \{\emptyset, \Omega\}$, $X_0 = E[X]$, and $\xi_0 = \begin{cases} 1, & X \geq X_0 \\ -1, & X < X_0 \end{cases}$. Define $\mathcal{G}_n$, X_n, ξ_n recursively by $\mathcal{G}_n = \sigma(\xi_0, \ldots, \xi_{n-1})$, $X_n = E[X|\mathcal{G}_n]$, $\xi_n = \begin{cases} 1, & X \geq X_n \\ -1, & X < X_n \end{cases}$.

Then $\mathcal{G}_n$ is generated by a partition $\mathcal{P}_n$ of Ω into 2^n events, each of which can be expressed as $A(x_0, \ldots, x_n)$. As each element in $\mathcal{P}_n$ is a union of two events in $\mathcal{P}_{n+1}$, we see that $\{X_n\}$ is binary splitting.

Also, Lévy's Forward Theorem gives $X_n \to E[X|\mathcal{G}_\infty]$ a.s. where $\mathcal{G}_\infty = \sigma\left(\bigcup_n \mathcal{G}_n\right)$.

Since X is square-integrable and $E\left[X_n^2\right] = E\left[E[X|\mathcal{G}_n]^2\right] \leq E\left[E\left[X^2|\mathcal{G}_n\right]\right] = E\left[X^2\right]$ for all n, it follows from Theorem 4.5 that X_n converges to $X_\infty = E[X|\mathcal{G}_\infty]$ in L^2 as well.

We'll be done upon showing that $X_\infty = X$ a.s. To see that this is so, we observe that

$$\lim_{n\to\infty} \xi_n (X - X_n) = |X - X_\infty| \text{ a.s.}$$

Indeed, if $X(\omega) = X_\infty(\omega)$, then this is clearly true. If $X(\omega) > X_\infty(\omega)$, then $X(\omega) \geq X_n(\omega)$ for all sufficiently large n, so $\xi_n(\omega) = 1$ eventually and the equality holds. Similarly, if $X(\omega) < X_\infty(\omega)$, then $\xi_n(\omega) = -1$ eventually.

Since $\xi_n \in \mathcal{G}_n$, we see that $E[\xi_n (X - X_n)] = E[\xi_n E[X - X_n|\mathcal{G}_n]] = 0$ for all n.

As $|\xi_n (X - X_n)| \leq |X_n| + |X| \leq |X_\infty| + 1 + |X|$ for large n, the DCT implies $E|X - X_\infty| = 0$. □

Skorokhod's embedding theorem is a simple consequence of the above observations.

Proof of Theorem 18.1. Let $\{X_n\}$ be a binary splitting martingale which converges to X a.s. and in L^2.

We know that if X is supported on $\{a, b\}$ for $a < 0 < b$, then $T_{a,b} = \inf\{t \geq 0 : B(t) \notin (a, b)\}$ gives the requisite stopping time.

Since, conditional on $A(x_0, \ldots, x_{n-1})$, X_n is supported on two such values, we can find $T_0 \leq T_1 \leq \cdots$ such that $B(T_n) =_d X_n$ and $E[T_n] = E\left[X_n^2\right]$.

As T_n is increasing, there is some stopping time T with $T_n \to T$ a.s. and

$$E[T] = \lim_{n\to\infty} E[T_n] = \lim_{n\to\infty} E\left[X_n^2\right] = E\left[X^2\right]$$

by dominated convergence.

Finally, $B(T_n)$ converges in distribution to X by construction, and $B(T_n)$ converges a.s. to $B(T)$ by continuity of Brownian paths, so we have $B(T) =_d X$ as required. □

Now that we have Skorokhod embedding at our disposal, we are ready to start proving

Theorem 18.2 (Donsker's invariance principle)**.** *On the space $C[0, 1]$, the sequence $\{S^n\}$ converges in distribution to a standard Brownian motion.*

The basic idea of the proof is to construct the random variables $X_1, X_2, \ldots$ on the same probability space as the Brownian motion in such a way that $\{S^n\}$ is close to a scaling of the Brownian motion with high probability.

Lemma 18.2. *Suppose that $\{B(t) : t \geq 0\}$ is a standard Brownian motion. Then for any random variable X with mean zero and variance one, there exists a sequence of stopping times $0 = T_0 \leq T_1 \leq T_2 \leq \cdots$ with respect to $\{\mathcal{F}^+(t)\}$ such that*

(1) *The sequence of random variables $\{B(T_n)\}_{n=0}^{\infty}$ has the distribution of the random walk whose increments are distributed as X.*
(2) *For all $\varepsilon > 0$, the sequence of functions $\{S^n\}_{n=0}^{\infty}$ constructed from this random walk satisfies*

$$\lim_{n\to\infty} P\left(\sup_{0\leq t\leq 1}\left|\frac{B(nt)}{\sqrt{n}} - S^n(t)\right| > \varepsilon\right) = 0.$$

Proof. By Skorokhod embedding, we can find a stopping time T_1 for $\{\mathcal{F}^+(t)\}$ with $E[T_1] = E\left[X^2\right] = 1$ and $B(T_1) =_d X$.

The strong Markov property shows that $\{B(T_1 + t) - B(T_1) : t \geq 0\}$ is a standard Brownian motion which is independent of $\mathcal{F}^+(T_1)$ and thus of $(T_1, B(T_1))$.

Accordingly, we can find T_2' with $E[T_2'] = 1$ and $B(T_2') =_d X$. Setting $T_2 = T_1 + T_2'$, we see that $E[T_2] = E[T_1] + E[T_2'] = 2$ and $B(T_2) = (B(T_1 + T_2') - B(T_1)) + B(T_1)$, which is distributed as the second step in a random walk with increment distribution $\mathscr{L}(X)$.

Continuing inductively gives a sequence $T_0 \leq T_1 \leq T_2 \leq \cdots$ where $E[T_n] = n$ and $S_n = B(T_n)$ is the embedded random walk.

To prove the second claim, write $W_n(t) = \frac{B(nt)}{\sqrt{n}}$ and define $A_n = \{|W_n(t) - S^n(t)| > \varepsilon \text{ for some } t \in [0,1]\}$. We must show that $P(A_n) \to 0$.

Let $k = k(t)$ be the unique integer with $\frac{k-1}{n} \leq t < \frac{k}{n}$. Since S^n is linear on $\left[\frac{k-1}{n}, \frac{k}{n}\right)$, we have

$$A_n \subseteq \left\{\left|W_n(t) - \frac{S_k}{\sqrt{n}}\right| > \varepsilon \text{ for some } t \in [0,1]\right\} \bigcup \left\{\left|W_n(t) - \frac{S_{k-1}}{\sqrt{n}}\right| > \varepsilon \text{ for some } t \in [0,1]\right\}.$$

Using $S_k = B(T_k) = \sqrt{n} W_n(T_k/n)$, we see that A_n is contained in

$$A_n^* = \{|W_n(t) - W_n(T_k/n)| > \varepsilon \text{ for some } t \in [0,1]\} \bigcup \{|W_n(t) - W_n(T_k/n)| > \varepsilon \text{ for some } t \in [0,1]\}.$$

Now for any fixed $0 < \delta < 1$, A_n^* is contained in

$$\{|W_n(t) - W_n(s)| > \varepsilon \text{ for some } s,t \in [0,2] \text{ with } |s-t| < \delta\} \bigcup \left\{\left|\frac{T_k}{n} - t\right| \vee \left|\frac{T_{k-1}}{n} - t\right| \geq \delta \text{ for some } t \in [0,1]\right\}.$$

Since $W_n(t) = \frac{B(nt)}{\sqrt{n}} =_d B(t)$, the probability of the first of these events is independent of n, and sample path continuity ensures that we can choose δ small enough to make this probability as small as we wish.

Thus we will be done upon showing that for any $0 < \delta < 1$,

$$P\left(\left|\frac{T_k}{n} - t\right| \vee \left|\frac{T_{k-1}}{n} - t\right| \geq \delta \text{ for some } t \in [0,1]\right) \to 0.$$

To prove this, we note that the strong law implies

$$\lim_{n\to\infty} \frac{T_n}{n} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (T_k - T_{k-1}) = 1 \text{ a.s.}$$

because the random variables $T_k' = T_k - T_{k-1}$ are i.i.d. with mean 1.

Now observe that any sequence of reals with $\frac{a_n}{n} \to 1$ must satisfy $\sup_{0\le k\le n} \frac{|a_k-k|}{n} \to 0$.

* Given $\varepsilon > 0$, we can choose N so that $\left|\frac{a_n}{n} - 1\right| < \varepsilon$ for $n \ge N$, and then choose $M > N$ so that $\max_{0\le k\le N} \left|\frac{a_k-k}{n}\right| < \varepsilon$ for $n \ge M$. Thus when $n \ge M$, we have

$$\max_{0\le k\le n} \left|\frac{a_k-k}{n}\right| = \max_{0\le k\le N} \left|\frac{a_k-k}{n}\right| \vee \max_{N<k\le n} \left|\frac{a_k-k}{n}\right| < \varepsilon \vee \max_{N<k\le n} \frac{k}{n}\left|\frac{a_k}{k} - 1\right| < \varepsilon.$$

Accordingly, we see that

$$\lim_{n\to\infty} P\left(\sup_{0\le k\le n} \left|\frac{T_k-k}{n}\right| \ge \delta\right) = 0.$$

Because $\frac{k(t)-1}{n} \le t < \frac{k(t)}{n}$, whenever $n > 2/\delta$, we have

$$\begin{aligned} P\left(\left|\frac{T_{k(t)}}{n} - t\right| \vee \left|\frac{T_{k(t)-1}}{n} - t\right| \ge \delta \text{ for some } t \in [0,1]\right) \\ \le P\left(\sup_{1\le k\le n} \left|\frac{T_k-(k-1)}{n}\right| \vee \left|\frac{T_{k-1}-k}{n}\right| \ge \delta\right) \\ \le P\left(\sup_{1\le k\le n} \left|\frac{T_k-k}{n}\right| \ge \frac{\delta}{2}\right) + P\left(\sup_{1\le k\le n} \left|\frac{T_{k-1}-(k-1)}{n}\right| \ge \frac{\delta}{2}\right) \to 0. \quad \square \end{aligned}$$

We are now able to give the

Proof of Theorem 18.2. Choose $T_0 \le T_1 \le \cdots$ as in Lemma 18.2 and note that Proposition 14.3 shows that the random functions $W_n \in C[0,1]$ given by $W_n(t) = \frac{B(nt)}{\sqrt{n}}$ are standard Brownian motions.

For any closed set $K \subseteq C[0,1]$, if we let denote its ε-neighborhood by

$$K[\varepsilon] = \{f \in C[0,1] : \|f-g\| \le \varepsilon \text{ for some } g \in K\},$$

then it is clear that

$$\begin{aligned} P(S_n \in K) &\le P(W_n \in K[\varepsilon]) + P(\|S_n - W_n\| > \varepsilon) \\ &= P(B \in K[\varepsilon]) + P(\|S_n - W_n\| > \varepsilon) \to P(B \in K[\varepsilon]) \end{aligned}$$

where B is a standard Brownian motion.

Since K is closed,

$$\lim_{\varepsilon\to 0} P(B \in K[\varepsilon]) = P\left(B \in \bigcap_{n\in\mathbb{N}} K\left[\frac{1}{n}\right]\right) = P(B \in K).$$

Therefore, $\limsup_{n\to\infty} P(S_n \in K) \le P(B \in K)$, showing that $S_n \Rightarrow B$ by the Portmanteau theorem. $\square$

One of the main uses of Donsker's theorem and the Skorokhod embedding theorem is to translate results about random walks into results about Brownian motions and conversely. Standard examples of this sort of reasoning are given by the arcsine laws and the law of the iterated logarithm.

For instance, the law of the iterated logarithm is easier to prove in the continuous setting of Brownian motion, and one derives the corresponding statement for random walk by embedding S_n in B.

Similarly, it is straightforward to prove the arcsine law for the last zero of Brownian motion, and then one can invoke Donsker's theorem to get the analogous statement for random walk. Conversely, combinatorial considerations make it relatively easy to prove an arcsine law for the time simple random walk spends above the x-axis, and one can use Donsker to get the statement about positive times of Brownian motion.

Moreover, Donsker's theorem is an invariance principle, meaning that the statement does not depend on the particulars of the increment distribution. Thus one can prove a result for simple random walk, use Donsker to establish its analogue for Brownian motion, and then convert it back to a result about random walk with arbitrary mean 0 variance 1 increments.

The following two examples are among the simplest illustrations of this line of reasoning.

Example 18.1. Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = 0$ and $E\left[X_1^2\right] = 1$, and set $S_n = \sum_{k=1}^n X_k$.
Define $S^n(t) = S(nt)/\sqrt{n}$ with $S(t) = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\left(S_{\lfloor t+1 \rfloor} - S_{\lfloor t \rfloor}\right)$ as before.

Let $h : \mathbb{R} \to \mathbb{R}$ be bounded and continuous, and define $\Lambda_h : C[0,1] \to \mathbb{R}$ by $\Lambda_h(f) = h(f(1))$. Then $\sup_{f \in C[0,1]} |\Lambda_h(f)| \leq \|h\|_\infty$and if $f_n \to f$ in $C[0,1]$, then $\Lambda_h(f_n) = h(f_n(1)) \to h(f(1)) = \Lambda_h(f)$, hence Λ_h is a bounded, continuous functional.

It follows from Theorem 18.2 that

$$E\left[h\left(\frac{S_n}{\sqrt{n}}\right)\right] = E\left[h\left(S^n(1)\right)\right] = E\left[\Lambda_h\left(S^n\right)\right] \to E\left[\Lambda_h(B)\right] = E\left[h(B(1))\right].$$

Since $h \in C_b(\mathbb{R})$ was arbitrary and $B(1) \sim \mathcal{N}(0,1)$, we have just proved the CLT!

Example 18.2. In the setting of the preceding example, let $M_n = \max_{0 \leq k \leq n} S_k$.

Suppose that $h : \mathbb{R} \to \mathbb{R}$ is bounded and continuous and define $\Gamma_h : C[0,1] \to \mathbb{R}$ by $\Gamma_h(f) = h\Big(\max_{0 \leq x \leq 1} f(x)\Big)$.
As before, it is clear that Γ_h is continuous and bounded.

Since $S(t)$ is the linear interpolation of S_n, we have

$$E\left[\Gamma_h\left(S^n\right)\right] = E\left[h\left(\max_{0 \leq x \leq 1} \frac{S(nx)}{\sqrt{n}}\right)\right] = E\left[h\left(\max_{0 \leq k \leq n} \frac{S_k}{\sqrt{n}}\right)\right].$$

Also, $E\left[\Gamma_h(B)\right] = E\left[h\Big(\max_{0 \leq t \leq 1} B(t)\Big)\right]$ and $\max_{0 \leq t \leq 1} B(t) =_d |B(1)|$ by Theorem 16.11, so Donsker's theorem implies

$$\lim_{n \to \infty} E\left[h\left(\frac{M_n}{\sqrt{n}}\right)\right] = \lim_{n \to \infty} E\left[\Gamma_h\left(S^n\right)\right] = E\left[\Gamma_h(B)\right] = E\left[h(|B(1)|)\right].$$

Thus the definition of weak convergence in terms of distribution functions shows that

$$\lim_{n \to \infty} P\left(M_n > x\sqrt{n}\right) = P(|B(1)| > x) = \sqrt{\frac{2}{\pi}} \int_x^\infty e^{-\frac{t^2}{2}}\, dt.$$

Homework 1

(1) Show that if $X = Y$ on $B \in \mathcal{G}$, then $E[X\,|\mathcal{G}] = E[Y\,|\mathcal{G}]$ a.s. on B.

(2) Suppose $X \geq 0$ and $E[X] = \infty$. Show that there is a unique $Y \in \mathcal{G}$ with $0 \leq Y \leq \infty$ so that

$$\int_A X dP = \int_A Y dP \text{ for all } A \in \mathcal{G}.$$

(Hint: Consider $X_n = X \wedge n$.)

(3) Prove the following conditional limit theorems.

(a) Fatou: If $X_1, X_2, ...$ are nonnegative, then $E\left[\liminf_n X_n\,|\mathcal{G}\right] \leq \liminf_n E\left[X_n\,|\mathcal{G}\right]$ a.s.

(b) DCT: If $X_n \to X$ a.s. and there is an integrable Z with $|X_n| \leq |Z|$, then $E[X_n\,|\mathcal{G}] \to E[X\,|\mathcal{G}]$ a.s.

(4) Give an example on $\Omega = \{a, b, c\}$ in which $E\left[E[X\,|\mathcal{F}_1]\,|\mathcal{F}_2\right] \neq E\left[E[X\,|\mathcal{F}_2]\,|\mathcal{F}_1\right]$.

(5) Suppose that $\mathcal{G}_1 \subseteq \mathcal{G}_2$ and $E[X^2] < \infty$. Show that

$$E\left[\left(X - E[X\,|\mathcal{G}_2]\right)^2\right] \leq E\left[\left(X - E[X\,|\mathcal{G}_1]\right)^2\right].$$

(6) Suppose that $E[X^2] < \infty$, and define $\mathrm{Var}\left(X\,|\mathcal{G}\right) = E[X^2\,|\mathcal{G}] - E[X\,|\mathcal{G}]^2$. Show that

$$\mathrm{Var}(X) = E\left[\mathrm{Var}\left(X\,|\mathcal{G}\right)\right] + \mathrm{Var}\left(E[X\,|\mathcal{G}]\right).$$

(7) Show that if $E[X\,|\mathcal{G}] = Y$ and $E[X^2] = E[Y^2] < \infty$, then $X = Y$ a.s.

(8) Suppose that X and Y have joint density $f(x,y) > 0$. Let

$$\mu(y, A) = \frac{\int_A f(x,y)dx}{\int f(x,y)dx}.$$

Show that $\mu(Y(\omega), A)$ is a r.c.d. for X given $\sigma(Y)$.

(9) Suppose that X and Y take values in a nice space $(S, \mathcal{S})$ and $\mathcal{G} = \sigma(Y)$. Show that there is a function $p : S \times \mathcal{S} \to [0,1]$ such that

(i) For each A, $p\,(Y(\omega), A)$ is a version of $P(X \in A\,|\mathcal{G}\,)$.

(ii) For a.e. ω, $A \to p\,(Y(\omega), A)$ is a probability measure on $(S, \mathcal{S})$.

Homework 2

(1) Let X_n be the position of simple random walk on $\mathbb{Z}$ at time n. That is, $X_n = \sum_{k=1}^n \xi_k$ where $\xi_1, \xi_2, \ldots$ are i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$. Show that $M_n = X_n^3 - 3nX_n$ is a martingale w.r.t. $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$.

(2) Suppose that $\xi_1, \xi_2, ...$ are independent with $E[\xi_i] = 0$ and $E[\xi_i^2] = \sigma_i^2 < \infty$, and set $S_n = \sum_{i=1}^n \xi_i$, $s_n^2 = \sum_{i=1}^n \sigma_i^2$. Show that $S_n^2 - s_n^2$ is a martingale w.r.t. $\mathcal{F}_n = \sigma(\xi_1, ..., \xi_n)$.

(3) Give an example of a (nonconstant) submartingale X_n such that X_n^2 is a supermartingale. Give an example of a martingale X_n with $X_n \to -\infty$ a.s.

(4) Let $(\Omega, \mathcal{F}, P)$ be $[0, 1)$ with the Borel sets and Lebesgue measure. Let $\mathcal{F}_n = \sigma\left(\left\{\left[\frac{j-1}{2^n}, \frac{j}{2^n}\right) : j = 1, ..., 2^n\right\}\right)$ and define $X_n = 2^n 1_{[0,2^{-n})}$. Show that $\{X_n\}$ is a (nonnegative) martingale w.r.t. $\{\mathcal{F}_n\}$. Does X_n converge in L^1?

(5) Suppose X_n^1 and X_n^2 are supermartingales with respect to $\mathcal{F}_n$, and N is a stopping time with $X_N^1 \geq X_N^2$. Show that $Y_n = X_n^1 1\{N > n\} + X_n^2 1\{N \leq n\}$ is a supermartingale

(6) Let X_n be a martingale with $X_0 = 0$ and $E[X_n^2] < \infty$. Show that for all $\lambda \geq 0$

$$P\left(\max_{1 \leq m \leq n} X_m \geq \lambda\right) \leq \frac{E[X_n^2]}{E[X_n^2] + \lambda^2}.$$

(Hint: For any $c \in \mathbb{R}$, $(X_n + c)^2$ is a submartingale.)

(7) Let $\varphi \geq 0$ be any function with $\frac{\varphi(x)}{x} \to \infty$ as $x \to \infty$. Show that $E[\varphi(|X_i|)] \leq C$ for all $i \in I$ implies $\{X_i\}_{i \in I}$ is uniformly integrable.

(8) Let $\xi_1, \xi_2, ...$ be i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$, and set $X_n = \sum_{k=1}^n \frac{\xi_k}{k}$. Show that X_n converges to an integrable random variable X with probability one. In other words, the random harmonic series is a.s. convergent.

(9) Suppose that $X_1, X_2, ...$ are i.i.d. picks from a density f which is either equal to f_0 or f_1, both of which are strictly positive on $\mathbb{R}$. Show that under the null hypothesis $f = f_0$, the test statistic $\Lambda_n = \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}$ converges a.s. as $n \to \infty$.

(10) Let $Z_1, Z_2, ...$ be i.i.d. standard normals, and let θ be an independent random variable with finite mean. Set $Y_n = Z_n + \theta$. In statistical terms, we have a sample from a normal population with unknown mean. The distribution of θ is called the prior distribution and $P(\theta \in \cdot\,|Y_1, ..., Y_n)$ is called the posterior distribution after n observations. Show that $E[\theta\,|Y_1, ..., Y_n] \to \theta$ a.s. (The Bayes estimate is consistent.)

Homework 3

(1) Show that if $\mathcal{F}_n \nearrow \mathcal{F}_\infty$ and $Y_n \to Y$ in L^1, then $E[Y_n \,|\mathcal{F}_n] \to E[Y \,|\mathcal{F}_\infty]$ in L^1.

(2) Give an example of a submartingale $\{X_n\}$ with $\sup_n E\,|X_{n+1} - X_n| < \infty$ and a stopping time N with $E[N] < \infty$ such that $\{X_{n\wedge N}\}$ is not uniformly integrable.

(3) Compute the expected number of tosses of a fair coin until the first occurrence of the patterns $THHTT$, $TTHHT$, and $THTHT$, respectively.

(4) Let $\mathcal{M} = \{f_i : S \to \mathbb{R}\}_{i\in I}$ be a family of bounded functions which is closed under multiplication, and let $\mathcal{C} = \sigma(f_i : i \in I)$ be the smallest σ-algebra on S that makes all of the f_i's measurable. Suppose that $\mathcal{H}$ is a vector space of bounded $\mathbb{R}$-valued functions on S satisfying

- (i) $\mathcal{M} \subseteq \mathcal{H}$
- (ii) $1 \in \mathcal{H}$
- (iii) If $h : S \to \mathbb{R}$ is bounded and there is a sequence of nonnegative functions in $\mathcal{H}$ that increase pointwise to h, then $h \in \mathcal{H}$.

Show that $\mathcal{H}$ contains all bounded functions which are measurable with respect to $\mathcal{C}$.

(5) Suppose that S is a countable set and $p : S \times S \to [0,1]$ satisfies $\sum_{t\in S} p(s,t) = 1$ for all $s \in S$. A *random mapping representation* of p is a function $f : S \times \Lambda \to S$, along with a Λ-valued random variable Z, satisfying $P(f(s,Z) = t) = p(s,t)$ for all $s,t \in S$.

(a) Give a random mapping representation for simple random walk on $\mathbb{Z}$.

(b) Show that if (f,Z) is a random mapping representation for p, $Z_1, Z_2, ...$ are i.i.d. with distribution $\mathscr{L}(Z)$, and $X_0 \sim \mu$, then the sequence $X_0, X_1, ...$ defined by $X_k = f(X_{k-1}, Z_k)$ for $k \in \mathbb{N}$ is a Markov chain with transition function p and initial distribution μ.

(c) Show that every Markov chain on a countable state space has a random mapping representation. (Hint: Let $Z \sim U(0,1)$ and consider the array $F_{i,j} = \sum_{k=1}^{j} p(s_i, s_k)$ where p is the transition function and $\{s_1, s_2, ...\}$ is an enumeration of the state space.)

(6) Give an example of a Markov chain X_n on a countable state space S and a measurable function g on S such that $g(X_n)$ is not a Markov chain. Can you give any conditions on $\{X_n\}$ and g which ensure that $g(X_n)$ is a Markov chain?

(7) Let $S = \{0, 1\}$ and $p = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$. Use induction to show that

$$P_\mu(X_n = 0) = \frac{\beta}{\alpha+\beta} + (1-\alpha-\beta)^n \left[\mu(\{0\}) - \frac{\beta}{\alpha+\beta}\right].$$

(8) Consider the following process: Two animals are mated and among their direct descendants two individuals of opposite sex are selected at random. These individuals are mated and the process continues. Suppose that each individual can be of one of three genotypes, AA, Aa, aa, and suppose that the type of offspring is determined by selecting a letter from each parent. With these rules, the pair of genotypes in the nth generation is a Markov chain with six states,

$$(AA, AA), (AA, Aa), (AA, aa), (Aa, Aa), (Aa, aa), (aa, aa).$$

Compute its transition probability.

(9) Let p be the transition matrix for simple random walk on the n-cycle ($X_k = X_{k-1} + \xi_k \pmod n$ where $\xi_1, \xi_2, \ldots$ are i.i.d. with $P(\xi_1 = 1) = P(\xi_1 = -1) = \frac{1}{2}$) for n odd. Find the smallest value of r so that $p^s(x, y) > 0$ for all $s \geq r$ and all $x, y \in \mathbb{Z}/n\mathbb{Z}$. What if n is even?

Homework 4

Unless otherwise noted, X_n is a Markov chain with countable state space S and transition probability p, and

$$\tau_R = \inf\{n \geq 0 : X_n \in R\},\ \tau_y = \tau_{\{y\}}$$
$$T_R = \inf\{n \geq 1 : X_n \in R\},\ T_y = T_{\{y\}}.$$

(1) Show that for any $x \in S$, $k \in \mathbb{N}$,

$$\sum_{m=0}^{n} P_x(X_m = x) \geq \sum_{m=k}^{n+k} P_x(X_m = x).$$

(2) Suppose that $C \subset S$ with $S \setminus C$ finite. Show that if $P_x(\tau_C < \infty) > 0$ for each $x \in S \setminus C$, then there exist $N < \infty$, $\varepsilon > 0$ such that $P_y(\tau_C > kN) \leq (1-\varepsilon)^k$ for all $y \in S$, $k \in \mathbb{N}$.

(3) Suppose that $A, B \subseteq S$ with $A \cap B = \emptyset$, $S \setminus (A \cup B)$ finite, and $P_x(\tau_{A\cup B} < \infty) > 0$ for all $x \in S \setminus (A \cup B)$. Let $h(x) = P_x(\tau_A < \tau_B)$.

(a) Show that

$$(\star) \quad h(x) = \sum_y p(x,y)h(y) \text{ for } x \notin A \cup B.$$

(b) Show that if h is any bounded function satisfying $(\star)$, then $h\left(X_{n\wedge\tau_{A\cup B}}\right)$ is a martingale.

(c) Conclude that $h(x) = P_x(\tau_A < \tau_B)$ is the unique solution to $(\star)$ that is 1 on A and 0 on B.

(4) f is said to be superharmonic if $f(x) \geq \sum_y p(x,y)f(y)$, or equivalently, if $f(X_n)$ is a supermartingale. Suppose p is irreducible. Prove that p is recurrent if and only if every nonnegative superharmonic function is constant.

(5) Suppose that p is irreducible and has a stationary distribution π. Define the *time reversal* of X_n to be the chain $\widetilde{X}_n$ with transition probabilities

$$\widetilde{p}(x,y) = \frac{\pi(y)p(y,x)}{\pi(x)}.$$

Show that π is stationary for $\widetilde{X}_n$, and for any $x_0, ..., x_t \in S$, we have

$$P_\pi(X_0 = x_0, ..., X_t = x_t) = P_\pi(\widetilde{X}_0 = x_t, ..., \widetilde{X}_t = x_0).$$

(6) We say that a state x is *essential* if $\rho_{xy} > 0$ implies $\rho_{yx} > 0$. Otherwise, x is called *inessential*. Show that if π is a stationary distribution, then $\pi(y) = 0$ for all inessential states y.

(7) Suppose that p is irreducible and positive recurrent. Show that $E_x[T_y] < \infty$ for all $x, y \in S$.

(8) Suppose that p is irreducible and has a stationary measure μ with $\sum_x \mu(x) = \infty$. Show that p is not positive recurrent.

(9) Give an example of a Markov chain with state space S and subsets $B, C \subseteq S$ such that B is irreducible but not closed and C is closed but not irreducible.

(10) Suppose that p is irreducible and let ν be any probability on S. Show that the transition function

$$q(x,y) = \begin{cases} p(x,y)\left(\frac{\nu(y)p(y,x)}{\nu(x)p(x,y)} \wedge 1\right), & y \neq x \\ 1 - \sum_{z \neq x} p(x,z)\left(\frac{\nu(z)p(z,x)}{\nu(x)p(x,z)} \wedge 1\right), & y = x \end{cases}$$

defines a reversible Markov chain with stationary distribution ν.

(11) Compute the expected number of moves it takes a knight to return to its initial position if it starts on the corner of a chessboard, assuming that there are no other pieces on the board and that each time it chooses a move at random from its legal moves.
(Consult the internet for any questions about chess.)

(12) Consider the following Markov chain on $\mathbb{Z}$. When the current state is $i > 0$, the chain moves to $i-1$ with probability p and to $i+1$ with probability $q = 1 - p < p$. When the current state is $j < 0$, the next state is $j+1$ with probability p and $j-1$ with probability q. From 0, the chain moves to ± 1 with equal probability. Compute the stationary distribution of this chain.

Homework 5

(1) Show that the state space of an irreducible Markov chain with period k can be uniquely decomposed as $S = S_1 \sqcup \cdots \sqcup S_k$ where $P(X_{n+1} \in S_{j+1} \,|X_n \in S_j) = 1$ (with the sums in the S indices taken $\bmod k$). Moreover, S cannot be partitioned into more than k sets having this relationship.

(2) Suppose that μ is a probability on a finite group G with support $\Sigma = \{g \in G : \mu(g) > 0\}$. The (right-invariant) random walk on G driven by μ has transition probabilities $p(g,h) = \mu(hg^{-1})$. We have seen that this chain is irreducible precisely when Σ generates G. Assuming irreducibility, show that the walk is aperiodic if and only if Σ is not contained in a coset of a proper normal subgroup of G.
(By definition, the period is the greatest common divisor of $\{m : e = s_m \cdots s_1$ for some $s_1, \ldots, s_m \in \Sigma\}$ where e denotes the identity in G).

(3) Show that if μ and ν are probabilities on a countable set S and we define w on $S \times S$ by

$$w(z,z) = \min\{\mu(z), \nu(z)\},$$
$$w(x,y) = \frac{(\mu(x) - w(x,x))\,(\nu(y) - w(y,y))}{1 - \sum_z w(z,z)},$$

then $(X,Y) \sim w$ is a coupling of μ and ν with $P(X \neq Y) = \|\mu - \nu\|_{TV}$.

(4) Let p be a transition probability for a Markov chain with countable state space S and stationary distribution π. We write μp^t for the distribution of X_t when $X_0 \sim \mu$, and p_x^t for the distribution of X_t when $X_0 = x$. Let $\mathcal{P}$ be the collection of probability measures on S. Show that

(a) $\sup_{\mu \in \mathcal{P}} \left\|\mu p^t - \pi\right\|_{TV} = \sup_{x \in S} \left\|p_x^t - \pi\right\|_{TV}$.

(b) $\sup_{x,y \in S} \left\|p_x^t - p_y^t\right\|_{TV} \geq \sup_{x \in S} \left\|p_x^t - \pi\right\|_{TV}$.

(5) Let μ and ν be probabilities on a countable set S and suppose that p is a transition probability for a chain with state space S. Prove that

$$\|\mu p - \nu p\|_{TV} \leq \|\mu - \nu\|_{TV}\,.$$

In particular, this shows that if π is a stationary distribution for p, then $\left\|p_x^{n+1} - \pi\right\|_{TV} \leq \|p_x^n - \pi\|_{TV}$.

(6) Lazy random walk on the n-cycle is defined by $p(x,x) = \frac{1}{2}$, $p(x,x+1) = p(x,x-1) = \frac{1}{4}$ where all additions are modulo n. As an irreducible random walk on a finite group, the stationary distribution, π, is uniform on $\mathbb{Z}/n\mathbb{Z}$.

Set $X_0 = x$, $Y_0 = y$ and define X_k, Y_k as follows: Let $U_1, U_2, \ldots$ and $V_1, V_2, \ldots$ be i.i.d. uniform on $\{-1, 1\}$, independent of each other. While $X_{k-1} \neq Y_{k-1}$, if $U_k = 1$, set $X_k = X_{k-1} + V_k \pmod n$, $Y_k = Y_{k-1}$, and if $U_k = -1$, set $X_k = X_{k-1}$, $Y_k = Y_{k-1} + V_k \pmod n$. If $X_{k-1} = Y_{k-1}$, set $X_k = Y_k = X_{k-1} + \frac{1}{2}(U_k + 1)\, V_k \pmod n$.

In words, at each time step, we flip a fair coin to decide whether to move the first chain or the second according to ordinary simple random walk on the cycle. Once the two chains meet, they couple and evolve together ever after.

Use the coupling lemma and Exercise 4 to show that $\|p_x^t - \pi\|_{TV} \leq \frac{1}{4c}$ whenever $t \geq cn^2$.

(7) Consider the following method of shuffling a deck of n cards: At each stage, choose a card uniformly at random and place it at the top of the deck. Use a coupling argument to show that the deck is completely mixed once every card has been chosen at least once. Conclude that the mixing time is $O(n \log(n))$.

(8) Suppose that X and Y are independent normals with mean 0 and variance σ^2. Show that $X + Y$ and $X - Y$ are independent normals with mean 0 and variance $2\sigma^2$.
(Hint: The standard Gaussian distribution is rotationally invariant.)

Homework 5.5

(1) Let $(\Omega, \mathcal{F}, P, T)$ be a probability preserving dynamical system. Recall that an event $A \in \mathcal{F}$ is called invariant if the symmetric difference of A and $T^{-1}A$ has probability zero. A random variable X is invariant if $X \circ T = X$ a.s.
Show that $\mathcal{I} = \{A \in \mathcal{F} : A \text{ is invariant}\}$ is a sub-σ-field of $\mathcal{F}$, and X is measurable with respect to $\mathcal{I}$ if and only if X is invariant.

(2) Show that T is ergodic if and only if for every $A \in \mathcal{F}$ with $P(A) > 0$, we have $\bigcup_{n=0}^{\infty} T^{-n}A = \Omega$ (up to null sets).

(3) Show that T is ergodic if and only if for every $A, B \in \mathcal{F}$ with $P(A), P(B) > 0$, there is a $j \in \mathbb{N}$ such that $P(A \cap T^{-j}B) > 0$

(4) Let $(\Omega, \mathcal{F}, P)$ be $[0, 1)$ with the Borel sets and Lebesgue measure. We saw in class that if $d \in \mathbb{N}$, then the map $T_d x = dx \pmod 1$ is probability preserving (and ergodic).
Show that that if $\beta = \frac{1+\sqrt{5}}{2}$, then $T_\beta x = \beta x \pmod 1$ does not preserve Lebesgue measure, but does preserve the probability Q defined by

$$Q(B) = \int_B g(x)dx \text{ with } g(x) = \begin{cases} \frac{1}{\beta^{-1}+\beta^{-3}}, & 0 \le x < \beta^{-1} \\ \frac{1}{\beta(\beta^{-1}+\beta^{-3})}, & \beta^{-1} \le x < 1 \end{cases}.$$

(Hint: $\beta^{-1} = \beta - 1$, $\beta^2 = \beta + 1$, and $\{[0, a) : a \in [0, 1)\}$ generates $\mathcal{F}$.)

(5) Let $(\Omega, \mathcal{F}, P, T)$ be a probability preserving dynamical system. Show that T is ergodic if and only if

$$\frac{1}{n}\sum_{k=0}^{n-1} P\left(U \cap T^{-k}V\right) \to P(U)P(V)$$

for all $U, V \in \mathcal{F}$.

(6) Show that if $f \in L^p$, $1 \le p < \infty$, then $\frac{1}{n}f^n \to E[f\,|\mathcal{I}]$ in L^p.

Homework 6

(1) Let $\{B_t\}_{t\geq 0}$ be a standard Brownian motion and let $a < 0 < b$. Define $T(x,y) = \inf\{t \geq 0 : B_t \in \{x,y\}\}$. Show that $E[T(a,b)] = a^2 E\left[T\left(1, \frac{b}{a}\right)\right]$, hence the expected exit time from a symmetric interval $[-b,b]$ is a constant multiple of b^2.

(2) Suppose that $\{B_t\}_{t\in[0,T]}$ is a Brownian motion on $[0,T]$. Show that the time reversed process $\{B_{T-t} - B_T\}_{t\in[0,T]}$ is a standard Brownian motion on $[0,T]$.

(3) Using the Lévy construction of Brownian motion given in class, show that if $f : [0,1] \to \mathbb{R}$ is continuous with $f(0) = 0$ and $\{B(t) : t \geq 0\}$ is a standard Brownian motion, then for any $\varepsilon > 0$,

$$P\left(\sup_{0\leq t\leq 1} |B(t) - f(t)| < \varepsilon\right) > 0.$$

(4) Let $\{B_t\}_{t\geq 0}$ be a standard Brownian motion. Prove that

$$\sup_{0\leq s<t\leq 1} \frac{|B_t - B_s|}{|t-s|^\gamma} = \infty \text{ a.s.}$$

whenever $\gamma \geq \frac{1}{2}$.

(5) Consider a (not necessarily nested) sequence of partitions $0 = t_0^{(n)} \leq t_1^{(n)} \ldots \leq t_{k(n)}^{(n)} = t$ with mesh converging to 0.

(a) Show that

$$\lim_{n\to\infty} \sum_{j=1}^{k(n)} \left(B\left(t_j^{(n)}\right) - B\left(t_{j-1}^{(n)}\right)\right)^2 \to t \text{ in } L^2.$$

(We say that Brownian motion has *quadratic variation* $V_B^{(2)}(t) = t$.)

(b) Show that if the sequence of partitions satisfies $\sum_{n=1}^{\infty} \sum_{j=0}^{k(n)} \left(t_j^{(n)} - t_{j-1}^{(n)}\right)^2 < \infty$, then the convergence in part (a) is almost sure.
(An example is partitioning $[0,1]$ with $t_j^{(n)} = \frac{j}{2^n}$, $j = 0, 1, ..., k(n) = 2^n$.)

(c) Argue that $\sum_{j=1}^{k(n)} \left|B\left(t_j^{(n)}\right) - B\left(t_{j-1}^{(n)}\right)\right| \to \infty$ a.s.

(6) A *standard Brownian bridge* is a Gaussian process $\{X(t) : 0 \leq t \leq 1\}$ with continuous paths, mean 0 and covariance $\mathrm{Cov}\,(X(s), X(t) =) s(1-t)$ for $0 \leq s \leq t \leq 1$. If $\{B(t) : t \geq 0\}$ is a standard Brownian motion, verify that the following processes are Brownian bridges.

(a) $X_1(t) = B(t) - tB(1)$

(b) $X_2(t) = (1-t)B\left(\frac{t}{1-t}\right) 1_{[0,1)}(t)$

(7) Let $\{B(t) : t \geq 0\}$ be a standard Brownian motion and let T be a stopping time with $E[T] < \infty$. Define a sequence of stopping times by $T_1 = T$, $T_n = T(B_n) + T_{n-1}$ where $T(B_n)$ is the same function as T but associated with the Brownian motion $B_n(t) = B(t + T_{n-1}) - B(T_{n-1})$.

(a) Show that $\lim_{n\to\infty} \frac{B(T_n)}{n} = 0$ a.s.

(b) Show that $B(T)$ is integrable.

(c) Show that $\lim_{n\to\infty} \frac{B(T_n)}{n} = E\,[B(T)]$. (Combined with (a), this gives Wald's lemma.)