Stat212 Lecture Notes

James Bailie

July 5, 2021

Abstract

These are my lecture notes for Stat212, a second graduate course in probability lectured by Prof. Subhabrata Sen in Spring 2021 at Harvard. All errors are my own. Sections marked add-on were not in the lecture and were added by me at a later point. Some diagrams are courtesy of Subhabrata Sen.

Contents

1 Lecture 26/1

1.1 Background

Definition 1.1. A *stochastic process* is a set of random variables $\{X_t : t \in \pi\}$ indexed by π. The *index set* π is usually unidimensional ($\mathbb{N}, \mathbb{R}^{\geq 0}$) and typically indexes time or 'system complexity' but it can be multidimensional (e.g. a spatial stochastic process). The *state space* is the smallest set $\mathcal{S}$ such that $X_t(\omega) \in \mathcal{S}$ for all t and ω, except possibly ω in a set of measure zero.

The aim of this course is to understand the long term behaviour of stochastic processes; we will learn how to predict, model and do inference on stochastic processes.

1.1.1 Measure theory

We require a small amount of measure theory so we can work on concrete foundations. But we will only touch on this once and then move on.

Set up: Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a measurable function. (Throughout we use the notation $\mathcal{B}_{\mathbb{R}}$ to denote the Borel σ-algebra on the real line $\mathbb{R}$.) We say that X is a *random variable*.

Definition 1.2. The *law* of a random variable X is the function $P(X \in \cdot) : \mathcal{B}_{\mathbb{R}} \to [0, 1]$ defined by

$$P(X \in A) = \mathbb{P}\left(\left\{\omega : \omega \in X^{-1}(A)\right\}\right),$$

for $A \in \mathcal{B}_{\mathbb{R}}$.

Remark 1.3. The law of a random variable is a probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

The law completely describes the random variable. "If you know the law, you can forget about the underlying definition of the random variable in terms of the probability space Ω."

Example 1.4. Let $\Omega = [0, 1]$, $\mathcal{F} = B_{[0,1]}$ and $\mathbb{P}$ be the Lebesgue measure. Define $X(\omega) = \omega$. Then X is a Unif$([0, 1])$ random variable.

From herein, it is not necessary to define the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ in order to get a Unif$([0, 1])$ random variable. If we

want a Unif$([0,1])$ random variable Z, we can specify Z by specifying that it has the same law as X. Moreover, we can now construct any other random variable from its law (or just its CDF) by using the Probability Integral Transform (PIT). "We only have to do this formalism once!"

Example 1.5. What about a countable number of random variables? Is "let $X_1, X_2, \ldots$ be iid Unif$([0,1])$ random variables" a well-defined statement? That is, can we construct a countably infinite collection of random variables $X_i : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ such that any finite number of the X_i's are independent? Yes, if we have one $U \sim$ Unif$([0,1])$ then we can construct a countable number of random variables. If $U = 0.a_1a_2a_3\ldots$ is the binary expansion of U then define

$$\begin{aligned} X_1 &= 0.a_1a_3a_5\ldots \\ X_2 &= 0.a_2a_6a_{10}\ldots \\ X_3 &= 0.a_4a_8a_{12}\ldots \\ &\vdots \end{aligned}$$

Example 1.6. "Let $\{X_t : t \in [0,1]\}$ be iid Unif$([0,1])$ random variables". Is this a legal statement? It turns out that we can't construct an uncountable collection of random variables from a single random variable. We need stronger machinery – the Kolmogorov existence criteria. While this is an important foundation for the course (since later we will examine Brownian motion which is an uncountable collection of random variables – in fact we will construct Brownian motion using only a countably infinite number of standard Gaussian random variables – in essence this proves that we can construct an uncountable collection of random variables), it suffices to know that we can construct an uncountable number of random variables; we will not think about this further.

1.1.2 Asymptotic properties

Let $\{X_n : n \geq 1\}$ be a collection of random variables with n unbounded above. (n usually denotes time or the 'system' size/complexity.) Recall the following definitions of stochastic convergence:

1. *Convergence in probability*: $X_n \xrightarrow{P} X$ if for all $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| < \epsilon) \to 0,$$

as $n \to \infty$.

2. *Convergence in distribution*: $X_n \xrightarrow{d} X$ if the CDF F_n of X_n converges pointwise to the CDF of X, for all points $x \in C(X)$. ($C(X)$ is the continuity points of X's CDF.)

3. *Convergence almost surely*: $X_n \xrightarrow{a.s.} X$ if

$$\mathbb{P}\left(\{\omega : X_n(\omega) \to X(\omega)\}\right) = 1.$$

(That is, X_n converges pointwise to X on a set of measure 1.)

Note that for definitions 1. and 3., we require that X_n and X are defined on the same probability space. If X is a constant, then all three definitions are equivalent.

Example 1.7. An example in modern research: Random matrices. Define $\boldsymbol{W} = (W_{ij}) \in \mathbb{R}^{n \times n}$ with

1. $\{W_{ij} : i < j\} \overset{iid}{\sim} \mathcal{N}(0, 1)$,

2. $\{W_{ii} : i \in [n]\} \overset{iid}{\sim} \mathcal{N}(0, 2)$ independent of W_{ij} for $i \neq j$,

3. W symmetric.

Let X_n be the largest eigenvalue of $W/\sqrt{n}$. Then $X_n \xrightarrow{P} 2$, even though there is no closed form expression for the largest eigenvalue in terms of the elements W_{ij}. This is an analogue of the LLN.

1.2 Martingales

Definition 1.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. $\{\mathcal{F}_n : n \in \mathbb{N}\}$ is a *filtration* if $\{\mathcal{F}_n\}$ is a sequence of nested σ-algebras – that is, for all $n \in \mathbb{N}$,

1. $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \ldots \subset \mathcal{F}$, and

2. $\mathcal{F}_n$ is a σ-algebra.

$\{X_n : n \in \mathbb{N}\}$ is *adapted* to a filtration $\{\mathcal{F}_n\}$ if X_n remains measurable when $\mathcal{F}$ is restricted to $\mathcal{F}_n$ – that is, for all $A \in \mathcal{B}_{\mathbb{R}}$, the event $\{X_n \in A\} \in \mathcal{F}_n$.

An adapted sequence $\{(X_n, \mathcal{F}_n) : n \in \mathbb{N}\}$ is a *martingale* if

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n. \tag{1}$$

Replacing the equality in equation (1) with $\geq$ gives the definition of a sub-martingale, and $\leq$ gives the definition of a super-martingale.

Proposition 1.9 (add-on)**.** *Let $X_1, X_2, \ldots$ be a sequence of random variables and define $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Given $M_1, M_2, \ldots \in L^1$, $\{(M_n, \mathcal{F}_n) : n \in \mathbb{N}\}$ is a martingale if and only if*

(i) M_n is a function of $X_1, \ldots, X_n$ for all n; and

(ii) $\mathbb{E}[M_{n+1}|X_1, \ldots, X_n] = M_n$.

The proof of Proposition 1.9 is left as an exercise.

2 Lecture 28/1

2.1 More on martingales

Proposition 2.1 (add-on)**.** *Let $X \in L^1$ and $\{\mathcal{F}_n : n \in \mathbb{N}\}$ be a filtration. Define*

$$Z_n = \mathbb{E}[Z|\mathcal{F}_n].$$

Then $\{Z_n : n \in \mathbb{N}\}$ is a martingale (called a Doob martingale*) with respect to the filtration $\{\mathcal{F}_n\}$.*

Theorem 2.2 (Martingale Convergence Theorem)**.** *Suppose $\{(X_n, \mathcal{F}_n) : n \in \mathbb{N}\}$ is a martingale with*

$$\sup_{n \geq 1} \mathbb{E}X_n^+ < \infty.$$

Then there exists a random variable X_∞ such that $X_n \xrightarrow{a.s.} X_\infty$.

Notation: Given a function X, define $X^+ = \max(0, X)$.

This theorem gives us a condition for convergence, yet it doesn't say anything about the limit object X_∞. What can we say about X_∞? Do we know anything about its moments? We would hope that

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X_\infty, \tag{2}$$

yet this is not true in general. A sufficient condition for equation (2) is that $\{X_n\}$ is L^1-convergent.

This motivates us to investigate L^1-convergence of martingales. We will see that L^1-convergence holds if $\{X_n\}$ satisfies a particular property called uniform integrability.

2.2 Uniform integrability

Motivation: Let

$$X \in L^1(\Omega, \mathcal{F}, \mathbb{P}) := \{X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_\mathbb{R}) | X \text{ is measurable and } \mathbb{E}|X| < \infty\}.$$

Then $\lim_{c\to\infty} \mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right] = 0$ using the dominated convergence theorem. (Why? The dominating function is $|X|$ and we know that $|X|\mathbb{1}\{|X| \geq c\} \to 0$ pointwise as $c \to \infty$.)

Now consider $S \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$. The convergence of $\mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right]$ may not be uniform for $X \in S$, so that

$$\sup_{X\in S} \mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right]$$

may not go to zero as $c \to \infty$.

Definition 2.3. $S \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$ is *uniform integrable* if for all $\epsilon > 0$, there exists $c > 0$ such that

$$\sup_{X\in S} \mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right] < \epsilon,$$

or equivalently $\lim_{c\to\infty} \sup_{X\in S} \mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right] = 0$.

The property is uniform in S since ϵ and c are uniform across all $X \in S$.

Remark 2.4. If S is uniform integrable then S is L^1 bounded – that is, there exists $M \in \mathbb{R}$ such that $\sup_{X \in S} \mathbb{E}|X| < M$. The converse is not true. (As a counterexample, construct a sequence $\{X_n\}$ of L^1 bounded random variables whose tails are increasing: $\Omega = [0,1]$ with Lebesgue measure and $X_n = n\mathbb{1}_{[0,\frac{1}{n}]}$ suffices.) The proofs of these statements are left as exercises.

Lemma 2.5. *$S \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$ is uniformly integrable if and only if*

(a) S is L^1 bounded (i.e. $\sup_{X \in S} \mathbb{E}|X| < \infty$); and

(b) For all $\epsilon > 0$, there exists $\delta > 0$ such that for all $X \in S$ and $A \in \mathcal{F}$ with $\mathbb{P}(A) < \delta$, we have

$$\mathbb{E}\left[|X|\mathbb{1}_A\right] < \epsilon.$$

Condition (b) is called uniform absolute continuity. Intuitively it ensures that the contribution of $|X|$ coming from negligible sets A is uniformly bounded for all $X \in S$.

Proof. Suppose S is uniformly integrable. Then (a) holds by Remark 2.4. To prove (b),

$$\begin{aligned}
\mathbb{E}\left[|X|\mathbb{1}_A\right] &= \mathbb{E}\left[|X|\mathbb{1}\{A \cap \{|X| \le c\}\}\right] + \mathbb{E}\left[|X|\mathbb{1}\{A \cap \{|X| > c\}\}\right] \\
&\le c\mathbb{P}(A) + \mathbb{E}\left[|X|\mathbb{1}\{|X| > c\}\right] \\
&\le \epsilon/2 + \epsilon/2,
\end{aligned}$$

by choosing δ small enough and c large enough.

In the opposite direction,

$$\mathbb{P}\left(|X| > c\right) \le \frac{\mathbb{E}|X|}{c} \le \frac{M}{c},$$

where the first inequality is Markov's and M is the (uniform) L^1 bound of S. Fix $\epsilon > 0$ and choose $\delta > 0$ given by (b). Then choose c large enough so that $M/c < \delta$. Then apply (b) with $A = \{|X| > c\}$ so that

$$\mathbb{E}\left[|X|\mathbb{1}\{|X| > c\}\right] < \epsilon,$$

for all $X \in S$. □

Question: what are some uniformly integrable collections of random variables? Obviously, any finite set of L^1 random variables is uniformly integrable. The following Lemma gives another example.

Lemma 2.6. *Suppose $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Then*

$$S = \{\mathbb{E}[X|\mathcal{C}] : \mathcal{C} \subset \mathcal{F} \text{ is a } \sigma\text{-algebra}\},$$

is uniformly integrable.

Proof.

$$\begin{aligned}
\mathbb{E}\Big[|\mathbb{E}(X|\mathcal{C})|\mathbb{1}\big\{|\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] &\leq \mathbb{E}\Big[\mathbb{E}(|X||\mathcal{C})\mathbb{1}\big\{|\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&= \mathbb{E}\Big[|X|\mathbb{1}\big\{|\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&= \mathbb{E}\Big[|X|\mathbb{1}\big\{|X| > d, |\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\quad + \mathbb{E}\Big[|X|\mathbb{1}\big\{|X| \leq d, |\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\leq \mathbb{E}\Big[|X|\mathbb{1}\big\{|X| > d, |\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\quad + d\mathbb{P}\Big[|\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\leq \mathbb{E}\Big[|X|\mathbb{1}\big\{|X| > d, |\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\quad + d\frac{\mathbb{E}\left[|\mathbb{E}(X|\mathcal{C})|\right]}{c} \\
&\leq \mathbb{E}\Big[|X|\mathbb{1}\big\{|X| > d, |\mathbb{E}(X|\mathcal{C})| \geq c\big\}\Big] \\
&\quad + d\frac{\mathbb{E}[|X|]}{c},
\end{aligned}$$

where the first line is Jensen's inequality, the second line is Adam's law, the second last line is Markov's inequality and the last line follows from Jensen's inequality and Adam's law.

Given $\epsilon > 0$, choose δ large enough such that the first term is at most $\epsilon/2$. Then choose c large enough such that the second term is at most $\epsilon/2$. □

Intuition of the above lemma: the condition expectations are mostly governed by X. So we can control everything in S by one random variable X – i.e. we have uniformity.

The following Lemma gives one useful method for checking uniform integrability.

Lemma 2.7. *If $S \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$ is L^2-bounded (that is, there exists $M > 0$ such that $\sup_{X \in S} \mathbb{E}X^2 \leq M$), then S is uniformly integrable.*

Proof.

$$\begin{aligned}\mathbb{E}\left[|X|\mathbb{1}\{|X| \geq c\}\right] &\leq \sqrt{\mathbb{E}(X^2)\mathbb{P}(|X| \geq c)} \\ &\leq \sqrt{\mathbb{E}X^2 \frac{\mathbb{E}X^2}{c^2}} \\ &\leq \frac{M}{c},\end{aligned}$$

where the first inequality is Cauchy-Schwartz, the second Chebychev's and the third follows from L^2-boundedness. □

Remark 2.8. The same proof works if S is L^p bounded for some $p > 1$. (Use Hölder's inequality instead of Cauchy-Schwarz.)

2.2.1 Uniform integrability and L^1 convergence

Uniform integrability is a useful concept since it is a sufficient condition for L^1 convergence.

Theorem 2.9. *Let $X_n \in L^1$ for all $n \in \mathbb{N}$ and suppose $X_n \xrightarrow{a.s.} X$. Then the following statements are equivalent:*

i) $\{X_n\}$ *is uniform integrable.*

ii) $X \in L^1$ *and* $X_n \xrightarrow{L^1} X$.

iii) $X \in L^1$ *and* $\mathbb{E}|X_n| \to \mathbb{E}|X|$.

Proof. "i) $\Rightarrow$ ii)":

$$\mathbb{E}|X| \leq \liminf_{n\to\infty} \mathbb{E}|X_n| \leq M < \infty,$$

where the first inequality is Fatou's lemma and the second inequality follows from L^1 boundedness of $\{X_n\}$. So $X \in L^1$.

Define $X_n^c = X_n \mathbb{1}\{|X_n| < c\}$ and $X^c = X\mathbb{1}\{|X| < c\}$. By the triangle inequality,

$$|X_n - X| \leq |X_n^c - X^c| + |X_n|\mathbb{1}\{|X_n| \geq c\} + |X|\mathbb{1}\{|X| \geq c\}.$$

We know that $|X_n^c - X^c| \to 0$ almost surely (by assumption) and $|X_n^c - X^c| \leq 2c$. Hence DCT says $\mathbb{E}|X_n^c - X^c| \to 0$ for all $c > 0$.

Uniform integrability gives us $\sup_{n\geq 1} \mathbb{E}\left[|X_n|\mathbb{1}\{|X_n| \geq c\}\right] \to 0$ as $c \to \infty$. DCT then also says $\mathbb{E}\left[|X|\mathbb{1}\{|X| > c\}\right] \to 0$ as $c \to \infty$. Thus,

$$\limsup_{n\to\infty} \mathbb{E}|X_n - X| = 0.$$

"ii) $\Rightarrow$ iii)":

$$|\mathbb{E}|X_n| - \mathbb{E}|X|| \leq \mathbb{E}||X_n| - |X|| \leq \mathbb{E}|X_n - X| \to 0,$$

where the first inequality is Jensen's and the second is the reverse triangle.

"iii) $\Rightarrow$ i)": We have

$$\mathbb{E}\left(|X_n|\mathbb{1}\{|X_n| \geq c\}\right) = \mathbb{E}|X_n| - \mathbb{E}\left(|X_n|\mathbb{1}\{|X_n| < c\}\right).$$

We know that the first term $\mathbb{E}|X_n| \to \mathbb{E}|X|$. For the second term, note

$$|X_n|\mathbb{1}\{|X_n| < c\} \xrightarrow{\text{a.s.}} |X|\mathbb{1}\{|X| < c\},$$

(by assumption). Then we can apply the bounded convergence theorem to show the second term converges to $\mathbb{E}\left(|X|\mathbb{1}\{|X| < c\}\right)$. Hence

$$\mathbb{E}\left(|X_n|\mathbb{1}\{|X_n| \geq c\}\right) \to \mathbb{E}|X| - \mathbb{E}\left(|X|\mathbb{1}\{|X| < c\}\right) = \mathbb{E}\left(|X|\mathbb{1}\{|X| \geq c\}\right).$$

Since $X \in L^1$, we know $\mathbb{E}\left(|X|\mathbb{1}\{|X| \geq c\}\right) < \epsilon$ for large enough c. Hence

$$\mathbb{E}\left(|X_n|\mathbb{1}\{|X_n| \geq c\}\right) \to \epsilon,$$

and so for large enough n,

$$\mathbb{E}\left(|X_n|\mathbb{1}\{|X_n| \geq c\}\right) < 2\epsilon.$$

We have got a uniform bound for all but finitely many n. Hence $\{X_n\}$ is uniformly integrable. □

2.2.2 Uniform integrability and martingales

Theorem 2.10. *Let $\{(X_n, \mathcal{F}_n) : n \in \mathbb{N}\}$ be a martingale. Define $M = \sup_{n\in\mathbb{N}} \mathbb{E}|X_n|$ (allowing for the possibility that $M = \infty$) and*

$$X_\infty(\omega) = \begin{cases} \lim_{n\to\infty} X_n(\omega) & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases}$$

Then the following statements are equivalent:

1) $M < \infty$ and $X_n \to X_\infty$ almost surely and in L^1.

2) $\mathbb{E}[X_\infty|\mathcal{F}_n] = X_n$ (so $\{X_n\}$ is a Doob martingale).

3) $M < \infty$ and $\mathbb{E}|X_\infty| = \lim_{n\to\infty} \mathbb{E}|X_n|$.

4) $\{X_n\}$ is uniformly integrable.

If $M < \infty$ (in fact we only need $\sup_{n\in\mathbb{N}} \mathbb{E}X_n^+ < \infty$), then $X_n \to X_\infty$ a.s. by the martingale convergence theorem. This theorem tells us that if a martingale is uniformly integrable then it converges a.s. and in L^1. Moreover, the only uniformly integrable martingales are Doob martingales.

3 Lecture 2/2

3.1 Proof of Theorem 2.10

Proof of Theorem 2.10. "1) ⇒ 2)": We know that $\mathbb{E}[X_{n+k}\mathbb{1}_A] = \mathbb{E}[X_n\mathbb{1}_A]$ for all $A \in \mathcal{F}_n$ by the martingale property (and the tower law). Moreover $X_{n+k}\mathbb{1}_A \to X_\infty\mathbb{1}_A$

almost surely and in L^1 as $k \to \infty$. L^1 convergence says

$$\mathbb{E}[X_\infty \mathbb{1}_A] = \lim_{k\to\infty} \mathbb{E}[X_{n+k}\mathbb{1}_A] = \mathbb{E}[X_n \mathbb{1}_A].$$

So $\mathbb{E}[(X_\infty - X_n)\mathbb{1}_A] = 0$ for all $A \in \mathcal{F}_n$.Yet this is precisely the definition of conditional expectation; so $\mathbb{E}[X_\infty|\mathcal{F}_n] = X_n$ almost surely.

"2) $\Rightarrow$ 4)": By Lemma 2.6, a Doob martingale is uniformly integrable.

"4) $\Rightarrow$ 3)": Since $\{X_n\}$ is uniformly integrable, $\{X_n\}$ is L^1 bounded (Lemma 2.5). Thus $M < \infty$. This implies $X_n \xrightarrow{a.s.} X_\infty$ by the martingale convergence theorem. 3) then follows by Theorem 2.9.

"3) $\Rightarrow$ 1)": $X_n \xrightarrow{a.s.} X_\infty$ by the bounded convergence theorem. L^1 convergence then follows by Theorem 2.9. □

3.2 Martingales and L^p convergence ($p > 1$)

Theorem 3.1. *Let $\{(X_n, \mathcal{F}_n) : n \in \mathbb{N}\}$ be a martingale and $p > 1$. Suppose that $\{X_n\}$ is L^p-bounded:*

$$\sup_{n\in\mathbb{N}} \|X_n\|_p < \infty,$$

where $\|\cdot\|_p$ is the L^p norm. Then $X_n \to X_\infty$ almost surely and in L^p.

We need the following lemma to prove this Theorem.

Lemma 3.2. *Let $\{X_n\}$ be an L^p bounded martingale and define $X^* := \sup|X_n|$. Then $X^* \in L^p$.*

We will defer the proof of this lemma.

Proof of Theorem 3.1. $\{X_n\}$ is L^p bounded implies that $\{X_n\}$ is uniformly integrable by Remark 2.8. This implies that $X_n \xrightarrow{a.s.} X_\infty$ by Theorem 2.10 and so

$$|X_n - X_\infty|^p \xrightarrow{a.s.} 0.$$

Now

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

where the first inequality follows from the triangle inequality and bounding $|X_n| + |X_\infty| \leq 2\max(|X_n|, |X_\infty|)$. Assuming Lemma 3.2, $X_n \xrightarrow{L^p} X_\infty$ by the dominated convergence theorem. □

3.3 Proof of Lemma 3.2 using Doob's inequalities

Lemma 3.3 (Doob's maximal inequality)**.** *If* $\{X_n : n \in \mathbb{N}\}$ *is a submartingale and* $\lambda > 0$*, then*

$$\lambda \mathbb{P}\left[\max_{0\leq j\leq n} X_j \geq \lambda\right] \leq \mathbb{E}\left[X_n \mathbb{1}\left\{\max_{0\leq j\leq n} X_j \geq \lambda\right\}\right] \leq \mathbb{E}(X_n^+) \leq \mathbb{E}|X_n|$$

for any $n \in \mathbb{N}$*.*

Doob's maximal inequality roughly means that "you can uniformly control the entire submartingale by controlling the martingale's endpoint".

Lemma 3.4 (Doob's L^p inequality)**.** *Let* U *and* V *be non-negative random variables and* $\lambda > 0$*. If* $\lambda \mathbb{P}(U \geq \lambda) \leq \mathbb{E}[V \mathbb{1}\{U \geq \lambda\}]$ *then*

$$\|U\|_p \leq \frac{p}{p-1}\|V\|_p.$$

Corollary 3.5 (of Doob's inequalities, add-on)**.** *If* $\{X_n : n \in \mathbb{N}\}$ *is a martingale, then*

$$\left\|\max_{0\leq i\leq n}|X_i|\right\|_p \leq \frac{p}{p-1}\|X_n\|_p.$$

Proof. The only thing to prove is that $\{|X_n| : n \in \mathbb{N}\}$ is a submartingale, which follows immediately from Jensen's inequality. □

Proof of Lemma 3.2 assuming Doob's inequalities. Define $U = \max_{1\leq j\leq n}|X_j|$ and $V = |X_n|$. Doob's inequalities imply that

$$\left\|\max_{1\leq j\leq n}|X_n|\right\|_p = \|U\|_p \leq \frac{p}{p-1}\|X_n\|_p \leq \frac{p}{p-1}M < \infty,$$

where the second last inequality follows from the assumption of L^p boundedness of $\{X_n\}$. Now we can use the monotone convergence theorem to replace $\max_{1\le j\le n}$ with $\sup_{n\ge 1}$, which gives

$$\|X^*\|_p \le \frac{p}{p-1} M,$$

as required. □

Proof of Doob's L^p inequality (Lemma 3.4). We will take as give the following statement: If $U \ge 0$ and $p > 1$ then

$$\mathbb{E}U^p = p\int_0^\infty \lambda^{p-1}\mathbb{P}(U \ge \lambda)d\lambda$$

(This is an extension of the "integration of the survival function rule".) Thus

$$\begin{aligned}
\mathbb{E}U^p &\le p\int_0^\infty \lambda^{p-2}\mathbb{E}\left(V\mathbb{1}\{U \ge \lambda\}\right)d\lambda \\
&= p\mathbb{E}\left(V\int_0^\infty \lambda^{p-2}\mathbb{1}\{U \ge \lambda\}\,d\lambda\right) \\
&= p\mathbb{E}\left(V\int_0^U \lambda^{p-2}d\lambda\right) \\
&= p\mathbb{E}\left(VU^{p-1}\right) \\
&\le \frac{p}{p-1}\|V\|_p\left\|U^{p-1}\right\|_q,
\end{aligned}$$

where the first line is by assumption, the second line uses Fubini's theorem and the final line uses Hölder's inequality with $q = \frac{p}{p-1}$. Now

$$\left\|U^{p-1}\right\|_q = \left[\mathbb{E}\left(U^{p-1}\right)^{p/(p-1)}\right]^{1/q} = \left[\mathbb{E}U^p\right]^{1-1/p}.$$

So we have

$$\mathbb{E}U^p \le \frac{p}{p-1}\|V\|_p\left[\mathbb{E}U^p\right]^{1-1/p}.$$

Consider three possible cases:

1. If $0 < \mathbb{E}U^p < \infty$ then we can divide through by $\left[\mathbb{E}U^p\right]^{1-1/p}$ and get the desired result: $\|U\|_p \le \frac{p}{p-1}\|V\|_p$.

2. If $\mathbb{E}U^p = 0$ then the desired result is trivial.

3. If $\mathbb{E}U^p = \infty$ then we need to show that $\|V\|_p = \infty$. We do this by working with truncated variables, as follows.

For $n \geq \lambda$, we have that the events $\{U \wedge n \geq \lambda\}$ and $\{U \geq \lambda\}$ are equal. Hence,

$$\lambda \mathbb{P}(U \wedge n \geq \lambda) \leq \mathbb{E}[V \mathbb{1}\{U \wedge n \geq \lambda\}].$$

Then $\|U \wedge n\|_p \leq \frac{p}{p-1}\|V\|_p$ by case 1. above. Take $n \to \infty$ and use the monotone convergence theorem to get $\|V\|_p = \infty$. □

Proof of Doob's maximal inequality (Lemma 3.3). Define the event

$$A_i = \{X_0 < \lambda, \ldots X_{i-1} < \lambda, X_i \geq \lambda\}.$$

(A_i is the event that the submartingale first crosses λ at time i.) Observe that the A_i's are disjoint and $\{\max_{0 \leq j \leq n} X_j \geq \lambda\} = \sum_{j=0}^n A_j$. Hence

$$\begin{aligned}
\lambda \mathbb{P}\left(\max_{0 \leq j \leq n} X_j \geq \lambda\right) &= \lambda \sum_{j=0}^n \mathbb{E}(\mathbb{1}_{A_j}) \\
&\leq \sum_{j=0}^n \mathbb{E}(X_j \mathbb{1}_{A_j}) \\
&\leq \sum_{j=0}^n \mathbb{E}(X_n \mathbb{1}_{A_j}) \\
&= \mathbb{E}\left(X_n \mathbb{1}\left\{\cup_{j=0}^n A_j\right\}\right) \\
&= \mathbb{E}\left(X_n \mathbb{1}\left\{\max_{0 \leq j \leq n} X_j \geq \lambda\right\}\right),
\end{aligned}$$

where the second inequality follows from the submartingale property $\mathbb{E}[X_n | \mathcal{F}_j] \geq X_j$ (and so by definition $\mathbb{E}[X_n \mathbb{1}_A] \geq \mathbb{E}[X_j \mathbb{1}_A]$ for any $A \in \mathcal{F}_j$). □

4 Lecture 4/2

4.1 Reverse martingales

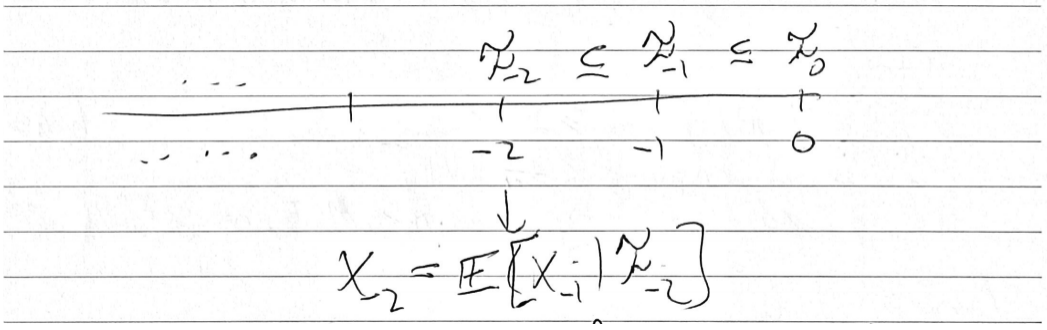
Definition 4.1. Let $\{(X_n, \mathcal{F}_n) : n \in \mathbb{Z}^{\leq 0}\}$ be an adapted sequence with $X_n \in L^1$ and

$$\mathcal{F}_0 \supseteq \mathcal{F}_{-1} \supseteq \mathcal{F}_{-2} \supseteq \dots.$$

$\{(X_n, \mathcal{F}_n)\}$ is a reverse martingale if

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

for all $n \in \mathbb{Z}^{<0}$.



Intuition for a reverse martingale: "as you go further fro the origin, you have less (rather than more) information."

Theorem 4.2. *Let $X_0 \in L^1$. Then $\{(X_n, \mathcal{F}_n) : n \in \mathbb{Z}^{\leq 0}\}$ is a reverse martingale if and only if*

$$X_n = \mathbb{E}(X_0|\mathcal{F}_n),$$

almost surely. In this case, as $n \to -\infty$,

$$\mathbb{E}[X_0|\mathcal{F}_n] \xrightarrow[a.s.]{L^1} \mathcal{E}[X_0|\mathcal{F}_{-\infty}],$$

where $\mathcal{F}_{-\infty} = \cap_{n \in \mathbb{Z}^{\leq 0}} \mathcal{F}_n$.

So all reverse martingales are so called 'Doob reverse martingales' – that is, reverse martingales are always conditional expectations of a random variable X_0. Further, reverse martingales always converge (a.s. and in L^1) and we know what they converge to. So reverse martingales behave a lot more nicely than martingales!

Proof. "$\Rightarrow$:" Use induction and the martingale property. The base case is $X_{-1} = \mathbb{E}[X_0|\mathcal{F}_{-1}]$ The step case uses the tower law of conditional expectation: Assume $X_{n+1} = \mathbb{E}[X_0|\mathcal{F}_{n+1}]$. Then

$$X_n = \mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}\big[[X_0|\mathcal{F}_{n+1}]|\mathcal{F}_n\big] = \mathbb{E}[X_0|\mathcal{F}_n],$$

by the tower law of conditional expectation (noting that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$).

"$\Leftarrow$:" Use the tower law of conditional expectation again

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = \mathbb{E}\big[\mathbb{E}[X_0|\mathcal{F}_{n+1}]|\mathcal{F}_n\big] = \mathbb{E}[X_0|\mathcal{F}_n] = X_n.$$

To prove almost sure convergence of reverse martingales, note that for all $n < 0$, the indexed set $\{(X_n, \mathcal{F}_n), \ldots, (X_0, \mathcal{F}_0)\}$ is a martingale. We will reuse ideas from the proof of the martingale convergence theorem. We have an upcrossing inequality

$$\mathbb{E}[U_n(a,b)] \leq \frac{1}{b-a}\mathbb{E}[(X_0 - a)_+],$$

where $U_n(a,b)$ is the number of times the reverse martingale goes from a to b in $X_{-n}, \ldots, X_0$. Since $X_0 \in L^1$ by assumption, we can recycle the proof of almost sure convergence of martingales. Hence there exists a random variable X_∞ such that $X_n \xrightarrow{\text{a.s.}} X_{-\infty}$.

We have seen that a Doob martingale is uniformly integrable. The same proof applies to $\{\mathbb{E}[X_0|\mathcal{F}_n] : n \in \mathbb{Z}^{\leq 0}\}$. Hence $X_n \xrightarrow{L^1} X_{-\infty}$.

All that remains is to establish what $X_{-\infty}$ is. We want to show that

$$X_{-\infty} = \mathbb{E}[X_0|\mathcal{F}_{-\infty}],$$

almost surely. It suffices to show that, for all $A \in \mathcal{F}_\infty$,

$$\mathbb{E}[X_{-\infty}\mathbb{1}_A] = \mathbb{E}[X_0\mathbb{1}_A]$$

The reverse martingale property implies that $\mathbb{E}[X_n \mathbb{1}_A] = \mathbb{E}[X_0 \mathbb{1}_A]$. (Here we need the fact that the sub-σ-algebras are nested $\mathcal{F}_n \subseteq F_{n+1}$, so that $A \in \mathcal{F}_{-\infty}$ is in $\mathcal{F}_n$.) And L^1 convergence says that $\mathbb{E}[X_n \mathbb{1}_A] \to \mathbb{E}[X_{-\infty} \mathbb{1}_A]$. □

Intuition for reverse martingales: "There is some X_0. As n decreases, you are conditioning on less and less information, to get $X_n = \mathbb{E}[X_0|\mathcal{F}_n]$. And in the limit, you are conditioning on the least information."

The motivation for introducing reverse martingales is so that we can reason about exchangeable sequences of random variables.

4.2 Exchangeable random variables

A set of random variables are exchangeable if their joint distribution is extremely symmetric, in the sense that it is invariant to reordering.

Definition 4.3. $\{X_n : n \in \mathbb{N}\}$ is exchangeable if, for all $m \geq 1$,

$$(X_1, \ldots, X_m) \sim \left(X_{\sigma(1)}, \ldots, X_{\sigma(m)}\right),$$

for any permutation $\sigma \in S_m$ of $\{1, \ldots, m\}$.

Iid-ness is often seen as a strong assumption, particularly for an infinite sequence of random variables. Exchangeability is seen as a natural relaxation of the iid assumption and is intuitively plausible.

Example 4.4.

1. iid sequences are exchangeable.

2. A special case of a more general phenomenon (de Finetti's theorem) which we will see later: Fix a probability distribution π on $[0, 1]$. Suppose $\theta \sim \pi$ and $X_1, X_2, \ldots | \theta \overset{iid}{\sim} \text{Bern}(\theta)$. Then

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^{\sum_i x_i} (1-\theta)^{n - \sum_i x_i} d\pi(\theta).$$

Since the RHS doesn't depend on the ordering of $(x_1, \ldots, x_n)$, the sequence is exchangeable. But the $X_i's$ are not necessarily independent – depending on the choice of π, the X_i's can be depend on each other.

We know $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\text{a.s.}} \theta$. We will see later (c.f. de Finetti's theorem) that "given the limiting proportion of 1's, the individual observations are iid with this probability".

Our goal for the next few lectures is to develop the machinery to state and prove de Finetti's theorem, which provides a characterisation of all exchangeable sequences. We saw in the previous example one way to construct an exchangeable sequence: Put a prior on the parameter θ and then draw iid observations from f_θ. Roughly, de Finetti's theorem states that, in fact, this is the only way to create exchangeable sequences. It is central to the Bayesian perspective since it says that "if you make the minimal assumption of exchangeability, then in fact you are assuming a prior and iid data".

4.2.1 The exchangeable σ-algebra

Definition 4.5. Define $\mathbb{R}^\infty = \mathbb{R} \times \mathbb{R} \times \ldots$ be the infinite-dimensional $\mathbb{R}$-space. Define the projection

$$\pi_i : \mathbb{R}^\infty \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto x_i.$$

Let

$$\mathcal{B}_{\mathbb{R}^\infty} = \sigma\big(\{\pi_i : i \in \mathbb{N}\}\big) = \sigma\big(\{\pi_i^{-1}(B) : i \in \mathbb{N}, B \in \mathcal{B}_\mathbb{R}\},$$

be the smallest σ-algebra such that every π_i is measurable.

We call $(\mathbb{R}^\infty, \mathcal{B}_{\mathbb{R}^\infty})$ *the sequence space*. We may also use the notation $(\mathbb{R}^\mathbb{N}, \mathcal{B}_\infty)$ in place of $(\mathbb{R}^\infty, \mathcal{B}_{\mathbb{R}^\infty})$.

Remark 4.6. $\{X_n : n \in \mathbb{N}\}$ is a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ if and only if the map

$$(\Omega, \mathcal{F}) \to (\mathbb{R}^\infty, \mathcal{B}_{\mathbb{R}^\infty})$$

$$\omega \mapsto \big(X_1(\omega), X_2(\omega), \dots\big),$$

is measurable. So a sequence of random variables is equivalent to a single $\mathbb{R}^\infty$-valued random variable.

Definition 4.7. The *exchangeable σ-algebra* $\mathcal{E}$ is a sub-σ-algebra of $\mathcal{B}_{\mathbb{R}^\infty}$ defined by

$$\mathcal{E} := \bigcap_{m\in\mathbb{N}} \mathcal{E}_m,$$

where

$$\mathcal{E}_m := \Big\{ A \in \mathcal{B}_{\mathbb{R}^\infty} : \text{ if } \boldsymbol{\omega} \in \mathbb{R}^\infty \text{ is in } A, \\ \text{then } \boldsymbol{\omega}_\sigma = (\omega_{\sigma(1)}, \dots, \omega_{\sigma(m)}, \omega_{m+1}, \dots) \text{ is also in } A, \text{ for all } \sigma \in S_m \Big\}.$$

Remark 4.8. For $\mathcal{E}$ to be well-defined, it is necessary to verify that $\mathcal{E}_m$ is a σ-algebra and $\mathcal{E}_{m+1} \subset \mathcal{E}_m$.

Intuition: σ-algebras are black box machines that you can ask yes/no questions. For the exchangeable σ-algebra, you can only ask questions that are invariant to the first m co-ordinates, for all m. So $\mathcal{E}$ is exactly the right object that contains all the 'relevant' information but forgets the ordering.

Lemma 4.9. *Suppose $\{X_n : n \in \mathbb{N}\}$ is an exchangeable sequence of random variables and $\phi : \mathbb{R}^l \to \mathbb{R}$ is bounded and measurable. For all $m \geq l$, define*

$$\hat{S}_m(\phi) = \frac{1}{(m)_l} \sum_{\underset{\sim}{i}} \phi\left(X_{i_1}, \dots, X_{i_l}\right),$$

(Notation: $\underset{\sim}{i}$ is some vector $(i_1, \dots, i_l)$ with i_j disjoint from $[m] = \{1, \dots, m\}$. $\sum_{\underset{\sim}{i}}$ is the summation over all such 'permutations' $\underset{\sim}{i}$ and $(m)_l = m(m-1)\cdots(m-l+1)$ is the number of $\underset{\sim}{i}$.) Then

$$\hat{S}_m(\phi) \xrightarrow[L^1]{a.s.} \mathbb{E}\left[\phi(X_1, \dots, X_l)|\mathcal{E}\right].$$

Think of $\hat{S}_m(\phi)$ as a U-statistic.

Proof. $\hat{S}_m(\phi)$ is $\mathcal{E}_m$-measurable since the function is symmetric in its co-ordinates. So

$$\begin{aligned}\hat{S}_m(\phi) &= \mathbb{E}\left[\hat{S}_m(\phi)|\mathcal{E}_m\right] \\ &= \frac{1}{(m)_l}\sum_{\underset{\sim}{i}} \mathbb{E}\left[\phi(X_{i_1},\ldots,X_{i_l})|\mathcal{E}_m\right] \\ &= \frac{1}{(m)_l}\sum_{\underset{\sim}{i}} \mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}_m\right] \\ &= \mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}_m\right],\end{aligned}$$

where the third line follows from the fact that the random variables are exchangeable in $\mathcal{E}_m$.

Now $\left\{(\hat{S}_{l-m}(\phi),\mathcal{E}_{l-n}) : n \in \mathbb{Z}^{\leq 0}\right\}$ is a reverse martingale. (Check this!) Thus,

$$\hat{S}_{l-n}(\phi) \xrightarrow[L^1]{\text{a.s.}} \mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right],$$

as $n \to -\infty$, by Theorem 4.2. □

5 Lecture 9/2

We are building up to stating and proving de Finetti's theorem, which roughly states: given the long term empirical distribution – that is the limiting set of unordered values – any exchangeable is iid. To make sense of this statement, we introduced the exchangeable σ-algebra. We still need to understand how probability measures play with the exchangeable σ-algebra $\mathcal{E}$. That is, we want to know what are the probabilities of events $E \in \mathcal{E}$.

5.1 The Hewitt-Savage zero-one law

The Hewitt-Savage zero-one law characterises how iid sequences 'play' with $\mathcal{E}$.

Theorem 5.1. *The exchangeable σ-algebra $\mathcal{E}$ of a sequence of iid random variables is trivial – that is, $\mathbb{P}(A) \in \{0,1\}$ for all $A \in \mathcal{E}$.*

"iid sequences do not assign non-trivial probabilities to events in the exchangeable σ-algebra."

add-on Unpacking this: $\mathcal{E}$ is a sub-σ-algebra of $\mathcal{B}_{\mathbb{R}^\infty}$, consisting of the events that are invariant to co-ordinate permutations. A sequence of random variables $\{X_n\}$ induces a probability measure on $\mathbb{R}^\infty$ (the law of $\{X_n\}$). The Hewitt-Savage zero-one law states that these probability measures are trivial on $\mathcal{E}$ if $\{X_n\}$ are iid.

Proof. Let $\phi : \mathbb{R}^l \to \mathbb{R}$ be bounded and measurable. Define $\hat{S}_m(\phi)$ as in Lemma 4.9.

We claim that

$$\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right] = \mathbb{E}\left[\phi(X_1, \ldots, X_l)\right], \tag{3}$$

for iid sequences $\{X_n\}$. Further we claim (3) implies

$$A \perp\!\!\!\perp A, \text{ for all } A \in \mathcal{E}. \tag{4}$$

Hence $\mathbb{P}(A)^2 = \mathbb{P}(A \cap A) = \mathbb{P}(A)$ and so $\mathbb{P}(A)$ is zero or one.

Proving (4) assuming (3): We will prove the stronger statement: Fix $G \in \mathcal{E}$; then for all $B \in \mathcal{E}$,

$$\mathbb{P}(G \cap B) = \mathbb{P}(G)\mathbb{P}(B). \tag{5}$$

We will prove (5) for $B \in \sigma(X_1, \ldots, X_l)$, for any l, and then appeal to the $\pi - \lambda$ theorem to extend (5) to $B \in \mathcal{E}$. If $B \in \sigma(X_1, \ldots, X_l)$ then $\mathbb{1}_B$ is a bounded measurable function ϕ of $X_1, \ldots, X_l$. Then

$$\begin{aligned}
\mathbb{P}(G \cap B) &= \mathbb{E}(\mathbb{1}_G \mathbb{1}_B) \\
&= \mathbb{E}\left(\mathbb{E}\left[\mathbb{1}_G \mathbb{1}_B|\mathcal{E}\right]\right) \\
&= \mathbb{E}\left(\mathbb{1}_G \mathbb{E}\left[\mathbb{1}_B|\mathcal{E}\right]\right) \\
&= \mathbb{E}\left(\mathbb{1}_G \mathbb{E}\left[\mathbb{1}_B\right]\right) \\
&= \mathbb{P}(G)\mathbb{P}(B),
\end{aligned}$$

where the second last line follows by (3).

All that remains is to prove (3). By Lemma 4.9,

$$\hat{S}_m(\phi) \xrightarrow[L^1]{\text{a.s.}} \mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right].$$

Intuition: $\hat{S}_m(\phi) = \frac{1}{(m)_l} \sum_{\underset{}{i}}^{m} \phi\left(X_{i_1}, \ldots, X_{i_l}\right)$ is an "average" of the ϕ statistics, across the first m random variables in the sequence. If $l = 1$ then $\hat{S}_m(\phi) = \frac{1}{m} \sum_{i=1}^{m} \phi(X_i)$ is exactly the average. The limiting average (i.e. $\hat{S}_m(\phi)$ as $m \to \infty$) doesn't care about the first few terms. Hence the limit $\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right]$ is independent of $X_1, \ldots, X_r$ for all r. This implies $\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right]$ is a constant, so it equals the unconditional expectation $\mathbb{E}\left[\phi(X_1, \ldots, X_l)\right]$.

The formal proof proceeds as follows: Fix $r < m$ and define

$$\hat{S}_{m,r}(\phi) = \frac{1}{(m)_l} \sum_{\substack{i_1,\ldots,i_l \\ i_j > r}} \phi(X_{i_1}, \ldots, X_{i_l}).$$

($\hat{S}_{m,r}(\phi)$ is the "average", leaving out the first $X_1, \ldots, X_r$. Our intuition tells us that this shouldn't change anything.)

(Left as an exercise:) Check that

$$\left|\hat{S}_m(\phi) - \hat{S}_{m,r}(\phi)\right| \leq \left(1 - \frac{(m-r)_l}{(m)_l} \|\phi\|_\infty\right) \leq \frac{c}{m} \xrightarrow{m\to\infty} 0.$$

Thus, for all $r \geq 1$,

$$\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right] = \lim_{m\to\infty} \hat{S}_m(\phi) = \lim_{m\to\infty} \hat{S}_{m,r}(\phi), \text{ surely.}$$

But $\hat{S}_{m,r}(\phi)$ doesn't depend on $X_1, \ldots, X_r$, which implies

$$\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right] \perp\!\!\!\perp \{X_1, \ldots, X_r\}, \text{ for all } r \geq 1.$$

In particular, this holds for $r = l$ and so

$$\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right] \perp\!\!\!\perp \phi(X_1, \ldots, X_l).$$

To prove (3), it suffices to show that $\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right]$ is a constant – i.e. that it has zero variance.

$$\begin{aligned}
\mathbb{E}^2\left(\phi(X_1, \ldots, X_l)\right) &= \mathbb{E}\left(\phi(X_1, \ldots, X_l)\right) \mathbb{E}\left(\mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right]\right) \\
&= \mathbb{E}\left(\phi(X_1, \ldots, X_l) \mathbb{E}\left[\phi(X_1, \ldots, X_l)|\mathcal{E}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left(\mathbb{E}\Big[\phi(X_1,\ldots,X_l)\mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\Big|\,\mathcal{E}\Big]\right)\\
&= \mathbb{E}\Big(\mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\Big)\\
&= \mathbb{E}\left(\mathbb{E}^2\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\right)
\end{aligned}
$$

where the second line holds by independence. Thus,

$$\mathrm{Var}\left(\mathbb{E}\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\right) = \mathbb{E}\left(\mathbb{E}^2\left[\phi(X_1,\ldots,X_l)|\mathcal{E}\right]\right) - \mathbb{E}^2\left[\phi(X_1,\ldots,X_l)\right] = 0. \quad \square$$

5.1.1 A strong law for U-statistics

This application demonstrates the power of some of the machinery we used in the proof of Theorem 5.1.

Let $\{X_n : n \in \mathbb{N}\}$ be an iid sequence and $\phi : \mathbb{R}^l \to \mathbb{R}$ be a bounded and measurable function. Define $\hat{S}_m(\phi)$ as in Lemma 4.9. Then

$$\hat{S}_m(\phi) \xrightarrow[L^1]{\text{a.s.}} \mathbb{E}\left[\phi(X_1,\ldots,X_l)\right],$$

from the proof of Theorem 5.1.

As an example consider $\phi(X_1, X_2) = (X_1 - X_2)^2$. To estimate the variance $\mathrm{Var}(X_1) = \sigma^2$, it is typical to use the U-statistic

$$T_n = \frac{1}{n(n-1)}\sum_{i\neq j}(X_i - X_j)^2.$$

(You can check that this is unbiased for $2\sigma^2$.) Our result gives us that

$$T_n \xrightarrow[L^1]{\text{a.s.}} \mathbb{E}\left[(X_1 - X_2)^2\right].$$

So T_n is strongly consistent!

5.2 De Finetti's theorem

Theorem 5.2. *If $\{X_n : n \in \mathbb{N}\}$ is an exchangeable sequence of random variables, then, conditional of $\mathcal{E}$, the sequence is iid – that is,*

$$\mathbb{E}\left[\prod_{k=1}^{l} f_k(X_k)\,\middle|\,\mathcal{E}\right] = \prod_{k=1}^{l}\mathbb{E}\left[f_k(X_k)|\mathcal{E}\right],$$

and

$$\mathbb{E}\left[f_1(X_1)|\mathcal{E}\right] = \mathbb{E}\left[f_1(X_l)|\mathcal{E}\right],$$

for all $l \in \mathbb{N}$ *and all measurable functions* $f_1, \ldots, f_l$.

When we condition on $\mathcal{E}$, intuitively we "have information on the set of unordered values of $\{X_n\}$". For example if X_n takes values zero and one, then this information is the long running proportion of zeroes to ones. More generally, conditioning on $\mathcal{E}$ is often equivalent to conditioning on the long term empirical PDF $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.

Lemma 5.3. *Suppose* $f : \mathbb{R}^{l-1} \to \mathbb{R}$ *and* $g : \mathbb{R} \to \mathbb{R}$ *are bounded and measurable. Define*

$$h_j(x_1, \ldots, x_l) = f(x_1, \ldots, x_{l-1})g(x_j) \text{ for } j = 1, \ldots, l.$$

Then

$$\hat{S}_m(h_l) = \frac{m}{m-l+1}\hat{S}_m(f)\hat{S}_m(g) - \frac{1}{m-l+1}\sum_{j=1}^{l-1}\hat{S}_m(h_j).$$

The proof of this lemma is left as an exercise.

Proof of Theorem 5.2. Fix l and take $m \to \infty$ in

$$\hat{S}_m(h_l) = \frac{m}{m-l+1}\hat{S}_m(f)\hat{S}_m(g) - \frac{1}{m-l+1}\sum_{j=1}^{l-1}\hat{S}_m(h_j).$$

Lemma 4.9 shows that the LHS

$$\hat{S}_m(h_l) \xrightarrow[L^1]{\text{a.s.}} \mathbb{E}\left[f(X_1, \ldots, X_{l-1})g(X_l)|\mathcal{E}\right]. \tag{6}$$

On the RHS, $\frac{m}{m-l+1} \to 1$ and $\frac{1}{m-l+1} \to 0$. Each of $\hat{S}_m(f)$, $\hat{S}_m(g)$ and $\hat{S}_m(h_j)$ converge to the corresponding conditional expectation as in (6); yet there are only a finite number of terms in the summation so

$$\frac{1}{m-l+1}\sum_{j=1}^{l-1}\hat{S}_m(h_j) \xrightarrow[L^1]{\text{a.s.}} 0.$$

Thus,

$$\mathbb{E}\left[f(X_1, \ldots, X_{l-1})g(X_l)|\mathcal{E}\right] = \mathbb{E}\left[f(X_1, \ldots, X_{l-1})|\mathcal{E}\right]\mathbb{E}\left[g(X_l)|\mathcal{E}\right].$$

By induction on l, we get conditional independence:

$$\mathbb{E}\left[\prod_{k=1}^{l} f_k(X_k)\middle|\mathcal{E}\right] = \prod_{k=1}^{l}\mathbb{E}\left[f_k(X_k)|\mathcal{E}\right].$$

All that remains is to check is that the X_n's are identically distributed. Yet this is true by the exchangeability assumption (regardless of whether we condition on $\mathcal{E}$ or not): For all $A \in \mathcal{E}$ (or even any event $A \in \mathcal{B}_{\mathbb{R}^\infty}$)

$$\mathbb{E}\left[g(X_1)\mathbb{1}_A\right] = \mathbb{E}\left[g(X_l)\mathbb{1}_A\right]. \qquad \square$$

5.3 High dimension analogues and graph limits

This section is a high-level glimpse of exchangeability's importance in modern research on network data.

Definition 5.4. A symmetric array $\{X_{ij} : i, j \in \mathbb{N}\}$ is *jointly exchangeable* if for all $m > 1$ and $\sigma \in S_m$,

$$\{X_{ij} : i \le m, j \le m\} \sim \{X_{\sigma(i)\sigma(j)} : i \le m, j \le m\}.$$

So a symmetric array is jointly exchangeable if its distribution is invariant to scrambling the rows and columns by the same permutation.

There is a representation theorem (an analogue to de Finetti) for exchangeable symmetric arrays.

Theorem 5.5 (Aldous '81, Hoover '79)**.** *The array $\{X_{ij} : i, j \in \mathbb{N}\}$ is symmetric and jointly exchangeable if and only if there exists $f : \mathbb{R}^4 \to \mathbb{R}$ and $U, U_i, U_j, U_{\{i,j\}} \overset{iid}{\sim}$ Unif$[0, 1]$ such that*

$$X_{ij} \sim f(U, U_i, U_j, U_{\{i,j\}}),$$

where and.

Notes:

1. $U_{\{i,j\}}$ doesn't depend on the order of i and j.

2. More explicitly, there are three sets of uniform random variables: 1. the singleton $\{U\}$; 2. the sequence $\{U_i : i \in \mathbb{N}\}$; and 3. the double index $\{U_{\{i,j\}} : i \leq j \in \mathbb{N}\}$.

3. U can be interpreted as the analogue of the limiting empirical distribution.

Why is this important for network data? A network is represented by a adjacency matrix and we assume that this matrix is the top left block of an infinite array. Exchangeability is a natural assumption here; all it says is the order of the vertices doesn't matter.

Since an adjacency matrix consists of zeroes and ones, specifying $\mathbb{P}(X_{ij} = 1)$ determines the entire distribution. Define graphons

$$W_u(x, y) = \lambda\left(\{z : f(u, x, y, z) = 1\}\right),$$

where λ is the Lebesgue measure. These are a non-parametric model for network data (since they specify the entire distribution of the array).

6 Lecture 11/2

6.1 Brownian motion

Brownian motion (BM) is the analogue of the Gaussian distribution to stochastic processes. The ideas that we will learn about BM will have wide applicability outside BM.

History: BM was first developed to model the trajectory of particles in a liquid. It also arose in physics (Einstein 1905) and finance (Bachelier 1900).

What are two properties that we would expect of a trajectory? We would expect it to be continuous and random. Given the centrality of Gaussian distributions, it might be model the randomness as Gaussian. This is exactly Brownian motion!

6.1.1 $\mathcal{C}([0,1])$-valued random variables

Definition 6.1. Define $\mathcal{C}([0,1]) = \{f : [0,1] \to \mathbb{R} \text{ continuous}\}$ to be the set of continuous real-valued functions with domain $[0,1]$. For $f, g \in \mathcal{C}([0,1])$, define

$$d_{\sup}(f,g) = \sup_{x\in[0,1]} |f(x) - g(x)|.$$

Facts (with proofs left as exercises): 1) $d_{\sup}$ is a metric on $\mathcal{C}([0,1])$. 2) $(\mathcal{C}([0,1]), d_{\sup})$ is a complete separable metric space. (Recall that in a complete metric space all Cauchy sequences converge and in a separable metric space, there exists a countable dense subset.)

Definition 6.2. Define the Borel σ-algebra $\mathcal{B}_{\mathcal{C}([0,1])}$ on $\mathcal{C}([0,1])$ to be the smallest σ-algebra containing all the open sets in $(\mathcal{C}([0,1]), d_{\sup})$. A $\mathcal{C}([0,1])$-*valued random variable* is a measurable function

$$B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{C}([0,1]), \mathcal{B}_{\mathcal{C}([0,1])}).$$

The law of a $\mathcal{C}([0,1])$-valued random variable is a probability distribution on $(\mathcal{C}([0,1]), \mathcal{B}_{\mathcal{C}([0,1])})$. As with real-valued random variables, if we know the law, then we don't need to consider the underlying probability space – all the information about the random variable is contained in the law.

6.1.2 Definitions of Brownian motion

Definition 6.3. *Standard Brownian motion* $\{B_t : t \in [0,1]\}$ is a $\mathcal{C}([0,1])$-valued random variable such that

i. $B_t \sim \mathcal{N}(0,t)$ for all $t \in [0,1]$ and $B_t - B_s \sim \mathcal{N}(0, t-s)$ for all $t \geq s \in [0,1]$;

ii. (Independent increments) For $0 \leq t_1 < \ldots < t_n \leq 1$, the random variables $B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent.

We use the term *standard* here since we have made two ad-hoc choices: 1) we've set the starting point $B_0 = 0$; and 2) we've set the scale to be 1, when we could

equally have defined $\text{Var}(B_t - B_s) = \alpha(t-s)$ for any $\alpha > 0$. We've also set the domain to be $[0,1]$; later on we will study BM in $[0,\infty)$ – in both these domains (and any others), we will say the BM is standard if 1) and 2) are satisfied.

Definition 6.4 (an equivalent definition of BM)**.** The *law of BM* (called the *Wierner measure*) is a probability distribution W on $(\mathcal{C}([0,1]), \mathcal{B}_{\mathcal{C}([0,1])}$ such that, given $\omega \sim W$,

$$\begin{aligned} X_t : \mathcal{C}([0,1]) &\to \mathbb{R} \\ \omega &\mapsto \omega(t) \end{aligned}$$

satisfies properties i. and ii. in the previous definition.

6.1.3 Existence and uniqueness of BM

Does there exist a law on $\mathcal{C}([0,1])$ satisfying the above definition? And is there exactly one such law? That is, can we talk about "the" Brownian motion?

Proof of uniqueness. Assume that there exist laws μ, ν on CZO satisfying Definition 6.4. We want to show that $\mu = \nu$. Draw $B_1 \sim \mu$ and $B_2 \sim \nu$. Define piecewise linear approximations B_i^k for $i \in \{1,2\}$ and $k = 1, 2, \ldots$:

$$B_i^k(t) = \begin{cases} B_i^k(j/k) & \text{if } t = j/k \text{ for some } j = 0, \ldots, k, \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

Observe that B_i^k is a $\mathcal{C}([0,1])$ random variable. Moreover, for all $k \geq 1$, $B_1^k \sim B_2^k$. Why? The joint distributions at the knots are the same since B_1 and B_2 satisfy i. and ii. of Definition 6.3. And the joint distribution at the knots completely determine the distribution of B_i^k, since outside the knots B_i^k is just (deterministic) linear interpolation.

We know that $B_i^k \xrightarrow{\text{a.s.}} B_i$ as $k \to \infty$ since $d_{\sup}(B_i^k, B_i) \to 0$ as $k \to \infty$ (using continuity of B_i). Let $f : \mathcal{C}([0,1]) \to \mathbb{R}$ be bounded and continuous. Since $B_1^k \sim B_2^k$, we have $\mathbb{E}[f(B_1^k)] = \mathbb{E}[f(B_2^k)]$. As f is bounded and continuous, we can swap the limit $k \to \infty$ with the expectation by DCT, so that

$$\mathbb{E}[f(B_1)] = \mathbb{E}[f(B_2)].$$

(We also needed continuity of f here to conclude $f(B_i^k) \to f(B_i)$.)

With some more work (i.e. establishing a portmanteau theorem for $\mathcal{C}([0,1])$-random variables) we get $B_1 \sim B_2$ – that is $\mu = \nu$. (Here we need the completeness and separability properties of $\mathcal{C}([0,1])$.) $\square$

This proof uses a standard idea: Use approximations that are linear interpolaters so that we need only establish agreement on a finite number of points. Then by continuity the approximations converge to the target object.

Proof of existence of BM (Lévy). We want to construct a measurable function $B : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{C}([0,1]), \mathbb{B}_{\mathcal{C}([0,1])})$ that satisfies the desired properties.

What should $(\Omega, \mathcal{F})$ be? It can be any measurable space such that we can build a sequence $Z_n : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ of iid standard Gaussian random variables. From the first lecture, we know that $(\Omega, \mathcal{F}) = ([0,1], \mathcal{B}_{[0,1]})$ will work.

The proof will proceed in three steps:

(i) Construct a sequence of $\mathcal{C}([0,1])$-valued random variables $B^{(1)}, B^{(2)}, \ldots$ ussing Z_n.

(ii) Show that $\{B^{(n)}\}$ is Cauchy and hence has a limit.

(iii) Show that the limit B has the desired properties. (Left as an exercise.)

Step (i): The sequence $\{B^{(n)}\}$ will again be linear interpolaters, built using Z_n at the knots. Define

$$B_t^{(1)} = tZ_1 = \begin{cases} 0 & \text{if } t = 0, \\ Z_1 & \text{if } t = 1, \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

This has the require BM properties at $t = 0, 1$. $B^(2)$ takes $B^(1)$ and adds a perturbation at $t = 1/2$ so that the values at $t = 0, 1/2, 1$ have the same joint distribution as BM. $B^(3)$ then perturbs at $t = 1/4, 3/4$ and so on.

General construction: Given the knots $B_0, B_{1/2^k}, \dots B_{(2^k-1)/2^k}, B_1$ of $B^{(k-1)}$, define new knots

$$B_{j/2^{k+1}} = \frac{1}{2}\left(B_{(j-1)/2^{k+1}} + B_{(j+1)/2^{k+1}}\right) + Z_{j/2^{k+1}},$$

for j odd, where $Z_{j/2^{k+1}}$ is a fresh $\mathcal{N}(0, 1/2^{k+2})$ random variable, independent of the past. Then define the new approximation $B_t^{(k)}$ as linear interpolation between all of these knots (new and given):

$$B_t^{(k)} = \begin{cases} B_{j/2^{k+1}} & \text{if } t = \frac{j}{2^{k+1}}, \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

Step (ii): We claim that

$$\mathbb{E}\left[\sum_{k=1}^{\infty} d_{\sup}\left(B^{(k+1)}, B^{(k)}\right)\right] < \infty, \tag{7}$$

which implies $\sum_{k=0}^{\infty} d_{\sup}\left(B^{(k+1)}, B^{(k)}\right) < \infty$ a.s. The tail probabilities must get arbitrarily small and via the triangle inequality, this in turn implies that

$$d_{\sup}\left(B^{(N)}, B^{(N+m)}\right) \le \sum_{k=N}^{N+m-1} d_{\sup}\left(B^{(k+1)}, B^{(k)}\right) \le \sum_{k=N}^{\infty} d_{\sup}\left(B^{(k+1)}, B^{(k)}\right)$$

gets arbitrarily small, for all m, as $N \to \infty$. Hence the sequence $\{B^{(k)}\}$ is Cauchy. We will continue this proof in the following lecture. All that remains is to establish (7). □

7 Lecture 16/2

7.1 Construction of Brownian motion (cont.)

Recall our progress on constructing Brownian motion: We started with iid Gaussians and used them to build processes $B_t^{(1)}, B_t^{(2)}, \dots \in \mathcal{C}([0,1])$ that match the Gaussian properties of Brownian motion at an increasing number of dyadic points and are linear interpolation away from these points. We built $B_t^{(k+1)}$ by adding Gaussian

perturbations halfway between the knots of $B_t^{(k)}$ with the perturbation magnitudes getting progressively smaller as k increased.

We need to show that this sequence $\{B_t^{(k)}\}_{k=1}^{\infty}$ is Cauchy in (the complete metric space) $\mathcal{C}([0,1])$ and hence has a limit. We have established that a sufficient condition for this is (7). Finally we will show that this limit B satisfies the properties of BM, hence proving the existence of BM. (This final step is left as an exercise: take any t, approximate it by the dyadics, then establish the Gaussian properties of B_t by taking the limit $B_{b(t,k)}^{(k)}$ (where $b(t,k)$ is the first k-th binary digits of t) and observing that the desired Gaussian properties are satisfied by $B_{b(t,k)}^{(k)}$.)

Proof of (7). By the MCT,

$$\mathbb{E}\left[\sum_{k=1}^{\infty} d_{\sup}\left(B^{(k)}, B^{(k+1)}\right)\right] = \sum_{k=1}^{\infty} \mathbb{E}\left[d_{\sup}\left(B^{(k)}, B^{(k+1)}\right)\right].$$

By the construction of $B^{(k+1)}$, the maximum distance between $B^{(k)}$ and $B^{(k+1)}$ must occur at $B^{(k+1)}$'s new knots. So

$$d_{\sup}\left(B^{(k)}, B^{(k+1)}\right) \leq \max_{\substack{0\leq j\leq 2^{k+1}\\ j \text{ odd}}} \left|Z_{j/2^{k+1}}\right|, \tag{8}$$

where $Z_{j/2^{k+1}} \sim \mathcal{N}(0, 1/2^{k+2})$.

We will take as given that, for $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ (which may be dependent),

$$\mathbb{E}\left[\max_{1\leq i\leq n} |X_i|\right] \leq 2\sigma\sqrt{2\log n}. \tag{9}$$

(This is proved in Homework 2.) Heuristically, the worst case is when the X_i are independent. In this case, it is easy to see that (9) holds. For correlated Gaussians, the maximum can't be (much) larger than the uncorrelated case.

Combining (8) and (9),

$$\mathbb{E}\left[d_{\sup}\left(B^{(k)}, B^{(k+1)}\right)\right] \leq C\frac{\sqrt{k}}{2^{k/2}},$$

where C is some universal constant. Observing $\sum_{k=1}^{\infty} \frac{\sqrt{k}}{2^{k/2}} < \infty$ completes the proof. □

7.2 An equivalent definition of Brownian motion

Definition 7.1. $\{G_t : t \in [0,1]\}$ is a *Gaussian process* if, for all k,

$$(G_{t_1}, \ldots, G_{t_k}) \sim \text{MVN},$$

where $0 \leq t_1 < \ldots < t_k \leq 1$.

Proposition 7.2. $\{B_t : t \in [0,1]\}$ *is standard BM if and only if*

1. $\{B_t : t \in [0,1]\}$ *is a Gaussian process with*

$$\begin{aligned} \mathbb{E}[B_t] &= 0 \\ \mathbb{E}[B_t B_s] &= s \wedge t. \end{aligned}$$

2. $t \mapsto B_t$ *is continuous almost surely.*

This is just a restatement of the first definition. The only (slightly) non-obvious fact needed to prove this Proposition is that for $t > s$,

$$\begin{aligned} \text{Cov}(B_t, B_s) &= \text{Cov}(B_s, B_s) + \text{Cov}(B_t - B_s, B_s) \\ &= s + \text{Cov}(B_t - B_s, B_s) \\ &= s \wedge t + \text{Cov}(B_t - B_s, B_s), \end{aligned}$$

and this equals $s \wedge t$ if and only if $B_t - B_s \perp\!\!\!\perp B_s$.

7.3 Brownian motion on $[0, \infty)$

To get BM on $[0, \infty)$ just glue together a bunch of iid BM on $[0, 1]$.

Definition 7.3. *Standard Brownian motion on* $[0, \infty)$ is a $\mathcal{C}([0, \infty))$-valued random variable $\{B_t : t \in \mathbb{R}^{\geq 0}\}$ such that

1. For all $0 \leq s < t < \infty$,

$$B_t - B_s \sim \mathcal{N}(0, t - s).$$

2. For all k and all $0 \leq t_1 < \ldots < t_k < \infty$,

$$B_{t_1}, B_{t_2} - B_{t_1}, \ldots, B_{t_k} - B_{t_{k-1}}$$

are independent.

Technically, before we can get BM on $[0, \infty)$, we need to first define $\mathcal{C}([0, \infty))$ as a complete metric space with a σ-algebra. We basically carry over everything from $\mathcal{C}([0, \infty))$.

Some relevant analysis facts (which we take as given):

1.
$$\mathcal{C}([0, \infty)) := \{f : [0\infty) \to \mathbb{R} \text{ continuous}\}.$$

2. Given $f, g \in \mathcal{C}([0, \infty))$, define $\|f - g\|_{[n,n+1)} := \sup_{x \in [n,n+1)]} |f(x) - g(x)|$ and

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

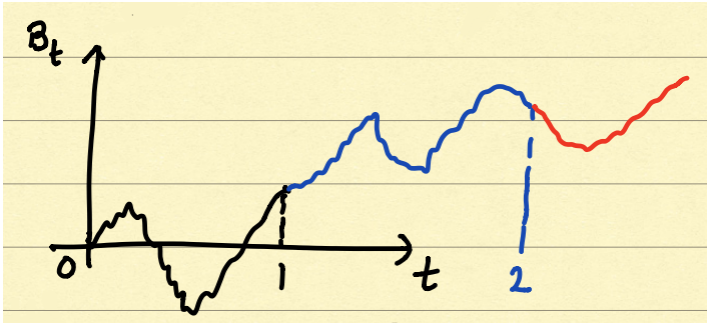
3. $(\mathcal{C}([0, \infty)), d)$ forms a complete, separable metric space.

4. Convergence: $f_n \to f \in (\mathcal{C}([0, \infty)), d)$ if and only if, for all m, f_n converges to g uniformly on the interval $[m, m+1)$. This type of convergence is called *uniform convergence on compacts*.

7.3.1 Proving existence and uniqueness of BM on $[0, \infty)$

Uniqueness: The proof is exactly the same as for BM on $[0, 1]$, but first restrict to a compact set $[0, n]$. Prove that the two Wiener laws are equal when restricted to $[0, n]$, for any n and then take the limit $n \to \infty$ to establish equality on $[0, \infty)$.

Existence: Take a bunch of iid BM (call them $B^1, B^2, \ldots$) on $[0, 1]$ and glue them together:



Formally: start on a probability space where we can construct $B^1, B^2, \ldots, \overset{iid}{\sim}$ sBM on $[0,1]$. (Note that if we can construct one copy of sBM then we can construct a countably infinite number of copies. See Example 1.5.) For any $t \in \mathbb{R}^{\geq 0}$, find k such that $k < t \leq k_1$ and define $B_t = B_1^1 + B_1^2 + \ldots + B_1^k + B_{t-k}^{k+1}$. Exercise: check that $\{B_t : t \in \mathbb{R}^{\geq 0}\}$ satisfies the properties of sBM.

7.3.2 Scaling properties of BM

Suppose $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ is sBM.

(i) If $a > 0$, then $X_t = \frac{1}{a} B_{a^2 t}$ is also sBM. (Exercise: prove this – check finite dimensional distributional properties and check continuity.)

(ii) $\{B_{t+s} - B_s : t \in \mathbb{R}^{\geq 0}\}$ is SBM for any $s \geq 0$. (We will prove this in Lemma 8.5.)

(iii)

$$X_t' = \begin{cases} 0 & \text{if } t = 0, \\ tB_{1/t} & \text{otherwise,} \end{cases}$$

is sBM on $[0, \infty)$.

Proof of (iii).

1. for all $0 \leq t_1 < \ldots < t_k < \infty$,

$$(X_{t_1}, \ldots, X_{t_k}) \sim \text{MVN},$$

with expectation $\mathbf{0}$.

2. $\mathbb{E}(X_t) = 0$.

3. $\mathbb{E}(X_t X_s) = ts\left(\frac{1}{t} \wedge \frac{1}{s}\right) = t \wedge s$.

4. Continuous trajectories: X_t is continuous at $t > 0$, by continuity of the transformation $tB_{1/t}$. For $t = 0$, we have already established that

$$\{X_t : t \in (0, \infty) \cap \mathbb{Q}\} \sim \{B_t : t \in (0, \infty) \cap \mathbb{Q}\},$$

where $\mathbb{Q}$ is the rationals. (This is saying that finite subsets from either side have the same joint distribution.) This implies that if $t \to 0$ on the rationals, then $X_t \to 0$ almost surely (by sBM continuity of B_t). Combine this with the fact that $t \mapsto X_t$ is continuous for all $t \geq 0$ and we have that $X_{t_n} \to 0$ for any sequence $t_n \to 0$ (not necessarily on the rationals) a.s. This is precisely the definition that $t \mapsto X_t$ is continuous at $t = 0$.

□

7.4 Nowhere differentiability of BM

We know that BM is a.s. continuous. Is it smoother? What is the extent of the smoothness of BM? Heuristically, we would expect that BM is jittery (i.e. not smooth) no matter how far we zoom in, because to construct BM we kept adding jitter at every dyadic point. So we would expect that BM is 'just barely' continuous. We will confirm this intuition.

Definition 7.4. Let $f : \mathbb{R} \to \mathbb{R}$. The *upper right derivative* of f is defined as

$$D^* f(t) = \limsup_{h \downarrow 0} \frac{f(t+h) - f(t)}{h}.$$

The *lower right derivative* of f is

$$D_* f(t) = \liminf_{h \downarrow 0} \frac{f(t+h) - f(t)}{h}.$$

If f is differentiable, then the lower and upper right derivatives agree.

Theorem 7.5 (Paley, Wiener and Zygmund, 1933)**.** *BM is nowhere differentiable a.s. In fact, with probability 1, all* $t \in [0,1]$ *satisfy* $D^*B(t) = \infty$ *or* $D_*B(t) = -\infty$*, or both.*

This is a strong statement: for any point t, the probability that $B(t)$ is differentiable at t is zero. Further, $B(t)$ is not differentiable at any point in a bounded interval with probability 1. That is, pick an interval I; then the probability that $B(t)$ is differentiable at some $t \in I$ is zero. And even more, $B(t)$ is not differentiable is the least smooth way – the lower and upper right derivatives after infinitely different.

Proof. Suppose for contradiction that with some non-zero probability, there exists $t_0 \in [0,1]$ such that

$$-\infty < D_*B(t_0) \leq B^*B(t_0) < \infty.$$

Let A denote the event where such a t_0 exists. We claim that this implies

$$\sup_{h \in [0,1]} \frac{|B(t_0+h) - B(t_0)|}{h} \leq M < \infty. \tag{10}$$

Why is this true? Near $h = 0$, the positive side $[B(t_0+h) - B(t_0)]_+$ is controlled by $D^*B(t_0)$ and the negative side $[B(t_0+h) - B(t_0)]_-$ is controlled by $D_*B(t_0)$. Away from $h = 0$ (i.e. $h > \epsilon$ for some ϵ), $1/h$ is bounded and $B(t)$ is continuous and hence bounded on the compact set $[t_0, t_0+1]$; this means the fraction is bounded away from $h = 0$.

We will contradict (10). Fix n and then choose k such that t_0 is in the dyadic interval $\left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)$. We will show that (10) implies

$$\left| B\left(\frac{k+j}{2^n}\right) - B\left(\frac{k+j-1}{2^n}\right) \right| \leq \frac{2j+1}{2^n} M, \tag{11}$$

for $j = 0, 1, 2$, provided that $k \leq 2^n - 2$.

Then we will show that

$$\mathbb{P}\left[\bigcup_{k=1}^{2^n-2} \Omega_{n,k} \text{ for infinitely many n} \right] = 0, \tag{12}$$

where

$$\Omega_{n,k} = \left\{ \left| B\left(\frac{k+j}{2^n}\right) - B\left(\frac{k+j-1}{2^n}\right) \right| \leq \frac{2j+1}{2^n} M, \text{ for } j = 0,1,2 \right\}.$$

(Note: there is nothing special about using the three intervals $j = 0, 1, 2$; it just turns out that we need three to be able to union bound the probability in (12).)

Yet (11) says that $\cup_{k=1}^{2^n-2} \Omega_{n,k}$ must happen for all n large enough. (n must be large enough so that t_0 doesn't fall in the last two dyadic intervals $\left[\frac{2^n-2}{2^n}, \frac{2^n-1}{2^n}\right]$ or $\left[\frac{2^n-1}{2^n}, 1\right]$.) Thus,

$$\mathbb{P}\left[\bigcup_{k=1}^{2^n-2} \Omega_{n,k} \text{ for infinitely many n} \right] \geq \mathbb{P}(A) > 0.$$

So we will have established a contradiction. We will fill in the details in the next lecture. □

8 Lecture 18/2

8.1 Proof of nowhere differentiability of BM (cont.)

Proof of Theorem 7.5. To establish (11):

$$\begin{aligned}
\left| B\left(\frac{k}{2^n}\right) - B\left(\frac{k-1}{2^n}\right) \right| &\leq \left| B\left(\frac{k}{2^n}\right) - B\left(t_0\right) \right| + \left| B\left(t_0\right) - B\left(\frac{k-1}{2^n}\right) \right| \\
&\leq M\left(\frac{k}{2^n} - t_0\right) + M\left(t_0 - \frac{k-1}{2^n}\right) \\
&= \frac{M}{2^n} \\
&< \frac{3M}{2^n},
\end{aligned}$$

where the second line follows from (10). Now look at

$$\begin{aligned}
\left| B\left(\frac{k+1}{2^n}\right) - B\left(\frac{k}{2^n}\right) \right| &\leq \left| B\left(\frac{k+1}{2^n}\right) - B\left(t_0\right) \right| + \left| B\left(t_0\right) - B\left(\frac{k}{2^n}\right) \right| \\
&\leq M\left(\frac{k+1}{2^n} - t_0 + \frac{k}{2^n} - t_0\right)
\end{aligned}$$

$$\leq M\left(\frac{3}{2^n} - 2t_0 + \frac{2k-2}{2^n}\right)$$
$$\leq \frac{3M}{2^n}$$

where the second last line follows since $t_0 > \frac{k-1}{2^n}$. Finally, we can also show that

$$\left|B\left(\frac{k+2}{2^n}\right) - B\left(\frac{k+1}{2^n}\right)\right| \leq \frac{5M}{2^n},$$

assuming (10) and that $k \leq 2^n - 2$.

Now we want to establish (12). The idea is that $B(t_0 + h) - B(t_0)$ is a Gaussian, so the probability that it is tiny (i.e. less than $\frac{3M}{2^n}$) on a lot of independent intervals is going to zero. Hence (12) must have zero probability. We need to use three increments rather than one for technical reasons. (Otherwise, when we do the union bound over the 2^n increments, we won't be able to send the sum of probabilities to zero.) We can use three increments (or as many as we like) since they are independent. By independence

$$\begin{aligned}
\mathbb{P}(\Omega_{n,k}) &= \prod_{j=0}^{2} \mathbb{P}\left[\left|B\left(\frac{k+j}{2^n}\right) - B\left(\frac{k+j-1}{2^n}\right)\right| \leq \frac{2j+1}{2^n}M\right] \\
&\leq \prod_{j=0}^{2} \mathbb{P}\left[\left|B\left(\frac{k+j}{2^n}\right) - B\left(\frac{k+j-1}{2^n}\right)\right| \leq \frac{5}{2^n}M\right] \\
&\leq \left(\mathbb{P}\left[|\mathcal{N}(0,1)| \leq \frac{5}{2^{n/2}}M\right]\right)^3 \\
&\leq \left([\text{density of } \mathcal{N}(0,1) \text{ at } 0] \times 2 \times \frac{5}{2^{n/2}}M\right)^3 \\
&= \frac{c}{2^{3n/2}},
\end{aligned}$$

for some constant c, where the third last line follows since $B\left(\frac{k+j}{2^n}\right) - B\left(\frac{k+j-1}{2^n}\right) \overset{iid}{\sim} \mathcal{N}(0, 1/2^n)$, for $j = 0, 1, 2$. (Since we end up with $2^{n/2}$ in the denominator and we are going to multiply this by 2^n to get the union bound, we see why we need to use three or more independent increments.) Then

$$\mathbb{P}\left(\bigcup_{k=1}^{2^n-2} \Omega_{n,k}\right) \leq 2^n \frac{c}{2^{3n/2}} = \frac{c}{2^{n/2}},$$

and

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\bigcup_{k=1}^{2^n-2} \Omega_{n,k} \right) < \infty.$$

By Borel-Cantelli,

$$\mathbb{P}\left[\bigcup_{k=1}^{2^n-2} \Omega_{n,k} \text{ for infinitely many } n \right] = 0.$$

This establishes (12), completing the proof. □

8.2 The Markov property of BM

Recall the definition of a discrete-time Markov chain:

Definition 8.1. $\{X_n : n \in \mathbb{N}\}$ is a *Markov chain* if

$$\mathbb{P}(X_{n+1} \in A | X_0, \ldots, X_n) = \mathbb{P}(X_{n+1} \in A | X_n),$$

for all events A.

To define a continuous-time Markov chain, we need to resolve two questions:

1. How do we condition on all of the past (which is typically an uncountable number of random variables)?
2. What does 'the next step' – i.e. the analogue of X_{n+1} – mean? How can we characterise "the future behaviour given the current state"?

Hopefully by now you would anticipate the answer to 1. is to construct an appropriate σ-algebra! First we need to generalise the notion of a filtration to continuous time.

Definition 8.2. $\{\mathcal{F}_t : t \in \mathbb{R}^{\geq 0}\}$ is a *filtration* if it is a collection of nested σ-algebras:

$$\mathcal{F}_s \subset \mathcal{F}_t,$$

if $s \leq t$.

Example 8.3. $\mathcal{F}_t = \sigma\left(\{B_s : 0 \leq s \leq t\}\right)$ is the smallest σ-algebra such that all B_s are measurable by $\mathcal{F}_t$. (Note, however, $\mathcal{F}_t$ is not generated by the uncountable number of $\mathbb{R}$-valued random variables B_s – it is generated by the single $\mathcal{C}\left([0,t]\right)$-valued random variable.) This σ-algebra can "tell you the random path of B up to time t". So to "condition on the past", we should condition on $\mathcal{F}_t$.

Answering question 2. is much harder. We will bypass a general answer and only answer it for BM by reformulating the Markov property:

Proposition 8.4. *$\{X_n : n \in \mathbb{N}\}$ is a Markov chain if and only if*

$$X_{n+1} \perp\!\!\!\perp (X_1, \ldots, X_{n-1}) \,|X_n,$$

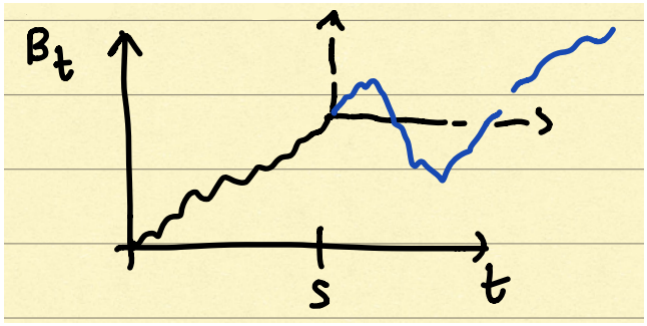
that is, X_{n+1} is conditionally independent of $(X_1, \ldots, X_{n-1})$, given X_n.

Lemma 8.5. *Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM. Fix $s > 0$ and define*

$$W_t = B_{t+s} - B_s.$$

Then $\{W_t : t \in \mathbb{R}^{\geq 0}\}$ is sBM independent of $\{B_t : 0 \leq t \leq s\}$.

How does it make sense to say that an uncountable set of random variables $\{W_t : t \in \mathbb{R}^{\geq 0}\}$ is independent of another uncountable set of random variables $\{B_t : 0 \leq t \leq s\}$? The key observations is that these are not sets of random variables – $\{W_t : t \in \mathbb{R}^{\geq 0}\}$ is a single $\mathcal{C}([0,\infty))$-valued random variable and $\{B_t : 0 \leq t \leq s\}$ is a single $\mathcal{C}\left([0,s]\right)$-valued random variable!



This lemma is a reformulation of the Markov property – "the future is independent of the past, given the present" – for BM. For other stochastic processes, stating the Markov property is not so nice; we would need more machinery to state the Markov property in general.

Proof of Lemma 8.5. We want to show: if A is a measurable subset of $\mathcal{C}([0,s])$ and B is a measurable subset of $\mathcal{C}([0,\infty))$, then

$$\mathbb{P}\left[(B_t)_{t=0}^s \in A, (W_t)_{t\geq 0} \in B\right] = \mathbb{P}\left[(B_t)_{t=0}^s \in A\right] \mathbb{P}\left[(Y_t)_{t\geq 0} \in B\right], \tag{13}$$

where $\{Y_t : t \in \mathbb{R}^{\geq 0}\}$ is sBM. But this is really hard to prove as we don't know what A and B look like. (What is a measurable subset of $\mathcal{C}([0,\infty))$??) Instead, we will do what we always do – approximate W_t and B_t by linear interpolaters; prove the desired result on the approximations; and take limits.

We can show: if $A_1, \ldots, A_k, B_1, \ldots, B_l$ are measurable subsets of $\mathbb{R}$; $0 \leq t_1 \leq \ldots \leq t_k \leq s$; and $0 \leq u_1 \leq \ldots \leq u_l < \infty$, then

$$\begin{aligned}
&\mathbb{P}\left[(B_{t_1}, \ldots, B_{t_k}) \in A_1 \times \ldots \times A_k, (W_{u_1}, \ldots, W_{u_l}) \in B_1 \times \ldots \times B_l\right] \\
&\quad = \mathbb{P}\left[(B_{t_1}, \ldots, B_{t_k}) \in A_1 \times \ldots \times A_k\right] \mathbb{P}\left[(W_{u_1}, \ldots, W_{u_l}) \in B_1 \times \ldots \times B_l\right].
\end{aligned}$$

Why? Use the fact that all the $(B_{t_1}, \ldots, B_{t_k}, W_{u_1}, \ldots, W_{u_l})$ are Gaussian and show their covariances are zero. This implies that if A_k is a measurable subset of $\mathbb{R}^k$ and B_l is a measurable subset of $\mathbb{R}^l$ then

$$\begin{aligned}
&\mathbb{P}\left[(B_{t_1}, \ldots, B_{t_k}) \in A_k, (W_{u_1}, \ldots, W_{u_l}) \in B_l\right] \\
&\quad = \mathbb{P}\left[(B_{t_1}, \ldots, B_{t_k}) \in A_k\right] \mathbb{P}\left[(Y_{u_1}, \ldots, Y_{u_l}) \in B_l\right].
\end{aligned}$$

Now we get (13) by the technique used in constructing BM:

1. Approximate B_t and W_t by linear interpolaters.
2. Show that the approximations $B_t^{(n)}, W_t^{(n)}$ are independent. (It suffices to check independence at the knots.)

3. We know that
$$\begin{bmatrix} B_t^{(n)} \\ W_t^{(n)} \end{bmatrix} \to \begin{bmatrix} B_t \\ W_t \end{bmatrix},$$
surely (by construction). The independence carries over in the limit. (Why? If $(X_n, Y_n) \xrightarrow{\text{a.s.}} (X, Y)$ then $\mathbb{P}(X_n \in A, Y_n \in B) = \mathbb{P}(X_n \in A)\mathbb{P}(Y_n \in B)$ implies $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ – just take limits of both sides.)

□

8.2.1 Right continuous filtrations

Definition 8.6. A filtration $\{\mathcal{F}_t^+ : t \in \mathbb{R}^{\geq 0}\}$ is *right continuous* if
$$\mathcal{F}_t^+ = \bigcap_{s>t} \mathcal{F}_s^+.$$

Example 8.7. Given a filtration $\{\mathcal{F}_t : t \in \mathbb{R}^{\geq 0}\}$, we can always construct a right continuous filtration by
$$\mathcal{F}_t^+ = \bigcap_{s>t} \mathcal{F}_s.$$
(Exercise: check that this is indeed right continuous.)

"We don't bother trying to check that a filtration is right continuous – we just construct a new one that is."

From now on, if we talk about the filtration for BM $\mathcal{F}_t = \sigma(\{B_s : 0 \leq s \leq t\})$, we will assume that it is right continuous. (Otherwise, replace it by its right continuous analogue.)

8.2.2 Rigorous definition of the Markov property of BM

Lemma 8.8 (Strengthening of Lemma 8.5)**.** *Let* $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ *be sBM. Fix* $s \geq 0$ *and define* $W_t = B_{t+s} - B_s$*. Then* $\{W_t : t \in \mathbb{R}^{\geq 0}\}$ *is independent of* $\mathcal{F}_s^+$*.*

This is what we will call *the Markov property of BM*. It is stronger than Lemma 8.5 since $\mathcal{F}_s^+ \supseteq \mathcal{F}_s = \sigma(\{B_t : 0 \leq t \leq s\})$ by construction.

Proof. For each $n \geq 1$, define $W_t^* = B_{t+s+1/n} - B_{s+1/n}$. Lemma 8.5 implies that

$$\{W_t^n : t \in \mathbb{R}^{\geq 0}\} \perp\!\!\!\perp \mathcal{F}_{s+1/n},$$

for all n. But $\mathcal{F}_s^+ \subset \mathcal{F}_{s+1/n}$, hence

$$\{W_t^n : t \in \mathbb{R}^{\geq 0}\} \perp\!\!\!\perp \mathcal{F}_s^+,$$

for all n. By continuity of trajectories,

$$\{W_t^n\} \xrightarrow{\text{a.s.}} \{W_t\}, \text{ as } n \to \infty.$$

Hence $\{W_t\} \perp\!\!\!\perp \mathcal{F}_s^+$. □

(We keep using the same idea: discretise and take limits!)

8.3 Stopping times

When we studied martingales, we examined their stopping times. We would like to do the same for BM.

Definition 8.9. A *stopping time* for the adapted sequence $\{(B_t, \mathcal{F}_t) : t \in \mathbb{R}^{\geq 0}\}$ is a non-negative random variable T, defined on the same probability space as B_t, with

$$\{T < t\} \in \mathcal{F}_t^+ \text{ for all } t \geq 0.$$

This is basically the same definition of a stopping time from Stat210, adapted to continuous time.

One reason why we use right continuous filtrations is the following Claim.

Claim 8.10. *If $\mathcal{F}_t^+$ is right continuous then*

$$\{T < t\} \in \mathcal{F}_t^+ \Leftrightarrow \{T \leq t\} \in \mathcal{F}_t^+ \text{ for all } t \geq 0.$$

For non-right continuous filtrations, we can have the left side but not the right! However, the right side always implies the left, regardless of whether the filtration is right continuous.

Proof. Suppose $\{T < t\} \in \mathcal{F}_t^+$ for all $t \geq 0$. We know

$$\{T \leq t\} = \bigcap_{n \geq 2/\epsilon} \{T < t + 1/n\},$$

with each $\{T < t + 1/n\} \in \mathcal{F}_{t+1/n}^+ \subset \mathcal{F}_{t+\epsilon}^+$. Hence $\{T \leq t\} \in \mathcal{F}_{t+\epsilon}^+$ for all $\epsilon > 0$. Thus by right continuity, $\{T \leq t\} \in \mathcal{F}_t^+$.

For the other direction, suppose $\{T \leq t\} \in \mathcal{F}_t^+$ for all $t \geq 0$. Then

$$\{T < t\} = \bigcup_{n \geq 1} \{T \leq t - 1/n\}.$$

Since each $\{T \leq t - 1/n\} \in \mathcal{F}_t^+$, their union is too. (Note that this direction doesn't rely on right continuity.) □

9 Lecture 23/2

9.1 Understanding right continuous filtrations

Working with right continuous filtrations makes life easier, but they make less intuitive sense. What is $\mathcal{F}_t^+$ exactly? $\mathcal{F}_t$ is easy to understand – ask anything about BM up to time t and $\mathcal{F}_t$ can answer. We know that $\mathcal{F}_t \subset \mathcal{F}_t^+$. But how much larger is $\mathcal{F}_t^+$? Intuitively, $\mathcal{F}_t^+$ gives the same – but infinitesimally more – information. The following Theorem makes this rigorous.

Theorem 9.1 (Blumenthal's zero-one law)**.** *If* $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ *is sBM and* $A \in \mathcal{F}_0^+$ *then* $\mathbb{P}(A) = \{0, 1\}$.

(Here $\mathbb{P}$ is the Wiener measure – the probability with respect to BM.)

B_0 is a constant, so $\sigma(B_0) = \mathcal{F}_0$ is the trivial σ-algebra $\{\Omega, \emptyset\}$. ($\mathcal{F}_0$ can't answer anything, sine all the information it has available is a constant.) Blumenthal's zero-one law says this basically extends to $\mathcal{F}_0^+$. $\mathcal{F}_0^+$ is bigger but it still can't answer non-trivial questions. This is a sort of "stochastic continuity".

Because BM is translation invariant, this idea carries over to $\mathcal{F}_t^+$ for $t > 0$.

Proof. As before, we want to show that $A \perp\!\!\!\perp A$ for all $A \in \mathcal{F}_0^+$. Use the Markov property: $\{B_t : t \geq 0\} \perp\!\!\!\perp \mathcal{F}_0^+$. But $A \in \mathcal{F}_0^+ \subset \sigma(\{B_t : t \geq 0\})$. So $A \perp\!\!\!\perp A$. □

So the evens in $\mathcal{F}_0^+ \cap (\mathcal{F}_0)^c$ still have probability zero or one. This can be really useful for thinking about BM locally around B_0, as in the following Lemma.

Lemma 9.2 (An application of Blumenthal's zero-one law)**.** *Suppose* $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ *is sBM. Let* $\tau = \inf\{t > 0 : B_t > 0\}$ *and* $\sigma = \inf\{t > 0 : B_t = 0\}$*. Then* $\mathbb{P}(\tau = 0) = \mathbb{P}(\sigma = 0) = 1$.

This lemma proves the heuristic that in any interval $I = (0, t]$ around 0, regardless of length, there exists $s \in I$ with $B_s > 0$ and $s' \in I$ with $B_{s'} < 0$. This follows the intuition that BM is very jagged. "B_t wiggles around zero with probability 1 infinitely often after starting at $B_0 = 0$."

Proof.

$$\{\tau = 0\} = \bigcap_{n=1}^{\infty} A_n = \bigcap_{n=1}^{\infty} \{\exists \epsilon \in (0, 1/n) : B_\epsilon > 0\}.$$

Each event A_n is in $\mathcal{F}_{1/n}^+$, so their intersection is in $\mathcal{F}_0^+$. Then Blumenthal's zero-one law gives $\mathbb{P}(\tau = 0) \in \{0, 1\}$. All we need now is to argue that $\mathbb{P}(\tau = 0) \neq 0$.

$$\{\tau = 0\} = \bigcap_{t>0} \{\tau \leq t\}.$$

So

$$\mathbb{P}(\tau = 0) = \lim_{t \to 0} \mathbb{P}(\tau \leq t).$$

Thus it suffices to uniformly bound the probabilities $\mathbb{P}(\tau \leq t)$ away from zero. But

$$\mathbb{P}(\tau \leq t) \geq \mathbb{P}(B_t > 0) = \frac{1}{2},$$

for all t since B_t is Gaussian with mean zero.

We can make exactly the same argument with $\tau' = \inf\{t > 0 : B_t < 0\}$. So $\mathbb{P}(\tau' = 0) = 1$. Then by continuity and the intermediate value theorem, any interval $I = (0, t]$ from zero, regardless of its length, must have $B_s = 0$ for $s \in I$. So $\mathbb{P}(\sigma = 0) = 1$. □

9.2 Strong Markov property

The (weak) Markov property states that BM refreshes after a fixed time t. The strong Markov property states that this remains true when t is replaced by a random variable T – as long as T is a stopping time. We need a new σ-algebra to formalise this notion.

Definition 9.3. Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM and $\mathcal{F}_t^+$ be the canonical right continuous filtration. Let T be a stopping time with respect to $\mathcal{F}_t^+$. The *stopped σ-algebra* is defined as

$$\mathcal{F}_T^+ = \left\{A \in \sigma\left(\{B_s : s \in \mathbb{R}^{\geq 0}\}\right) : A \cap \{T < t\} \in \mathcal{F}_t^+ \ \forall t > 0\right\}.$$

Remark 9.4. While T is random, there is nothing random about $\mathcal{F}_T^+$ – it is a σ-algebra. Intuitively, $\mathcal{F}_T^+$ knows information about what happened before T occur; it knows about the information implied by T stopping.

Theorem 9.5 (Strong Markov property)**.** *Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM. Suppose T is a stopping time with respect to the canonical right continuous filtration $\mathcal{F}_t^+$ and $\mathbb{P}(T < \infty) = 1$. Define $W_t = B_{T+t} - B_T$ to be the process beyond the stopping time. Then*

$$\{W_t : t \in \mathbb{R}^{\geq 0}\} \sim \text{sBM},$$

and $\{W_t\} \perp\!\!\!\perp \mathcal{F}_T^+$.

Proof. The technical challenge we need to overcome is that we are dealing with continuous-valued stopping times. So we can't write down $\mathbb{P}(T = t)$, unlike in discrete time processes. A very useful trick to overcome this is to 'do what you know': discretise the possible values of T so we can work with $\mathbb{P}(T = t)$; then argue that any continuous-valued T can be approximated by discrete T and finally take the limit of discrete T to get the result for continuous T. This strategy will work because 1) BM is continuous and 2) $\mathcal{F}_t^+$ is right continuous.

Assume that T takes values in a countable set $0 \leq t_1 < t_2 < \ldots$. Define $W_t^i = B_{t+t_i} - B_{t_i}$. Fix $A \in \mathcal{F}_T^+$ and $E \in \mathcal{B}_{\mathcal{C}([0,\infty))}$. We want to show independence between

A and $(W_t)_{t\geq 0} \in E$:

$$\begin{aligned}
\mathbb{P}\left((W_t)_{t\geq 0} \in E, A\right) &= \sum_{i=1}^{\infty} \mathbb{P}\left((W_t^i)_{t\geq 0} \in E, A, T = t_i\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}\left((W_t^i)_{t\geq 0} \in E, A \cap \{T = t_i\}\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}\left((W_t^i)_{t\geq 0} \in E\right) \mathbb{P}\left(A \cap \{T = t_i\}\right) \\
&= \mathbb{P}\left((B_t)_{t\geq 0} \in E\right) \sum_{i=1}^{\infty} \mathbb{P}\left(A \cap \{T = t_i\}\right) \\
&= \mathbb{P}\left((B_t)_{t\geq 0} \in E\right) \mathbb{P}\left(A\right),
\end{aligned}$$

where the third line follows since $(W_t)_{t\geq 0} \perp\!\!\!\perp \mathcal{F}_{t_i}^+$ by the Markov property while $A \cap \{T = t_i\} \in \mathcal{F}_{t_i}^+$ by definition of $\mathcal{F}_T^+$; and the fourth line since $(W_t^i)_{t\geq 0} \sim$ sBM for all i by the Markov property.

How do we generalise this argument to continuous-valued stopping times? Idea: forcibly discretise T: Given a stopping time T, define a sequence

$$T_n = \frac{m+1}{2^n} \text{ if } \frac{m}{2^n} \leq T < \frac{m+1}{2^n.}$$

(Choosing the right endpoint of the interval $[\frac{m}{2^n}, \frac{m+1}{2^n})$ is crucial so that T_n is still a stopping time.) Each T_n is a stopping time with respect to $\mathcal{F}_t^+$ and $T_n \downarrow T$ (check this!). Define

$$W_t^{(n)} = B_{T_n+t} - B_{T_n}.$$

From the discrete argument above, $(W_t^{(n)})_{t\geq 0} \perp\!\!\!\perp \mathcal{F}_{T_n}^+$ for all n.

We claim that if T and S are stopping times with $T \leq S$ always then $\mathcal{F}_T^+ \subset \mathcal{F}_S^+$. (The proof is left as an exercise.) This implies $\mathcal{F}_T^+ \subset \mathcal{F}_{T_n}^+$, so $(W_t^{(n)})_{t\geq 0} \perp\!\!\!\perp \mathcal{F}_T^+$ for all n. Yet by the continuity of BM $W_t^{(n)} \xrightarrow{\text{a.s.}} W_t$ as $n \to \infty$. Hence

$$\{W_t : t \in \mathbb{R}^{\geq 0}\} \perp\!\!\!\perp \mathcal{F}_T^+.$$

Further, for all $n \geq 1$, $\{W_t^{(n)}\} \sim$ sBM, so $\{W_t\} \sim$ sBM as well. □

Why is the strong Markov property better than the Markov property? For one, it allows us to prove the following result.

Claim 9.6 (An application of the strong Markov property)**.** *Let*

$$T = \inf\{t \in [0,1] : B_t = \max_{s\in[0,1]} B_s\}.$$

Then T is not a stopping time.

Intuitively the claim should be easy to see. Yet is difficult to prove from the definition of stopping times.

Proof. First we prove that $\mathbb{P}(T < 1) = 1$. Define $W_t = B_{1-t} - B_1$. We know that $\{W_t\} \sim$ sBM. Yet if $T = 1$ then $W_t < 0$ in a neighbourhood of $t = 0$. Yet this contradicts Lemma 9.2. so $\mathbb{P}(T = 1) = 0$. Now if T is a stopping time, then $W_t = B_{T+t} - B_T$ is sBM by the strong Markov property. But then again we would have $W_t < 0$ in a neighbourhood of $t = 0$. □

10 Lecture 25/2

10.1 Hitting times

Definition 10.1. Given $a > 0$ the *hitting time* of a is

$$T_a = \inf\{t \geq 0 : B_t > a\}.$$

Lemma 10.2. *A hitting time T_a is a stopping time with respect to $\mathcal{F}_t^+$.*

Proof. We want to show that $\{T_a \leq t\} \in \mathcal{F}_t^+$:

$$\{T_a \leq t\} = \bigcap_{s>t} \{T_a < s\} = \bigcap_{s>t} \bigcup_{r\in(0,s)\cap\mathbb{Q}} \{B_r > a\},$$

where the second equality relies on the continuity of BM and the density of the rationals in $\mathbb{R}$. The union $\bigcup_{r\in(0,s)\cap\mathbb{Q}}\{B_r > a\}$ is in $\mathcal{F}_s^+$. (We must use a countable union here – σ-algebras are not necessarily closed under uncountable unions.) Then by right continuity, the intersection $\bigcap_{s>t}\bigcup_{r\in(0,s)\cap\mathbb{Q}}\{B_r > a\}$ is in $\mathcal{F}_t^+$. □

We want to study the distribution of hitting times. As a start, we should show that we are dealing with finite random variables.

Claim 10.3.

$$\mathbb{P}(T < \infty) = 1,$$

for any hitting time T.

Proof. For any $n \in \mathbb{N}$, $B_n \sim X_1 + \ldots + X_n$ where $X_i \overset{iid}{\sim} \mathcal{N}(0,1)$. Then apply the law of iterated logarithm:

$$\limsup_{n\to\infty} \frac{B_n}{\sqrt{2n \log\log n}} \to 1 \text{ a.s.}$$

So there is an event A with probability 1 where B_n grows like $\sqrt{2n \log\log n}$. But $\sqrt{2n \log\log n}$ diverges so B_n must grow unbounded on A which implies $T < \infty$ on A. Hence

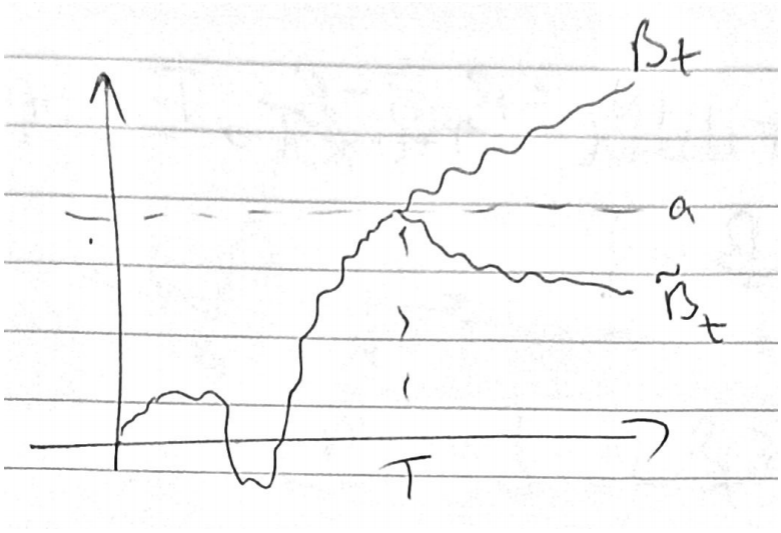
$$\mathbb{P}(T < \infty) \geq \mathbb{P}(A) = 1.$$

□

10.1.1 The reflection principle for BM

Let sBM be sBM and T be the hitting time of a. Then define

$$\tilde{B}_t = \begin{cases} B_t & \text{if } 0 \leq t \leq T, \\ 2B_T - B_t & \text{if } t > T. \end{cases} \tag{14}$$



Lemma 10.4. $\{\tilde{B}_t\}$ *is sBM.*

"Start with sBM, wait until a hitting time, reflect it around a after that – then marginally the reflected process is still sBM." Intuiively this should help us find the distribution of hitting times T since the reflective process is dependent on T. So if we can study $\tilde{B}_t$, then we might be able to use it to get information about T.

Proof. Use the strong Markov process. Define

$$W_t = B_{T+t} - B_t$$
$$U_t = B_{t\wedge T}.$$

We claim that $\{U_t\}$ and T are measurable with respect to $\mathcal{F}_T^+$. (The proof of this is left as an exercise.)

Define

$$\varphi : \mathcal{C}([0,\infty)) \times [0,\infty) \times \mathcal{C}([0,\infty)) \to \mathcal{C}([0,\infty))$$
$$(f,t,g) \mapsto \begin{cases} f(s) & \text{if } s \le t, \\ f(t) + g(s-t) & \text{if } s > t. \end{cases}$$

If we define $U_t = B_{t\wedge T}, W_t = B_{T+t} - B_T$ and $T = \inf\{t : B_t > a\}$ then $\varphi(U,T,W) = B$. On the other hand $\varphi(U,T,-W) = \tilde{B}$.

$W \perp\!\!\!\perp U, T$ since $W \perp\!\!\!\perp \mathcal{F}_T^+$ by the strong Markov property. Also, since $W \sim$ sBM, $W \sim -W$. Hence $\varphi(U, T, W) \sim \varphi(U, T, -W)$. (Why? If $X \sim Y$ and $X, Y \perp\!\!\!\perp Z$ then $(X, Z) \sim (Y, Z)$.) $\square$

10.1.2 The distribution of the running maximum

Let $M_t = \max_{0 \le s \le t} B_s$. What is the distribution of M_t? For a start, it is related to the distribution of T since

$$\{M_t > a\} = \{T_a \le t\}.$$

(This is like the count-time duality.) Also,

$$\begin{aligned}
\mathbb{P}[M_t \ge a] &= \mathbb{P}[M_t \ge a, B_t \ge a] + \mathbb{P}[M_t \ge a, B_t < a] \\
&= \mathbb{P}[B_t \ge a] + \mathbb{P}[M_t \ge a, B_t < a] \\
&= \mathbb{P}[B_t \ge a] + \mathbb{P}\left[\tilde{B}_t > a\right] \\
&= 2\mathbb{P}[B_t \ge a]
\end{aligned}$$

where $\tilde{B}_t$ is the reflected process (defined in (14)); and the third line follows from the fact that $\{M_t \ge a, B_t < a\} = \{\tilde{B}_t > a\}$. Since $B_t \sim \mathcal{N}(0, t)$,

$$M_t \sim \sqrt{t}|Z|, \tag{15}$$

where $Z \sim \mathcal{N}(0, 1)$.

We can get the distribution of T_a from M_t. Further we can talk about the first time hitting b after the BM has hit a by using reflection arguments again. What about understanding stopping times in general? One tool is to use the strong Markov property. We will soon introduced another: the martingale property of BM.

10.2 Continuous time martingales

Definition 10.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{\mathcal{F}_t : t \in \mathbb{R}^{\ge 0}\}$ be a filtration. An adapted sequence $\{(X_t, \mathcal{F}_t) : t \in \mathbb{R}^{\ge 0}\}$ is a *martingale* if

(i) $\mathbb{E}|X_t| < \infty$ for all $t \in \mathbb{R}^{\ge 0}$; and

(ii) $\mathbb{E}[X_t|\mathcal{F}_s] = X_s$ if $s \leq t$.

We need property (i) so that the conditional expectation in (ii) is well defined.

Example 10.6. Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM and $\{\mathcal{F}_t^+ : t \in \mathbb{R}^{\geq 0}\}$ be the canonical right continuous filtration. Then $\{(B_t, \mathcal{F}_t^+) : t \in \mathbb{R}^{\geq 0}\}$ is a martingale.

The proof follows from the Markov property of BM.

Proof. By construction, B_t is adapted to $\mathcal{F}_t^+$. For $s \leq t$, $B_t - B_s \perp\!\!\!\perp \mathcal{F}_s^+$. This implies $\mathbb{E}[B_t - B_s|\mathcal{F}_s^+] = 0$. (Why? If $X \perp\!\!\!\perp \mathcal{G}$ then $\mathbb{E}(X|G) = \mathbb{E}X$.) Hence $\mathbb{E}(B_t|\mathcal{F}_s^+) = B_s$. □

Most of the properties for discrete-time martingales carry over to countinuous time martingales (assuming some regularity conditions on the trajectories – usually continuity is sufficient). So we can get analogous results for BM as we did for discrete-time martingales in Stat210.

10.2.1 The optional stopping theorem

Theorem 10.7. *Assume that $\{X_t : t \in \mathbb{R}^{\geq 0}\}$ is a right continuous martingale (i.e. it has right continuous trajectories) adapted to a right continuous filtration $\mathcal{F}_t^+$. Let T be a stopping time with respect to $\mathcal{F}_t^+$ such that $T \leq c$ a.s. for some constant $c < \infty$. Suppose that*

$$\mathbb{E} \sup_{0 \leq t \leq c+\epsilon} |X_t| < \infty, \tag{16}$$

for some $\epsilon > 0$ (typically 1). Then

1. $\mathbb{E}X_T = \mathbb{E}X_0$;
2. *If S is another stopping time such that $S \leq T$ a.s., then $\mathbb{E}\left[X_t|\mathcal{F}_S^+\right] = X_S$.*

The assumption $T \leq c$ a.s. sounds strong but it is often easy to ensure. Establishing (16) is usually the hard part; but we need it so we can apply DCT. In the proof of Wald's first Lemma, we will show that we can weaken the assumptions of Theorem 10.7 to A) $\mathbb{E}T < \infty$ and B) $\mathbb{E}\sup_{0 \leq t \leq T+\epsilon}|X_t| < \infty$.

Proof. We will take as given the optional stopping theorem for discrete-time martingales.

We've only got one strategy – discretise – so we will use it again! Define $T_n = \frac{m+1}{2^n}$ if $\frac{m}{2^n} \leq T < \frac{m+1}{2^n}$. Then T_n is a stopping time with respect to $\{\mathcal{F}_0^+, \mathcal{F}_{1/2^n}^+, \mathcal{F}_{2/2^n}^+, \ldots\}$ and $T_n \in \{0, 1/2^n, 2/2^n, \ldots\}$.

So $\{(X_{m/2^n}, \mathcal{F}_{m/2^n}^+) : m \in \mathbb{N}\}$ is a discrete time martingale with $T_n \leq c + \epsilon$ a.s. for large enough n. Then the discrete time optional stopping theorem gives $\mathbb{E}X_{T_n} = \mathbb{X}_0$. Since $T_n \downarrow T$ and $\{X_t\}$ is right continuous,

$$X_{T_n} \xrightarrow{\text{a.s.}} X_T. \tag{17}$$

Finally apply DCT to show that (17) converges in L^1 too, using the bound $|X_{T_n}| \leq \sup_{0 \leq t \leq c+\epsilon} |X_t|$.

The proof of (ii) is left as an exercise: use the same strategy: discretise S and T, prove the result using the discrete time optional stopping theorem, then take the limit and use DCT. □

The canonical example of a right continuous processes is the Poisson point process (count of Poisson across time) (since they jump up unit intervals).

10.2.2 Wald's first lemma

Lemma 10.8. *Suppose* $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ *is sBM and* T *a stopping time with respect to the canonical right continuous filtration* $\mathcal{F}_t^+$ *of* $\{B_t\}$*. If* $\mathbb{E}T < \infty$ *then* $\mathbb{E}B_T = 0$*.*

Wald's first lemma turns out to be critical for sequential learning – it is used to construct the sequential probability ratio test (SPRT).

Proof. This proof uses the common trick of capping T to get a bounded stopping time, so that we can apply Theorem 10.7, and then taking the cap to infinity.

Fix $t \geq 0$. $T \wedge t$ is bounded a stopping time with respect to $\{\mathcal{F}_t^+\}$. We know that

$$\mathbb{E} \max_{0 \leq s \leq t+1} |B_s| < \infty,$$

(see Homework 2, Question 4). The optional stopping theorem then implies $\mathbb{E}B_{T\wedge t} = \mathbb{E}B_0 = 0$.

Now we need to take $t \to \infty$ so that we get $T \wedge t = T$. Since $\mathbb{E}T < \infty$, $T < \infty$ with probability 1. Hence $B_{T\wedge t} \to B_T$ a.s. as $t \to \infty$.

Define $M = \max_{0\le s\le T+1}|B_s|$. We have $M \ge |B_{T\wedge t}|$. We claim that

$$\mathbb{E}M < \infty, \tag{18}$$

which then implies we can use DCT to conclude $\mathbb{E}B_{T\wedge t} \to \mathbb{E}B_T$ as $t \to \infty$. Notice that we haven't used the fact that B_t is sBM yet – just that it is right continuous. This implies we can weaken the assumptions of Theorem 10.7 to A) $\mathbb{E}T < \infty$ and B) $\mathbb{E}\sup_{0\le t\le T+\epsilon}|B_t| < \infty$.

To prove (18), define $Z_i = \max_{t\in[i,i+1)}|B_t - B_i|$. Suppose T^* is any maximiser of $|B_t|$ on $[0, T+1]$. Then

$$\begin{aligned}
|B_{T^*}| &\le \left|B_{T^*} - B_{\lfloor T^*\rfloor}\right| + \left|B_{\lfloor T^*\rfloor} - B_{\lfloor T^*\rfloor-1}\right| + \ldots + |B_1 - B_0| \\
&\le \sum_{i=1}^{\lfloor T\rfloor} Z_i \\
&= \sum_{i=1}^{\infty} Z_i \mathbb{1}\{T \ge i\}\,.
\end{aligned}$$

Further

$$\begin{aligned}
\mathbb{E}M &= \mathbb{E}|B_{T^*}| \\
&\le \mathbb{E}\left[\sum_{i=1}^{\infty} Z_i \mathbb{1}\{T \ge i\}\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}\left[Z_i \mathbb{1}\{T \ge i\}\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}Z_i \mathbb{P}(T \ge i) \\
&= \mathbb{E}Z_1 \sum_{i=1}^{\infty} \mathbb{P}(T \ge i)
\end{aligned}$$

$$\leq \mathbb{E}Z_1\mathbb{E}T$$
$$< \infty$$

where the third line follows by MCT; the fourth line uses the strong Markov property with $Z_i \perp\!\!\!\perp \mathcal{F}_i^+$ and $\{T \geq i\} = \{T < i\}^c \in \mathcal{F}_i^+$; the fifth line since the Z_i's are iid; and the sixth line follows since

$$\mathbb{E}T = \int_0^\infty \mathbb{P}(T \geq t)dt \geq \int_0^\infty \mathbb{P}(T \geq \lceil t \rceil)dt = \sum_{i=1}^\infty \mathbb{P}(T \geq i). \qquad \square$$

11 Lecture 2/3

11.1 Martingale properties of BM (cont.)

11.1.1 Wald's second lemma

Lemma 11.1. *Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM and $\mathcal{F}_t^+$ be the canonical right continuous filtration. Suppose T is a stopping time with respect to $\mathcal{F}_t^+$ with $\mathbb{E}T < \infty$. Then*

$$\mathbb{E}B_T^2 = \mathbb{E}T.$$

Proof. This proof uses the technique of bounding T in time and space so that we can apply the optional stopping theorem, and then progressively lifting these restrictions. (This is called localisation.)

We can check that $\{(B_t^2 - t, \mathcal{F}_t^+) : t \in \mathbb{R}^{\geq 0}\}$ is a martingale. Let $T_n = \inf\{t > 0 : |B_t| = n\}$ and fix $s \in \mathbb{R}^{\geq 0}$. Then $T \wedge T_n \wedge s$ is a bounded stopping time.

We want to use the optional stopping theorem to conclude

$$\mathbb{E}B_{T\wedge T_n \wedge s}^2 = \mathbb{E}\left[T \wedge T_n \wedge s\right]. \tag{19}$$

But we need to verify that the domination random variable has finite expectation. That is, we want to show that

$$\mathbb{E} \sup_{0 \leq t \leq s+1} \left|B_t^2 - t\right| < \infty.$$

Since

$$\begin{aligned}\mathbb{E}\sup_{0\le t\le s+1}\left|B_t^2-t\right| &\le \mathbb{E}\sup_{0\le t\le s+1}\left|B_t^2\right|+s+1\\ &= \mathbb{E}\left[\left(\sup_{0\le t\le s+1}|B_T|\right)^2\right]+s+1,\end{aligned}$$

it suffices to show $\mathbb{E}\left[\left(\sup_{0\le t\le s+1}|B_T|\right)^2\right]$ is finite.

We can show that

$$\mathbb{P}\left[\sup_{0\le t\le s+1}|B_t|>\lambda\right]\le\frac{\mathbb{E}B_{s+1}^4}{\lambda^4}=\frac{3(s+1)^2}{\lambda^4}.$$

How? Similar to the question in Homework 2, but use Doob's maximal inequality with the submartingale $\{B_t^4\}$ instead of $\{B_t^2\}$. Then by integrating the survival function,

$$\mathbb{E}\left[\left(\sup_{0\le t\le s+1}|B_T|\right)^2\right]=\int_0^\infty\mathbb{P}\left[\left(\sup_{0\le t\le s+1}|B_t|\right)^2>\lambda\right]d\lambda<\infty.$$

(Note that a $\frac{1}{\lambda^2}$ tail bound is not strong enough since we need to square it to get a tail bound on $\left(\sup_{0\le t\le s+1}|B_t|\right)^2$.) Thus, we have established (19).

Now we need to take away the bounds on T. Since $T\wedge T_n\wedge s\uparrow T\wedge T_n$ as $s\to\infty$,

$$\mathbb{E}\left[T\wedge T_n\wedge s\right]=\mathbb{E}\left[T\wedge T_n\right],$$

by MCT. By continuity, $B^2_{T\wedge T_n\wedge s}\to B^2_{T\wedge T_n}$ and $\left|B^2_{T\wedge T_n\wedge s}\right|\le n$, so DCT implies

$$\mathbb{E}B^2_{T\wedge T_n\wedge s}\to\mathbb{E}B^2_{T\wedge T_n}.$$

Hence $\mathbb{E}B^2_{T\wedge T_n}=\mathbb{E}T\wedge T_n$.

Now, taking $n\to\infty$, then $T\wedge T_n\uparrow T$ a.s. and $B^2_{T\wedge T_n}\to B^2_T$ by continuity. Fatou's lemma implies

$$\begin{aligned}\mathbb{E}\left[B_T^2\right] &= \mathbb{E}\left[\liminf_{n\to\infty}B^2_{T\wedge T_n}\right]\\ &\le \liminf_{n\to\infty}\mathbb{E}\left[B^2_{T\wedge T_n}\right]\\ &= \liminf_{n\to\infty}\mathbb{E}\left[T\wedge T_n\right]\end{aligned}$$

$$= \mathbb{E}T,$$

where the final line uses MCT. Thus,

$$\mathbb{E}B_T^2 \leq \mathbb{E}T. \tag{20}$$

For the opposite direction,

$$\begin{aligned}
\mathbb{E}B_T^2 &= \mathbb{E}\left[(B_T - B_{T\wedge T_n} + B_{T\wedge T_n})^2\right] \\
&= \mathbb{E}\left[(B_T - B_{T\wedge T_n})^2\right] + \mathbb{E}\left[B_{T\wedge T_n}^2\right] + 2\mathbb{E}\left[(B_T - B_{T\wedge T_n})\, B_{T\wedge T_n}\right] \\
&= \mathbb{E}\left[(B_T - B_{T\wedge T_n})^2\right] + \mathbb{E}\left[B_{T\wedge T_n}^2\right] \\
&\geq \mathbb{E}\left[B_{T\wedge T_n}^2\right] \\
&= \mathbb{E}T \wedge T_n,
\end{aligned}$$

where the third line follows by observing

$$\begin{aligned}
\mathbb{E}\left[(B_T - B_{T\wedge T_n})\, B_{T\wedge T_n}\right] &= \mathbb{E}\left[\mathbb{E}\left((B_T - B_{T\wedge T_n})\, B_{T\wedge T_n} | \mathcal{F}_{T\wedge T_n}^+\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left((B_T - B_{T\wedge T_n})\, | \mathcal{F}_{T\wedge T_n}^+\right) B_{T\wedge T_n}\right] \\
&= 0,
\end{aligned}$$

with the second line uses the fact that $B_{T\wedge T_n}$ is $\mathcal{F}_{T\wedge T_n}^+$-measurable; and the third line uses (ii) of the optional stopping theorem which states that if $T \leq S$ then $\mathbb{E}\left[B_S | \mathcal{F}_T^+\right] = B_T$.

Thus, $\mathbb{E}B_T^2 \geq \mathbb{E}\left[T \wedge T_n\right]$ for all n and hence

$$\mathbb{E}B_T^2 \geq \mathbb{E}T, \tag{21}$$

since $\mathbb{E}T \wedge T_n \to \mathbb{E}T$ by MCT. Combining (20) and (21) completes the proof. □

11.1.2 Applications of Wald's lemmas

Why are Wald's Lemmas important? Here are two interesting application:

Example 11.2.

1. Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$ be sBM and define $T_1 = \inf\{t > 0 : B_t > 1\}$ to be the hitting time of 1. We have established that $\mathbb{P}[T_1 < \infty] = 1$ (by the law of iterated logarithm). We claim that $\mathbb{E}T_1 = \infty$! Why? Use Wald's first Lemma: Suppose that $\mathbb{E}T_1 < \infty$. Then $\mathbb{E}B_{T_1} = EB_0 = 0$. Yet $B_{T_1} = 1$ almost surely by continuity.

   Intuition for this result: We know that BM will hit 1 with probability one, but $\mathbb{E}T_1 = \infty$. This is because BM can go negative for an arbitrarily long excursion before becoming positive and hitting 1. So the tail probabilities $\mathbb{P}[T_1 > t]$ do not decay fast enough to ensure $\mathbb{E}T_1 < \infty$.

2. An analogue to Gambler's Ruin: Fix $a, b > 0$ and let $T = \inf\{t > 0 : B_t < -a \text{ or } B_t > b\}$. We showed in section that $\limsup_{n\to\infty} B_n/\sqrt{n} = \infty$ a.s. and $\limsup_{n\to\infty} B_n/\sqrt{n} = -\infty$ a.s. So $\mathbb{P}(T < \infty) = 1$. Now $\mathbb{E}T < \infty$ by Wald's lemmas: (Sketch proof): We need to show that $\mathbb{P}(T > t)$ decays fast. Use the strong Markov property. Suppose $T > k$ and $B_k = x \in (-a, b)$. We have

$$\mathbb{P}[(x + B_t) \text{ hits } -a \text{ or } b \text{ for } t \in [0, 1]] > 0,$$

   since B_t is Normal. Then, for $t \in [k, k + 1]$,

$$\begin{aligned}
\mathbb{P}[T > t] &\leq \left[\sup_{x\in(-a,b)} \mathbb{P}[(x + B_s) \text{ doesn't hit } -a \text{ or } b \text{ for } s \in [0, 1]]\right]^k \\
&= \left[1 - \inf_{x\in(-a,b)} \mathbb{P}[(x + B_s) \text{ hits } -a \text{ or } b \text{ for } s \in [0, 1]]\right]^k \\
&= [1 - \varepsilon]^k .
\end{aligned}$$

   So we have geometric decay. This is enough to prove $\mathbb{E}T < \infty$ by integrating the survival function. Then Wald's first lemma gives

$$-a\mathbb{P}(B_T = -a) + b\mathbb{P}(B_T = b) = \mathbb{E}B_T = 0,$$

   so that $\mathbb{P}(B_T = -a) = \frac{b}{a+b}$, which is analogous to the solution to the Gambler's Ruin. Further, Wald's second lemma gives

$$ab = a^2\mathbb{P}(B_T = -a) + b^2\mathbb{P}(B_T = b) = \mathbb{E}B_T^2 = \mathbb{E}T$$

11.2 Roadmap for coming lectures

We understand BM quite well now. But we still don't know why BM is important. In the coming lectures we will establish a CLT for iid stochastic processes. This result, named Donsker's theorem, shows that random walks (appropriately centred and scaled) converge in distribution (in $\mathcal{C}([0,\infty))^*$) to sBM.

12 Lecture 9/3

12.1 Weak convergence in $\mathcal{C}([0,1])$

Definition 12.1. Let (S,d) be a metric space. Suppose $\{X_n : n \in \mathbb{N}\}$ and X are S-valued random variables. We say that $X_n \xrightarrow{d} X$ if

$$\mathbb{E}f(X_n) \to \mathbb{E}f(X),$$

for all bounded continuous functions $f : S \to \mathbb{R}$.

We have seen this definition before in Assignment 2, where we showed that it generalises the definition of weak convergence for $\mathbb{R}$-valued random variables.

Example 12.2 (Special case of Donsker's theorem). We are interested in $S = \mathcal{C}([0,1])$ and $d(f,g) = \sup_{x\in[0,1]}|f(x) - g(x)|$. Next lecture we will state and prove Donsker's theorem. For now, we will describe a special case of Donsker's theorem: Let

$$X_i = \begin{cases} 1 & \text{w.p.}\frac{1}{2}, \\ -1 & \text{otherwise.} \end{cases}$$

(The general statement of the theorem only requires X_i's are iid with mean zero and variance 1.) Define $S_k = \sum_{i=1}^k X_i$ and

$$S^{(n)}(t) = \begin{cases} \frac{S_k}{\sqrt{n}} & \text{if } t = \frac{k}{n}, \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

*Recall that in Homework 2, we defined converge in distribution $X_n \xrightarrow{d} X$ in an arbitrary metric space to be $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$ for all bounded and continuous real-valued functions f.

Observe that for all $n \geq 1$, $S^{(n)}$ are $\mathcal{C}([0,1])$-valued random variables. Donsker's theorem states that

$$\{S^{(n)}(t) : 0 \leq t \leq 1\} \xrightarrow{d} \{B_t : 0 \leq t \leq 1\},$$

where $\{B_t : 0 \leq t \leq 1\} \sim$ sBM.

To visualise this process: imagine a simple symmetric random walk (SSRW). Take the walk up to length n and shrink the x-axis down to $[0,1]$ (so shrink by a factor of n). Shrink the y-axis by a factor of $\sqrt{n}$. Then in the limit this is standard Brownian motion! In general, this type of limits is called scaling limits. Scaling of n in the x-axis and $\sqrt{n}$ in the y-axis is common.

Why do we care about Donsker's theorem? What are the upshots of this result? Obviously, it is useful in giving us intuition and understanding for limiting processes. But more than that, Donsker's theorem is one of the most powerful results that we will see this semester. Combined with the continuous mapping theorem, it allows us to derive many strong results on the distributional convergence of many processes and random variables.

Here is one application. Define

$$\begin{aligned} \phi : \mathcal{C}([0,1]) &\to \mathbb{R} \\ f &\mapsto f(1). \end{aligned}$$

This is a continuous function. Apply the continuous mapping theorem to the result in Example 12.2 using ϕ to get

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} B_1 \sim \mathcal{N}(0,1).$$

So we have just derived the standard CLT!

In the rest of the lecture, we will be building machinery to prove Donsker's theorem. Specifically, we need to develop methods for proving weak convergence in $\mathcal{C}([0,1])$. We will see a specific strategy (based on coupling) next lecture in the proof of Donsker's theorem. We will also see a general strategy (based on tightness and finite dimensional convergence). First, we will find alternate ways to characterise convergence in distribution.

12.1.1 Portmanteau theorem

This is a slight detour into general theory above convergence in distribution.

Note that convergence in distribution is a property of the laws of the random variables, not the random variables itself. We do not care whether the random variables are defined on the same space (c.f. convergence in probability, a.s., in L_p etc.). So it is more convenient to consider convergence in distribution just as a certain type of convergence of probability measures defined on the metric space S:

Definition 12.3. Let (S, d) be a metric space and define $\mathcal{P}(S)$ to be the set of all (tsoa) probability measures on S. Suppose $\{\mu_n : n \in \mathbb{N}\} \subset \mathcal{P}(S)$ and $\mu \in \mathcal{P}(S)$. We say that $\mu_n \xrightarrow{d} \mu$ if

$$\int f d\mu_n \to \int f d\mu,$$

for all $f : S \to \mathbb{R}$ bounded and continuous.

This is nothing new – it's just a restatement of Definition 12.1.

Theorem 12.4 (Portmanteau)**.** *The following are equivalent:*

(i) $\mu_n \xrightarrow{d} \mu$*;*

(ii) $\int f d\mu_n \to \int f d\mu$*, for all* $f : S \to \mathbb{R}$ *bounded and uniformly continuous;*

(iii) for all $C \subset S$ *closed,*

$$\limsup_{n\to\infty} \mu_n(C) \leq \mu(C);$$

(iv) for all $G \subset S$ *open,*

$$\liminf_{n\to\infty} \mu_n(G) \geq \mu(G);$$

(v) $\mu_n(A) \to \mu(A)$ *for all* A *measurable with* $\mu(\delta A) = 0$ *(where* δA *is the boundary of* A*).*

This is a number of equivalent characterisations that we can use to prove weak convergence. (v) is closest in spirit to the original $\mathbb{R}$-valued random variable weak convergence, which is defined as convergence of CDFs (at continuity points of X).

Proof. "(i) $\Rightarrow$ (ii)" is trivial.

"(ii) $\Rightarrow$ (iii)": Use the trick where we approximate $\mathbb{1}_C$ by a continuous function. Given $C \subset S$ closed, define

$$C^\epsilon = \{x : d(x, C) < \epsilon\}.$$

Fix $\mu \in \mathcal{P}(S)$. For any $\eta > 0$, there exists $\epsilon > 0$ such that

$$\mu(C^\epsilon) < \mu(C) + \eta. \tag{22}$$

Why? As C is closed, $C = \bigcap_{m \in \mathbb{N}} C^{1/n}$, but

$$\mu\left(C^{1/n}\right) \downarrow \mu(C),$$

since $C^{1/n} \downarrow C$. This immediately implies that, given $\eta > 0$, we can choose n large enough such that $\mu\left(C^{1/n}\right) < \mu(C) + \eta$. This proves (22).

Define

$$g(x) = \frac{d\left(x, (C^\epsilon)^c\right)}{d(x, C) + d\left(x, (C^\epsilon)^c\right)}.$$

Then

$$g(x) \in \begin{cases} \{1\} & \text{if } x \in C, \\ \{0\} & \text{if } x \in (C^\epsilon)^c, \\ [0, 1] & \text{otherwise.} \end{cases}$$

g is sandwich between $\mathbb{1}_C$ and $\mathbb{1}_{C^\epsilon}$ (importantly, it is an approximation of $\mathbb{1}_C$). We can check that $g : S \to [0, 1]$ is uniformly continuous.

Fix $C \subset S$ closed. Then $\mu_n(C) \leq \int g d\mu_n$ by construction. So

$$\limsup_{n \to \infty} \mu_n(C) \leq \lim_{n \to \infty} \int g d\mu_n = \int g d\mu \leq \mu(C^\epsilon) \leq \mu(C) + \eta,$$

for ϵ small, where the equality follows by (ii) and the last inequality follows by (22). Send ϵ (and hence η) to zero to complete the proof.

"(iii) $\Leftrightarrow$ (iv)" is straightforward (just take complements).

"(iii), (iv) $\Rightarrow$ (v)": Fix A with $\mu(\delta A) = 0$. Since $\int A \subset A \subset \bar{A}$ (where $\bar{A}$ is the closure of A), we know $\mu_n(A) \leq \mu_n(\bar{A})$ and

$$\limsup_{n \to \infty} \mu_n(A) \leq \limsup_{n \to \infty} \mu_n(\bar{A}) \leq \mu(\bar{A}) = \mu(A),$$

where the second inequality follows by (iii) and the last equality follows by the assumption $\mu(\delta A) = 0$.

As $\mu_n(A) \geq \mu(\int A)$, we can find $\liminf_{n\to\infty} \mu_n(A) \geq \mu(A)$ by similar reasoning. Combine these two results to get (v).

"(v) $\Rightarrow$ (i)": Let $f : S \to \mathbb{R}$ be bounded and continuous. WLOG, assume $0 \leq f \leq 1$. (Otherwise shift and scale f.)

Look at the level sets $A_y = \{x : f(x) > y\}$. Observe that $\delta A_y \subset \{x : f(x) = y\}$. Only countably many y can satisfy $\mu(\{x : f(x) = y\}) > 0$. Hence

$$\mu_n(A_y) \to \mu(A_y),$$

for all but countably many y. Then

$$\int f d\mu_n = \int_0^1 \mu_n(A_y) dy \to \int_0^1 \mu(A_y) dy = \int f d\mu,$$

where the convergence $\to$ is an application of DCT since $\mu_n(A_y) \xrightarrow{\text{a.s.}} \mu(A_y)$. □

What does the Portmanteau Theorem give us? It provides some intuition on what weak convergence means. Yet it is still hard to check any of these equivalent characterisations in practise. We need specific, easily checkable conditions to verify weak convergence. These conditions require regularity conditions on the metric space S (e.g. complete and separable). So this is where we give up on general theory and specialise to particular metric spaces.

Example 12.5.

1. When $S = \mathbb{R}$, it suffices to check convergence A) on closed sets of the form $\{(-\infty, x] : x \in \mathbb{R}\}$; B) of characteristic functions.

2. When $S = \mathbb{R}^d$ we have the Crámer-Wald device.

3. What about when $S = \mathcal{C}([0, 1])$? How to prove $\mu_n \to \mu$? Two strategies:

   (a) (This holds for any metric space S): *Skorohod's representation:* On some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, construct random variables X_n and X such that a) $X_n \sim \mu_n$; b) $X \sim \mu$; and c) $d(X_n, X) \xrightarrow{d} 0$. Then $\mu_n \xrightarrow{d} \mu$.

(b) For this strategy, we will fill in the details in the coming lectures and for now only illustrate the idea using $\mathbb{R}$-valued (non-random) sequences: Given $\{x_n : n \in \mathbb{N}\} \subset R$, suppose $\{x_n\}$ is bounded (a compactness assumption) and x^* is a limit point[†] of $\{x_n\}$ (a 'uniqueness of limits' assumption) These two conditions imply that $x_n \to x$.

This strategy is not that useful for sequences in $\mathbb{R}$. But it is easily generalisable to S-valued random variables, where S is complete and separable. (Use Prokhorov's theorem to generalise.)

13 Lecture 11/3

13.1 Donsker's theorem

Theorem 13.1 (Donsker's)**.** *Let* $X_1, \ldots \overset{iid}{\sim} F$ *with* $\mathbb{E}X_1 = 0$ *and* $\mathbb{E}X_1^2 = 1$*. Define* $S_k = X_1 + \ldots + X_k$ *and*

$$S^{(n)}(t) = \begin{cases} \frac{S_k}{\sqrt{n}} & \text{if } t = \frac{k}{n}, \\ \textit{linear interpolation} & \textit{otherwise.} \end{cases}$$

Then

$$\{S^{(n)}(t) : 0 \le t \le 1\} \xrightarrow{d} \{B(t) : 0 \le t \le 1\},$$

in $\mathcal{C}([0,1])$*, where* $\{B(t) : 0 \le t \le 1\}$ *is sBM.*

We will prove Donsker's theorem using Skorohod's representation theorem:

Theorem 13.2 (Skorohod's representation)**.** *Given* $X \sim F$ *with* $\mathbb{E}X = 0$ *and* $\mathbb{E}X^2 = 1$*, one can construct (on some probability space)*

$$\left\{B_t : t \in \mathbb{R}^{\ge 0}\right\} \sim \text{sBM}$$
$$(U, V) \perp\!\!\!\perp \left\{B_t : t \in \mathbb{R}^{\ge 0}\right\},$$

[†]i.e. x^* is a limit of a subsequence

with $U \leq 0 \leq V$ surely such that

$$B_T \sim X,$$

where $T = \inf\{t > 0 : B_t \notin (U, V)\}$. Further, $\mathbb{E}T = \mathbb{E}X^2 = 1$.

As a special case of Skorohod's representation, consider

$$X = \begin{cases} 1 & \text{with probability} \frac{1}{2}, \\ -1 & \text{otherwise.} \end{cases}$$

Then $(U, V) = (-1, 1)$ surely and by gambler's ruin (Example 11.2) and Wald's second lemma (Lemma 11.1), we have $B_T \sim X$ and $\mathbb{E}T = \mathbb{E}X^2$.

The proof of Skorohod's representation theorem is constructive.

Proof of Theorem 13.2. Since $\mathbb{E}X = 0$,

$$\mathbb{E}\left[X \mathbb{1}\{X > 0\}\right] = \mathbb{E}\left[-X \mathbb{1}\{X < 0\}\right] = c > 0,$$

where the last inequality follows since $\mathbb{E}X^2 = 1$ implies $\mathbb{P}\left[X = 0\right] < 1$.

Construct a new probability distribution on $(-\infty, 0) \times (0, \infty) \cup \{(0, 0)\}$ with probability measure ν given by

$$\begin{aligned} \nu\left(\{0, 0\}\right) &= \mathbb{P}\left[X = 0\right] \\ \nu(A) &= \frac{1}{c} \int \int_A (v - u) d\mu(u) d\mu(v) \\ &= \frac{1}{c} \mathbb{E}_{X_1, X_2 \overset{iid}{\sim} X} \left[(X_1 - X_2) \mathbb{1}\{X, 1X_2 \in A\}\right], \end{aligned} \tag{23}$$

where $A \subset (-\infty, 0) \times (0, \infty)$ measurable and μ is the law of X. Exercise: check that ν is a probability measure.

Consider any probability space where we can construct $(U, V) \sim \nu$ and

$$\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\} \sim \text{sBM},$$

independently. (For example, two copies of $[0, 1]$ with the Borel sets will suffice.) Define

$$T = \inf\{t \geq 0 : B_t \notin (U, V)\}. \tag{24}$$

We claim that $B_T \sim X$. It suffices to show that, for all bounded, continuous[‡] $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}\phi(B_T) = \mathbb{E}\phi(X).$$

How do we prove this? Basically, condition on (U, V) and use independence of B_t and (U, V): Given $U = u$ and $V = v$,

$$B_T = \begin{cases} u & \text{with probability} \frac{v}{v-u}, \\ v & \text{with probability} \frac{-u}{v-u}, \end{cases}$$

by the gambler's ruin argument (Example 11.2). Then

$$\mathbb{E}\left[\phi(B_T)|U, V\right] = \phi(U)\frac{V}{V-U} + \phi(V)\frac{-U}{V-U}.$$

Thus,

$$\mathbb{E}\phi(B_T) = \mathbb{E}\left[\phi(U)\frac{V}{V-U} + \phi(V)\frac{-U}{V-U}\right] = \mathbb{E}\phi(X),$$

where the last equality is left as an exercise. So $B_T \sim X$. Further

$$\mathbb{E}(T|U, V) = -UV$$

by Wald's second lemma (Lemma 11.1). Therefore,

$$\mathbb{E}T = -\mathbb{E}UV = \mathbb{E}X^2 = 1,$$

where the second last equality is left as an exercise. □

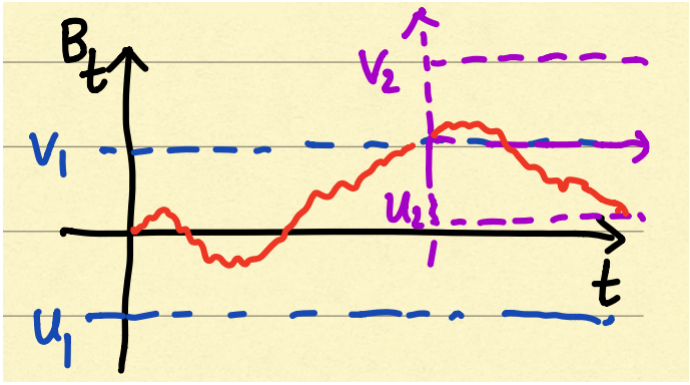
Proof of Donsker's theorem (Theorem 13.1). Embed the process $\{S_k : k \in \mathbb{N}\}$ into Brownian motion B_t: Define T_1 such that $B_{T_1} \sim S_1$. By the strong Markov property, $B_{t-T_1} - B_{T_1}$ is also sBM (for $t \geq T_1$), independent of $\{B_t : 0 \leq t \leq T_1\}$. So we can construct $T_2 > T_1$ such that $B_{T_2} \sim S_2$. Continue this process to construct $T_1 < T_2 < \ldots$ such that

$$(B_{T_1}, B_{T_2}, \ldots) \sim (S_1, S_2, \ldots).$$

[‡]We don't actually need to require that ϕ is continuous, only that it is measurable

More explicitly, let $(U_1, V_1), (U_2, V_2), \ldots \overset{iid}{\sim} \nu$, where ν is given in (23). Define

$$\begin{aligned}
\Delta_1 &= \inf\{t \geq 0 : B_t \notin (U_1, V_1)\}, \\
T_{n-1} &= \Delta_1 + \ldots + \Delta_{n-1}, \\
\Delta_n &= \inf\{t \geq 0 : B_{T_{n-1}+t} - B_{T_{n-1}} \notin (U_n, V_n)\}, \\
X'_n &= B_{T_n} - B_{T_{n-1}}.
\end{aligned}$$



Lemma:

$$\begin{aligned}
\Delta_1, \Delta_2, \ldots &\overset{iid}{\sim} T \\
X'_1, X'_2, \ldots &\overset{iid}{\sim} X_1,
\end{aligned}$$

where T is defined in (24) and X_1 is given in the Theorem statement.

Proof of the lemma: Given $(U_1, V_1), (U_2, V_2), \ldots$, $T_1, T_2, \ldots$ is an increasing sequence of stopping times. The strong Markov property implies $\Delta_1, \Delta_2, \ldots$ are independent and that $X'_1, X'_2, \ldots$ are also independent. Further Δ_i and X'_i are functions of (U_i, V_i), given T_{i-1} and $\{B_t\}$. This implies that $\Delta_i \sim \Delta_j$ and $X'_i \sim X'_j$. Finally, Skorohod's representation implies that $X'_1 \sim X$ and $\Delta_1 \sim T$. This proves the lemma.

This lemma implies

$$\{B_{T_n} : n \in \mathbb{N}\} \sim \{S_n : n \in \mathbb{N}\},$$

where $S_n = X_1 + \ldots + X_n$. This is great but not so applicable: we might need to wait a long time to see B_{T_n}. We need to shrink the process to $[0, 1]$. Also, once we have shrunk the process, we will see that we can ignore the BM between B_{T_n} and $B_{T_{n+1}}$.

Construct

$$W^{(n)}(t) = \frac{1}{\sqrt{n}} B_{nt}$$

$$B^{(n)}(t) = \begin{cases} \frac{B_{T_k}}{\sqrt{n}} & \text{if } t = \frac{k}{n}, \\ \text{linear interpolation} & \text{otherwise.} \end{cases}$$

Observe that $W^{(n)}(t) \sim \text{sBM}$ for all n and that $B^{(n)} \sim S^{(n)}$. So we will have finished the proof if we can show

$$d_{\sup}\left(B^{(n)}, W^{(n)}\right) \xrightarrow{P} 0. \tag{25}$$

More explicitly, using (25) we can apply a Slutsky-type argument to show that

$$S^{(n)} \sim B^{(n)} = \left(B^{(n)} - W^{(n)}\right) + W^{(n)} \xrightarrow{d} 0 + \text{sBM}.$$

To make this argument rigorous, fix $F \subset \mathcal{C}([0,1])$ closed. Define

$$F_\epsilon = \{g : d_{\sup}(g, F) < \epsilon\}.$$

Then

$$\mathbb{P}\left(B^{(n)} \in F_\epsilon\right) \leq \mathbb{P}\left(W^{(n)} \in F_{2\epsilon}\right) + \mathbb{P}\left(d_{\sup}(B^{(n)}, W^{(n)}) > \epsilon\right),$$

and

$$\begin{aligned} \limsup_{n\to\infty} \mathbb{P}\left(B^{(n)} \in F_\epsilon\right) &\leq \limsup_{n\to\infty} \mathbb{P}\left(W^{(n)} \in F_{2\epsilon}\right) + 0 \\ &= \mathbb{P}\left[\text{sBM} \in F_{2\epsilon}\right], \end{aligned}$$

by (25). This gives

$$\limsup_{n\to\infty} \mathbb{P}\left(B^{(n)} \in F\right) \leq \limsup_{n\to\infty} \mathbb{P}\left(B^{(n)} \in F_\epsilon\right) \leq \mathbb{P}\left[\text{sBM} \in F_{2\epsilon}\right].$$

Take $\epsilon \to 0$ and since F is closed and probability continuous,

$$\limsup_{n\to\infty} \mathbb{P}\left(B^{(n)} \in F\right) \leq \mathbb{P}\left(\text{sBM} \in F\right).$$

This completes the proof using Portmanteau theorem, modulo proving (25).

Proof of (25): Let $t \in \left[\frac{k}{n}, \frac{k+1}{n}\right]$. Then we can write

$$\begin{aligned} t &= \alpha\frac{k}{n} + (1-\alpha)\frac{k+1}{\alpha} \\ B^{(n)}(t) &= \alpha\frac{B_{T_k}}{\sqrt{n}} + (1-\alpha)\frac{B_{T_{k+1}}}{\sqrt{n}}, \end{aligned}$$

for some $\alpha \in [0,1]$. This allows us to bound

$$\begin{aligned} \left|W^{(n)}(t) - B^{(n)}(t)\right| &= \left|\frac{1}{\sqrt{n}}B_{nt} - \left(\alpha\frac{B_{T_k}}{\sqrt{n}} + (1-\alpha)\frac{B_{T_{k+1}}}{\sqrt{n}}\right)\right| \\ &\le \frac{\alpha}{\sqrt{n}}|B_{nt} - B_k| + \frac{1-\alpha}{\sqrt{n}}|B_{nt} - B_{k+1}| \\ &\quad + \frac{\alpha}{\sqrt{n}}|B_{T_k} - B_k| + \frac{1-\alpha}{\sqrt{n}}\left|B_{T_{k+1}} - B_{k+1}\right| \\ &\le \frac{1}{\sqrt{n}}\max_{0\le k\le n-1}\sup_{k\le s\le k+1}\left(|B_s - B_k| + |B_s - B_{k+1}|\right) \\ &\quad + \frac{1}{\sqrt{n}}\max_{0\le k\le n}|B_{T_k} - B_k|, \end{aligned}$$

where the last line follows by taking the worst-case over all $t \in [0,1]$. (The above calculations may seem unintuitive, but the basic idea is that the first line would be a lot easier to bound if we could replace the random T_k and T_{k+1} with fixed k and $k+1$. So we do that, and hope that we can bound $|B_{T_k} - B_k|$ since $\mathbb{E}T_k = k\mathbb{E}\Delta_1 = k$.) Since the last line is not a function of t,

$$\begin{aligned} d_{\sup}\left(W^{(n)}, B^{(n)}\right) &\le \frac{1}{\sqrt{n}}\max_{0\le k\le n-1}\sup_{k\le s\le k+1}\left(|B_s - B_k| + |B_s - B_{k+1}|\right) \\ &\quad + \frac{1}{\sqrt{n}}\max_{0\le k\le n}|B_{T_k} - B_k|. \end{aligned}$$

So we just need to show that the two terms

$$\begin{aligned} T_1 &= \frac{1}{\sqrt{n}}\max_{0\le k\le n-1}\sup_{k\le s\le k+1}\left(|B_s - B_k| + |B_s - B_{k+1}|\right), \\ T_2 &= \frac{1}{\sqrt{n}}\max_{0\le k\le n}|B_{T_k} - B_k|, \end{aligned}$$

on the RHS go to zero in probability. Intuitively, $\sup_{k\le s\le k+1}(|B_s - B_k| + |B_s - B_{k+1}|)$ is independent of the other k. So $\sqrt{n}T_1$ is the maximum over n independent Gaussians. We know from Homework 2 that $\sqrt{n}T_1 \approx \sqrt{\log n}$. So $T_1 \xrightarrow{P} 0$. We will make this rigorous and complete the proof next lecture. □

14 Lecture 18/3

14.1 Completing the proof of Donsker's theorem

Recall that we need to prove

(i)
$$\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} (|B_s - B_k| + |B_s - B_{k+1}|) \xrightarrow{P} 0$$

(ii)
$$\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| \xrightarrow{P} 0$$

Proof of (ii). Recall that $T_k = \Delta_1 + \ldots + \Delta_k$ where $\mathbb{E}\Delta_i = \mathbb{E}X_i^2 = 1$. The SLLN gives $\frac{T_n}{n} \xrightarrow{\text{a.s.}} 1$. This implies

$$\max_{1\le k\le n} \frac{|T_k - k|}{n} \xrightarrow{\text{a.s.}} 0. \tag{26}$$

To prove (26) use the following number theory result: If $\{a_n : n \ge 1\} \subset \mathbb{R}$ is a real-valued sequence with $a_n \ge 0$ and $\frac{a_n}{n} \to 1$ then

$$\max_{1\le k\le n} \frac{|a_k - k|}{n} \to 0.$$

The proof of this result is left as an exercise.

Intuitively, T_k is close to k, so by continuity, B_{T_k} should be close to B_k.

Formally,

$$\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| > \epsilon\right] \le \mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| > \epsilon, \max_{1\le k\le n} \frac{|T_k - k|}{n} \le \delta\right]$$

$$+ \mathbb{P}\left[\max_{1\le k\le n} \frac{|T_k - k|}{n} > \delta\right]$$

Define $f_n(\delta) = \mathbb{P}\left[\max_{1\le k\le n} \frac{|T_k-k|}{n} > \delta\right]$. Then for all $\delta > 0$, $f_n(\delta) \to 0$ as $n \to \infty$ by (26). For the first term,

$$\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| > \epsilon, \max_{1\le k\le n} \frac{|T_k - k|}{n} \le \delta\right] \le \mathbb{P}\left[\max_{\substack{0\le a,b\le 2\\ |a-b|\le\delta}} |B_a - B_b| > \epsilon\right], \quad (27)$$

where $\{B_t\}$ is sBM. To see this, note that $\frac{|T_k-k|}{n} \le \delta$ implies that $\frac{T_n}{n} \le 1 + \delta \le 2$ for small enough δ. Also $\frac{T_k}{n} \le 2$ since $T_k \le T_n$. Finally, observe that

$$\left\{\frac{1}{\sqrt{n}} B_{nt} : 0 \le t \le 2\right\} \sim \text{sBM on } [0, 2n].$$

Use this with $a = \frac{T_k}{n}$ and $b = \frac{k}{n}$ to get (27).

Importantly, the RHS of (27) is free of n. This implies we can send δ to zero to get (27) to zero while sending $n \to \infty$ to get $f_n(\delta)$ to zero. More formally, as $\delta \to 0$,

$$\max_{\substack{0\le a,b\le 2\\ |a-b|\le\delta}} |B_a - B_b| \xrightarrow{\text{a.s.}} 0,$$

by uniform continuity. Hence

$$\begin{aligned}\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| > \epsilon\right] &\le \mathbb{P}\left[\max_{\substack{0\le a,b\le 2\\ |a-b|\le\delta^*}} |B_a - B_b| > \epsilon\right] + f_n(\delta^*)\\ &\le \epsilon + f_n(\delta^*)\end{aligned}$$

for small enough δ^*. Yet for large enough n, $f_n(\delta^*) \le \epsilon$, so that

$$\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n} |B_{T_k} - B_k| > \epsilon\right] \le 2\epsilon,$$

for large enough n. Yet ϵ is arbitrary, so the LHS must converge to zero. □

The key idea for this proof was to use rescaling to remove the dependence on n.

Proof of (i).

$$\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} (|B_s - B_k| + |B_s - B_{k+1}|) \le \frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} |B_s - B_k| + \frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} |B_s - B_{k+1}|$$

Since we can flip BM at time $k+1$, the two terms on the RHS have the same distribution. So it suffices to prove that the first term goes to zero in probability.

Observe that

$$\sup_{k\le s\le k+1} |B_s - B_k| \le \left| \sup_{k\le s\le k+1} (B_s - B_k) \right| + \left| \sup_{k\le s\le k+1} (B_k - B_s) \right|,$$

so that

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} |B_s - B_k| > \epsilon \right] &\le \mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \left| \sup_{k\le s\le k+1} (B_s - B_k) \right| > \epsilon/2 \right] \\
&\quad + \mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \left| \sup_{k\le s\le k+1} (B_k - B_s) \right| > \epsilon/2 \right] \\
&= 2\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \left| \sup_{k\le s\le k+1} (B_s - B_k) \right| > \epsilon/2 \right] \\
&\le \frac{4}{\sqrt{n}\epsilon} \mathbb{E}\left[\max_{0\le k\le n-1} \left| \sup_{k\le s\le k+1} (B_s - B_k) \right| \right] \qquad (28)
\end{aligned}$$

by Markov's inequality. By the Markov property and (15),

$$\sup_{k\le s\le k+1} (B_s - B_k) \overset{iid}{\sim} |Z| \text{ where } \mathcal{N}(0,1),$$

for $k = 0, \dots, n-1$. We then claim that

$$\mathbb{E}\left[\max_{0\le k\le n-1} |Z_i| \right] \le C\sqrt{\log n}, \qquad (29)$$

for some constant C and $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$. The proof is similar to the bound on $\mathbb{E}\left[\max_{1\le i\le n} Z_i\right]$ proved in Homework 2. Thus,

$$\mathbb{P}\left[\frac{1}{\sqrt{n}} \max_{0\le k\le n-1} \sup_{k\le s\le k+1} |B_s - B_k| > \epsilon \right] \le \frac{4}{\sqrt{n}\epsilon} \mathbb{E}\left[\max_{0\le k\le n-1} \left| \sup_{k\le s\le k+1} (B_s - B_k) \right| \right]$$

$$\leq \frac{4C\sqrt{\log n}}{\epsilon\sqrt{n}}$$
$$\to 0.$$

Proof of (29): Integrate the survival function:

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq k\leq n}|Z_i|\right] &= \int_0^\infty \mathbb{P}\left[\max_{1\leq k\leq n}|Z_i| > t\right] dt \\
&= \int_0^{\sqrt{2\log n}} \mathbb{P}\left[\max_{1\leq k\leq n}|Z_i| > t\right] dt + \int_{\sqrt{2\log n}}^\infty \mathbb{P}\left[\max_{1\leq k\leq n}|Z_i| > t\right] dt \\
&\leq \sqrt{2\log n} + n\int_{\sqrt{2\log n}}^\infty \mathbb{P}\left[|Z| > \frac{t}{n}\right] dt \\
&\leq \sqrt{3\log n},
\end{aligned}$$

where the second last line follows by bounding the integrand of the first term by 1 and using a union bound on the second term; and the final line follows since $\Phi(x) \leq \frac{1}{2}\exp(-x^2/2)$. □

This completes the proof of Donsker's theorem. This proof is very specific to scaled random walks. It is hard to do the embedding of $\{S_k : k \in \mathbb{N}\}$ into Brownian motion B_t otherwise. We want a more robust, generally-applicable technique for proving convergence in distribution in $\mathcal{C}([0,1])$.

14.2 A general strategy for proving weak convergence

We foreshadowed this a few lectures ago.

Lemma 14.1. *Let (S,d) be a complete, separable metric space. Suppose $\{X_n : n \in \mathbb{N}\}$ is a sequence of S-valued random variables and X a S-valued random variable. Suppose that*

(i) Every subsequence $\{X_{n_k} : k \in \mathbb{N}\}$ has a weakly convergent subsequence $\{X_{n_{k_l}} : l \in \mathbb{N}\}$;

(ii) If any subsequence $\{X_{n_k} : k \in \mathbb{N}\}$ converges in distribution, then $X_{n_k} \xrightarrow{d} X$;

Then $X_n \xrightarrow{d} X$.

The fist condition can be considered as a formulation of compactness.

Proof. Suppose not. Then there exists a subsequence $\{X_{n_k} : k \in \mathbb{N}\}$, a bounded and continuous function $h : S \to \mathbb{R}$ and $\epsilon > 0$ such that

$$|\mathbb{E}\left[h\left(X_{n_k}\right)\right] - \mathbb{E}\left[h(X)\right]| > \epsilon,$$

for all k. (This is exactly the negation of convergence in distribution.)

Conditions (i) and (ii) imply that there exists a further subsequence $\{X_{n_{k_l}} : l \in \mathbb{N}\}$ which converges weakly to X. Hence

$$\left|\mathbb{E}\left[h\left(X_{n_{k_l}}\right)\right] - \mathbb{E}\left[h(X)\right]\right| < \epsilon,$$

for large enough l. □

This Lemma is a step towards proving weak convergence. Yet we still need strategies for proving conditions (i) and (ii).

14.2.1 Proving condition (ii)

The following Lemma states that finite dimensional distributional convergence implies condition (ii).

Lemma 14.2. *Let* $S = \mathcal{C}([0,1])$ *in the setup of the previous Lemma. Suppose that for all* $k \geq 1$, *and all* $0 \leq t_1 \leq \ldots \leq t_k \leq 1$,

$$(X_n(t_1), \ldots, X_n(t_k)) \xrightarrow{d} (X(t_1), \ldots, X(t_k)). \tag{30}$$

Then condition (ii) of Lemma 14.1 holds – that is: If any subsequence $\{X_{n_k} : k \in \mathbb{N}\}$ *converges in distribution, then* $X_{n_k} \xrightarrow{d} X$.

(30) is called finite dimensional (distributional/weak) convergence.

This result is extremely useful, since it transfers conditions for convergence in a infinite-dimensional, difficult-to-handle space CZO into a condition on convergence in $\mathbb{R}^k$. And we have many powerful tools for proving convergence in $\mathbb{R}^k$.

This Lemma holds for other metric spaces, beyond $\mathcal{C}([0,1])$, yet it requires S to be a continuous function space.

Proof. Suppose $\{X_{n_k} : k \in \mathbb{N}\}$ converges weakly to Y. Then

$$(X_{n_k}(t_1), \ldots, X_{n_k}(t_l)) \xrightarrow[k\to\infty]{d} (Y(t_1), \ldots, Y(t_l)),$$

by CMT. Yet this implies $X \sim Y$ by the interpolation argument we have used before. (Crucially, this requires continuity of X and Y.) □

14.2.2 Proving condition (i) - tightness

We want to prove that every subsequence has a further convergent subsequence.

Example 14.3. Consider

1. $X_n \sim \mathcal{N}(0, \sigma^2)$ where $\{\sigma_n^2\}_{n=1}^{\infty}$ is bounded.
2. $X_n = x_n$ with probability 1, where $\{x_n\}_{n=1}^{\infty}$ is a real-valued sequence diverging to infinity.

Since $\{\sigma_n^2\}_{n=1}^{\infty}$ is bounded, there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ such that $\sigma_{n_k}^2 \to \sigma_*^2$ (by the Heine-Borel theorem). Thus $X_{n_k} \xrightarrow{d} \mathcal{N}(0, \sigma_*^2)$.

Whereas for 2., we can never extract convergent subsequences as they are always going off to infinity.

The insight is that in example 1., we can bound X_n in probability, uniformly in n – that is,

$$\mathbb{P}[X_n \in [-M, M]] > 1 - \epsilon.$$

On the other hand, we cannot do this for example 2. This turns out to be the crucial property for proving (i). We formalise this with the following defintion.

Definition 14.4. A sequence $\{X_n : n \in \mathbb{N}\}$ of S-valued random variables is *tight* if, for all $\epsilon > 0$, there exists a compact set K_ϵ such that

$$\mathbb{P}[X_n \in K_\epsilon] > 1 - \epsilon,$$

for all $n \geq 1$.

The definition of tightness intuitively captures the idea that "the measure gets trapped in compact sets". It is connected to condition (i) via a deep result:

Theorem 14.5 (Prokhorov)**.** *Let (S, d) be a separable metric space and suppose $\{X_n : n \in \mathbb{N}\}$ is a sequence of S-valued random variables. Then $\{X_n : n \in \mathbb{N}\}$ satisfies condition (i) of Lemma 14.1 – that is every subsequence has a further convergent subsequence – if and only if $\{X_n : n \in \mathbb{N}\}$ is tight.*

Prokhorov's theorem gives a characterisation of compactness in terms of tightness. This result will not be proved in lectures.

14.2.3 Summary

To recap, we can prove weak convergence in $S = \mathcal{C}([0, 1])$ by showing tightness and finite dimensional convergence. Finite dimensional convergence is usually straightforward to establish using existing tools from Stat210 (Portmanteau, characteristic functions, etc.). For tightness in $S = \mathcal{C}([0, 1])$, there exists explicit, easily checkable conditions; we won't cover these in lectures, instead see Section 8, Parts 4 and 5.

15 Lecture 23/3

15.1 General stochastic processes

Recall our proof of why sBM exists. We showed this by Levy's construction, which utilised very specific properties of BM. What if we are interested in constructing other stochastic processes? We will see that Kolmogorov's existence theorem will help us here. First, we need to understand what is meant by the term 'stochastic process'.

Definition 15.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A collection $\{X_t : t \in \Pi\}$ is a stochastic process if, for all $t \in \Pi$,

$$X_t : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}}),$$

is measurable.

Put simply, a stochastic process is any collection of random variables defined on the same space.

Π is referred to as the *index set*, and it generally conveys the notion of time (or space). Typically, $\Pi = \mathbb{N}$ (discrete time), $\Pi = \mathbb{R}^{\geq 0}$ (continuous time) or $\Pi = [0, 1]$.

We required that each X_t is $\mathbb{R}$-valued. This is not necessary – but all the X_t's must share a common codomain, called the *state space*.

We found that thinking of sBM as a $\mathcal{C}([0, 1])$-valued random variable was very useful. This can be carried across more generally.

If $\{X_t : t \in [0, 1]\}$ is a stochastic process with index set $[0, 1]$ and state space $\mathbb{R}$, then, for all $\omega \in \Omega$, the function

$$t \mapsto X_t(\omega)$$

(called the trajectory map) lives in $\mathbb{R}^{[0,1]} = \{f : [0, 1] \to \mathbb{R}\}$. So can we think of a general stochastic process as a $\mathbb{R}^{[0,1]}$-valued random variable? Yes, although we first need to build an appropriate σ-algebra on $\mathbb{R}^{[0,1]}$ in order to do this.

Definition 15.2. For all $t \in [0, 1]$, define the evaluation map

$$\begin{aligned} \pi_t : \mathbb{R}^{[0,1]} &\to \mathbb{R} \\ f &\mapsto f(t). \end{aligned}$$

Define the *Borel σ-algebra* (or cyclindrical σ-algebra) on $\mathbb{R}^{[0,1]}$ to be

$$\mathcal{B}\left(\mathbb{R}^{[0,1]}\right) = \sigma\left(\{\pi_t : t \in [0, 1]\}\right),$$

the smallest σ-algebra such that each evaluation map π_t is measurable.

So $\mathcal{B}\left(\mathbb{R}^{[0,1]}\right)$ is the smallest σ-algebra so that, for each $t \in [0, 1]$ and $X : \Omega \to \mathbb{R}^{[0,1]}$,

$$\omega \mapsto \pi_t(X(\omega)),$$

is a $\mathbb{R}$-valued random variable.

We can similarly define the cylindrical σ-algebra on any function space, $\{f : X \to Y\}$, provided that there is a Borel σ-algebra on Y.

15.1.1 Constructing stochastic processes

Proposition 15.3. *$\{X_t : t \in [0,1]\}$ is a stochastic process if and only if*

$$(X_t)_{t\in[0,1]} : (\Omega, \mathcal{F}) \to (\mathbb{R}^{[0,1]}, \mathcal{B}(\mathbb{R}^{[0,1]})),$$

is measurable.

This proposition generalises from $\mathbb{R}^{[0,1]}$ to any function space for which we can define a cylindrical σ-algebra.

Why is this proposition useful? It shows an equivalence between two questions: 1) Does there exist a specific stochastic process (i.e. a process satisfying some given properties)? 2) Can we construct an appropriate distribution on $(\mathbb{R}^{[0,1]}, \mathcal{B}(\mathbb{R}^{[0,1]}))$? Previously, we would need to construct infinitely many – one for each X_t – distributions on $\mathbb{R}$. Now we just need to construct a single distribution. This is equivalent, but conceptually it is easier to create a single random variable rather than infinitely many. So this proposition is a step forward in proving the existence of stochastic processes.

Example 15.4 (Poisson process). A stochastic process $\{N_t : t \in \mathbb{R}^{\geq 0}\}$ satisfying

() $N_t \in \mathbb{N}$ for all t;

(i) $N_0 = 0$;

(ii) $N_t \sim \text{Pois}(\lambda t)$;

(iii) $N_{t+s} - N_s \sim N_t$;

(iv) $N_t - N_s \perp\!\!\!\perp N_r - N_q$, for $q \leq r < s \leq t$; and

(v) $t \mapsto N_t$ is right continuous and non-decreasing.

How can we show that there is a stochastic process which satisfies conditions (i)-(iv)? Proposition 15.3 says that to answer this question in the affirmative, we need only construct some distribution on $(\mathcal{N}^{[0,\infty)}, \mathcal{B}(\mathbb{N}^{[0\infty)}))$.

Conditions (i)-(iv) state the finite dimensional properties of the process. Is there some way of telling whether a process with certain finite dimensional properties exists or not?

Example 15.5. As a second example, can we construct a stochastic process $\{X_t : t \in [0,1]\}$ with the following properties?

1. for all $0 \leq t_1 \leq \ldots \leq t_k \leq 1$,
$$(X_{t_1}, \ldots, X_{t_k}) \sim \text{MVN};$$
2. $X_t \sim \mathcal{N}(0,1)$ for all t,
3. for all $s < t$,
$$\begin{bmatrix} X_s \\ X_t \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 2 & \rho \\ \rho & 2 \end{bmatrix}\right)$$

As in Example 15.4, conditions 1.-3. specify the finite dimensional properties of the process. Yet it is impossible to construct a stochastic process satisfying conditions 1.-3., since 3. implies $\text{Var}(X_t) = 2$ while 2. implies $\text{Var}(X_t) = 1$ – the finite dimensional (aka marginal) properties are not consistent.

So consistency of the marginals is a necessary condition for existence of a stochastic process. Kolmogorov's existence theorem states that it is also a sufficient condition.

Definition 15.6. A set of marginals
$$\{Q_{t_1,\ldots,t_k} : k \geq 1, 0 \leq t_1 \leq \ldots \leq t_k \leq 1\},$$
where $Q_{t_1,\ldots,t_k}$ is a probability distribution on $\mathbb{R}^k$, is *consistent* if
$$Q_{t_1,\ldots,t_{k-1}}(A) = Q_{t_1,\ldots,t_k}(A \times \mathbb{R}),$$
for all $k \geq 1, 0 \leq t_1 \leq \ldots \leq t_k \leq 1$ and $A \in \mathcal{B}_{\mathbb{R}^{k-1}}$.

Theorem 15.7 (Kolmogorov's existence (or extension) theorem)**.** *If* $\{Q_{t_1,\ldots,t_k} : 0 \leq t_1 \leq \ldots \leq t_k \leq 1\}$ *is consistent then there exists a probability measure* P *on* $(\mathbb{R}^{[0,1]}, \mathcal{B}(\mathbb{R}^{[0,1]}))$ *such that*
$$P\left(\{f \in \mathbb{R}^{[0,1]} : (f(t_1), \ldots, f(t_k)) \in A\}\right) = Q_{t_1,\ldots,t_k}(A),$$
for all $k \geq 1$*, all* $0 \leq t_1 \leq \ldots \leq t_k \leq 1$ *and* $A \in \mathcal{B}_{\mathbb{R}^k}$.

Kolmogorov's existence theorem shows that there always exists a stochastic process which satisfies any given consistent, finite dimensional distributional specification. The proof of this Theorem is deep and not covered here.

Note that the Theorem does not provide a uniqueness result – in general the resulting stochastic process is not unique without enforcing further constraints (typically continuity).

Example 15.8. Does there exist $\{X_t : 0 \leq t \leq 1\}$ such that, for all $k \geq 1$ and $0 \leq t_1 \leq \ldots \leq t_k \leq 1$,

$$(X_{t_1}, \ldots, X_{t_k}) \sim \mathcal{N}(\mathbf{0}, \Sigma_{t_1,\ldots,t_k}), \tag{31}$$

where $\Sigma_{t_1,\ldots,t_k} \in \mathbb{R}^{k\times k}$ with $\Sigma_{t_i,t_j} = t_i \wedge t_j$? Yes, because sBM satisfies (31). But also because we can check that $\Sigma_{t_1,\ldots,t_k}$ is positive semi-definite; (31) is consistent and then apply Kolmogorov's existence theorem.

Is the process in the above example Brownian motion? Not necessarily, since you can have many processes satisfying (31).

Example 15.9. Let $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\} \sim$ sBM and $U \sim \text{Unif}([0,1])$ independent of $\left\{B_t : t \in \mathbb{R}^{\geq 0}\right\}$. Define

$$Y_t = \begin{cases} B_t & \text{if } t \neq U, \\ 0 & \text{otherwise.} \end{cases}$$

Then marginally Y_t has the same finite dimensional distributions as sBM!

Kolmogorov's existence theorem is a very powerful black box. But it says nothing about uniquness- there can be many processes satisfying any given consistent marginals, since $\mathbb{R}^{[0,1]}$ is so big. Continuity is critical for the uniqueness of sBM.

Given this, one might view Kolmogorov's existence theorem in a negative light. But there is an easy fix.

15.1.2 Modifications and continuous stochastic processes

Definition 15.10. A stochastic process $\{Y_t : t \in [0,1]\}$ is called a *modification* of $\{X_t : t \in [0,1]\}$ if, for all $t \in [0,1]$,

$$\mathbb{P}[X_t \neq Y_t] = 0.$$

Note that the for all quantifier is outside the probability. So we are not requiring that $\{X_t\}$ and $\{Y_t\}$ are equal everywhere almost surely, but that they are equal at any given point t almost surely.

It is straightforward to see that a modification of $\{X_t\}$ has the same finite dimensional distribution as $\{X_t\}$.

Theorem 15.11 (Kolmogorov-Chentsov)**.** *Suppose $\{X_t : t \in [0,1]\}$ is a stochastic process. Assume that there exists $\alpha, \beta > 0$ and $0 \leq c < \infty$ such that*

$$\mathbb{E}|X_t - X_s|^\alpha \leq c|t-s|^{1+\beta}, \tag{32}$$

for all $s, t \in [0,1]$. Then there exists a continuous modification of X_t.

The proof of this Theorem is not covered in class. By continuous modification, we mean a modification of X_t with a.e. continuous trajectories:

$$\mathbb{P}\left(\{\omega : t \mapsto Y_t(\omega) \text{ is continuous }\}\right) = 1.$$

(32) can be thought of as a stochastic continuity (or more specifically, Hölder continuity) condition.

Since the condition (32) can be checked easily, this Theorem – combined with Kolmogorov's existence theorem – gives a general and powerful strategy for constructing continuous stochastic processes: first prove consistency of the finite dimensional distributions; then appeal to Kolmogorov's existence theorem; show (32) holds; and finally apply Kolmogorov-Chentsov's theorem. This strategy gives an alternate construction of sBM (which we will see next lecture).

16 Lecture 25/3

16.1 An alternate proof of the existence of BM

This proof utilises the Kolmogorov existence and Kolmogorov-Chenstov theorems from the previous lecture.

Claim 16.1. *There exists a stochastic process* $\{B_t : t \in [0,1]\}$ *with continuous trajectories such that, for all* $k \geq 1$ *and* $0 \leq t_1 \leq \ldots \leq t_k \leq 1$,

$$(X_{t_1}, \ldots, X_{t_k}) \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{33}$$

where $\Sigma_{ij} = t_i \wedge t_j$.

Proof. Part (i): Fix $k \geq 1$ and $0 \leq t_1 \leq \ldots \leq t_k \leq 1$. We need to check that Σ is positive semi-definite. Let

$$Q_{t_1,\ldots,t_k} \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

as given in (33). Then the collection

$$\{Q_{t_1,\ldots,t_k} : k \geq 1, 0 \leq t_1 \leq \ldots \leq t_k \leq 1\},$$

is consistent, by properties of MVN. The Kolmogorov existence theorem implies that there exists a stochastic process $\{Y_t : t \in [0,1]\}$ with finite dimensional distributions $\{Q_{t_1,\ldots,t_k}\}$.

Part (ii): Since $Y_t - Y_s \sim \mathcal{N}(0, t-s)$, for $t \geq s$,

$$\mathbb{E}|Y_t - Y_s|^4 = 3(t-s)^2.$$

Hence $\{Y_t : t \in [0,1]\}$ satisfies the Kolmogorov-Chentsov condition. So there exists a continuous modification $\{X_t : t \in [0,1]\}$ of $\{Y_t : t \in [0,1]\}$. Since $\{X_t : t \in [0,1]\}$ is continuous and satisfies the finite dimensional distribution properties (33), it is sBM. □

This proof exhibits a widely-applicable template for constructing stochastic processes with continuous trajectories.

16.2 Stochastic integrals

16.2.1 Introduction to the Ito integral

As motivation, let $\{Y_t\}$ be the (one-dimensional) position of a particle moving in a liquid medium. Let $v : \mathbb{R} \to \mathbb{R}$ be the position-dependent velocity – i.e. $v(x)$

is the velocity of the particle when it is at the point x. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the position-dependent noise-strength. We have the discrete-time model:

$$Y_{t+\epsilon} = Y_t + \epsilon v(t) + \sqrt{\epsilon}\sigma(Y_t)\xi_t,$$

for some small ϵ where $\xi_t \sim \mathcal{N}(0,1)$. The second term in the RHS is the movement due to the known velocity and the third term is the stochastic movement. If $Y_0 = 0$, then

$$Y_1 = \epsilon \sum_{i=1}^{1/\epsilon} v(Y_{i\epsilon}) + \sqrt{\epsilon} \sum_{i=1}^{1/\epsilon} \sigma(Y_{i\epsilon})\xi_{i\epsilon}.$$

What happens as ϵ approaches zero? We know that

$$\epsilon \sum_{i=1}^{1/\epsilon} v(Y_{i\epsilon}) \xrightarrow{\epsilon \to 0} \int_0^1 v(Y_s)ds,$$

assuming absolute continuity of v. But how can we understand the 'integral' of the second, stochastic term? It includes a random component $\xi_{i\epsilon}$. We would like to say something along the lines of

$$\sqrt{\epsilon} \sum_{i=1}^{1/\epsilon} \sigma(Y_{i\epsilon})\xi_{i\epsilon} \xrightarrow{\epsilon \to 0} \int_0^1 \sigma(Y_s)dB_s,$$

where the RHS is a 'stochastic integral', which is meant to capture the cumulative effect to the particle's location due to random movement. This notation should be reminiscent of Riemann-Stieltjes integration – in fact, the stochastic integral is a generalisation of the R.S. integral which allows for integrating some functions of unbounded variation (for example, Brownian motion – we showed that it is of unbounded variation in section 4) by making use of weaker forms of limits (use the L^2 limit rather than the pointwise limit).

Consider the specal case $\sigma(x) = 1$. Then

$$\left\{ \sqrt{\epsilon} \sum_{i}^{t/\epsilon} z_{i\epsilon} : t \in [0,1] \right\} \xrightarrow[\epsilon \to 0]{d} \{B_t : t \in [0,1]\}.$$

Hence, any reasonable definition of stochastic integration should have

$$\int_0^t dB_s = B_t.$$

16.2.2 Formal construction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. (Unlike in previous concepts, the underlying probability space is crucial for defining the stochastic integral.) Let $\{B_t : t \geq 0\} \sim$ sBM on $(\Omega, \mathcal{F}, \mathbb{P})$. Henceforth, we will work with a filtration $\{\mathcal{F}_t : t \in \mathbb{R}^{\geq 0}\}$ such that

(i) $\{B_t : t \in \mathbb{R}^{\geq 0}\}$ is adapted to $\{\mathcal{F}_t : t \in \mathbb{R}^{\geq 0}\}$;

(ii) $\{(B_t, \mathcal{F}_t) : t \in \mathbb{R}^{\geq 0}\}$ satisfies the strong Markov property;

(iii) for all $t > 0$, $\mathcal{F}_t$ is complete – that is, if $N \in \mathcal{F}$ with $\mathbb{P}(N) = 0$ then $N \in \mathcal{F}_t$.

Completeness says that the σ-algebra contains all of the null sets of $(\mathcal{F}, \mathbb{P})$.

We know that the canonical right continuous filtration satisfies properties (i) and (ii).

Two natural questions immediately arise: 1. Does a filtration satisfying (i)-(iii) even exist? 2. Why do we need property (iii)?

We will delay the answer to question 2.: later we will want that

$$\int_0^t f_s dB_s,$$

is measurable with respect to $\mathcal{F}_t$. Completeness will be useful to prove this. For now it suffices to say that it is just a technical condition.

In answering question 1., it turns out that assuming completeness is not such a big deal: Let $\mathcal{F}_t^+$ be the canonical right continuous filtration with respect to $\{B_t : t \in \mathbb{R}^{\geq 0}\}$. Define the collection of null sets,

$$\mathcal{N} = \{N \in \mathcal{F} : \mathbb{P}(N) = 0\}.$$

Then

$$\mathcal{F}_t = \sigma\left(\mathcal{F}_t^+ \cup \mathcal{N}\right),$$

satisfies properties (i)-(iii). Henceforth assume that this is the filtration we are working with, when we discuss stochastic integration.

To define $\int_0^\infty f_s dB_s$, what is the "right" class of possible integrands $\{f : s \in \mathbb{R}^{\geq 0}\}$? We want a generally applicable integration theory, but we can't allow the collection of integrable functions to be too big, otherwise the integral will lose its nice properties.

It turns out that the "right" class is the progressively measurable processes (with finite L^2 norm[§]):

Definition 16.2. A stochastic process

$$\{X(t,\omega) : t \in \mathbb{R}^{\geq 0}, \omega \in \Omega\},$$

is *progressively measurable* if, for all $t > 0$,

$$\begin{aligned} [0,t] \times \Omega &\to \mathbb{R} \\ (s,\omega) &\mapsto X(s,\omega), \end{aligned}$$

is measurable with respect to $\mathcal{B}_{[0,1]} \otimes \mathcal{F}_t$[¶].

What are some examples of progressively measurable processes?

Lemma 16.3. *Any process $\{X_t : t \in \mathbb{R}^{\geq 0}\}$ which is adapted to $\mathcal{F}_t$ and is either left or right continuous, is progressively measurable.*

Proof sketch. Assume $\{X_t : t \in \mathbb{R}^{\geq 0}\}$ is right continuous WLOG. Fix $T > 0$ and divide $[0,T]$ into intervals of length $T/2^n$. Define

$$X_n(s,\omega) = X\left(\frac{(k+1)T}{2^n}, \omega\right),$$

if $\frac{kT}{2^n} < s \leq \frac{(k+1)T}{2^n}$. (We choose the right end-point of the interval for right continuity.)

Check that for each $n \geq 1$,

$$X_n : [0,T] \times \Omega \to \mathbb{R},$$

[§] We will define precisely what we mean by this later.

[¶] $\mathcal{B}_{[0,1]}$ is the Borel σ-algebra on $[0,1]$ and $\otimes$ denotes the product σ-algebra.

is measurable with respect to $\mathcal{B}_{[0,1]} \otimes \mathcal{F}_t$. The intuition for this is that X_n is just piecewise constatn in time, so it must be measurable. By right continuity, $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$. This implies that

$$\begin{aligned} [0,T] \times \Omega &\to \mathbb{R} \\ (s,\omega) &\mapsto X(s,\omega), \end{aligned}$$

is also measurable with respect to $\mathcal{B}_{[0,1]} \otimes \mathcal{F}_t$. □

16.2.3 Defining the stochastic integral

The construction of the stochastic integral is very similar to the InSiPoD definition of the Lebesgue intregral.

Step 1: Restrict to "simple functions":

$$H(t,\omega) = \sum_{i=1}^{k} A_i(\omega) \mathbb{1}_{(t_i, t_{i+1}]}(t), \tag{34}$$

where $0 \leq t_1 \leq \ldots \leq t_{k+1}$ and A_i is $\mathcal{F}_{t_i}$-measurable.

Step 2: Define the stochastic integral for "simple functions" (34) by

$$\int_0^\infty H_s dB_s := \sum_{i=1}^{k} A_i \left(B_{t_{i+1}} - B_{t_i} \right).$$

add-on We could replace B_s with another stochastic process and then continue the procedure below. In this way we can define the stochastic integral for general processes, not just Brownian motion. However, this general integral would not satisfy the nice properties as the Brownian motion integral, such as Ito isometry (Theorem 17.5).

Step 3: Take limits: We will prove that, given any progressively measurable H with finite L^2 norm, there exists a sequence $\{H_n\}_{n=1}^{\infty}$ satisfying (34) such that

$$\mathbb{E} \int_0^\infty \left(H_n(s) - H(s) \right)^2 ds \to 0,$$

and then define the stochastic integral

$$\int_0^\infty H(s)dB_s = \lim_{n\to\infty} \int_0^\infty H_n(s)dB_s.$$

The limit is in the L^2 sense, but we need to make this precise. We also need to precisely define what we mean by H having finite L^2 norm. Finally, we must make sure that this is well-defined (i.e. independent of the choice of the approximating sequence $\{H_n\}_{n=1}^\infty$).

17 Lecture 30/3

The reference for today and last lecture's material is [MP, Chp. 7].

17.1 Defining the stochastic integral (cont.)

Recall that we need to establish a number of lemmas in order to lay the groundwork for a proper definition of the stochastic integral: 1) That progressively measurable processes are approximable by step processes; 2) that the approximating step processes have L^2-convergent stochastic integrals; and 3) that the L^2-limit is independent of the approximating step processes. The following result shows that we can approximate progressively measurable processes by step processes.

Lemma 17.1. *Let $\{H(s,\omega) : s \in \mathbb{R}^{\geq 0}, \omega \in \Omega\}$ be a progressively measurable stochastic process. Assume that*

$$\mathbb{E}\int_0^\infty H^2(s)ds < \infty. \tag{35}$$

Then there exists a sequence $\{H_n : n \in \mathbb{N}\}$ of progressively measurable step processes with

$$\mathbb{E}\int_0^\infty \left[H_n(s) - H(s)\right]^2 ds \xrightarrow{n\to\infty} 0.$$

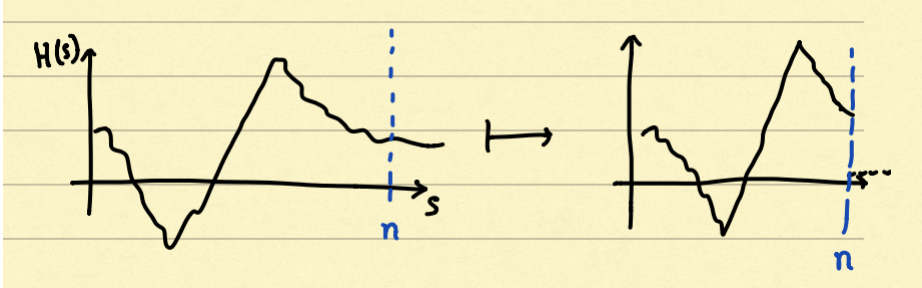
Given a stochastic process $X(t,\omega)$ on $[0,\infty)$, we define its L^2-norm as

$$\mathbb{E}\int_0^\infty X^2(t)dt = \int_\Omega \int_0^\infty X^2(t,\omega)dtd\mu(\omega).$$

So (35) is assuming that H has finite L^2-norm, while the conclusion of Lemma 17.1 is that H_n approximates H in the L^2 sense.

Proof sketch. Step 1: "Truncation step": Truncate H to time n. That is, define

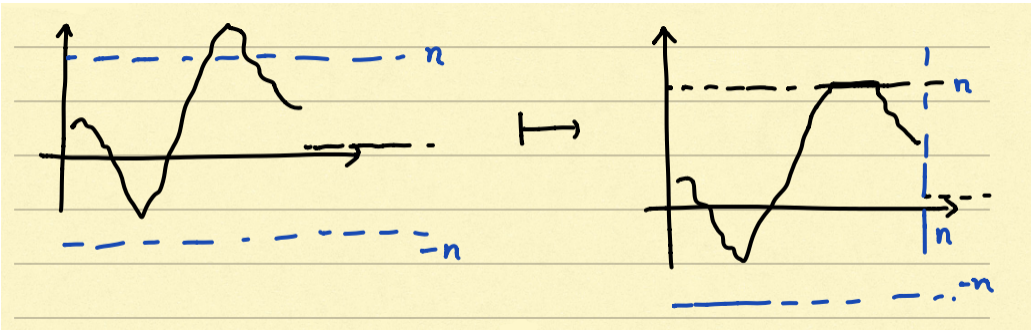
$$\tilde{H}_n(s) = \begin{cases} H(s) & \text{if } s \le n, \\ 0 & \text{otherwise.} \end{cases}.$$



(Note that the trajectories need not be continuous, despite the fact that they are drawn as continuous.)

Step 2: "Localisation step": Localise the process on the y-axis. That is, define

$$H_n^*(s) = \begin{cases} \tilde{H}_n(s) & \text{if } -n \le s \le n, \\ -n & \text{if } \tilde{H}_n(s) < -n, \\ n & \text{if } \tilde{H}_n(s) > n. \end{cases}$$



Importantly, due to L_2 boundedness, we can approximate H by H_n^* in the L^2 sense.

Step 3: Now approximate H_n^* by continuous, progressively measurable functions:

$$\hat{H}_n(s,\omega) = n \int_{s-\frac{1}{n}}^{s} H_n^*(t,\omega)dt.$$

$\hat{H}_n$ averages the process H_n^* over smaller and smaller intervals.

Facts (proof left as an exercise): $\hat{H}_n$ is bounded and continuous. It is also progressively measurable for all n (to prove this, use the fact that the integral is over $[s-1/n, n]$).

By the Lebesgue differentiation theorem, $\hat{H}_n(s) \xrightarrow{\text{a.s.}} H(s)$ for all s. Further

$$\mathbb{E}\int_0^\infty \left[H(s,\omega) - \hat{H}_n(s,\omega)\right]^2 ds \to 0,$$

as $n \to \infty$ by the DCT and assumption (35).

Step 4: Approximate $\hat{H}_n$ by progressively measurable step processes H_n. □

Lemma 17.2. *Let H be a progressively measurable step process with*

$$\mathbb{E}\int_0^\infty H^2(s)ds < \infty.$$

Then

$$\mathbb{E}\left[\left(\int_0^\infty H(s)dB_s\right)^2\right] = \mathbb{E}\int_0^\infty H^2(s)ds. \tag{36}$$

(36) is called the Ito isometry. Later we will prove it holds for any progressively measurable process with finite L^2 norm (Theorem 17.5). It says that the L^2 norm of the stochastic integral (which is itself a random variable) is equal to the L^2 norm of the stochastic process.

Proof. Given the definition of the stochastic integral for step processes, we can just verify this property directly. Let

$$H(t,\omega) = \sum_{i=1}^{k} A_i(\omega)\mathbb{1}\{t \in (a_i, a_{i+1}]\},$$

where the A_i are $\mathcal{F}_{a_i}$-measurable. We defined the stochastic integral

$$\int_0^\infty H(s)dB_s = \sum_{i=1}^k A_i \left(B_{a_{i+1}} - B_{a_i}\right)^2.$$

Hence

$$\mathbb{E}\left[\left(\int_0^\infty H(s)dB_s\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^k A_i^2 \left(B_{a_i+1} - B_{a_i}\right)^2\right] + \sum_{i\neq j} \mathbb{E}\left[A_i A_j \left(B_{a_{i+1}} - B_{a_i}\right)\left(B_{a_{j+1}-a_j}\right)\right]. \quad (37)$$

For the first term, use the tower law,

$$\begin{aligned}\mathbb{E}\left[A_i^2 \left(B_{a_i+1} - B_{a_i}\right)^2\right] &= \mathbb{E}\left[A_i^2 \mathbb{E}\left[\left(B_{a_i+1} - B_{a_i}\right)^2 \middle| \mathcal{F}_{a_i}\right]\right] \\ &= \mathbb{E}\left[A_i^2\right](a_{i+1} - a_i),\end{aligned}$$

where the first line follows since A_i is $\mathcal{F}_{a_i}$-measurable and the second line since $B_{a_{i+1}} - B_{a_i} \perp\!\!\!\perp \mathcal{F}_{a_i}$. For the second term of (37), assume WLOG that $i < j$,

$$\begin{aligned}\mathbb{E}\left[A_i A_j \left(B_{a_{i+1}} - B_{a_i}\right)\left(B_{a_{j+1}-a_j}\right)\right] &= \mathbb{E}\left[A_i A_j \left(B_{a_{i+1}} - B_{a_i}\right) \mathbb{E}\left[B_{a_{j+1}} - B_{a_j} \middle| \mathcal{F}_{a_j}\right]\right] \\ &= 0,\end{aligned}$$

where the first line follows since $A_i A_j \left(B_{a_{i+1}} - B_{a_i}\right)$ is $\mathcal{F}_{a_j}$-measurable and the second line since $\mathbb{E}\left[B_{a_{j+1}} - B_{a_j}\right] = 0$. Therefore,

$$\mathbb{E}\left[\left(\int_0^\infty H(s)dB_s\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^k A_i^2 \left(a_{i+1} - a_i\right)\right] = \mathbb{E}\int_0^\infty H^2(s)ds. \qquad \square$$

Lemma (17.2) is important since it allows us to establish that the Ito integral is well defined (i.e. independent of the choice of the approximating step processes), via the following corollary.

Corollary 17.3. *If $\{H_n : n \in \mathbb{N}\}$ is a sequence of progressively measurable step functions such that*

$$\mathbb{E}\int_0^\infty \left[H_n(s) - H_m(s)\right]^2 ds \to 0,$$

as $m, n \to \infty$*, then*

$$\mathbb{E}\left[\left(\int_0^\infty H_n(s)dB_s - \int_0^\infty H_m(s)dB_s\right)^2\right] \to 0,$$

as $m, n \to \infty$*.*

Proof. $H_n - H_m$ is also a progressively measurable step function for all m, n. Apply (17.2). □

17.1.1 The Ito integral construction

We are now ready to formally define the Ito integral.

Given any progressively measurable process H satisfying

$$\mathbb{E}\int_0^\infty H^2(s)ds < \infty, \tag{38}$$

let $\{H_n : n \in \mathbb{N}\}$ be a sequence of progressively measurable step processes with

$$\mathbb{E}\int_0^\infty \left(H_n(s) - H(s)\right)^2 ds \to 0. \tag{39}$$

(The existence of such a sequence follows from Lemma 17.1.) Define the Ito integral of H as

$$\int_0^\infty H(s)dB_s = \lim_{n\to\infty}\int_0^\infty H_n(s)dB_s,$$

where the limit is in the L^2 sense. More specifically, we will show that the sequence of random variables $\int_0^\infty H_n(s)dB_s$ converges in L_2 and we define $\int_0^\infty H(s)dB_s$ to be the limit of this sequence.

There are two questions to answer to make this definition proper:

(i) Why does $\int_0^\infty H_n(s)dB_s$ converge in L^2?

(ii) Why is the limit independent of the approximating sequence?

Very succinctly, the answers to these two questions follow from Corollary (17.3) and the fact that L^2 spaces are complete.

Proposition 17.4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Define*

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) = \{[X] : \mathbb{E}X^2 < \infty\},$$

where $[X]$ is the equivalence class of the random variable $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ defined by the relation $X \sim Y$ if $E\,(X - Y)^2 = 0$. Then $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is a complete metric space.

The proof is omitted.

The following Theorem is nothing new; it just collects together results from earlier.

Theorem 17.5.

1. *A sequence of random variables*

$$\left\{\int_0^\infty H_n(s)dB_s : n \in \mathbb{N}\right\},$$

satisfying (39) converges in L^2.

2. *$\int_0^\infty H(s)dB_s$ is well defined (i.e. independent of the approximating sequence) when H is progressively measurable and has finite L^2 norm (38).*

3.
$$\mathbb{E}\left[\left(\int_0^\infty H(s)dB_s\right)^2\right] = \mathbb{E}\int_0^\infty H^2(s)ds. \tag{40}$$

(40) is called the Ito isometry and it shows that the L^2 norm is preserved by the stochastic integral.

Proof. 1. We know that

$$\mathbb{E}\int_0^\infty (H(s) - H_n(s))^2\, ds \to 0, \tag{41}$$

by construction. By the triangle inequality,

$$\mathbb{E}\int_0^\infty (H_m(s) - H_n(s))^2\, ds \to 0,$$

as $m, n \to \infty$. Corollary 17.3 then implies

$$\mathbb{E}\left[\left(\int_0^\infty H_m(s)dB_s - \int_0^\infty H_n(s)dB_s\right)^2\right] \to 0,$$

as $m, n \to \infty$. So the sequence of approximating integrals in Cauchy in L^2. By Proposition 17.4, the sequence must converge.

2. Let $\{H_n : n \in \mathbb{N}\}$ and $\{H'_n : n \in \mathbb{N}\}$ be two approximating sequences. Then

$$\mathbb{E}\int_0^\infty \left[H_n(s) - H'_n(s)\right]^2 \to 0,$$

by the triangle inequality, using property (41) of H_n and H'_n. Again Corollary 17.3 implies that the L^2 distance between the two sequences of stochastic integrals goes to zero:

$$\mathbb{E}\left[\left(\int_0^\infty H_n(s)dB_s - \int_0^\infty H'_n(s)dB_s\right)^2\right] \to 0,$$

and so the L^2 limits of the sequences must agree. (This finally establishes the results required for properly defining the Ito integral.)

3. We have the Ito isometry property for step processes (Lemma 17.2). This property is preserved when taking L^2 limits: By (41),

$$\mathbb{E}\left[\int_0^\infty H_n^2(s)ds\right] \to \mathbb{E}\int_0^\infty H^2(s)ds,$$

while

$$\mathbb{E}\left[\left(\int_0^\infty H_n(s)dB_s\right)^2\right] \to \mathbb{E}\left[\left(\int_0^\infty H(s)dB_s\right)^2\right],$$

by L^2 convergence of the Ito integral. Yet Lemma (17.2) says that

$$\mathbb{E}\left[\int_0^\infty H_n^2(s)ds\right] = \mathbb{E}\left[\left(\int_0^\infty H_n(s)dB_s\right)^2\right],$$

so their limits must be equal too. □

17.1.2 Future directions

We want to define the definite integral $\int_0^t H(s)dB_s$ – recall from section 16.2.1 that the initial motivating example for the stochastic integral was in terms of a definite integral. We would like to study the definite integral as a function of t, or equivalently as a stochastic process in t.

We also want to be able to evaluate stochastic integrals. We have proven the existence of the integrals, but this isn't much use on its own. We want a calculus for explicit evaluation of stochastic integrals. Ito's lemma provides this.

18 Lecture 1/4

18.1 The definite Ito integral

Definition 18.1. Fix $t \in \mathbb{R}^{\geq 0}$. Given a progressively measurable process $H(s,\omega)$ with finite L^2 norm on $[0,t]$,

$$\mathbb{E}\int_0^t H^2(s,\omega)ds < \infty,$$

define

$$H^{(t)}(s,\omega) = H(s,\omega)\mathbb{1}_{s\leq t}.$$

The *definite integral* of $H(s,\omega)$ from 0 to t, is defined as

$$\int_0^t H(s,\omega)dB_s = \int_0^\infty H^{(t)}(s,\omega)dB_s.$$

(To make this definition proper, we need to check that $H^{(t)}$) is progressively measurable.)

Now,

$$\left\{\int_0^t H(s)dB_s : t \in \mathbb{R}^{\geq 0}\right\}, \tag{42}$$

is a stochastic process. We want this process to have continuous trajectories (for many applications – e.g. the location of a particle in a liquid medium – we would expect continuity) – yet nothing we have done so far would require (42) to have continuous trajectories. Instead we need to construct a continuous modification of (42).

Theorem 18.2. *If $\{H(s,\omega) : s \in \mathbb{R}^{\geq 0}, \omega \in \Omega\}$ is progressively measurable and*

$$\mathbb{E}\int_0^t H^2(s,\omega)ds < \infty,$$

for all $t \geq 0$, then there exists an almost-sure-continuous modification of

$$\left\{\int_0^t H(s)dB_s : t \in \mathbb{R}^{\geq 0}\right\}.$$

Moreover, this process is a martingale and

$$\mathbb{E}\int_0^t H(s)dB_s = 0, \tag{43}$$

for all $t \geq 0$.

To be clear, the first conclusion of this Theorem is that there exists $\{M_t : t \in \mathbb{R}^{\geq 0}\}$ such that

$$\mathbb{P}\left[M_t \neq \int_0^t H(s)dB_s\right] = 0,$$

for all $t \geq 0$ and $t \mapsto M_t(\omega)$ is continuous almost surely.

Proof. As usual, we use the one proof technique that we know: prove the result for step functions and then approximate. Suppose that

$$H(s) = \sum_{i=1}^{k} A_i \mathbb{1}\{s \in (a_i, a_{i+1}]\},$$

where A_i is $\mathcal{F}_{a_i}$-measurable. Then

$$\begin{aligned}\int_0^t H(s)dB_s &= \int_0^\infty H(s)\mathbb{1}_{s\leq t}dB_s \\ &= \int_0^\infty \sum_{i=1}^{k} A_i \mathbb{1}\{s \in (a_i \wedge t, a_{i+1} \wedge t]\}\, dB_s \\ &= \sum_{i=1}^{k} A_i \left(B_{a_{i+1}\wedge t} - B_{a_i \wedge t}\right).\end{aligned}$$

(i) $t \mapsto \sum_{i=1}^{k} A_i \left(B_{a_{i+1}\wedge t} - B_{a_i \wedge t}\right)$ is continuous a.s. (since sBM has continuous trajectories).

(ii) $\int_0^t H(s)dB_s = \sum_{i=1}^k A_i \left(B_{a_{i+1}\wedge t} - B_{a_i \wedge t}\right)$ is $\mathcal{F}_t$-measurable.

(iii)

$$\mathbb{E}\left[A_i \left(B_{a_{i+1}\wedge t_2} - B_{a_i \wedge t_2}\right) \middle| \mathcal{F}_{t_1}\right] = A_i \left(B_{a_{i+1}\wedge t_1} - B_{a_i \wedge t_1}\right).$$

(ii) and (iii) prove that $\{\int_0^t H(s)dB_s\}$ is a martingale. Substituting $t = 0$ into (ii) gives (43). Thus the Theorem holds for progressively measurable step processes.

Our goal now is to extend this result to general progressively measurable processes with $\mathbb{E}\int_0^\infty H^2(s)ds < \infty$.

Fix $T > 0$. We will only prove the Theorem for the process

$$\left\{\int_0^t H(s)dB_s : 0 \le t \le T\right\}. \tag{44}$$

We know that there exists a sequence $\{H_n : n \in \mathbb{N}\}$ of progressively measurable step processes such that

$$\mathbb{E}\int_0^T (H_n(s) - H(s))^2 \to 0.$$

Define

$$M_n(t) = \int_0^t H_n(s)dB_s,$$

for $0 \le t \le T$. For each n, $M_n(t)$ is continuous, adapted and a martingale.

The proof idea is first to show that $M_n(t)$ converges in $C([0,T])$ to some limit $L(t)$ then the continuous modification of (44) is this limit $L(t)$.

We know that (i) for all $0 \le t \le T$,

$$M_n(t) \xrightarrow[n\to\infty]{L^2} \int_0^t H(s)dB_s,$$

and (ii) for all $n \in \mathbb{N}$, $t \mapsto M_n(t)$ is continuous – or equivalently, M_n is a $C([0,T])$-valued random variable.

To show that $M_n(t)$ converges in $C([0,T])$, it suffices to show that it is Cauchy. We prove this via Doob's maximal and L^p inequalities (Corollary 3.5, relying on the fact that $M_n(t) - M_{n'}(t)$ is a martingale in t):

$$\mathbb{E}\left[\left(\sup_{0\le t\le T} |M_n(t) - M_{n'}(t)|\right)^2\right] \le 4\mathbb{E}\left[(M_n(T) - M_{n'}(T))^2\right]$$

$$= 4\mathbb{E}\int_0^T \left(H_n(s) - H_{n'}(s)\right)^2 ds$$
$$\to 0, \tag{45}$$

as $n, n' \to \infty$; where the second line uses the Ito isometry. Then by Markov's inequality,

$$\mathbb{P}\left[\sup_{0\le t\le T} |M_n(t) - M_{n'}(t)| > \alpha^{-k}\right] \le 4\alpha^{2k}\mathbb{E}\int_0^T \left(H_n(s) - H_{n'}(s)\right)^2 ds$$
$$\le \left(\frac{\alpha^2}{4}\right)^k,$$

where the second line follows for $n, n' > n_k$ where the sequence $\{n_k : k \ge 1\}$ is defined such that

$$4\mathbb{E}\int_0^T \left(H_n(s) - H_{n'}(s)\right)^2 ds \le 2^{-2k-1},$$

for all $n, n' > n_k$. (Such a subsequence exists by (45).)

Fix m. Choose $\alpha \in (1, 2)$ and use Borel-Cantelli to get

$$\mathbb{P}\left[\limsup_{k\to\infty}\left\{\sup_{0\le t\le T}\left|M_{n_{k+1}}(t) - M_{n_{k+m}}(t)\right| > \alpha^{-k}\right\}\right] = 0.$$

This proves that $\{M_{n_k}(t) : 0 \le t \le T\}$ is Cauchy in $C([0,T])$ almost surely. The fact that we have a Cauchy subsequence (rather than proving $\{M_n\}$ is Cauchy) is okay: Define $M(t) = \lim_{k\to\infty} M_{n_k}(t) \in C([0,T])$. $M(t)$ is our candidate continuous modification.

Observe that for all $t \in [0,T]$,

$$M(t) = \lim_{k\to\infty}\int_0^t H_{n_k}(s)dB_s,$$

almost surely. However, we also know

$$\int_0^t H_{n_k}(s)dB_s \xrightarrow{L^2} \int_0^t H(s)dB_s.$$

The only way that both these convergences can happen is if the limits are equal almost surely:

$$\mathbb{P}\left[M(t) = \int_0^t H(s)dB_s\right] = 1,$$

for all $t \in [0, T]$. This proves that $M(t)$ is a modification.

All that remains to prove is that $M(t)$ is a martingale:

$$\begin{aligned}
\mathbb{E}\left[M(t_2)|\mathcal{F}_{t_1}\right] &= \mathbb{E}\left[\lim_{k\to\infty}\int_0^{t_2} H_{n_k}(s)dB_s \middle| \mathcal{F}_{t_1}\right] \\
&= \lim_{k\to\infty}\mathbb{E}\left[\int_0^{t_2} H_{n_k}(s)dB_s \middle| \mathcal{F}_{t_1}\right] \\
&= \lim_{k\to\infty}\int_0^{t_1} H_{n_k}(s)dB_s \\
&= \int_0^{t_1} H(s)dB_s \\
&= M(t_1),
\end{aligned}$$

where the first line follows by a.s. convergence; the second since $\int_0^t H_{n_k}(s)dB_s$ converges in L_2, so it also converges in L_1 (and in conditional expectation); the third since $\int_0^t H_{n_k}(s)dB_s$ is a martingale; and the second last by L^2 convergence. □

18.2 The Ito lemma

The Ito lemma gives us a tool for evaluating a stochastic integral.

Theorem 18.3 (Ito lemma 1)**.** *Suppose* $f : \mathbb{R} \to \mathbb{R} \in C^2$[‖] *and*

$$\mathbb{E}\int_0^t \left[f'(B_s)\right]^2 ds < \infty, \tag{46}$$

for some $t > 0$*. Then a.s. for all* $0 \le s \le t$*,*

$$f(B_s) = f(B_0) + \int_0^s f'(B_u)dB_u + \frac{1}{2}\int_0^s f''(B_u)du. \tag{47}$$

Compare (47) with the Riemann integral version (aka the fundamental theorem of calculus): $f(s) = f(0) + \int_0^s f'(u)du$. The third term on the RHS of (47) is sometimes called Ito's correction term for this reason.

[‖] C^2 denotes the set of twice continuously-differentiable functions.

What is meant by the equality in (47)? For each fixed s, the LHS and the RHS are both random variables. (47) means that these random variables are equal with probability 1 – i.e., for all $s \in [0,t]$,

$$\mathbb{P}\left(f(B_s) = f(B_0) + \int_0^s f'(B_u)dB_u + \frac{1}{2}\int_0^s f''(B_u)du\right) = 1. \tag{48}$$

But (47) is expressing a stronger notion than this – it is stating that, for a compact interval $[0,t] \subset \mathbb{R}^{\geq 0}$ with t satisfying (46),

$$\mathbb{P}\left(\forall s \in [0,t], f(B_s) = f(B_0) + \int_0^s f'(B_u)dB_u + \frac{1}{2}\int_0^s f''(B_u)du\right) = 1. \tag{49}$$

We call this 'equal as stochastic processes'. However, since the LHS and RHS are both continuous processes, equality on the rationals – that is, (48) for all $s \in [0,t] \cap \mathbb{Q}$ – implies (49). (Why? Use the facts that the rationals are dense – so we can approximate $s \in [0,t]$ by rationals – and countable – so we can move the $\forall$ quantifier within the probability.)

Proof. By the above discussion, we need only prove (48) holds for all $s \in [0,t]$. We will show the result for $f \in C^\infty$ with bounded derivatives of all orders. Let $0 = t_1 \leq \ldots \leq t_k = s$. By Taylor expansion of f around $f(B_{t_i})$,

$$\begin{aligned} f(B_s) - f(B_0) &= \sum_{i=0}^{k-1} f(B_{t_{i+1}}) - f(B_{t_i}) \\ &= \sum_{i=0}^{k-1} \left[f'(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right) + \frac{1}{2}f''(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right)^2 \right. \\ &\qquad \left. + \frac{1}{6}f'''(\theta_i)\left(B_{t_{i+1}} - B_{t_i}\right)^3 \right], \end{aligned}$$

where θ_i is between B_{t_i} and $B_{t_{i+1}}$. Then we know

$$\begin{aligned} &\left| f(B_s) - f(B_0) - \sum_{i=0}^{k-1} f'(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right) - \frac{1}{2}\sum_{i=0}^{k-1} f''(B_{t_i})\left(B_{t_{i+1}} - B_{t_i}\right)^2 \right| \\ &\qquad\qquad \leq \frac{1}{6}\sum_{i=0}^{k-1} |f'''(\theta_i)| \left|B_{t_{i+1}} - B_{t_i}\right|^3. \end{aligned}$$

Observe that

$$T_1 := \sum_{i=0}^{k-1} f'(B_{t_i}) \left(B_{t_{i+1}} - B_{t_i}\right) = \int_0^t \sum_{i=0}^{k-1} f'(B_{t_i}) \mathbb{1}\{u \in (t_i, t_{i+1}]\} \, dB_u.$$

Check that

$$\mathbb{E} \int_0^t \left(\sum_{i=0}^{k-1} f'(B_{t_i}) \mathbb{1}\{u \in (t_i, t_{i+1}]\} - f'(B_u) \right)^2 du \to 0.$$

as $k \to \infty$, since f is bounded and smooth. By Ito isometry, this implies

$$T_1 \xrightarrow[k\to\infty]{L^2} \int_0^t f'(B_s) dB_s.$$

This gives L^2 convergence but we want a.s. convergence. To get a.s. convergence, we use the following fact: convergence in probability implies that there exists a subsequence that converges a.s.

So we have a.s. convergence along a subsequence of the partitions of $[0, s]$. From herein, we will work with this subsequence. (Later in the proof, we will take further subsequences; each time we do this, we will discard the original sequence, and just work with the subsequence.)

We claim that

$$T_2 = \sum_{i=0}^{k-1} f''(B_{t_i}) \left(B_{t_{i+1}} - B_{t_i}\right)^2 \xrightarrow[k\to\infty]{L^2} \int_0^t f''(B_s) ds, \tag{50}$$

and hence converges a.s. along a subsequence. Further, we claim

$$T_3 = \sum_{i=0}^{k-1} |f'''(\theta_i)| \left|B_{t_{i+1}} - B_{t_i}\right|^3 \xrightarrow{\text{a.s.}} 0, \tag{51}$$

along a subsequence. We will show (50) and (51) in the next lecture and this will (basically) complete the proof of the Ito lemma. □

19 Lecture 6/4

19.1 Proof of the Ito lemma (cont.)

Recall that to complete the proof of the Ito lemma, we need to show (50) and (51).

Claim 19.1.

$$T_2 - \sum_{i=0}^{k-1} f''(B_{t_i})(t_{i+1} - t_i) \xrightarrow{L^2} 0.$$

This claim establishes (50), since

$$\sum_{i=0}^{k-1} f''(B_{t_i})(t_{i+1} - t_i) \xrightarrow{\text{a.s.}} \int_0^t f''(B_s)ds,$$

by definition of Riemann integration.

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=0}^{k-1} f''(B_{t_i})(B_{t_{i+1}} - B_{t_i})^2 - \sum_{i=0}^{k-1} f''(B_{t_i})(t_{i+1} - t_i)\right)^2\right] \\
&\qquad = \mathbb{E}\left[\left(\sum_{i=0}^{k-1} f''(B_{t_i})\left[(B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i)\right]\right)^2\right] \qquad (52) \\
&\qquad = \mathbb{E}\left[\sum_{i=0}^{k-1} \left(f''(B_{t_i})\right)^2 \left[(B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1} - t_i)\right]^2\right] \\
&\qquad \leq 2\|f''\|_\infty^2 \mathbb{E}\left[\sum_{i=0}^{k-1} (B_{t_{i+1}} - B_{t_i})^4 + (t_{i+1} - t_i)^2\right] \\
&\qquad = 2\|f''\|_\infty^2 \sum_{i=0}^{k-1} 3(t_{i+1} - t_i)^2 + (t_{i+1} - t_i)^2 \\
&\qquad = C \sum_{i=0}^{k-1} (t_{i+1} - t_i)^2 \\
&\qquad \leq C \max_{0 \leq i \leq k-1} (t_{i+1} - t_i) \sum_{i=0}^{k-1} (t_{i+1} - t_i) \\
&\qquad \leq C \max_{0 \leq i \leq k-1} (t_{i+1} - t_i) t \\
&\qquad \xrightarrow{k \to \infty} 0,
\end{aligned}$$

where the third line follows by verifying that the cross terms (when expanding the square) vanish by the weak Markov property; $\|\cdot\|_\infty$ in the fourth line is the sup-norm

and the factor 2 arises by the inequality $(a-b)^2 \leq 2(a^2+b^2)$; and $C = 8\|f''\|_\infty^2$ in the sixth line. □

Proof of (51).

$$\begin{aligned} T_3 &\leq \|f'''\|_\infty \sum_{i=0}^{k-1} \left|B_{t_{i+1}} - B_{t_i}\right|^3 \\ &\leq \|f'''\|_\infty \max_{0\leq i\leq k-1} \left|B_{t_{i+1}} - B_{t_i}\right| \sum_{i=0}^{k-1} \left(B_{t_{i+1}} - B_{t_i}\right)^2 \end{aligned}$$

Two observations will complete the proof:

1. $\max_{0\leq i\leq k-1}\left|B_{t_{i+1}} - B_{t_i}\right| \xrightarrow{\text{a.s.}} 0$, by uniform continuity (on $[0,s]$) of Brownian motion; and
2. $\sum_{i=0}^{k-1}\left(B_{t_{i+1}} - B_{t_i}\right)^2 \xrightarrow{L^2} t$. (This can be checked by similar arguments to the proof of Claim 19.1: write $\mathbb{E}\left[\left(\sum_{i=0}^{k-1}(B_{t_{i+1}} - B_{t_i})^2 - (t_{i+1}-t_i)\right)^2\right]$ and then follow (52), since the cross-terms also vanish in this case.) This implies that $\sum_{i=0}^{k-1}\left(B_{t_{i+1}} - B_{t_i}\right)^2 \xrightarrow{\text{a.s.}} t$ along a subsequence.

□

19.2 Ito lemma version 2

Theorem 18.3 is the simplest version of the Ito lemma. A more complex version is presented below. The proof for this version is basically the same as for Theorem 18.3 but requires more cumberson notation.

Theorem 19.2 (Ito lemma 2)**.** *For* $\varphi : [0,\infty)\times\mathbb{R}\to\mathbb{R} \in C_{1,2}$[**]*,*

$$\varphi(t,B_t) = \varphi(0,B_0) + \int_0^t \varphi_1(s,B_s)ds + \int_0^t \varphi_2(s,B_s)dB_s + \frac{1}{2}\int_0^t \varphi_{22}(s,B_s)ds,$$

where

[**] $C_{1,2} = \left\{g(t,x) : \frac{\partial g}{\partial t}, \frac{\partial^2 g}{\partial x^2} \text{ bounded}\right\}$.

1. $\varphi_1(s,x) = \frac{\partial \varphi}{\partial s}(s,x)$;

2. $\varphi_2(s,x) = \frac{\partial \varphi}{\partial x}(s,x)$; *and*

3. $\varphi_{22}(s,x) = \frac{\partial^2 \varphi}{\partial x^2}(s,x)$.

Plug in a function $\varphi(t,x)$ which is constant in t and we get back the first version of the Ito lemma (Theorem 18.3).

Given a function $\varphi(t,x)$, we will also use notation $\partial_t \varphi$ for $\frac{\partial \varphi}{\partial t}$; $\partial_x \varphi$ for $\frac{\partial \varphi}{\partial x}$ and ∂^2_{xx} for $\frac{\partial^2 \varphi}{\partial x^2}$.

19.2.1 Applications

1. If $\varphi(t,x) = tx$, then $\partial_t \varphi = x$, $\partial_x \varphi = t$ and $\partial^2_{xx}\varphi = 0$. The Ito lemma gives

$$tB_t = \int_0^t B_s ds + \int_0^t s dB_s.$$

   This is "integration by parts".

2. If $\varphi(t,x) = x^2 - t$, then $\partial_t \varphi = -1$, $\partial_x \varphi = 2x$ and $\partial^2_{xx}\varphi = 2$. We get

$$B_t^2 - t = -\int_0^t ds + 2\int_0^t B_s dB_s + \int_0^t ds,$$

   and thus

$$\int_0^t B_s dB_s = \frac{B_t^2 - t}{2}.$$

   Compare this with the Riemann integral $\int x dx = \frac{1}{2}x^2$. This shows that $B_t^2 - t$ must be a martingale, since stochastic integrals are martingales. This is an example of how the Ito lemma can be used to evaluate integrals.

19.2.2 Solving stochastic differential equations

The Ito lemma can also be used to find solutions to stochastic differential equations (SDEs).

Recall our original motivation for stochastic integrals: we wanted to construct a differential equation describing a particle's movement in a liquid medium. Now we can finally resolve this problem.

Does there exist a stochastic process $\{Y_t : t \in \mathbb{R}^{\geq 0}\}$ such that

$$dY_t = \mu Y_t dt + \sigma Y_t dB_t,$$

(in differential form)? This is shorthand for the equation

$$Y_t = Y_0 + \int_0^t \mu Y_s ds + \int_0^t \sigma Y_s dB_s,$$

(integral form).

As is the standard method for ODEs, we solve this SDE by first guessing the general form of the solution. Assuming $Y_t = f(t, B_t)$[††], Ito's lemma states that

$$dY_t = \partial_t f(t, B_t)dt + \partial_x f(t, B_t)dB_t + \frac{1}{2}\partial^2_{xx} f(t, B_t)dt.$$

Now match coefficients of dB_t and dt:

1. $\partial_x f(t, x) = \sigma f(t, x)$, so $f(t, x) = A(t)e^{\sigma x}$;
2. $\partial_t f(t, x) + \frac{1}{2}\partial^2_{xx} f(t, x) = \mu f(t, x)$. Combining with 1., this implies $A'(t) + \frac{1}{2}\sigma^2 A(t) = \mu A(t)$, so that $A(t) = Be^{(\mu - \frac{1}{2}\sigma^2)t}$, where B is some constant.

By the initial condition $B = Y_0$ so that

$$Y_t = f(t, B_t) = Y_0 \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma B_t\right].$$

Y_t is called *geometric Brownian motion*.

There are two natural questions that arise from this discussion? Is such a process Y_t unique? That is, when are SDE solutions unique? Also – instead of having to guess

[††]add-on This might appear to be the most general functional form for Y_t but it actually is not! Why? Y_t is a function of B_t but not B_s for $s \leq t$. To solve other SDEs, we may need to use functional forms which include terms like $\int_0^t b(s)dB_s$, where $b(s)$ is some deterministic form. (We know that in this case $\int_0^t b(s)dB_s$ is a Gaussian process – see section 11). $\int_0^t b(s)dB_s$ depends on B_s for all $s \leq t$. For details, see section 12.

the solution case-by-base – are there general conditions that guarantee the existence of SDE solutions? We will prove uniqueness and provide sufficient conditions for existence in the next lecture.

20 Lecture 8/4

20.1 Stochastic differential equations (SDEs)

20.1.1 Existence and uniqueness of solutions

Does a given SDE have at least one solution? And if so, is the solution unique?

Theorem 20.1. *Suppose* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a probability space,* $\{\mathcal{F}_t : t \in \mathbb{R}^{\geq 0}\}$ *is a complete filtration, and* $\{B_t : t \in \mathbb{R}^{\geq 0}\}$ *is sBM adapted to* $\mathcal{F}_t$*. Let* $a : \mathbb{R} \to \mathbb{R}$ *and* $\sigma : \mathbb{R} \to \mathbb{R}^{>0}$ *be Lipschitz and* ξ *be a bounded*[‡‡] *random variable adapted to* $\mathcal{F}_0$*.*

Then the SDE

$$dX_t = a(X_t)dt + \sigma(X_t)dB_t, \text{ and } X_0 = \xi, \tag{53}$$

has a unique solution.

This theorem does not capture the most general scenario. That is, the assumptions can be weakened. The purpose here is to give the flavour of this topic, while keeping the exposition simple.

The coefficient $a(X_t)$ of dt is called the *drift*. The coefficient $\sigma(X_t)$ for dB_t is called the *diffusion*. So basically, as long as the drift and diffusion are bounded in a finite interval of time, the SDE has a unique solution.

Recall that (53) means that there exists $\{X_t : t \in \mathbb{R}^{\geq 0}\}$ such that

$$X_t = \int_0^t a(X_s)ds + \int_0^t \sigma(X_s)dB_s + \xi.$$

Proof. As usual, the proof for uniqueness is easier than the proof for existence. We prove existence first.

[‡‡]That is, there exists some $M > 0$ such that $\mathbb{P}[|\xi| \leq M] = 1$.

Fix $T > 0$ and exhibit a solution on $[0, T]$. (Since T is arbitrary, this is sufficient.) The proof idea is to construct a sequence $\{X_t^k : t \in [0,T]\}_{k=1}^{\infty}$ of $C([0,T])$-valued random variables that is Cauchy and approximately solves the SDE. Then we will show that $X_t = \lim_{k\to\infty} X_t^k$ – where the limit is in terms of the function space $C([0,T])$ – is a solution to the SDE.

We use Picard iteration to construct the approximation sequence $\{X_t^k : t \in [0,T]\}_{k=1}^{\infty}$. Start by defining $X_t^0 = \xi$ (X_t^0 is constant in time) and then iteratively define

$$X_t^k = \xi + \int_0^t a(X_s^{k-1})ds + \int_0^t \sigma(X_s^{k-1})dB_s.$$

(Why is this well defined? We need to check that we can integrate $a(X_s^{k-1})$ and $\sigma(X_s^{k-1})$. For the first integral, $a(X_s^{k-1})$ has continuous trajectories since a is Lipschitz and X_s^{k-1} is the sum of integrals and hence is also continuous. For the second integral, one can check that $\mathbb{E}\int_0^T \left[\sigma(X_s^{k-1})\right]^2 ds < \infty$ by induction and σ-Lipschitz.)

I claim that $\{X_t^k : t \in [0,T]\}$ is Cauchy (as $C([0,T])$)-valued random variables) with probability 1:

$$X_t^{k+1} - X_t^k = \int_0^t a(X_s^k) - a(X_s^{k-1})ds + \int_0^t \sigma(X_s^k) - \sigma(X_s^{k-1})dB_s = Y_t^k + Z_t^k.$$

Roughly, we want to show that Y_t^k and Z_t^k go to zero in probability as $k \to \infty$. For the first term,

$$\begin{aligned}\left|Y_t^k\right| &\le \int_0^t \left|a(X_s^k) - a(X_s^{k-1})\right|ds \\ &\le \int_0^T \left|a(X_s^k) - a(X_s^{k-1})\right|ds.\end{aligned}$$

The second term Z_k^t is a martingale in t for all k. Thus,

$$\begin{aligned}\mathbb{P}\left[\max_{0\le t\le T}\left|X_t^{k+1} - X_t^k\right| > \epsilon\right] &\le \mathbb{P}\left[\int_0^T \left|a(X_t^k) - a(X_t^{k-1})\right|dt > \epsilon/2\right] + \mathbb{P}\left[\max_{0\le t\le T}\left|Z_t^k\right| > \epsilon/2\right] \\ &\le \frac{4}{\epsilon^2}\mathbb{E}\left[\left(\int_0^T \left|a(X_t^k) - a(X_t^{k-1})\right|dt\right)^2\right] + \frac{4}{\epsilon^2}\mathbb{E}\left[{Z_T^k}^2\right] \\ &\le \frac{4TA^2}{\epsilon^2}\mathbb{E}\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt + \frac{4A^2}{\epsilon^2}\mathbb{E}\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt\end{aligned}$$

$$= \frac{4(T+1)A^2}{\epsilon^2}\mathbb{E}\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt, \tag{54}$$

where the second line follows by Markov's and Doob's maximal inequalities; A is the maximum Lipschitz constant of a and σ – that is, $|a(x) - a(y)| \leq A|x-y|$ and $|\sigma(x) - \sigma(y)| \leq A|x-y|$ – and the third line follows by the following reasoning: The second term can be bounded

$$\begin{aligned}
\mathbb{E}\left[Z_T^{k\,2}\right] &= \mathbb{E}\left[\int_0^T \sigma(X_t^k) - \sigma(X_t^{k-1})dB_t\right]^2 \\
&= \mathbb{E}\int_0^T \left(\sigma(X_t^k) - \sigma(X_t^{k-1})\right)^2 dt \\
&\leq A^2\mathbb{E}\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt,
\end{aligned} \tag{55}$$

where the second line follows by the Ito isometry and the third by the Lipschitz property. The first term can similarly be bounded:

$$\begin{aligned}
\mathbb{E}\left[\left(\int_0^T \left|a(X_t^k) - a(X_t^{k-1})\right|dt\right)^2\right] &\leq A^2\mathbb{E}\left[\left(\int_0^T \left|X_t^k - X_t^{k-1}\right|dt\right)^2\right] \\
&\leq A^2T\mathbb{E}\left[\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt\right],
\end{aligned} \tag{56}$$

where the first line follows by the Lipschitz property and the second line follows by Cauchy-Schwarz (supposedly). (I don't understand how to derive the second line, although the result, without the factor T, follows easily by Jensen's inequality. The rest of the proof also flows through without the factor T.)

We control the RHS of (54) using Gronwall's lemma: We have, by definition,

$$X_t^{k+1} - X_t^k = \int_0^t a(X_s^k) - a(X_s^{k-1})ds + \int_0^t \sigma(X_s^k) - \sigma(X_s^{k-1})dB_s.$$

Thus,

$$\begin{aligned}
\mathbb{E}\left(X_t^{k+1} - X_t^k\right) &\leq 2\mathbb{E}\left[\left(\int_0^t a(X_s^k) - a(X_s^{k-1})ds\right)^2 + \left(\int_0^t \sigma(X_s^k) - \sigma(X_s^{k-1})dB_s\right)^2\right] \\
&\leq 2A^2t\mathbb{E}\int_0^t \left(X_s^k - X_s^{k-1}\right)^2 ds + 2A^2\mathbb{E}\int_0^t \left(X_s^k - X_s^{k-1}\right)^2 ds
\end{aligned}$$

$$
\begin{aligned}
&\leq 2A^2(1+T)\mathbb{E}\int_0^t \left(X_s^k - X_s^{k-1}\right)^2 ds \\
&= 2A^2(1+T)\int_0^t \mathbb{E}\left(X_{t_1}^k - X_{t_1}^{k-1}\right)^2 dt_1 \\
&\leq \left(2A^2(1+T)\right)^2 \int_0^t \int_0^{t_1} \mathbb{E}\left(X_{t_2}^{k-1} - X_{t_2}^{k-1}\right)^2 dt_2 dt_1 \\
&\leq \ldots \leq \left(2A^2(1+T)\right)^k \int_0^t \int_0^{t_1} \cdots \int_0^{t_{k-1}} \mathbb{E}\left(X_{t_k}^1 - X_{t_k}^0\right)^2 dt_k \ldots dt_1,
\end{aligned} \tag{57}
$$

where the second line follows by the same logic as in (55) and (56); and the fourth line follows by Fubini's theorem (which we will justify later, by induction). By definition,

$$
\begin{aligned}
\mathbb{E}\left[\left(X_s^1 - X_s^0\right)^2\right] &= \mathbb{E}\left[\left(\int_0^s a(\xi)dt + \int_0^s \sigma(\xi)dB_t\right)^2\right] \\
&= \mathbb{E}\left[\left(sa(\xi) + B_s\sigma(\xi)\right)^2\right] \\
&\leq 2s^2\mathbb{E}\left[a^2(\xi)\right] + 2\mathbb{E}\left[B_s^2\right]\mathbb{E}\left[\sigma^2(\xi)\right] \\
&\leq 2s^2\mathbb{E}\left[a^2(\xi)\right] + 2s\mathbb{E}\left[\sigma^2(\xi)\right],
\end{aligned}
$$

where the second line follows by Ito's lemma; and the third line by independence of B_s and ξ, since ξ is $\mathcal{F}_0$ adapted. Plugging this into (57), we obtain

$$
\mathbb{E}\left(X_t^{k+1} - X_t^k\right) \leq \left[2A^2(1+T)\right]^k \left[2\mathbb{E}\left[a^2(\xi)\right]\frac{t^{k+2}}{(k+2)!} + 2\mathbb{E}\left[\sigma^2(\xi)\right]\frac{t^{k+1}}{(k+1)!}\right]. \tag{58}
$$

Finally, plugging (58) into (54),

$$
\begin{aligned}
\mathbb{P}\left[\max_{0\leq t\leq T}\left|X_t^{k+1} - X_t^k\right| > \epsilon\right] &\leq \frac{4(T+1)A^2}{\epsilon^2}\mathbb{E}\int_0^T \left(X_t^k - X_t^{k-1}\right)^2 dt \\
&\leq \frac{2}{\epsilon^2}\left[2A^2(1+T)\right]^k \left[2\mathbb{E}\left[a^2(\xi)\right]\frac{T^{k+2}}{(k+2)!} + 2\mathbb{E}\left[\sigma^2(\xi)\right]\frac{T^{k+1}}{(k+1)!}\right] \\
&\leq \frac{4}{\epsilon^2}\frac{\left[2A^2T(1+T)\right]^k}{k!}C_T,
\end{aligned}
$$

using Fubini's theorem again, where C_T is some positive, finite constant ($\mathbb{E}\left[a^2(\xi)\right]$ and $\mathbb{E}\left[\sigma^2(\xi)\right]$ are bounded since ξ is bounded a.s.).

Now define

$$\epsilon_k = \left[\frac{\left(2A^2T(1+T)\right)^k}{k!}\right]^{1/3},$$

so that $\epsilon_k \to 0$ yet,

$$\sum_{k=1}^{\infty} \mathbb{P}\left[\max_{0\le t\le T}\left|X_t^{k+1} - X_t^k\right| > \epsilon_k\right] \le 4C_T \sum_{k=1}^{\infty}\left[\frac{\left(2A^2T(1+T)\right)^k}{k!}\right]^{1/3} < \infty,$$

by noting that $k!$ grows much faster than e^k. Then Borel-Cantelli says

$$\mathbb{P}\left[\limsup_{k\to\infty}\left\{\max_{0\le t\le T}\left|X_t^{k+1} - X_t^k\right| > \epsilon_k\right\}\right] = 0.$$

Hence $\{X_t^k : t \in [0,T]\}_{k=1}^{\infty}$ is Cauchy with probability 1. Set X_t to be the limit in $C([0,T])$ and verify that it satisfies the SDE. □

The above is a typical proof: our only reliable friend is completeness, which we use to give us a limit of our approximations.

Proof of uniqueness (the second part of Theorem 53). Suppose there are two solutions,

$$X_t = \xi + \int_0^t a(X_s)ds + \int_0^t \sigma(X_s)dB_s,$$
$$Y_t = \xi + \int_0^t a(Y_s)ds + \int_0^t \sigma(Y_s)dB_s.$$

Using analogous arguments to the previous proof,

$$\phi(t) = \mathbb{E}\left(X_t - Y_t\right)^2 \le 2A^2(1+t)\int_0^t \mathbb{E}\left(X_s - Y_s\right)^2 ds. \tag{59}$$

By Lipschitz of a, σ and boundedness of ξ,

$$\mathbb{E}X_t^2, \mathbb{E}Y_t^2 \le C,$$

for all $t \in [0,T]$ and some constant C. Thus, $\phi(t) \le 2\mathbb{E}X_t^2 + 2\mathbb{E}Y_t^2 \le 4C$. Yet using (59),

$$\phi(t) \le 2A^2(1+T)\int_0^t \phi(s)ds$$

$$\leq \ldots \leq \left(2A^2(1+T)\right)^n \frac{t^n}{n!} 4C$$
$$\leq \frac{[2A^2(1+T)T]^n}{n!} 4C$$
$$\xrightarrow{n\to\infty} 0.$$

Hence $\mathbb{E}\left(X_t - Y_t\right)^2 = 0$. This implies

$$\mathbb{P}\left(X_t \neq Y_t\right) = 0 \ \forall t \in [0,T].$$

By continuity of X_t and Y_t (with countability and density of the rationals),

$$\mathbb{P}\left[X_t \neq Y_t \text{ for some } t \in [0,T]\right] = 0. \qquad \square$$

21 Lecture 13/4

21.1 Concentration inequalities

We will now switch gears and move to a topic which has become extremely useful in modern research – concentration inequalities.

21.1.1 Motivation

Consider the law of large numbers. Given iid $X_1, X_2, \ldots$ with $\mathbb{E}|X_1| < \infty$,

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mathbb{E}X_1,$$

where $\bar{X}_n = \frac{1}{n}\left[X_1 + \ldots + X_n\right]$. There are two key properties driving the LLN: 1) $\bar{X}_n$ is a function of many independent random variables $X_1, \ldots, X_n$; and 2) each random variable X_i has a "small" contribution (we will make this precise later) since we scale X_i by $\frac{1}{n}$. 1) and 2) imply that the sample mean "concentrates" around a deterministic value.

Two questions naturally arise: (i) Can we be more quantitative? What is the rate of convergence? That is, how large must n be so that $\bar{X}_n$ is close to $\mathbb{E}X_1$? (The limit equation is only theoretically useful since we only ever have access to finite samples.) (ii) Can this be generalised to other functions (beyond the sample mean)?

21.1.2 Set-up

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and let $Y = f(X_1, \ldots, X_n)$ (generally we will consider X_i's iid). We are interested in 'deviation bounds': statements of the form

$$\mathbb{P}\left[|Y - \mathbb{E}Y| > \epsilon\right] \leq C_\epsilon,$$

or (more crudely) variance bounds:

$$\mathrm{Var}(Y) \leq C,$$

(which we can convert to deviation bounds through Chebychev's inequality).

Our aim is to develop widely-applicable tools to establish deviation or variance bounds for general functions of many independent random variables.

21.1.3 Further motivation

As an example, consider the random matrix $\boldsymbol{M} = (M_{ij})$ with $M_{ij} = M_{ji}$ and

$$\{M_{ij} : i \leq j\} \overset{iid}{\sim} \mathcal{N}(0, 1).$$

Let $\lambda_{\max}(M)$ be the largest eigenvalue of M. In many modern scenarios, we are interested in the concentration inequality

$$\mathbb{P}\left[|\lambda_{\max}(M) - \mathbb{E}\lambda_{\max}(M)| > \epsilon\right] = ?$$

We currently do not have tools for determining this.

These are the types of questions that motivate the study of concentration inequalities.

Returning to the example of the sample mean, we can derive concentration inequalities for certain distributions. For example, if $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, 1)$ then $\bar{X}_n \sim \mathcal{N}(0, 1/n)$. We can write precise deviation bounds:

$$\mathbb{P}\left[\left|\bar{X}_n\right| > \epsilon\right] \leq 2e^{-n\epsilon^2/2}.$$

(This comes from Mill's ratio bound to the tail probability of a Gaussian.) This proves $\bar{X}_n = O_P\left(\frac{1}{\sqrt{n}}\right)$.

As another example, if $X_1, \ldots, X_n \overset{iid}{\sim} \text{Expo}(1)$ then $\sum_{i=1}^n X_i \sim \text{Gamma}(n, 1)$ and this can be used to derive similar concentration bounds.

So we can get deviation bounds for the sample mean in some specific scenarios. But these results very crucially depend on the assumptions of the distribution and on the functional form of the sample mean. Generalising these types of results to other functions (beyond the sample mean) and to other families of distributions is very hard.

21.1.4 The bounded differences inequalities

Theorem 21.1. *Let $X_1, \ldots, X_n$ be independent and $f : \mathbb{R}^n \to \mathbb{R}$ satisfy*

1. $\mathbb{E}|f(X_1, \ldots, X_n)| < \infty$*; and*

2. $$\sup_{\substack{x_1,\ldots,x_n \in \mathbb{R} \\ x_i' \in \mathbb{R}}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \tag{60}$$

   for some constant c_i and all $i = 1, \ldots, n$.

Then

$$\mathbb{P}\left[|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| > t\right] \leq 2\exp\left[-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right]. \tag{61}$$

In 2., we are looking at the maximum change in f that results from changing one co-ordinate of f. In the theorem, we assume that this change is bounded.

We hope that the bounds c_i are small, so that the concentration bound (61) is small. If the c_i's are small comparative to t^2 – i.e. $t \ll \sqrt{\sum_{i=1}^n c_i^2}$ – then the probability (61) is small.

This crystallises the intuition behind the earlier exposition on the sample mean (assuming that the random variables X_i are a.s. bounded). The assumption of bounded differences implies that each X_i doesn't contribute much to the function.

The connection of this result to the rest of the course is that martingales are crucial for the proof.

One shortcoming with Theorem 21.1 is that the worst-case difference (as in (60)) is hard to bound, even when, in probability, the difference is likely to be small. That

is, Theorem 21.1 does not capture the right behaviour if c_i is much greater than the typical difference $f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})$ (where $\boldsymbol{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$ with $X_i \sim X_i'$ independently.

The proof of the bounded differences inequalities relies on Hoeffding's lemma:

Lemma 21.2 (Hoeffding's)**.** *Suppose X is a random variable with $\mathbb{E}X = 0$ and $a \leq X \leq b$ a.s. for some constants a and b. Then*

$$\mathbb{E}\left[e^{tX}\right] \leq \exp\left[\frac{t^2(b-a)^2}{8}\right].$$

We will prove this Lemma later.

Proof of Theorem 21.1. Write f for $f(\boldsymbol{X})$ and $f = \mathbb{E}f = \sum_{i=1}^n V_i$, where

$$V_i = \mathbb{E}\left[f|X_1, \ldots, X_i\right] - \mathbb{E}\left[f|X_1, \ldots, X_{i-1}\right],$$

are called the martingale differences.

We use a Chernoff-bound idea (which is a common proof technique for concentration inequalities):

$$\begin{aligned}
\mathbb{P}\left[f - \mathbb{E}f > t\right] &= \mathbb{P}\left[\sum_{i=1}^n V_i > t\right] \\
&= \mathbb{P}\left[e^{s\sum_{i=1}^n V_i} > e^{st}\right] \\
&\leq e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^n V_i}\right],
\end{aligned}$$

for any $s > 0$ where the third line follows by Markov's inequality. Hence

$$\mathbb{P}\left[f - \mathbb{E}f > t\right] \leq \inf_{s>0}\left\{e^{-st}\mathbb{E}\left[e^{s\sum_{i=1}^n V_i}\right]\right\}. \tag{62}$$

For $1 \leq i \leq n$, we have

$$\mathbb{E}V_i = \mathbb{E}\left[\mathbb{E}(f|X_1, \ldots, X_i) - \mathbb{E}(f|X_1, \ldots, X_{i=1})\right] = 0,$$

and $L_i \leq V_i \leq U_i$ where

$$L_i = \inf_x \left\{\mathbb{E}\left[f(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n)|X_1, \ldots, X_{i-1}\right] - E\left[f|X_1, \ldots, X_{i-1}\right]\right\},$$

$$U_i = \sup_x \left\{ \mathbb{E}\left[f(X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n) | X_1, \ldots, X_{i-1} \right] - E\left[f | X_1, \ldots, X_{i-1} \right] \right\}.$$

L_i and U_i are random variables and functions solely of $X_1, \ldots, X_{i-1}$. So L_i and U_i are deterministic given $X_1, \ldots, X_{i-1}$. Finally, $U_i - L_i \leq c_i$ by assumption. So we can apply Hoeffding's lemma:

$$\begin{aligned}
\mathbb{E}\left[e^{s \sum_{i=1}^n V_i} \right] &= \mathbb{E}\left[\mathbb{E}\left(e^{s \sum_{i=1}^n V_i} | X_1, \ldots, X_{n-1} \right) \right] \\
&= \mathbb{E}\left[e^{s \sum_{i=1}^{n-1} V_i} \mathbb{E}\left(e^{s V_n} | X_1, \ldots, X_{n-1} \right) \right] \\
&\leq \mathbb{E}\left[e^{s \sum_{i=1}^{n-1} V_i + s^2 c_n^2/8} \right] \\
&\leq \ldots \leq e^{s^2 \sum_{i=1}^n c_i^2/8}.
\end{aligned}$$

(Note that we haven't used independence of the X_i's – there are more general statements of the bounded differences inequality which does not assume independence.) Plugging this into (62),

$$\begin{aligned}
\mathbb{P}\left[f - \mathbb{E}f > t \right] &\leq \inf_{s>0} \exp\left[-st + \frac{s^2}{8} \sum_{i=1}^n c_i^2 \right] \\
&= \exp\left[-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right],
\end{aligned}$$

after some calculus. The proof for bounding the other tail:

$$\mathbb{P}\left[\mathbb{E}f - f > t \right] = \mathbb{P}\left[f - \mathbb{E}f < -t \right] \leq \exp\left[-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right],$$

is basically the same (just work with $s < 0$ and get the same bound). A union bound argument then gives the rsult. □

Proof of Hoeffding's lemma (Lemma 21.2). Hoeffding's lemma is essentially a fact about convexity:

$$e^{tX} \leq \frac{X_a}{b-a} e^{tb} + \frac{b - X}{b-a} e^{ta},$$

because $x \mapsto e^{tx}$ is convex.

Since $\mathbb{E}X = 0$,

$$\mathbb{E}\left[e^{tX} \right] \leq \frac{-a}{b-a} e^{tb} + \frac{b}{b-a} e^{ta}$$

$$= e^{g(u)},$$

where $u = t(b-a)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$ and $\gamma = -\frac{a}{b-a}$. (The calculation of this is left as an exercise.)

Observe that $g(0) = 0 = g'(0)$ and $g''(u) \leq \frac{1}{4}$ for all $u > 0$. So "the growth near zero is quadratic and small", which allows us to bound $g(u)$. More formally,

$$g(u) = g(0) + g'(0) + \frac{u^2}{2} g''(\xi) \leq u^2/8,$$

for some $0 \leq \xi \leq u$. Thus,

$$\mathbb{E}\left[e^{tX}\right] \leq \exp\left[u^2/8\right] = \exp\left(\frac{t^2(b-a)^2}{8}\right). \qquad \square$$

21.1.5 Application: max cuts for random graphs

We will demonstrate the bounded differences inequality with an application: Let $G_n \sim G(n, cn)$ be an Erdős-Rényi graph. (Sample G_n uniformly at random from all graphs on n vertices with cn edges.)

Define $\text{Maxcut}(G_n)$ to be the maximum number of edges between two disjoint sets of vertices. Then

$$\mathbb{P}\left[|\text{Maxcut}(G_n) - \mathbb{E}\left[\text{Maxcut}(G_n)\right]| > n\epsilon\right] \leq 2\exp\left[-\frac{2n^2\epsilon^2}{nc}\right],$$

since each edge can change the Maxcut value by at most 1. (Maxcut is a function f of the adjacency matrix and so the differences in f by changing one co-ordinate are determined by adding or removing a single edge.)

22 Lecture 20/4

22.1 The Efron-Stein inequality

Recall the bounded differences inequality: If $f : \mathbb{R}^n \to \mathbb{R}$ is a function of independent $X_1, \ldots, X_n$ and f doesn't change "too much" as we perturb one co-ordinate, then

$f \approx \mathbb{E}f$. A limitation with this result is that it looks at the worst case influences (i.e. the worst case changes in f from changes in one co-ordinate). We want a result that only requires bounding the typical influences/differences.

Theorem 22.1. *Let $X_1, \ldots, X_n, X_1', \ldots, X_n'$ independent with $X_i \sim X_i'$. Write*

$$\boldsymbol{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n).$$

Then

$$\operatorname{Var}[f(\boldsymbol{X})] \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)^2\right].$$

Think of $\boldsymbol{X}_i'$ as a fresh iid copy of $\boldsymbol{X}$. So $\mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)^2\right]$ looks like the typical (in terms of expectation) difference from drawing a fresh X_i.

The Efron-Stein inequality bounds the variance; a deviation bound follows from Chebychev's inequality.

Proof. Write $\boldsymbol{X}^{[i]} = (X_1', \ldots, X_i', X_{i+1}, \ldots, X_n)$. We will use a telescoping argument:

$$\begin{aligned}
\operatorname{Var}[f(\boldsymbol{X})] &= \mathbb{E}f^2(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})f(\boldsymbol{X}') \\
&= \mathbb{E}\left[f(\boldsymbol{X})\left(f(\boldsymbol{X}) - f(\boldsymbol{X}')\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[f(\boldsymbol{X})\left(f(\boldsymbol{X}^{[i-1]}) - f(\boldsymbol{X}^{[i]})\right)\right].
\end{aligned}$$

We can flip X_i and X_i' by independence:

$$f(\boldsymbol{X})\left(f(\boldsymbol{X}^{[i-1]}) - f(\boldsymbol{X}^{[i]})\right) \sim f(\boldsymbol{X}^{(i)})\left(f(\boldsymbol{X}^{[i]}) - f(\boldsymbol{X}^{[i-1]})\right),$$

since $X_i \sim X_i'$ independent of X_j, X_j'. Thus,

$$\begin{aligned}
\operatorname{Var}[f(\boldsymbol{X})] &= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)\left(f(\boldsymbol{X}^{[i-1]}) - f(\boldsymbol{X}^{[i]})\right)\right] \\
&\leq \frac{1}{2} \sum_{i=1}^{n} \sqrt{\mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)^2\right] \mathbb{E}\left[\left(f(\boldsymbol{X}^{[i-1]}) - f(\boldsymbol{X}^{[i]})\right)^2\right]} \\
&= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)^2\right],
\end{aligned}$$

where the second line follows by Cauchy-Schwarz. □

This proof technique – using an interpolation-type argument – is extremely useful in many different contexts, especially in modern probability research.

When is the E.S. bound tight? The only point where the argument is loose, is in the application of Cauchy-Schwarz. This observation leads to a rich and emerging field that explores what it means if the bound is loose. See [Cha14].

22.1.1 The jackknife

The E.S. inequality first arose in solving problems in non-parameter statistics. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} F$ and we are interested in estimating a functional $\theta(F)$ of F. (As concrete examples, $\theta(F) = \mathbb{E}_F(X)$ or $\theta(F) = \mathrm{Var}_F(X)$.) Suppose we have an estimator $\hat{\theta}_n = f_n(X_1, \ldots, X_n)$. Typically we want to understand the bias $B(\hat{\theta}_n) = \mathbb{E}\hat{\theta}_n - \theta$ and variance $\mathrm{Var}(\hat{\theta}_n)$. Yet, on the face of it, the bias and variance are hard to estimate: we don't know the underlying probability distribution and even if we did, $\hat{\theta}_n$ could be a really complicated function of the X_i's.

This motivates the following important question: Is there a straightforward way to estimate the bias and variance of an estimator $\hat{\theta}_n$ in this general, non-parametric setting?

The jackknife is one possible method: Let $\hat{\theta}_n^{(i)} = f_{n-1}(X^{(i)})$, where

$$X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n),$$

is the *i-th jackknife sample*. ($X^{(i)}$ drops the i-th point from the sample.)

The jackknife estimate of the variance is then given by $\sum_{i=1}^n \left(\hat{\theta}_n^{(i)} - \hat{\theta}_n\right)^2$.

Lemma 22.2 (Efron-Stein)**.**

$$\mathrm{Var}\left(\hat{\theta}_n\right) \leq \sum_{i=1}^n \mathbb{E}\left[\left(\hat{\theta}^{(i)} - \hat{\theta}\right)^2\right].$$

This result is very useful in practice, since it gives an upper bound to the variance without requiring consideration of the analytical form of $\hat{\theta}_n$.

Proof.

$$\mathrm{Var}\left(\hat{\theta}_n\right) \leq \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(f_n(X_1, \ldots, X_n) - f_n(X_1, \ldots, X_i', \ldots, X_n)\right)^2\right]$$

$$= \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}\left(f_n | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\right)\right]$$
$$\leq \sum_{i=1}^{n} \mathbb{E}\left[\left(f_n(X_1, \ldots, X_n) - f_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\right)^2\right],$$

where the second line follows by adding and subtracting $\mathbb{E}f_n$ inside the square; and the final line follows from the inequality

$$\mathrm{Var}(Y) = \mathbb{E}\left[(Y - \mathbb{E}Y)^2\right] \leq \mathbb{E}\left[(X - c)^2\right],$$

for all constants c. □

22.1.2 Applications of the E.S. inequality

1. Consider the scenario where we have a function $f : \mathcal{X}^n \to \mathbb{R}$ with bounded differences. That is, assume

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i, \tag{63}$$

for all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$.

Lemma 22.3. *If f has the bounded differences property (63), then*

$$\mathrm{Var}(f) \leq \frac{1}{4} \sum_{i=1}^{n} c_i^2.$$

Proof. Using the E.S. inequality,

$$\mathrm{Var} f \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[\left(f(\boldsymbol{X}) - f(\boldsymbol{X}^{(i)})\right)^2\right]$$
$$= \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}\left(f | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\right)\right]$$
$$\leq \frac{1}{4} \sum_{i=1}^{n} c_i^2,$$

where the second line follows by adding and subtracting $\mathbb{E}f$ inside the square; and the third line by the fact that, conditioning on $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$, the function

f can vary by at most c. (Subject to this constraint, the variance is maximised whne the distribution of f has equal point masses at $-\frac{c_i}{2}$ and $\frac{c_i}{2}$. The variance of this distribution is $\frac{c_i^2}{4}$.) $\square$

2. Kernel density estimation: Given $X_1, \ldots, X_n \overset{iid}{\sim} \varphi$ (where φ is the PDF), we are interested in estimating φ based on the data. The idea is that $\varphi(x)$ can be estimated by how many of the $X_i's$ are observed in a small window around x. This can be made more sophisticated by weighting X_i's contribution to $\varphi(x)$ by the distance between x and X_i.

We formalise this using a function K which satisfies $K \geq 0$ and $\int K(x)dx = 1$. Such a K is called a *kernel function*. One may think of K as the weighting function. Then define

$$\hat{\varphi}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right),$$

where h_n is the length of the window (called the *bandwidth*) around x. (h_n is a smoothing parameter since it determines how many X_i's contribute to estimating $\hat{\varphi}_n(x)$.) We divide by nh_n since if the X_i's were uniform then we would expect nh_n observations in the region $(x - h_n/2, x + h_n/2)$.

This is an example of non-parametric density estimation. The E.S. inequality can be used to understand the statistical properties/performance of $\hat{\varphi}_n$.

One performance metric is the L_1 loss:

$$L_n = \int |\hat{\varphi}_n(x) - \varphi(x)|dx.$$

L_n is a random variable since $\hat{\varphi}_n$ is a random function. We could study $\mathbb{E}L_n$ instead of L_n directly; but for this to be valid, we would need to show that L_n behaves like its expectation – that is, L_n's random fluctuations around $\mathbb{E}L_n$ are small.

Let $\hat{\varphi}_n^{(i)}$ be the density estimate based on $X_1, \ldots, X_i', \ldots, X_n$. Then

$$\begin{aligned}\left| \int |\varphi(x) - \hat{\varphi}_n(x)|dx - \int \left|\varphi(x) - \hat{\varphi}_n^{(i)}(x)\right|dx \right| &\leq \int \left|\hat{\varphi}_n(x) - \hat{\varphi}_n^{(i)}(x)\right|dx \\ &= \int \left| \frac{1}{nh} \left[K\left(\frac{x - X_i}{h_n}\right) - K\left(\frac{x - X_i'}{h_n}\right) \right] \right| dx\end{aligned}$$

$$\leq \int \frac{1}{nh_n}\left[K\left(\frac{x-X_i}{h_n}\right)+K\left(\frac{x-X_i'}{h_n}\right)\right]dx$$
$$=\frac{2}{n},$$

where the first line uses the triangle inequality; and the final line follows by change of variables using $\int K(x)dx = 1$. The E.S. inequality implies that

$$\mathrm{Var}(L_n) \leq \frac{C}{n}, \tag{64}$$

for some constant C.

Using independent arguments (not related to concentration),

$$\mathbb{E}L_n = \mathbb{E}\left[\int |\varphi(x) - \hat{\varphi}_n(x)|dx\right] \gg \frac{1}{\sqrt{n}}. \tag{65}$$

Thus,

$$\frac{L_n}{\mathbb{E}L_n} = \frac{\int |\varphi(x) - \hat{\varphi}_n(x)|dx}{\mathbb{E}\left[\int |\varphi(x) - \hat{\varphi}_n(x)|dx\right]} \xrightarrow{P} 1,$$

by Chebychev's inequality. (Use the lower bound (65) on the denominator and the upper bound (64) on the numerator.)

Thus, fluctuations of L_n are on a much smaller scale than $\mathbb{E}L_n$, so it is enough to understand $\mathbb{E}L_n$, instead of studying L_n directly.

22.1.3 Summary

The Efron-Stein inequality allows us to derive robust upper bounds on complicated functions of many independent random variables. The upper bound will be tight if the function doesn't change too much when we refresh each co-ordinate.

23 Lecture 22/4

23.1 Concentration for Lipschitz functions of Gaussians

Theorem 23.1. *Let $\boldsymbol{g} = (g_i)_{i=1}^n \sim \mathcal{N}(\mathbf{0}, I_n)$ and $F : \mathbb{R}^n \to \mathbb{R}$ be Lipschitz. Then, for all $t > 0$,*

$$\mathbb{P}\left[|F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g})| > t\right] \leq 2\exp\left(-\frac{t^2}{4\|F\|_{\text{Lip}}^2}\right),$$

where

$$\|F\|_{\text{Lip}} = \inf\left\{L > 0 : |F(x) - F(y)| \leq L\|x - y\|_2\right\},$$

and $\|\cdot\|_2$ is the L_2 norm.

"Take any Lipschitz function of iid Gaussians and it concentrates on the $O(1)$ (in terms of n) scale." ($\|F\|_{\text{Lip}}$ is usually $O(1)$ in terms of n.)

This theorem is useful and widely applicable. (Such a result is a rarity.)

The proof of Theorem 23.1 relies on a non-trivial bound on the MGF of F:

Lemma 23.2.

$$\mathbb{E}\left[\exp\left(\lambda(F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g}))\right)\right] \leq e^{\lambda^2 L^2},$$

if $L^2 \geq \|F\|_{\text{Lip}}$.

$e^{\lambda^2 L^2}$ looks like the MGF of $\mathcal{N}(0, L)$. So the intuition is that $F(\boldsymbol{g}) \sim \mathcal{N}\left(\mathbb{E}F(\boldsymbol{g}), L\right)$.

Proof of Theorem 23.1, assuming Lemma 23.2.

$$\begin{aligned}
\mathbb{P}\left[F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g}) \geq t\right] &= \mathbb{P}\left[e^{\lambda[F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g})]} \geq e^{\lambda t}\right] \\
&\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda(F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g}))}\right] \\
&\leq e^{-\lambda t + \lambda^2 L^2},
\end{aligned}$$

where the second line uses Markov's and the third relies on Lemma 23.2.

Now minimise the RHS with respect to λ. After some algebra, we get $\lambda = \frac{t}{2L^2}$ and $L = \|F\|_{\text{Lip}}$.

This controls the upper tail at $\exp\left(-\frac{t^2}{4\|F\|_{\text{Lip}}^2}\right)$. Apply the same argument to $-F$ and use a union bound to get the result. □

23.1.1 Proof of Lemma 23.2

This is where the meat of the problem is. The motivation for showing this proof is that it will illustrate high dimensional properties of Gaussians which are counterintuitive. It will also present some useful proof techniques.

We need the following two results:

Lemma 23.3 (Gaussian integration by parts)**.** *If $\boldsymbol{g} \sim \mathcal{N}(0, \Sigma)$, then*

$$\mathbb{E}\left[g_1 F(\boldsymbol{g})\right] = \sum_{l=1}^{n} \mathbb{E}\left[g_1 g_l\right] \mathbb{E}\left[\frac{\partial F}{\partial x_l}(\boldsymbol{g})\right],$$

if either $\mathbb{E}\left|\frac{\partial F}{\partial x_l}\right| < \infty$ or $\mathbb{E}\left[g_1 g_l\right] = 0$.

Aside: the univariate case

$$\mathbb{E}gF(g) = \sigma^2 \mathbb{E}F'(g),$$

for $g \sim \mathcal{N}(0, \sigma^2)$, is called Stein's identity or Stein's lemma.

Lemma 23.4 (Gaussian interpolation)**.** *Suppose $\boldsymbol{X} = (X_i)_{i=1}^n \perp\!\!\!\perp \boldsymbol{Y} = (Y_i)_{i=1}^n$ are mean zero Gaussian vectors (with covariances Σ_X and Σ_Y). Let $a_{ij} = \mathbb{E}X_i X_j$ and $b_{ij} = \mathbb{E}Y_i Y_j$. Define*

$$\boldsymbol{Z}(t) = \sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y}.$$

Then $\mathbb{E}\boldsymbol{Z}(t) = \boldsymbol{0}$ and

$$\mathbb{E}\left[Z_i(t) Z_j(t)\right] = t a_{ij} + (1-t) b_{ij},$$

for all $t \in [0, 1]$.

If $F : \mathbb{R}^n \to \mathbb{R}$ is sufficiently well-behaved (i.e. the conditions for EDI are satisfied and the derivatives of F exist and don't grow too fast, so that $\mathbb{E}\left|\frac{\partial F}{\partial x_l}\right| < \infty$), and $f(t) = \mathbb{E}F(\boldsymbol{Z}(t))$, then

$$f'(t) = \frac{1}{2}\sum_{i,j}(a_{ij} - b_{ij})\mathbb{E}\frac{\partial^2 F}{\partial x_i \partial x_j}\left(\boldsymbol{Z}(t)\right).$$

$\boldsymbol{Z}(t)$ should be viewed as some "path" between X and Y. It is a linear combination of Gaussians, and hence is itself Gaussian. Its covariance is a linear interpolation between $\boldsymbol{X}$ and $\boldsymbol{Y}$'s covariance.

Proof of Lemma 23.2, assuming Lemma 23.4. Assume that F is differentiable and $\|\nabla F(\boldsymbol{x})\|_2 \leq L$, for all $\boldsymbol{x} \in \mathbb{R}^n$. (If F is Lipschitz and differentiable then we get this bound for free.) If F is not differentiable, then we can smooth F.

Let $\boldsymbol{g}, \boldsymbol{g}^{(1)}, \boldsymbol{g}^{(2)} \overset{iid}{\sim} \mathcal{N}(0, I_n)$ and define

$$f(t) = \mathbb{E} \exp\left[\lambda\left(F\left[\sqrt{t}g^{(1)} + \sqrt{1-t}\boldsymbol{g}\right] - F\left[\sqrt{t}\boldsymbol{g}^{(2)} + \sqrt{1-t}\boldsymbol{g}\right]\right)\right].$$

There are two immediate, but key, observations:

1. $f(0) = 1$;
2. Defining $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{g}^{(1)} \\ \boldsymbol{g}^{(2)}\end{bmatrix}$ and $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{g} \\ \boldsymbol{g}\end{bmatrix}$, we have $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, I_{2n})$ and

$$\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} I_n & I_n \\ I_n & I_n \end{bmatrix}\right),$$

with $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$. Set

$$\begin{aligned} G : \mathbb{R}^{2n} &\to \mathbb{R} \\ (x_1, \ldots, x_{2n}) &\mapsto \exp\left[\lambda\left(F(x_1, \ldots, x_n) - F(x_{n+1}, \ldots, x_{2n})\right)\right]. \end{aligned}$$

Then $f(t) = \mathbb{E}\left[G\left(\sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y}\right)\right]$.

Differentiating in t, using Lemma 23.4,

$$\begin{aligned} f'(t) &= \frac{1}{2}\sum_{i,j}(a_{ij} - b_{ij})\mathbb{E}\left[\frac{\partial^2 G}{\partial x_i \partial x_j}(\boldsymbol{Z}(t))\right] \\ &= -\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial^2 G}{\partial x_i \partial x_{i+n}}(\boldsymbol{Z}(t))\right], \end{aligned} \tag{66}$$

since $a_{ij} - b_{ij} \neq 0$ if and only if $j \neq i + n$. (Why? Look at $(a_{ij}) = I_{2n}$ and $(b_{ij}) = \begin{bmatrix} I_n & I_n \\ I_n & I_n \end{bmatrix}$.)

Next observe that

$$\frac{\partial^2 G}{\partial x_i \partial x_{i+n}}(x_1, \ldots, x_{2n}) = -\lambda^2 G \frac{\partial F}{\partial x_i}(x_1, \ldots, x_n) \frac{\partial F}{\partial x_i}(x_{n+1}, \ldots, x_{2n}),$$

so that

$$\begin{aligned} -\sum_{i=1}^{n} \frac{\partial^2 G}{\partial x_i \partial x_{i+n}}(x_1, \ldots, x_{2n}) &= \lambda^2 G \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(x_1, \ldots, x_n) \frac{\partial F}{\partial x_i}(x_{n+1}, \ldots, x_{2n}) \\ &\leq \lambda^2 G \|\nabla F(x_1, \ldots, x_n)\|_2 \times \|\nabla F(x_{n+1}, \ldots, x_{2n})\|_2 \\ &\leq \lambda^2 G L^2, \end{aligned}$$

by Cauchy-Schwarz and the bound $\|\nabla F\|_2 \leq L$. Plugging this into (66), we obtain

$$\begin{aligned} f'(t) &\leq \lambda^2 L^2 \mathbb{E}\left[G(\sqrt{t}\boldsymbol{X} + \sqrt{1-t}\boldsymbol{Y})\right] \\ &= \lambda^2 L^2 f(t). \end{aligned}$$

So we have obtained a differential equation. (This is crucial for the proof!) Thus,

$$\frac{d}{dt}\left(f(t)e^{-\lambda^2 L^2 t}\right) = \left(f'(t) - \lambda^2 L^2 f(t)\right) e^{-\lambda^2 L^2 t} \leq 0.$$

So $f(1)e^{-\lambda^2 L^2} \leq f(0) = 1$ and hence

$$f(1) \leq e^{\lambda^2 L^2}. \tag{67}$$

Also, by Jensen's inequality

$$\begin{aligned} f(1) &= \mathbb{E} \exp\left(\lambda\left[F(\boldsymbol{g}^{(1)}) - F(\boldsymbol{g}^{(2)})\right]\right) \\ &\geq \mathbb{E}_{\boldsymbol{g}^{(1)}} \exp\left(\lambda\left[F(\boldsymbol{g}^{(1)}) - \mathbb{E}_{\boldsymbol{g}^{(2)}} F(\boldsymbol{g}^{(2)})\right]\right) \\ &= \mathbb{E} \exp\left(\lambda\left[F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g})\right]\right), \end{aligned} \tag{68}$$

where $\mathbb{E}_W$ denotes the expectation is with respect to the random variable W. Observe that this is exactly the MGF that we wanted. Combining (68) with (67), we get the desired result

$$\mathbb{E} \exp\left(\lambda\left[F(\boldsymbol{g}) - \mathbb{E}F(\boldsymbol{g})\right]\right) \leq e^{\lambda^2 L^2}. \qquad \square$$

We still need to prove Lemmas 23.3 and 23.4.

Proof of Lemma 23.4. By EDI and then chain rule,

$$f'(t) = \sum_{i=1}^{n} \mathbb{E}\left[Z_i'(t) \frac{\partial F}{\partial x_i}(\boldsymbol{Z}(t)) \right].$$

We want to apply Gaussian IBP (Lemma 23.3). We know that $(Z_i'(t), Z_1(t), \ldots, Z_n(t))$ is MVN. Consider the function

$$F_i(x_1, \ldots, x_{n+1}) = \frac{\partial F}{\partial x_i}(x_2, \ldots, x_{n+1}).$$

By IBP,

$$\begin{aligned}\mathbb{E}\left[Z_i'(t) \frac{\partial F}{\partial x_i}(\boldsymbol{Z}(t)) \right] &= \mathbb{E}\left[Z_i' Z_i'\right] \mathbb{E}\left[\frac{\partial F_i}{\partial x_1}(Z_i'(t), Z_1(t), \ldots, Z_n(t)) \right] \\ &\quad + \sum_{j=1}^{n} \mathbb{E}\left[Z_i' Z_j\right] \mathbb{E}\left[\frac{\partial F_i}{\partial x_{j+1}}(Z_i'(t), Z_1(t), \ldots, Z_n(t)) \right].\end{aligned}$$

(Note that $\frac{\partial F_i}{\partial x_{j+1}}$ is the partial derivative of F_i with respect to the $(j+1)$-th coordinate.) Since F_i is constant in x_1, $\frac{\partial F_i}{\partial x_1} = 0$. For the other terms on the RHS,

$$\begin{aligned}\mathbb{E}\left[Z_i' Z_j\right] &= \mathbb{E}\left[\left(\frac{1}{2\sqrt{t}} X_i - \frac{1}{2\sqrt{1-t}} Y_i \right) \left(\sqrt{t} X_j + \sqrt{1-t} Y_j \right) \right] \\ &= \frac{1}{2}\left(a_{ij} - b_{ij}\right),\end{aligned}$$

since $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$. Thus,

$$\mathbb{E}\left[Z_i'(t) \frac{\partial F}{\partial x_i}(\boldsymbol{Z}(t)) \right] = \frac{1}{2} \sum_{j=1}^{n} \left(a_{ij} - b_{ij}\right) \mathbb{E}\left[\frac{\partial F_i}{\partial x_{j+1}}(Z_i'(t), Z_1(t), \ldots, Z_n(t)) \right]. \qquad \square$$

We leave the proof of Lemma 23.3 until next lecture.

23.1.2 Application

Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, I_n)$ and $F(\boldsymbol{X}) = \max_{1 \le i \le n} X_i$. Since F is 1-Lipschitz,

$$\mathbb{P}\left[\left| \max_{1 \le i \le n} X_i - \mathbb{E}\left[\max_{1 \le i \le n} X_i \right] \right| > t \right] \le 2e^{-t^2/4}. \tag{69}$$

This suggest that $\mathrm{Var}\left(\max_{1\le i\le n} X_i\right) = O(1)$. This is false – $\mathrm{Var}\left(\max_{1\le i\le n} X_i\right) \to 0$! So in this example, the concentration inequality (Theorem 23.1) is not even tight!

But the same result (69) applies for MVN with any covariance matrix. If the covariance is a matrix of ones, then

$$\mathrm{Var}\left(\max_{1\le i\le n} X_i\right) = \mathrm{Var}(X_1) = O(1).$$

So the bound (Theorem 23.1) is tight – we can't do any better than this if we want to look at MVN with arbitrary covariance.

24 Lecture 27/4

24.1 Gaussian concentrations (cont.)

24.1.1 Proof of Lemma 23.3

Proof of 23.3. Begin in the univariate setting: $g \sim \mathcal{N}(0, \sigma^2)$. It suffices to prove

$$\mathbb{E}gF(g) = \sigma^2 \mathbb{E}F'(g),$$

assuming

$$\mathbb{E}|F'(g)| < \infty. \tag{70}$$

We need only prove this for $\sigma = 1$ – or $z \sim \mathcal{N}(0,1)$ – and then use a change of variables $g = \sigma z$.

We have that

$$\mathbb{E}F'(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} F'(u) e^{-u^2/2} du + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} F'(u) e^{-u^2/2} du. \tag{71}$$

Observe that

$$e^{-u^2/2} = \begin{cases} -\int_{-\infty}^{u} x e^{-x^2/2} dx & \text{if } u \le 0, \\ \int_{u}^{\infty} x e^{-x^2/2} dx & \text{if } u > 0. \end{cases} \tag{72}$$

The required result follows by plugging in (72) into (71) and interchanging the integrals (by Fubini, using assumption (70)).

Now consider the multivariate setting. We will use a typical proof strategy: Regress out the effect of g_1 (i.e. remove the correlation of g_1 and g_l), allowing us to apply the univariate result.

Suppose that $\nu^2 = \mathbb{E}g_1^2$. Define $g_l' = g_l - \lambda_l g_1$ where $\lambda_l = \nu^{-2}\mathbb{E}g_1 g_l$, so that

$$\mathbb{E}g_1 g_l' = \mathbb{E}g_1 g_l - \nu^{-2}\mathbb{E}g_1 g_l \mathbb{E}g_1^2 = 0,$$

and hence $g_l' \perp\!\!\!\perp g_1$ for all l.

Denote $\mathbb{E}_1$ to be the expectation with respect to g_1 only (i.e. condition on $g_2, \ldots, g_n$). We have that

$$\mathbb{E}_1 \left[g_1 F(\boldsymbol{g}) \right] = \mathbb{E}_1 \left[g_1 F(\boldsymbol{g}' + g_1 \boldsymbol{\lambda}) \right].$$

Conditioning on $g_2, \ldots, g_n$, $\boldsymbol{g}'$ is a constant, so that $F(\boldsymbol{g}' + g_1\boldsymbol{\lambda})$ is a function of g_1 only. So we can apply univariate Gaussian IBP:

$$\mathbb{E}_1 \left[g_1 F(\boldsymbol{g}) \right] = \nu^2 \mathbb{E}_1 \left[\left. \frac{\partial F(\boldsymbol{g}' + y\boldsymbol{\lambda})}{\partial y} \right|_{y=g_1} \right].$$

By the chain rule,

$$\begin{aligned} \left. \frac{\partial F(\boldsymbol{g}' + y\boldsymbol{\lambda})}{\partial y} \right|_{y=g_1} &= \sum_{l=1}^{n} \lambda_l \frac{\partial F}{\partial x_l}(\boldsymbol{g}' + g_1\boldsymbol{\lambda}) \\ &= \sum_{l=1}^{n} \lambda_l \frac{\partial F}{\partial x_l}(\boldsymbol{g}). \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}_1 \left[g_1 F(\boldsymbol{g}) \right] &= \nu^2 \sum_{l=1}^{n} \lambda_l \mathbb{E}_1 \left[\frac{\partial F}{\partial x_l}(\boldsymbol{g}) \right] \\ &= \sum_{l=1}^{n} \mathbb{E}\left[g_1 g_l \right] \mathbb{E}_1 \left[\frac{\partial F}{\partial x_l}(\boldsymbol{g}) \right]. \end{aligned}$$

Taking the expectation with respect to $g_2, \ldots, g_n$ completes the proof. □

24.1.2 Applications of Gaussian concentration

First application – supremum of linear functionals: Let $g \sim \mathcal{N}(0, I_n)$ and $A \subset \mathbb{R}^n$ bounded. Consider $F(x) = \sum_{a \in A} \langle a, x \rangle$. This is Lipschitz:

$$\begin{aligned} |F(x) - F(y)| &\leq \sup_{a \in A} |\langle a, x - y \rangle| \\ &\leq \sup_{a \in A} \|a\| \|x - y\| \\ &= \left(\sup_{a \in A} \|a\| \right) \|x - y\|, \end{aligned}$$

by the Cauchy-Schwarz inequality, where $(\sup_{a \in A} \|a\|) < \infty$ since A is bounded.

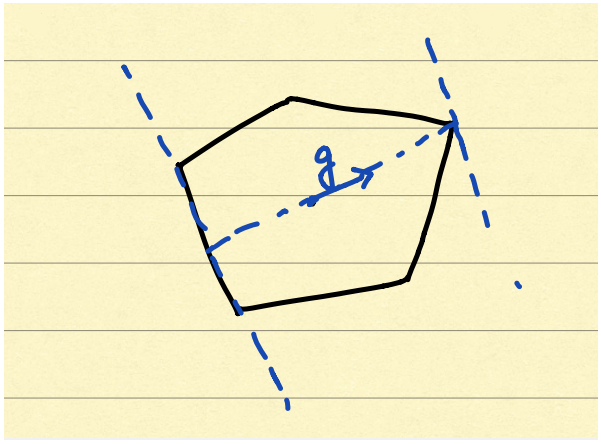
Hence,

$$\mathbb{P}\left[\left| \sup_{a \in A} \langle a, g \rangle - \mathbb{E}\left[\sup_{a \in A} \langle a, g \rangle \right] \right| > t \right] \leq 2 \exp\left(- \frac{t^2}{4 \sup_{a \in A} \|a\|^2} \right). \tag{73}$$

But, why is $\sup_{a \in A} \langle a, g \rangle$ interesting? It has a geometric significance: How can we determine the "width" of A, particularly when A is irregular? Assuming that A contains the origin, then $\frac{1}{\|g\|} \sup_{a \in A} \langle a, g \rangle$ is the distance from the origin to the boundary of A in the direction g. Similarly, $-\frac{1}{\|g\|} \inf_{a \in A} \langle a, g \rangle$ is the distance from the origin to the boundary of A in the direction $-g$. Thus,

$$\frac{1}{\|g\|} \left[\sup_{a \in A} \langle a, g \rangle - \inf_{a \in A} \langle a, g \rangle \right]$$

is the width of A in the direction g. Choosing a Gaussian $g \sim \mathcal{N}(0, I_n)$ is like choosing a direction at random. So $\sup_{a \in A} \langle a, g \rangle$ is like the width of A in a random direction g.



It turns out that determining the width of sets is relevant in many applications. So we need to understand $\sup_{a\in A}\langle a, g\rangle$. The concentration inequality (73) says that it suffices to study its expected value $\mathbb{E}\left[\sup_{a\in A}\langle a, g\rangle\right]$ – called the Gaussian width of A. This can be easy or hard to estimate, depending on A.

Second application – concentration of Gaussian maxima. Let $\boldsymbol{X} \sim (\mathbf{0}, \Sigma)$ be n-dimensional with $\max_{1\le i\le n}\Sigma_{ii} \le 1$. We can write $\boldsymbol{X} = \Sigma^{1/2}g$ where $g \sim \mathcal{N}(\mathbf{0}, I_n)$. Hence $X_i = \langle a_i, g\rangle$ for some a_i. We know that $\mathrm{Var}X_i = \|a_i\|^2 \le 1$ for all i.

Applying our previous result (73),

$$\mathbb{P}\left[\left|\max_{1\le i\le n} X_i - \mathbb{E}\left[\max_{1\le i\le n} X_i\right]\right| \ge t\right] \le 2\exp\left[-\frac{t^2}{4}\right].$$

This result holds for any covariance Σ with $\max_{1\le i\le n}\Sigma_{ii} \le 1$. The RHS is small if $t \to \infty$. This suggests that $\max_{1\le i\le n} X_i$ has fluctuations on $O(1)$ (of n) scale. This bound is not tight for iid Gaussians, but this bound is tight for arbitrary covariance structure. (In particular, the bound is tight when $X_1 = \ldots = X_n = g \sim \mathcal{N}(0, 1)$.) In many applications,

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \gg O(1). \tag{74}$$

SO the concentration inequality says that the fluctuations are small relative to the expected value. So it suffices to study the expected value (under assumption (74)).

24.2 Review and further directions

Recall what we have covered in the course:

1. Limit theorems for martingales:
   (a) L_1 and L_p convergence;
   (b) The role of UI, reverse martingales and exchangeability.
2. Brownian motion:
   (a) the strong Markov property;
   (b) the reflection principle;

(c) Wald's lemma.

3. Donsker's theorem (why Brownian motion is the equivalent of Gaussianity in the world of stochastic processes).

4. General stochastic theory:

   (a) The existence of stochastic processes and continuous modifications;

   (b) The Kolmogorov extension theorem;

   (c) The Kolmogoriv-Chentsov criteria (continuity of moments imply continuous modification).

5. Stochastic integration:

   (a) SDEs and their solutions.

6. Concentration inequalities:

   (a) This topic was far removed from the rest of the semester. It provided a glimpse into modern probability and how it supports high dimensional inference.

   (b) Yet it has links to the rest of the course: Concentration inequalities try to understand the cumulative effect of many small effects, like Donsker's theorem.

What else is 'out there'? What are some related and important topics which we didn't cover?

1. Poisson processes and stochastic integration with respect to point processes:

   (a) The Poisson process is a canonical process, like Brownian motion;

   (b) Stocastic integration with respect to point processes has a very similar theory to integration with respect to Brownian motion.

2. Lèvy processes and stable laws:

(a) What model should be used (instead of Brownian motion) when we do not want to assume continuous trajectories? For example, we may want to model stock prices with discontinuous catastrophes. Typically, a Lèvy process is used.

(b) Lèvy processes can be represented as the sum of a Brownian motion, a Poisson process and some deterministic jumps.

3. The Gaussian free field:

   (a) Brownian motion is like choosing a continuous function at random. What if we wanted to sample a random continuous surface? This would be a Gaussian free field.

   (b) The study of the properties of these random surfaces is an active research area.

4. Gaussian fluctuations:

   (a) The CLT states that $\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$ if $X_1, \ldots, X_n$ are iid with $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = 1$. Donsker's theorem states that

$$\left\{ \left(\frac{k}{n}, \frac{S_k}{\sqrt{n}} \right) : k \leq n \right\} + \text{ linear interpolation } \xrightarrow[n\to\infty]{d} \text{sBM}.$$

   What about general functions (not just rescaled sums)? What functions exhibit Gaussian limits? Two important results in this field are Stein's method and the second order Poincarè inequalities. They try to show that if a function doesn't depend too heavily on any single co-ordinate then it should have Gaussian fluctuations.

5. Probability in high dimensions:

   (a) Research in this area in expanding rapidly and often intersects with geometry.

References

[Cha14] S. Chatterjee. *Superconcentration and Related Topics*. Springer Monographs in Mathematics. Springer, New York, 2014.

[MP] P. Mörters and Y. Peres. *Brownian Motion*. Number 30 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK ; New York, 2010.