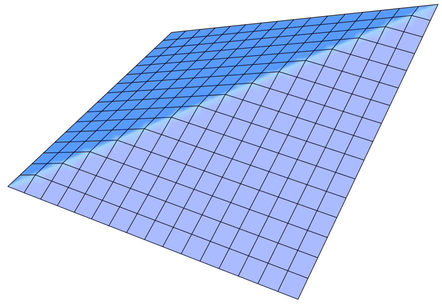

MAT 1601H: Graduate Probability II

Prof. Bálint Virág • Winter 2019 • University of Toronto
TR 9:00–11:00AM in BA 6183

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is `https://rob-zimmerman.github.io/coursenotes`.

Contents

Lecture 1 — Tuesday, January 08, 2019

1. Limit Theorems and Weak Convergence

Central Limit Theorem

At the beginning of the nineteenth century, Pierre-Simon Laplace rediscovered Abraham de Moivre's version of the central limit theorem. De Moivre had introduced the normal approximation in the context of games of chance; Laplace remarked, with some understatement, that "this doctrine is far from encouraging play."

Let $X_1, X_2, \dots$ be independent random variables satisfying

$$\mathbb{P}(X_n = 1) = \mathbb{P}(X_n = -1) = \frac{1}{2} \qquad \text{and} \qquad S_n = \sum_{i=1}^{n} X_i.$$

The strong and weak laws of large numbers give $S_n/n \to 0$ almost surely and in probability. More generally, $S_n/n^p \to 0$ in probability if and only if $p > 1/2$. The critical scale is therefore $\sqrt{n}$. At this scale the probability histogram converges to the bell-shaped curve shown in Figure 1; remarkably, this behaviour persists under assumptions far more general than the present symmetric random walk model.

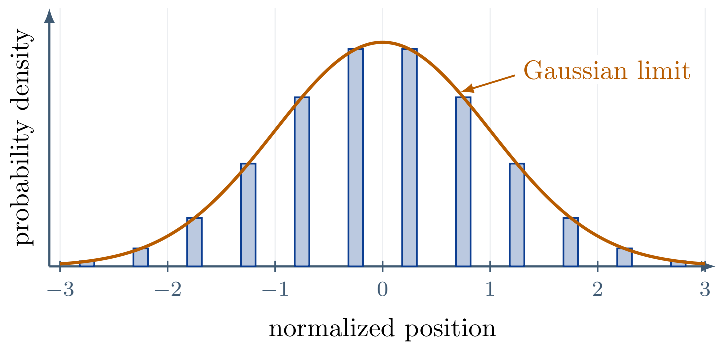

FIGURE 1

To identify the limiting curve, consider $\mathbb{P}(S_{2n} = 2k)$; the even time $2n$ avoids parity distractions. Suppose that $2k/\sqrt{2n} \to x$. Heuristically, the probability of the lattice point corresponding to x should be approximately the bin width, of order $1/\sqrt{n}$, multiplied by a limiting density evaluated at x. We now determine that density directly.

We will use Stirling's formula,

$$n! \sim \sqrt{2\pi n}\, e^{-n} n^n,$$

where the notation $\sim$ means that the ratio of the two sides converges to 1 as $n \to \infty$.

Lemma 1.1 Suppose that $a_n \to 0$, that $1 + a_n > 0$ for all sufficiently large n, and that $a_n b_n \to x$. Then

$$(1 + a_n)^{b_n} \longrightarrow e^x.$$

Proof. The standard limit $\log(1 + a_n)/a_n \to 1$ follows from a Taylor expansion or l'Hôpital's rule. Consequently,

$$b_n \log(1 + a_n) = a_n b_n \, \frac{\log(1 + a_n)}{a_n} \longrightarrow x.$$

Exponentiating gives the result. $\square$

Let H and T denote the numbers of $+1$ and -1 increments, respectively, among the first $2n$ steps.

The event $S_{2n} = 2k$ is equivalent to

$$H - T = 2k, \qquad H + T = 2n, \qquad \text{and} \qquad H = n + k.$$

It follows that

$$\begin{aligned}
\mathbb{P}(S_{2n} = 2k) &= \binom{2n}{n+k}\frac{1}{2^{2n}} \\
&= \frac{1}{2^{2n}}\frac{(2n)!}{(n+k)!(n-k)!} \\
&\sim \frac{1}{2^{2n}\sqrt{2\pi}}\frac{\sqrt{2n}}{\sqrt{n+k}\sqrt{n-k}}\frac{(2n)^{2n}}{(n+k)^{n+k}(n-k)^{n-k}} \\
&= \frac{1}{\sqrt{\pi n}}\frac{1}{\sqrt{1-(k/n)^2}}\frac{1}{(1+k/n)^{n+k}(1-k/n)^{n-k}}.
\end{aligned}$$

The denominator in the final factor can be decomposed as

$$(1+k/n)^{n+k}(1-k/n)^{n-k} = (1+k/n)^k(1-k/n)^{-k}(1-k^2/n^2)^n.$$

The assumption $2k/\sqrt{2n} \to x$ is equivalent to $k/\sqrt{n} \to x/\sqrt{2}$, and therefore $k^2/n \to x^2/2$. Applying Lemma 1.1 to each factor gives

$$(1+k/n)^k \to e^{x^2/2}, \qquad (1-k/n)^{-k} \to e^{x^2/2}, \qquad \text{and} \qquad (1-k^2/n^2)^n \to e^{-x^2/2}.$$

Thus the entire denominator converges to $e^{x^2/2}$, while $\sqrt{1-(k/n)^2} \to 1$. We have thus proved the following local limit theorem:

Theorem 1.2 If $n \to \infty$ and $2k/\sqrt{2n} \to x$, then

$$\mathbb{P}(S_{2n} = 2k) \sim \frac{1}{\sqrt{\pi n}}e^{-x^2/2}.$$

Definition 1.3 The function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

is the **standard Gaussian density**. In this notation, Theorem 1.2 says that

$$\mathbb{P}(S_{2n} = 2k) \sim \varphi(x)\sqrt{\frac{2}{n}}.$$

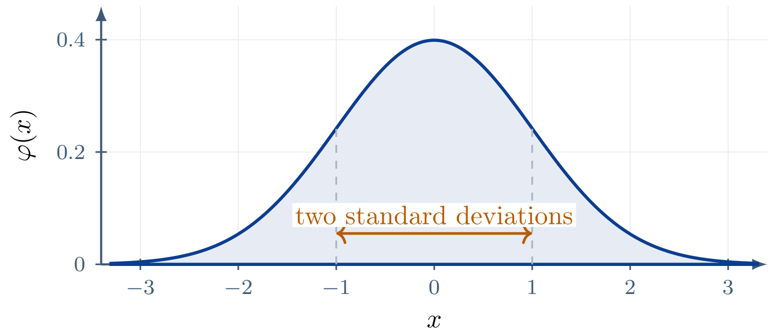

FIGURE 2

The graph of φ appears in Figure 2. To pass from the point probabilities in Theorem 1.2 to interval probabilities, fix $a < b$. The admissible values of $2k/\sqrt{2n}$ are separated by $\sqrt{2/n}$, so the following sum is a Riemann sum:

$$\mathbb{P}\left(a \le \frac{S_{2n}}{\sqrt{2n}} \le b\right) = \sum_{a\sqrt{2n} \le 2k \le b\sqrt{2n}} \mathbb{P}(S_{2n} = 2k) \longrightarrow \int_a^b \varphi(x)\,\mathrm{d}x.$$

Thus, we have the following:

Theorem 1.4 (De Moivre–Laplace central limit theorem) For every $a < b$,

$$\mathbb{P}\left(a \le \frac{S_n}{\sqrt{n}} \le b\right) \longrightarrow \int_a^b \varphi(x)\,\mathrm{d}x \quad \text{as } n \to \infty.$$

Remark 1.5 The preceding calculation treats even values of n. The odd case requires only a parity adjustment and yields the same limit.

Weak Convergence of Distribution Functions

For a real random variable Y, the function $x \mapsto \mathbb{P}(Y \le x)$ records its distribution. For example,

$$F_n(x) = \mathbb{P}\left(\frac{S_n}{\sqrt{n}} \le x\right)$$

is the distribution function of $S_n/\sqrt{n}$. If

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\,\mathrm{d}t,$$

then Theorem 1.4 implies that $F_n(x) \to \Phi(x)$ for every x. For discontinuous limiting distribution functions, however, pointwise convergence at every point is too much to require.

Definition 1.6 A sequence of distribution functions (F_n) **converges weakly** to a distribution function F if $F_n(x) \to F(x)$ at every continuity point x of F.

For a simple illustration, let

$$F(x) = \mathbb{1}_{\{0 \le x\}}$$

be the distribution function of the random variable that is identically zero, and define

$$F_n(x) = F(x - 1/n) = \mathbb{1}_{\{1/n \le x\}}.$$

These distributions should converge to the point mass at zero. Nevertheless, $F_n(0) = 0$ for every n, whereas $F(0) = 1$. Thus pointwise convergence fails precisely at the jump of F, and Definition 1.6 excludes that point.

If μ_n and μ are probability measures on $\mathbb{R}$, with respective distribution functions F_n and F, then $\mu_n \Rightarrow \mu$ means that F_n converges weakly to F. Likewise,

$$X_n \xrightarrow{d} X$$

means that the law of X_n converges weakly to the law of X; this is called **convergence in distribution**.

Distribution of Maxima

Let $X_1, X_2, \ldots$ be independent and identically distributed exponential random variables with rate 1, and set

$$M_n = \max_{1 \le k \le n} X_k.$$

We seek centering constants a_n and scaling constants $b_n > 0$ for which $(M_n - a_n)/b_n$ has a nondegenerate limiting distribution. Independence gives

$$\mathbb{P}(M_n \le y) = \prod_{k=1}^{n} \mathbb{P}(X_k \le y) = (1 - e^{-y})^n \quad (y \ge 0).$$

Consequently,

$$\mathbb{P}\left(\frac{M_n - a_n}{b_n} \le x\right) = \left(1 - e^{-xb_n - a_n}\right)^n.$$

By Lemma 1.1, a nontrivial limit arises when $ne^{-xb_n - a_n}$ converges to a finite, nonzero function of x. Taking $a_n = \log n$ and $b_n = 1$ gives

$$\mathbb{P}(M_n - \log n \le x) = \left(1 - \frac{e^{-x}}{n}\right)^n \longrightarrow e^{-e^{-x}}.$$

The limiting distribution function $x \mapsto e^{-e^{-x}}$ is the **standard Gumbel distribution**. Thus the centred maximum $M_n - \log n$ converges in distribution to a standard Gumbel random variable.

Exercise 1.7 Show that convergence in probability implies convergence in distribution.

Solution. Suppose that $X_n \xrightarrow{\mathbb{P}} X$, and let x be a continuity point of the distribution function F_X. For every $\varepsilon > 0$,

$$\{X_n \le x\} \subseteq \{X \le x + \varepsilon\} \cup \{|X_n - X| > \varepsilon\},$$

and

$$\{X \le x - \varepsilon\} \subseteq \{X_n \le x\} \cup \{|X_n - X| > \varepsilon\}.$$

Consequently,

$$F_X(x - \varepsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x + \varepsilon).$$

Letting $\varepsilon \downarrow 0$ and using continuity at x gives $F_{X_n}(x) \to F_X(x)$. Hence X_n converges to X in distribution. $\square$

Remark 1.8 Convergence in probability requires X_n and X to be defined on a common probability space. Convergence in distribution depends only on their laws, so the random variables need not share a probability space.

Lecture 2 — Thursday, January 10, 2019

Suppose that X_n has density f_n and X has density f, so that

$$\mathbb{P}(X_n \in B) = \int_B f_n(x)\,\mathrm{d}x$$

for every Borel set $B \subseteq \mathbb{R}$.

Theorem 2.1 (Scheffé's theorem) If $f_n(x) \to f(x)$ for almost every x, then $X_n \xrightarrow{d} X$. In fact,

$$\sup_{B \in \mathcal{B}_{\mathbb{R}}} |\mathbb{P}(X_n \in B) - \mathbb{P}(X \in B)| \longrightarrow 0.$$

However, the converse is false.

Proof. For every Borel set B,

$$|\mathbb{P}(X_n \in B) - \mathbb{P}(X \in B)| = \left| \int_B (f_n(x) - f(x))\,\mathrm{d}x \right| \leq \int_{\mathbb{R}} |f_n(x) - f(x)|\,\mathrm{d}x.$$

Write $u_n = (f_n - f)^+$ and $v_n = (f_n - f)^- = (f - f_n)^+$. Since both f_n and f integrate to 1,

$$\int_{\mathbb{R}} u_n(x)\,\mathrm{d}x - \int_{\mathbb{R}} v_n(x)\,\mathrm{d}x = \int_{\mathbb{R}} (f_n(x) - f(x))\,\mathrm{d}x = 0.$$

Therefore,

$$\int_{\mathbb{R}} |f_n(x) - f(x)|\,\mathrm{d}x = 2\int_{\mathbb{R}} v_n(x)\,\mathrm{d}x.$$

We have $v_n(x) \to 0$ almost everywhere and $0 \leq v_n(x) \leq f(x)$. The dominated convergence theorem now shows that the final integral tends to 0, uniformly in the choice of B. Taking $B = (-\infty, x]$ proves convergence of the distribution functions at every x, and hence convergence in distribution. $\square$

Example 2.2 Let $X_1, \ldots, X_{2n+1}$ be independent random variables, each uniformly distributed on $[0, 1]$, and denote their order statistics by

$$X_{(1)} \leq \cdots \leq X_{(2n+1)}.$$

Determine the limiting distribution of the sample median $X_{(n+1)}$.

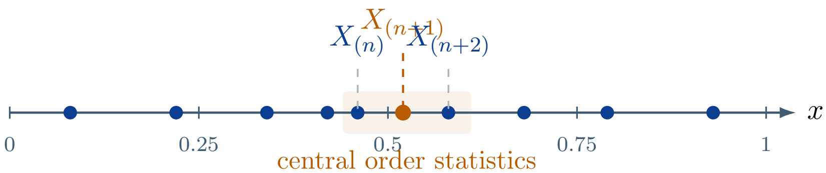

FIGURE 3

The central points are illustrated in Figure 3.

Solution. The density of $X_{(n+1)}$ is

$$f_{X_{(n+1)}}(x) = (2n+1)\binom{2n}{n} x^n (1-x)^n, \quad 0 < x < 1.$$

Indeed, one observation lies near x, while exactly n of the remaining $2n$ observations lie to its left and n lie to its right. Define

$$Y_n = \left(X_{(n+1)} - \frac{1}{2}\right)\sqrt{8n}.$$

The substitution $x = 1/2 + y/\sqrt{8n}$ gives the density

$$\begin{aligned} f_{Y_n}(y) &= (2n+1)\binom{2n}{n}\left(\frac{1}{2} + \frac{y}{\sqrt{8n}}\right)^n \left(\frac{1}{2} - \frac{y}{\sqrt{8n}}\right)^n \frac{1}{\sqrt{8n}} \\ &= \left(\binom{2n}{n}\frac{1}{2^{2n}}\right)\frac{2n+1}{\sqrt{8n}}\left(1 - \frac{y^2}{2n}\right)^n \\ &\longrightarrow \frac{1}{\sqrt{2\pi}} e^{-y^2/2}. \end{aligned}$$

Here $\binom{2n}{n}2^{-2n} \sim 1/\sqrt{\pi n}$, and Lemma 1.1 handles the final factor. The Scheffé theorem therefore yields

$$Y_n \xrightarrow{d} \mathcal{N}(0,1).$$

$\square$

Exercise 2.3 Let p_n and p be probability mass functions on $\mathbb{Z}$. Prove that if $p_n(k) \to p(k)$ for every $k \in \mathbb{Z}$, then the corresponding distribution functions converge weakly. This is the discrete analogue of Theorem 2.1.

Solution. Fix $\varepsilon > 0$. Choose M so large that

$$\sum_{|k|>M} p(k) < \varepsilon.$$

Pointwise convergence on the finite set $\{-M, \dots, M\}$ gives

$$\sum_{|k|\leq M} |p_n(k) - p(k)| \longrightarrow 0$$

and

$$\sum_{|k|>M} p_n(k) = 1 - \sum_{|k|\leq M} p_n(k) \longrightarrow 1 - \sum_{|k|\leq M} p(k) < \varepsilon.$$

It follows that

$$\sum_{k\in\mathbb{Z}} |p_n(k) - p(k)| \longrightarrow 0.$$

For every $x \in \mathbb{R}$,

$$|F_n(x) - F(x)| \leq \sum_{k\in\mathbb{Z}} |p_n(k) - p(k)|,$$

so the distribution functions converge uniformly, and therefore weakly. $\square$

Example 2.4 Fix $\lambda > 0$, and let $X_n \sim \mathrm{Bin}(n, \lambda/n)$. Show that $X_n \xrightarrow{d} X$, where $X \sim \mathrm{Poisson}(\lambda)$.

Solution. For each fixed nonnegative integer k,

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&\longrightarrow \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}$$

The limiting values form the probability mass function of a Poisson random variable with parameter λ. The discrete version of the Scheffé theorem therefore gives the claimed convergence in distribution. □

Lecture 3 — Tuesday, January 15, 2019

Skorohod Representation in $\mathbb{R}$

Recall from Definition 1.6 that $F_n \Rightarrow F$ means $F_n(x) \to F(x)$ at every continuity point of F. The Scheffé theorem gives one sufficient condition for this convergence. The next result provides a particularly useful coupling of weakly convergent distributions.

Theorem 3.1 (Skorohod representation theorem on $\mathbb{R}$) Suppose that $F_n \Rightarrow F$. Then there exist random variables X_n and X, defined on a common probability space, such that X_n has distribution function F_n, X has distribution function F, and $X_n \to X$ almost surely.

We will prove Theorem 3.1 later; for now, we will use it to derive some consequences of weak convergence. One application concerns convergence of moments. Write

$$\overline{m}_{k,n} = \int_{\mathbb{R}} |x|^k \, \mathrm{d}F_n(x)$$

for the kth absolute moment and

$$m_{k,n} = \int_{\mathbb{R}} x^k \, \mathrm{d}F_n(x)$$

for the kth moment.

Proposition 3.2 Let $k \geq 0$ be an integer and let $\ell > k$. Suppose that $F_n \Rightarrow F$ and that there is a constant b such that $\overline{m}_{\ell,n} \leq b$ for every n. Then

$$m_{k,n} \longrightarrow m_k = \int_{\mathbb{R}} x^k \, \mathrm{d}F(x).$$

Proof. By Theorem 3.1, we may realize $X_n \sim F_n$ and $X \sim F$ on a common probability space so that $X_n \to X$ almost surely. Fatou's lemma gives

$$\mathbb{E}\left[|X|^\ell\right] \le \liminf_{n\to\infty} \mathbb{E}\left[|X_n|^\ell\right] \le b.$$

Moreover, for every $R > 0$,

$$\mathbb{E}\left[|X_n|^k \mathbb{1}_{\{|X_n|>R\}}\right] \le \frac{1}{R^{\ell-k}}\mathbb{E}\left[|X_n|^\ell\right] \le \frac{b}{R^{\ell-k}}.$$

Thus (X_n^k) is uniformly integrable. Since $X_n^k \to X^k$ almost surely, their expectations converge, and hence $m_{k,n} \to m_k$. □

Definition 3.3 For a distribution function F, define its **generalized inverse** by

$$F^{-1}(u) = \sup\{y \in \mathbb{R} : F(y) < u\}, \quad 0 < u < 1.$$

The geometry of this definition is shown in Figure 4.

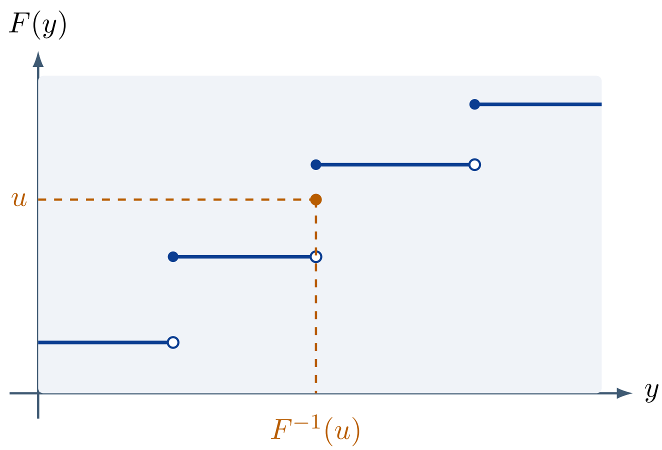

FIGURE 4

Lemma 3.4 (Inverse transform sampling) If $U \sim \text{Unif}[0,1]$, then $F^{-1}(U)$ has distribution function F.

Proof. The definition of the generalized inverse gives

$$F^{-1}(U) \le y \qquad \text{if and only if} \qquad U \le F(y).$$

Therefore,

$$\mathbb{P}(F^{-1}(U) \le y) = \mathbb{P}(U \le F(y)) = F(y).$$

□

For example, if

$$F(y) = \begin{cases} 1 - e^{-y}, & y \ge 0, \\ 0, & y < 0, \end{cases}$$

then $F^{-1}(u) = -\log(1-u)$. Hence $-\log(1-U)$, and equivalently $-\log(U)$, has an exponential distribution with rate 1.

Definition 3.5 We call $u \in (0,1)$ a **plateau value** of F if $F(a) = F(b) = u$ for some $a < b$.

Lemma 3.6 If $F_n \Rightarrow F$ and u is not a plateau value of F, then

$$F_n^{-1}(u) \longrightarrow F^{-1}(u).$$

Proof. Set $q = F^{-1}(u)$, and fix $\varepsilon > 0$. Choose continuity points a, b of F such that

$$q - \varepsilon < a < q < b < q + \varepsilon.$$

Because u is not a plateau value, these points may be chosen so that $F(a) < u < F(b)$. Weak convergence gives $F_n(a) < u < F_n(b)$ for all sufficiently large n. By the definition of the generalized inverse,

$$a < F_n^{-1}(u) \le b.$$

Thus $\left|F_n^{-1}(u) - q\right| < \varepsilon$ for all sufficiently large n. □

Proof of Theorem 3.1. Distinct plateau values of F correspond to disjoint nonempty open intervals, each of which contains a rational number. Hence F has at most countably many plateau values. Let U be uniformly distributed on $[0,1]$, and define

$$X_n = F_n^{-1}(U) \qquad \text{and} \qquad X = F^{-1}(U).$$

By Lemma 3.4, these random variables have the required distribution functions. With probability 1, the value of U is not among the countably many plateau values of F. On that event, Lemma 3.6 gives $X_n \to X$. Thus the convergence is almost sure. □

The phrases **weak convergence**, **convergence in distribution**, and **convergence in law** are synonymous on $\mathbb{R}$. This probabilistic notion agrees with the following analytic criterion:

Proposition 3.7 $F_n \Rightarrow F$ if and only if, for every bounded continuous function $g : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}} g(x)\,\mathrm{d}F_n(x) \longrightarrow \int_{\mathbb{R}} g(x)\,\mathrm{d}F(x).$$

Proof. *Forward implication.* By Theorem 3.1, choose $X_n \sim F_n$ and $X \sim F$ on a common probability space such that $X_n \to X$ almost surely. Continuity gives $g(X_n) \to g(X)$ almost surely, and boundedness permits an application of the bounded convergence theorem. Therefore,

$$\int g\,\mathrm{d}F_n = \mathbb{E}[g(X_n)] \longrightarrow \mathbb{E}[g(X)] = \int g\,\mathrm{d}F.$$

Reverse implication. Assume convergence of the integrals for every bounded continuous g. Fix a continuity point a of F and $\varepsilon > 0$. Define

$$g_{a,\varepsilon}(x) = \begin{cases} 1, & x \le a, \\ 1 - (x-a)/\varepsilon, & a < x < a+\varepsilon, \\ 0, & x \ge a+\varepsilon. \end{cases}$$

This continuous approximation to the indicator of $(-\infty, a]$ is illustrated in Figure 5.

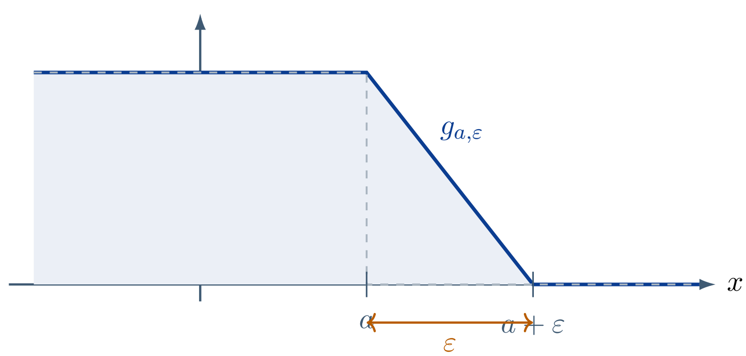

FIGURE 5

Writing $g_c = \mathbb{1}_{(-\infty,c]}$, we have

$$g_a \le g_{a,\varepsilon} \le g_{a+\varepsilon}.$$

The corresponding integral inequalities are summarized by the following diagram:

$$
\begin{array}{ccc}
\displaystyle\int g_{a+\varepsilon}\,\mathrm{d}F_n = F_n(a+\varepsilon) & & \displaystyle\int g_{a+\varepsilon}\,\mathrm{d}F = F(a+\varepsilon) \\
\geq & & \geq \\
\displaystyle\int g_{a,\varepsilon}\,\mathrm{d}F_n & \longrightarrow & \displaystyle\int g_{a,\varepsilon}\,\mathrm{d}F \\
\geq & & \geq \\
\displaystyle\int g_{a}\,\mathrm{d}F_n = F_n(a) & & \displaystyle\int g_{a}\,\mathrm{d}F = F(a)
\end{array}
$$

By hypothesis, the middle integral in the left column converges to the middle integral in the right column. Hence

$$\limsup_{n\to\infty} F_n(a) \leq \int g_{a,\varepsilon}\,\mathrm{d}F \leq F(a+\varepsilon).$$

Letting $\varepsilon \downarrow 0$ and using continuity at a gives $\limsup_n F_n(a) \leq F(a)$. Applying the same argument to $g_{a-\varepsilon,\varepsilon}$ yields

$$\liminf_{n\to\infty} F_n(a) \geq \int g_{a-\varepsilon,\varepsilon}\,\mathrm{d}F \geq F(a-\varepsilon).$$

Again sending $\varepsilon \downarrow 0$ gives $\liminf_n F_n(a) \geq F(a)$. Therefore $F_n(a) \to F(a)$, as required. $\square$

Notation 3.8 If S is a topological space equipped with its Borel σ-algebra and μ_n, μ are probability measures on S, then $\mu_n \Rightarrow \mu$ means that

$$\int_S g\,\mathrm{d}\mu_n \longrightarrow \int_S g\,\mathrm{d}\mu$$

for every bounded continuous function $g : S \to \mathbb{R}$.

Proposition 3.7 allows for a simpler proof of Exercise 1.7:

Lemma 3.9 If $X_n \xrightarrow{P} X$, then $X_n \Rightarrow X$.

Proof. Every subsequence of (X_n) has a further subsequence that converges almost surely to X. Along such a further subsequence, $g(X_{n_k}) \to g(X)$ almost surely for every bounded continuous g, and the bounded convergence theorem gives

$$\mathbb{E}[g(X_{n_k})] \longrightarrow \mathbb{E}[g(X)].$$

Thus every subsequence of the real sequence $(\mathbb{E}[g(X_n)])$ has a further subsequence converging to $\mathbb{E}[g(X)]$. The entire sequence must therefore have this limit. The conclusion now follows from Proposition 3.7. $\square$

Characteristic Functions

Definition 3.10 A measurable function $Z : \Omega \to \mathbb{C}$ is a **complex random variable**. When its real and imaginary parts are integrable, its expectation is defined by

$$\mathbb{E}[Z] = \mathbb{E}[\operatorname{Re} Z] + i\mathbb{E}[\operatorname{Im} Z].$$

Definition 3.11 For a real random variable X, its **characteristic function** is the function $\varphi_X : \mathbb{R} \to \mathbb{C}$ defined by

$$\varphi_X(\lambda) = \mathbb{E}\left[e^{i\lambda X}\right] = \mathbb{E}[\cos(\lambda X)] + i\mathbb{E}[\sin(\lambda X)].$$

This expectation always exists because $\left|e^{i\lambda X}\right| = 1$. If F is the distribution function of X, then equivalently

$$\varphi_F(\lambda) = \int_{\mathbb{R}} e^{i\lambda x} \, \mathrm{d}F(x).$$

The Lévy continuity theorem, which we will prove later, states that $X_n \Rightarrow X$ if and only if $\varphi_{X_n}(\lambda) \to \varphi_X(\lambda)$ for every $\lambda \in \mathbb{R}$. This is one reason characteristic functions are so effective in limit theorems.

Proposition 3.12 Let $\varphi = \varphi_X$. Then the following properties hold:

1. For every $\lambda \in \mathbb{R}$, $|\varphi(\lambda)| \leq 1$. Indeed,

$$|\varphi(\lambda)| = \left|\mathbb{E}[e^{i\lambda X}]\right| \leq \mathbb{E}[\left|e^{i\lambda X}\right|] = 1.$$

2. The function φ is continuous. To see this, suppose that $\lambda_n \to \lambda$. Then $e^{i\lambda_n X} \to e^{i\lambda X}$

almost surely, and the bounded convergence theorem gives $\varphi(\lambda_n) \to \varphi(\lambda)$.

3. If $X \stackrel{d}{=} -X$, so that X is symmetric about 0, then φ is real-valued. In fact,

$$\overline{\varphi(\lambda)} = \mathbb{E}[e^{-i\lambda X}] = \mathbb{E}[e^{i\lambda X}] = \varphi(\lambda).$$

The converse is also true, although its proof requires the inversion theory developed below.

4. For every $\lambda \in \mathbb{R}$,

$$\overline{\varphi(\lambda)} = \mathbb{E}[e^{-i\lambda X}] = \varphi(-\lambda).$$

5. If X and Y are independent, then $\varphi_{X+Y} = \varphi_X \varphi_Y$. Indeed,

$$\varphi_{X+Y}(\lambda) = \mathbb{E}[e^{i\lambda X} e^{i\lambda Y}] = \mathbb{E}[e^{i\lambda X}]\mathbb{E}[e^{i\lambda Y}] = \varphi_X(\lambda)\varphi_Y(\lambda).$$

6. For $a, b \in \mathbb{R}$,

$$\varphi_{aX+b}(\lambda) = \mathbb{E}[e^{i\lambda(aX+b)}] = e^{i\lambda b}\varphi_X(a\lambda).$$

Example 3.13 Let X take the values 1 and -1, each with probability $1/2$. If $X_1, X_2, \ldots$ are independent copies of X and

$$S_n = \sum_{i=1}^{n} X_i,$$

compute the characteristic functions of X, S_n, and $S_n/\sqrt{n}$.

Solution. Directly from the definition,

$$\varphi_X(\lambda) = \frac{e^{i\lambda} + e^{-i\lambda}}{2} = \cos\lambda.$$

Independence and part (5) of Proposition 3.12 give

$$\varphi_{S_n}(\lambda) = \cos^n \lambda \qquad \text{and} \qquad \varphi_{S_n/\sqrt{n}}(\lambda) = \cos^n\left(\frac{\lambda}{\sqrt{n}}\right).$$

The expansion $\cos u = 1 - u^2/2 + o(u^2)$ shows that

$$\cos^n\left(\frac{\lambda}{\sqrt{n}}\right) \longrightarrow e^{-\lambda^2/2}.$$

This is the characteristic function version of the simple random walk central limit theorem. □

Lecture 4 — Thursday, January 17, 2019

Characteristic functions are especially useful for sums of independent random variables: a convolution of distributions becomes an ordinary product of functions by part Proposition 3.12(5).

Exercise 4.1 If $X \sim \mathcal{N}(0,1)$, then

$$\mathbb{P}(X \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x.$$

Use integration by parts to prove that $\varphi_X(\lambda) = e^{-\lambda^2/2}$.

Solution. Let $f(x) = (2\pi)^{-1/2} e^{-x^2/2}$. Since $f'(x) = -xf(x)$, differentiation under the integral sign and integration by parts give

$$\begin{aligned}
\varphi_X'(\lambda) &= i \int_{\mathbb{R}} x e^{i\lambda x} f(x)\,\mathrm{d}x \\
&= -i \int_{\mathbb{R}} e^{i\lambda x} f'(x)\,\mathrm{d}x \\
&= -i \left(\Big[e^{i\lambda x} f(x) \Big]_{-\infty}^{\infty} - i\lambda \int_{\mathbb{R}} e^{i\lambda x} f(x)\,\mathrm{d}x \right) \\
&= -\lambda \varphi_X(\lambda).
\end{aligned}$$

The boundary term vanishes because $f(x) \to 0$ exponentially fast. Together with $\varphi_X(0) = 1$, this differential equation has the unique solution $\varphi_X(\lambda) = e^{-\lambda^2/2}$. □

Example 4.2 If X is uniformly distributed on $[0,1]$, then

$$\varphi_X(\lambda) = \int_0^1 e^{i\lambda x}\,\mathrm{d}x = \frac{e^{i\lambda} - 1}{i\lambda}.$$

If Y is uniformly distributed on $[-1,1]$, then $Y \stackrel{d}{=} 2X-1$, and hence

$$\varphi_Y(\lambda) = e^{-i\lambda}\varphi_X(2\lambda) = \frac{e^{-i\lambda}(e^{2i\lambda}-1)}{2i\lambda} = \frac{\sin\lambda}{\lambda}.$$

The value at $\lambda = 0$ is defined by continuity and equals 1. Although Y is bounded, its characteristic function is not absolutely integrable; the significance of this distinction will appear in the inversion theorem.

Example 4.3 If X has an exponential distribution with rate 1, then

$$\varphi_X(\lambda) = \int_0^\infty e^{i\lambda x}e^{-x}\,\mathrm{d}x = \int_0^\infty e^{(i\lambda-1)x}\,\mathrm{d}x = \frac{1}{1-i\lambda}.$$

The boundary term at infinity vanishes because $e^{i\lambda x}$ has modulus 1, whereas $e^{-x}\to 0$.

Example 4.4 If X has a Poisson distribution with parameter $\nu > 0$, then

$$\begin{aligned}\varphi_X(\lambda) &= \sum_{k=0}^\infty e^{i\lambda k}e^{-\nu}\frac{\nu^k}{k!}\\ &= e^{-\nu}\sum_{k=0}^\infty \frac{(\nu e^{i\lambda})^k}{k!}\\ &= \exp\bigl(\nu(e^{i\lambda}-1)\bigr).\end{aligned}$$

Example 4.5 If X has a binomial distribution with parameters n and p, write

$$X = \sum_{i=1}^n X_i,$$

where the X_i are independent Bernoulli random variables with parameter p. Then

$$\varphi_{X_i}(\lambda) = pe^{i\lambda} + 1 - p = 1 + p(e^{i\lambda}-1),$$

and therefore

$$\varphi_X(\lambda) = \Bigl(1 + p(e^{i\lambda}-1)\Bigr)^n.$$

Remark 4.6 If X has density f, then φ_X is the Fourier transform of f:

$$\varphi_X(\lambda) = \int_{\mathbb{R}} e^{i\lambda x} f(x)\,\mathrm{d}x.$$

Formally, we expect the inverse relation

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-i\lambda x}\varphi(\lambda)\,\mathrm{d}\lambda.$$

Even when the probability measure μ has no density, we can still seek a formula for $\mu((a,b))$ in terms of its characteristic function. The elementary calculation

$$\int_a^b e^{-i\lambda x}\,\mathrm{d}x = -\frac{e^{-i\lambda x}}{i\lambda}\bigg|_a^b = \frac{e^{-i\lambda a} - e^{-i\lambda b}}{i\lambda}$$

suggests the appropriate inversion kernel.

Lecture 5 — Tuesday, January 22, 2019

Theorem 5.1 (Lévy inversion formula) Let μ be a probability measure on $\mathbb{R}$ with characteristic function φ, and let $a < b$. Then

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-i\lambda a} - e^{-i\lambda b}}{i\lambda}\varphi(\lambda)\,\mathrm{d}\lambda = \mu((a,b)) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}).$$

In particular, the characteristic function determines the probability of every interval, and hence determines the probability measure uniquely.

Proof. Define the **sine integral** by

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}\,\mathrm{d}t.$$

The Dirichlet integral famously gives $\mathrm{Si}(x) \to \pi/2$ as $x \to \infty$, and Si is bounded on $\mathbb{R}$. For fixed T, Fubini's theorem applies because

$$\left|\frac{e^{-i\lambda a} - e^{-i\lambda b}}{i\lambda}\right| = \left|\int_a^b e^{-i\lambda y}\,\mathrm{d}y\right| \le b - a.$$

Using $\varphi(\lambda) = \int_{\mathbb{R}} e^{i\lambda x}\,\mathrm{d}\mu(x)$, we obtain

$$\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-i\lambda a}-e^{-i\lambda b}}{i\lambda}\varphi(\lambda)\,\mathrm{d}\lambda = \int_{\mathbb{R}} K_T(x)\,\mathrm{d}\mu(x),$$

where

$$K_T(x) = \frac{1}{2\pi}\int_{-T}^{T}\frac{e^{i\lambda(x-a)}-e^{i\lambda(x-b)}}{i\lambda}\,\mathrm{d}\lambda.$$

For $\theta \in \mathbb{R}$, set

$$R_T(\theta) = \int_{-T}^{T}\frac{e^{i\lambda\theta}}{i\lambda}\,\mathrm{d}\lambda.$$

The cosine term is odd in λ and integrates to zero, while the sine term is even. For $\theta > 0$, the change of variables $u = \theta\lambda$ gives

$$R_T(\theta) = \int_{-T}^{T}\frac{\sin(\theta\lambda)}{\lambda}\,\mathrm{d}\lambda = 2\int_{0}^{\theta T}\frac{\sin u}{u}\,\mathrm{d}u = 2\,\mathrm{Si}(\theta T).$$

By oddness, the formula for every θ is

$$R_T(\theta) = 2\,\mathrm{Si}(|\theta|T)\,\mathrm{sgn}(\theta).$$

It follows that K_T is bounded uniformly in both T and x. Moreover,

$$K_T(x) \longrightarrow \begin{cases} 1, & x \in (a,b), \\ 1/2, & x \in \{a,b\}, \\ 0, & x \notin [a,b]. \end{cases}$$

Indeed, outside $[a,b]$ the two sine integral limits cancel; inside (a,b) they have opposite signs and add to π; at an endpoint exactly one term contributes $\pi/2$. The bounded convergence theorem therefore gives

$$\lim_{T\to\infty}\int_{\mathbb{R}} K_T(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}}\lim_{T\to\infty} K_T(x)\,\mathrm{d}\mu(x) = \mu((a,b)) + \frac{1}{2}\mu(\{a\}) + \frac{1}{2}\mu(\{b\}).$$

□

Let us now turn to the precise relationship between characteristic functions and weak convergence:

Theorem 5.2 (Lévy continuity theorem)

1. If $\mu_n \Rightarrow \mu$, then $\varphi_{\mu_n}(t) \to \varphi_\mu(t)$ for every $t \in \mathbb{R}$.
2. Suppose that $\varphi_n(t) \to \varphi(t)$ for every $t \in \mathbb{R}$, where φ_n is the characteristic function of μ_n and φ is continuous at 0. Then there exists a probability measure μ such that $\varphi = \varphi_\mu$ and $\mu_n \Rightarrow \mu$.

Part (1) follows immediately from Proposition 3.7, applied separately to $\cos(tx)$ and $\sin(tx)$. The reverse direction requires a compactness argument.

Theorem 5.3 (Helly's selection theorem) If (F_n) is a sequence of distribution functions, then there exist a subsequence (F_{n_k}) and a right-continuous, nondecreasing function $F : \mathbb{R} \to [0, 1]$ such that

$$F_{n_k}(x) \longrightarrow F(x)$$

at every continuity point x of F.

The limiting function in Theorem 5.3 need not be a distribution function, because it may fail to have the correct limits at infinity. For example, the distribution functions

$$F_n(x) = \mathbb{1}_{\{n \le x\}} \qquad \text{and} \qquad G_n(x) = \mathbb{1}_{\{-n \le x\}}$$

converge pointwise to the constant functions 0 and 1, respectively.

Remark 5.4 The right-continuous, nondecreasing functions from $\mathbb{R}$ to $[0, 1]$ can be viewed as distribution functions on the extended real line $\mathbb{R} \cup \{-\infty, +\infty\}$. The two examples above place all mass at $+\infty$ and $-\infty$, respectively.

Proof of Theorem 5.3. Enumerate the rational numbers as $q_1, q_2, \ldots$. Since every sequence in $[0, 1]$ has a convergent subsequence, first choose a subsequence along which $F_n(q_1)$ converges. From it choose a further subsequence along which $F_n(q_2)$ converges, and continue inductively. The diagonal subsequence, still denoted (F_{n_k}), converges at every rational point. Define $G : \mathbb{Q} \to [0, 1]$ by

$$G(q) = \lim_{k\to\infty} F_{n_k}(q).$$

The function G is nondecreasing: if $q < r$, then

$$G(q) = \lim_{k\to\infty} F_{n_k}(q) \leq \lim_{k\to\infty} F_{n_k}(r) = G(r).$$

For $x \in \mathbb{R}$, set

$$F(x) = \inf\{G(q) : q \in \mathbb{Q},\ q > x\}.$$

This function is nondecreasing. It is also right-continuous. Indeed, if $x_m \downarrow x$, then $F(x_m) \geq F(x)$. Conversely, given $\varepsilon > 0$, choose a rational $q > x$ with $G(q) < F(x) + \varepsilon$. For all sufficiently large m, $x_m < q$, and hence

$$F(x_m) \leq G(q) < F(x) + \varepsilon.$$

Thus $F(x_m) \to F(x)$.

It remains to prove convergence at every continuity point x of F. Fix $\varepsilon > 0$. By continuity, choose rational numbers $q_- < x < q_+$ so that

$$F(x) - \varepsilon < G(q_-) \leq G(q_+) < F(x) + \varepsilon.$$

Monotonicity gives

$$F_{n_k}(q_-) \leq F_{n_k}(x) \leq F_{n_k}(q_+).$$

Passing to lower and upper limits yields

$$F(x) - \varepsilon \leq \liminf_{k\to\infty} F_{n_k}(x) \leq \limsup_{k\to\infty} F_{n_k}(x) \leq F(x) + \varepsilon.$$

Letting $\varepsilon \downarrow 0$ proves $F_{n_k}(x) \to F(x)$. $\square$

Lecture 6 — Thursday, January 24, 2019

Definition 6.1 A sequence (μ_n) of probability measures on $\mathbb{R}$ is **tight** if for every $\varepsilon > 0$, there exists $K > 0$ such that

$$\mu_n([-K, K]^c) < \varepsilon$$

for every n. (If μ_n has distribution function F_n, then we also say that (F_n) is tight.)

Equivalently, it is enough that this inequality hold for all sufficiently large n, because the finitely many remaining probability measures can be absorbed by increasing K.

Tightness prevents probability mass from escaping to infinity. It identifies exactly when the limit furnished by Theorem 5.3 is a genuine distribution function.

Lemma 6.2 Suppose that $F_n(x) \to F(x)$ at every continuity point of a right-continuous, nondecreasing function $F : \mathbb{R} \to [0, 1]$. Then F is a distribution function if and only if (F_n) is tight.

Proof. *Forward implication.* Suppose that F is a distribution function. Fix $\varepsilon > 0$, and choose $K > 0$ such that $-K$ and K are continuity points of F and

$$F(-K) + 1 - F(K) < \varepsilon.$$

For all sufficiently large n, convergence at $-K$ and K gives

$$F_n(-K) + 1 - F_n(K) < 3\varepsilon.$$

After increasing K to handle the finitely many remaining indices, the sequence is tight.

Reverse implication. Suppose that (F_n) is tight. Fix $\varepsilon > 0$, and choose $K > 0$, with $-K$ and K continuity points of F, such that

$$F_n(-K) + 1 - F_n(K) < \varepsilon$$

for every n. To justify this choice, let μ_n be the probability measure with distribution function F_n. First use tightness to find $K_0 > 0$ such that

$$\mu_n([-K_0, K_0]^c) < \varepsilon$$

for every n, and then choose a continuity point $K > K_0$ for which $-K$ is also a continuity point. Since

$$F_n(-K) + 1 - F_n(K) \le \mu_n([-K_0, K_0]^c),$$

the required inequality follows. Passing to the limit gives

$$F(-K) + 1 - F(K) \le \varepsilon.$$

As $\varepsilon \downarrow 0$, this forces

$$\lim_{x \to -\infty} F(x) = 0 \qquad \text{and} \qquad \lim_{x \to +\infty} F(x) = 1.$$

Together with right-continuity and monotonicity, these are precisely the remaining requirements for F to be a distribution function. □

Theorem 6.3 (Prokhorov's theorem on $\mathbb{R}$) Every tight sequence of probability measures on $\mathbb{R}$ has a weakly convergent subsequence.

Proof. Apply Theorem 5.3 and then Lemma 6.2. □

Lemma 6.4 Let μ be a probability measure with characteristic function φ. For every $u > 0$,

$$\mu\left(\left[-\frac{2}{u}, \frac{2}{u}\right]^c\right) \leq \frac{1}{u}\int_{-u}^{u} (1 - \varphi(\lambda))\, \mathrm{d}\lambda.$$

The integral on the right is real and nonnegative.

Proof.

$$\begin{aligned}\int_{-u}^{u} (1 - \varphi(\lambda))\, \mathrm{d}\lambda &= \int_{\mathbb{R}}\int_{-u}^{u} (1 - e^{i\lambda x})\, \mathrm{d}\lambda\, \mathrm{d}\mu(x) \\ &= 2\int_{\mathbb{R}} \left(u - \frac{\sin(ux)}{x}\right) \mathrm{d}\mu(x) \\ &= 2u\int_{\mathbb{R}} \left(1 - \frac{\sin(ux)}{ux}\right) \mathrm{d}\mu(x).\end{aligned}$$

If $|x| \geq 2/u$, then

$$\left|\frac{\sin(ux)}{ux}\right| \leq \frac{1}{2} \qquad \text{and} \qquad 1 - \frac{\sin(ux)}{ux} \geq \frac{1}{2}.$$

Consequently,

$$\int_{-u}^{u} (1 - \varphi(\lambda))\, \mathrm{d}\lambda \geq 2u\int_{|x|\geq 2/u} \frac{1}{2}\, \mathrm{d}\mu(x) = u\,\mu\left(\left[-\frac{2}{u}, \frac{2}{u}\right]^c\right),$$

which proves the claim. □

Proof of part (2) of Theorem 5.2. By Lemma 6.4,

$$\mu_n\left(\left[-\frac{2}{u}, \frac{2}{u}\right]^c\right) \leq \frac{1}{u}\int_{-u}^{u} (1 - \varphi_n(\lambda))\, \mathrm{d}\lambda.$$

For fixed u, the bounded convergence theorem gives

$$\frac{1}{u}\int_{-u}^{u}(1-\varphi_n(\lambda))\,\mathrm{d}\lambda \longrightarrow \frac{1}{u}\int_{-u}^{u}(1-\varphi(\lambda))\,\mathrm{d}\lambda.$$

Because $\varphi(0)=1$ and φ is continuous at 0, the final average tends to 0 as $u \downarrow 0$. Hence (μ_n) is tight. By Theorem 6.3, every subsequence has a further subsequence converging weakly to some probability measure μ. Part (1) then shows that the characteristic function of any such subsequential limit is φ. The inversion formula makes this limit unique. Therefore the full sequence converges weakly to μ, and $\varphi = \varphi_\mu$. □

Lecture 7 — Tuesday, January 29, 2019

Remark 7.1 The tail bound in Lemma 6.4,

$$\mu\left(\left\{x : |x| \geq \frac{2}{u}\right\}\right) \leq \frac{1}{u}\int_{-u}^{u}(1-\varphi(\lambda))\,\mathrm{d}\lambda,$$

shows directly how the tails of μ are controlled by the behaviour of φ near 0.

Theorem 7.2 If φ is integrable on $\mathbb{R}$, then μ has the continuous density

$$f(y) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-iyt}\varphi(t)\,\mathrm{d}t.$$

Proof. By the inversion formula and the bound

$$\left|\frac{e^{-iat}-e^{-ibt}}{it}\right| = \left|\int_a^b e^{-iyt}\,\mathrm{d}y\right| \leq b-a,$$

the dominated convergence theorem gives

$$\mu((a,b)) + \frac{\mu(\{a\})+\mu(\{b\})}{2} = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-iat}-e^{-ibt}}{it}\varphi(t)\,\mathrm{d}t,$$
$$\left|\mu((a,b)) + \frac{\mu(\{a\})+\mu(\{b\})}{2}\right| \leq \frac{b-a}{2\pi}\int_{\mathbb{R}}|\varphi(t)|\,\mathrm{d}t.$$

Letting $b \downarrow a$ shows that $\mu(\{a\}) = 0$, so μ has no atoms. Therefore, for $a < b$,

$$\mu((a,b)) = \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_a^b e^{-iyt}\,\mathrm{d}y\right)\varphi(t)\,\mathrm{d}t = \int_a^b\left(\frac{1}{2\pi}\int_{\mathbb{R}} e^{-iyt}\varphi(t)\,\mathrm{d}t\right)\mathrm{d}y = \int_a^b f(y)\,\mathrm{d}y,$$

where Fubini's theorem is justified by the integrability of φ. The function f is continuous by the dominated convergence theorem, so it is the desired density. □

The Taylor series

$$e^{ix} = \sum_{j=0}^{\infty}\frac{(ix)^j}{j!}$$

has a remainder satisfying, for real x,

$$\left|e^{ix} - \sum_{j=0}^{m}\frac{(ix)^j}{j!}\right| \le \min\left\{\frac{|x|^{m+1}}{(m+1)!}, \frac{2|x|^m}{m!}\right\}.$$

Substituting $x = \lambda X$, taking expectations, and absorbing constants depending only on m into c_m, we obtain

$$\left|\varphi(\lambda) - \sum_{j=0}^{m}\frac{(i\lambda)^j}{j!}\mathbb{E}[X^j]\right| \le c_m|\lambda|^m\mathbb{E}\left[\min\{|\lambda||X|^{m+1}, |X|^m\}\right].$$

If $\mathbb{E}[|X|^m] < \infty$, the expectation on the right tends to 0 by the dominated convergence theorem, since the integrand converges pointwise to 0 and is bounded by $|X|^m$.

Proposition 7.3 If $\mathbb{E}[|X|^m] < \infty$, then

$$\varphi(\lambda) = \sum_{j=0}^{m}\frac{(i\lambda)^j}{j!}\mathbb{E}[X^j] + o(|\lambda|^m)$$

as $\lambda \to 0$.

In particular, if $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}(X) < \infty$, then

$$\varphi(\lambda) = 1 + i\mu\lambda - \frac{\mu^2+\sigma^2}{2}\lambda^2 + o(\lambda^2).$$

Theorem 7.4 (Lindeberg–Lévy central limit theorem) Let $X_1, X_2, \ldots$ be independent and identically distributed, with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] = \sigma^2 < \infty$. Then

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Proof. By Proposition 7.3,

$$\varphi_{X_1}(\lambda) = 1 - \frac{\sigma^2\lambda^2}{2} + o(\lambda^2).$$

If $Y_n = (X_1 + \cdots + X_n)/\sqrt{n}$, then independence gives

$$\varphi_{Y_n}(\lambda) = \varphi_{X_1}^n\left(\frac{\lambda}{\sqrt{n}}\right) = \left(1 - \frac{\sigma^2\lambda^2}{2n} + o(1/n)\right)^n \longrightarrow e^{-\sigma^2\lambda^2/2}.$$

The limiting function is the characteristic function of $\mathcal{N}(0, \sigma^2)$, so Theorem 5.2 proves the claim. □

Lemma 7.5 If f is twice differentiable at x, then

$$f''(x) = \lim_{h\to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

Proof. Taylor expansion in the two directions gives

$$f(x \pm h) = f(x) \pm f'(x)h + \frac{f''(x)}{2}h^2 + o(h^2).$$

Adding the two expansions cancels the linear terms and proves the formula. □

Theorem 7.6 If

$$\limsup_{h\to 0} \frac{\varphi(-h) - 2\varphi(0) + \varphi(h)}{h^2} > -\infty,$$

then $\mathbb{E}[X^2] < \infty$. The quotient is always nonpositive, and the hypothesis is slightly weaker than the existence of $\varphi''(0)$. Moreover,

$$\lim_{h\to 0} \frac{2 - \varphi(-h) - \varphi(h)}{h^2} = \mathbb{E}[X^2].$$

Proof.

$$\frac{\varphi(-h) - 2\varphi(0) + \varphi(h)}{h^2} = -\mathbb{E}\left[\frac{2 - 2\cos(hX)}{h^2}\right].$$

Since the integrand is nonnegative, Fatou's lemma and Lemma 7.5 imply

$$\begin{aligned} -\limsup_{h\to 0} \frac{\varphi(-h) - 2\varphi(0) + \varphi(h)}{h^2} &= \liminf_{h\to 0} \mathbb{E}\left[\frac{2 - 2\cos(hX)}{h^2}\right] \\ &\geq \mathbb{E}\left[\liminf_{h\to 0} \frac{2 - 2\cos(hX)}{h^2}\right] \\ &= \mathbb{E}[X^2]. \end{aligned}$$

The left-hand side is finite by hypothesis, so the second moment is finite. Now

$$0 \leq \frac{2 - 2\cos(hX)}{h^2} \leq X^2$$

and the integrand converges almost surely to X^2. The dominated convergence theorem therefore gives the asserted limit. $\square$

The Pólya criterion, proved below, characterizes a useful family of characteristic functions: a continuous, even function that equals 1 at the origin, is convex and nonnegative on $[0, \infty)$, and tends to 0 at infinity is a characteristic function. The triangular examples in Figures 6 and 7 motivate the proof.

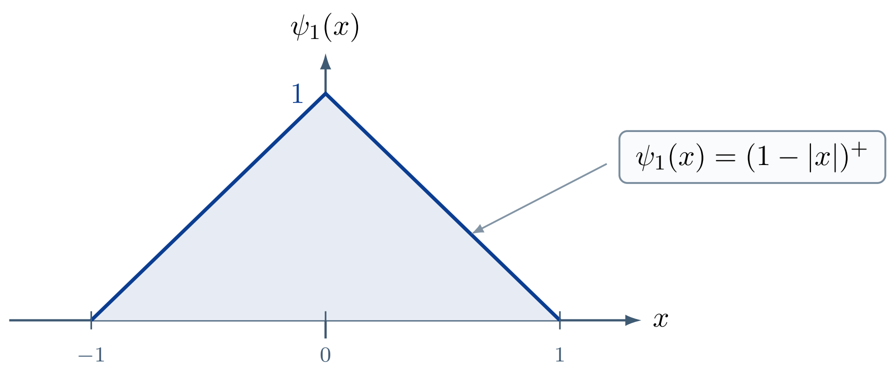

FIGURE 6

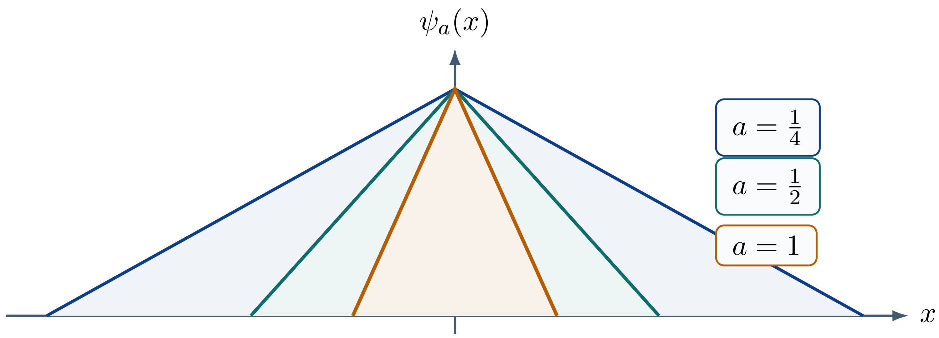

FIGURE 7

The functions in Figure 7 have the form $\psi_a(x) = (1-|ax|)^+$. Each is a characteristic function, and the proof of the Pólya criterion represents a general ψ as a mixture of these triangular functions.

Example 7.7 Let Y and Y' be independent random variables uniformly distributed on $[-1/2, 1/2]$. Find the density and characteristic function of $Y+Y'$, and use Fourier inversion to construct a random variable whose characteristic function is $(1-|t|)^+$.

Solution. More generally, if Y is uniformly distributed on $[a, b]$, then

$$\varphi_Y(\lambda) = \frac{1}{b-a}\int_a^b e^{i\lambda y}\,\mathrm{d}y = \frac{e^{ib\lambda} - e^{ia\lambda}}{(b-a)i\lambda}.$$

For $Y, Y' \sim \mathrm{Unif}[-1/2, 1/2]$, convolution gives the triangular density

$$t(y) = (1-|y|)^+,$$

shown in Figure 8.

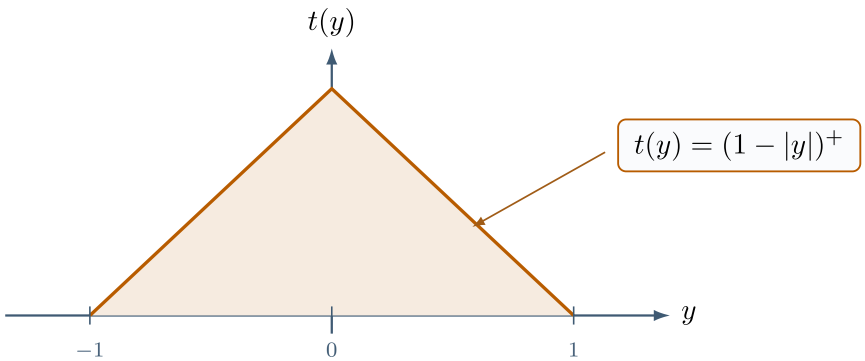

Figure 8

By independence,

$$\varphi_{Y+Y'}(\lambda) = \left(\frac{e^{i\lambda/2} - e^{-i\lambda/2}}{i\lambda}\right)^2 = \frac{2 - 2\cos\lambda}{\lambda^2}.$$

This function is integrable. The Fourier inversion formula therefore gives

$$t(y) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{2 - 2\cos\lambda}{\lambda^2} e^{-iy\lambda}\,\mathrm{d}\lambda.$$

At $y = 0$, this identity becomes

$$t(0) = \frac{1}{2\pi}\int_{\mathbb{R}} \frac{2 - 2\cos\lambda}{\lambda^2}\,\mathrm{d}\lambda = 1.$$

It follows that

$$q(\lambda) = \frac{1}{2\pi}\frac{2 - 2\cos\lambda}{\lambda^2}$$

is itself a probability density. If Q has density q, then its characteristic function is $\varphi_Q(t) = (1-|t|)^+$. Thus Fourier inversion has exchanged the triangular density and the normalized squared-sinc density. $\square$

Lecture 8 — Thursday, January 31, 2019

Theorem 8.1 (Pólya's criterion) Let $\varphi : \mathbb{R} \to \mathbb{R}$ satisfy the following conditions:

1. φ is continuous and $\varphi(t) = \varphi(-t)$ for every t;
2. $\varphi(0) = 1$;
3. φ is convex on $(0, \infty)$;
4. $\varphi(t) \to 0$ as $t \to \infty$.

Then φ is a characteristic function. More precisely, there exists a probability measure ν on $(0, \infty)$ such that

$$\varphi(t) = \int_0^\infty \left(1 - \frac{|t|}{s}\right)^+ \mathrm{d}\nu(s).$$

Remark 8.2 The calculation above shows that a random variable with density $(1 - \cos x)/(\pi x^2)$ has characteristic function $(1 - |t|)^+$. Scaling therefore shows that every triangular function appearing in the mixture in Theorem 8.1 is a characteristic function.

Proof. Let φ'_+ be the right derivative of φ on $[0, \infty)$. Convexity guarantees that this derivative exists everywhere, is right-continuous, and is nondecreasing.

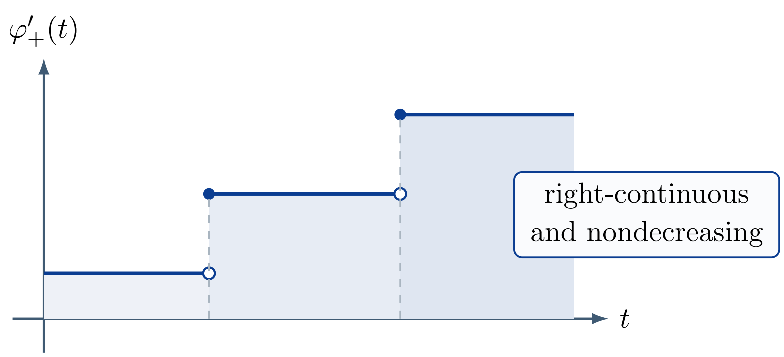

Figure 9

Define a measure η on $(0, \infty)$ by

$$\eta((a, b]) = \varphi'_+(b) - \varphi'_+(a),$$

as suggested by Figure 9. Define ν by the size bias

$$\nu(A) = \int_A s \, \mathrm{d}\eta(s).$$

Since $\varphi(t) \to 0$ and φ is convex, $\varphi'_+(t) \to 0$. Consequently,

$$\eta((a, \infty)) = -\varphi'_+(a) = \int_a^\infty \frac{1}{s} \, \mathrm{d}\nu(s).$$

For $t \geq 0$, the fundamental theorem of calculus and Tonelli's theorem give

$$\varphi(t) = -\int_t^\infty \varphi'_+(s) \, \mathrm{d}s = \int_t^\infty \int_s^\infty \frac{1}{r} \, \mathrm{d}\nu(r) \, \mathrm{d}s = \int_t^\infty \left(1 - \frac{t}{r}\right) \mathrm{d}\nu(r) = \int_0^\infty \left(1 - \frac{t}{r}\right)^+ \mathrm{d}\nu(r).$$

Evenness extends this representation to every t:

$$\varphi(t) = \int_0^\infty \left(1 - \frac{|t|}{r}\right)^+ \mathrm{d}\nu(r).$$

At $t = 0$,

$$1 = \varphi(0) = \nu((0, \infty)),$$

so ν is a probability measure. If ν has finite support, the integral is a convex combination of triangular characteristic functions and is therefore a characteristic function. For general ν, choose finitely supported probability measures $\nu_n \Rightarrow \nu$. For each fixed t, the integrand is bounded and continuous as a function of $r > 0$, so

$$\int_0^\infty \left(1 - \frac{|t|}{r}\right)^+ \mathrm{d}\nu_n(r) \longrightarrow \varphi(t).$$

These approximating functions are characteristic functions, and φ is continuous at 0. The Lévy continuity theorem now proves that φ is a characteristic function. □

Remark 8.3 The proof illustrates a general advantage of weak convergence: arbitrary probability measures can often be approximated by finitely supported measures, for which calculations are elementary.

Example 8.4 Let $\varphi(x) = e^{-|x|^\alpha}$, where $0 < \alpha \leq 1$.

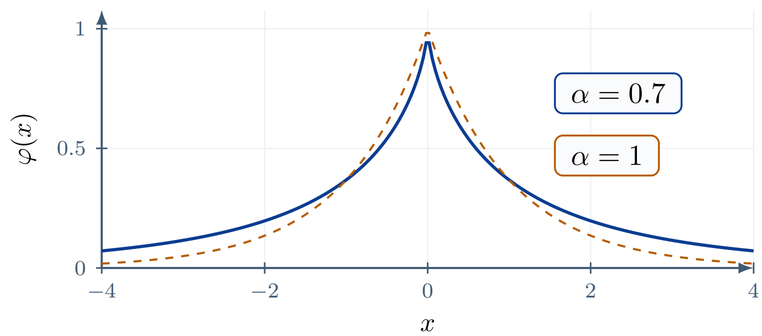

FIGURE 10

The shape is illustrated in Figure 10. The Pólya criterion shows that φ is a characteristic function. Let X_1 and X_2 be independent random variables having this characteristic function. Then

$$\varphi_{X_1+X_2}(x) = \varphi_{X_1}(x)\varphi_{X_2}(x) = e^{-2|x|^\alpha} = \varphi_{X_1}(2^{1/\alpha}x) = \varphi_{2^{1/\alpha}X_1}(x).$$

The inversion formula therefore gives

$$X_1 + X_2 \overset{d}{=} 2^{1/\alpha}X_1.$$

For comparison, if Y_1, Y_2 are independent standard Gaussian random variables, then $Y_1 + Y_2 \overset{d}{=} \sqrt{2}\,Y_1$.

The central limit theorem identifies the standard Gaussian law as a fixed point of convolution followed by the normalization $x \mapsto x/\sqrt{2}$. The preceding example exhibits analogous fixed points with other scaling exponents.

Remark 8.5 The space of finite variance distributions can be visualized as a set in which repeated normalized convolution drives distributions toward the $\mathcal{N}(0,1)$ fixed point, as shown in Figure 11.

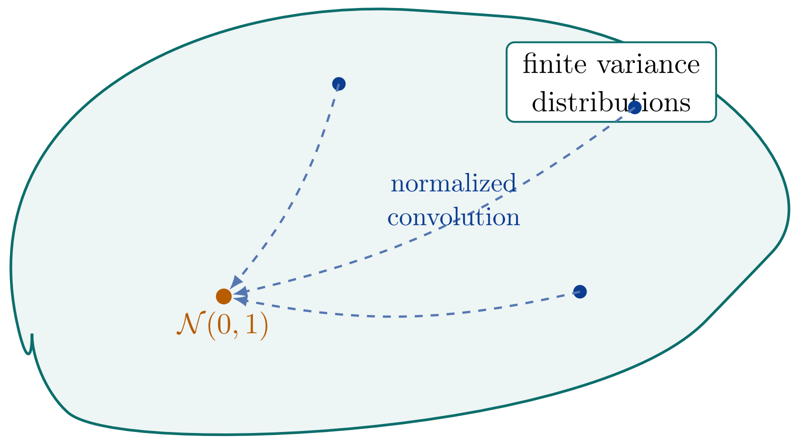

Figure 11

Definition 8.6 A random variable Y has a **stable law** if, for every positive integer k, there exist constants $a_k > 0$ and $b_k \in \mathbb{R}$ such that, whenever $Y_1, \dots, Y_k$ are independent copies of Y,

$$\frac{\sum_{i=1}^k Y_i - b_k}{a_k} \stackrel{d}{=} Y_1.$$

Example 8.7 Gaussian and Cauchy laws are stable. Stable laws are natural candidates for limits in generalized central limit theorems; among nondegenerate stable laws, the Gaussian laws are precisely those with finite variance.

Lecture 9 — Tuesday, February 05, 2019

Proofs of the Central Limit Theorem

The characteristic function proof of the central limit theorem has a useful geometric interpretation. Suppose that $X_1, X_2, \dots$ are independent and identically distributed, with mean 0 and variance 1, and write

$$S_n = X_1 + \cdots + X_n.$$

If $\varphi(\lambda) = \mathbb{E}[e^{i\lambda X_1}]$, then the central limit theorem is equivalent, by the Lévy continuity theorem, to

$$\varphi_{S_n/\sqrt{n}}(\lambda) = \varphi\left(\frac{\lambda}{\sqrt{n}}\right)^n \longrightarrow e^{-\lambda^2/2} \quad \text{for every } \lambda \in \mathbb{R}.$$

For intuition, temporarily suppose that φ is real and positive near the origin, and set $\Psi = \log \varphi$. Then

$$\log \varphi_{S_n/\sqrt{n}}(\lambda) = n\Psi\left(\frac{\lambda}{\sqrt{n}}\right) \longrightarrow -\frac{\lambda^2}{2}.$$

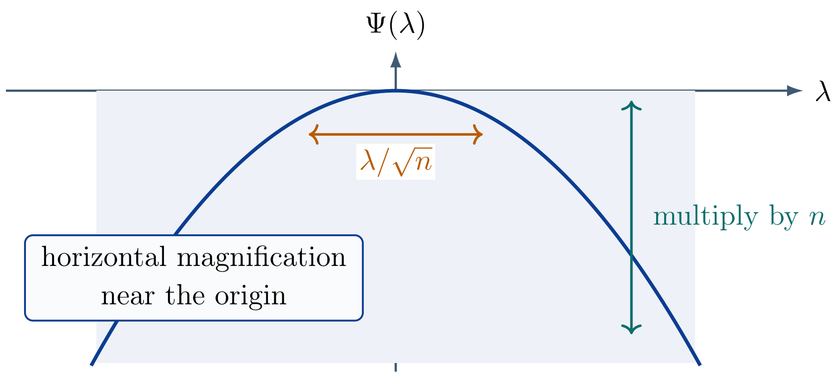

FIGURE 12

As illustrated in Figure 12, the map $\lambda \mapsto n\Psi(\lambda/\sqrt{n})$ magnifies the behavior of Ψ near the origin. The quadratic function $-\lambda^2/2$ is unchanged by this rescaling because

$$2\left(-\frac{(\lambda/\sqrt{2})^2}{2}\right) = -\frac{\lambda^2}{2}.$$

This fixed point property suggests looking for other limiting laws and other normalizations. If $\Psi(\lambda) = -|\lambda|^\alpha$, then the analogous identity

$$2\Psi\left(\frac{\lambda}{2^\gamma}\right) = \Psi(\lambda)$$

holds precisely when $\gamma = 1/\alpha$.

Definition 9.1 Let $\alpha > 0$. A probability distribution is called **strictly symmetric α-stable** if either, and hence both, of the following equivalent conditions holds:

1. Its characteristic function φ satisfies

$$\varphi\left(\frac{\lambda}{2^{1/\alpha}}\right)^2 = \varphi(\lambda) \quad \text{for every } \lambda \in \mathbb{R}.$$

2. If X_1 and X_2 are independent random variables with the given distribution, then

$$\frac{X_1 + X_2}{2^{1/\alpha}} \overset{d}{=} X_1.$$

The equivalence follows directly from the product rule for characteristic functions. In particular, whenever

$$\varphi_\alpha(\lambda) = e^{-|\lambda|^\alpha}$$

is a characteristic function, it defines a strictly symmetric α-stable law. For example, the distribution $\mathcal{N}(0,2)$ has characteristic function $e^{-\lambda^2}$ and is therefore strictly symmetric 2-stable.

Proposition 9.2 If $\alpha > 2$, then $\lambda \mapsto e^{-|\lambda|^\alpha}$ is not the characteristic function of a nondegenerate random variable.

Proof. Suppose that X had this characteristic function. By Theorem 7.6,

$$\mathbb{E}[X^2] = \lim_{h\to 0} \frac{2 - \varphi_X(h) - \varphi_X(-h)}{h^2} = \lim_{h\to 0} \frac{2(1 - e^{-|h|^\alpha})}{h^2} = 2 \lim_{h\to 0} |h|^{\alpha-2} = 0.$$

Thus $X = 0$ almost surely, whose characteristic function is identically equal to 1. This contradicts $\varphi_X(\lambda) = e^{-|\lambda|^\alpha}$ for $\lambda \neq 0$. □

Proposition 9.3 If $0 < \alpha \leq 1$, then $\lambda \mapsto e^{-|\lambda|^\alpha}$ is a characteristic function.

Proof. For $\lambda > 0$, differentiation gives

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} e^{-\lambda^\alpha} = e^{-\lambda^\alpha}\left(\alpha^2 \lambda^{2\alpha-2} - \alpha(\alpha-1)\lambda^{\alpha-2}\right) \geq 0.$$

Hence the function is convex on $[0,\infty)$. It is also even, continuous, equal to 1 at the origin, and converges to 0 at infinity. Pólya's criterion, Theorem 8.1, therefore shows that it is a characteristic function. □

Definition 9.4 A random variable X has the **standard Cauchy distribution** if

$$\mathbb{P}(X \in A) = \int_A \frac{1}{\pi(1+x^2)}\,\mathrm{d}x$$

for every Borel set $A \subseteq \mathbb{R}$.

The standard Cauchy characteristic function is $\varphi_X(\lambda) = e^{-|\lambda|}$. Consequently, if $X_1, \ldots, X_n$ are independent standard Cauchy random variables, then

$$\frac{X_1 + \cdots + X_n}{n} \overset{d}{=} X_1.$$

This stability is consistent with the failure of the usual law of large numbers: a standard Cauchy random variable does not have a finite first moment. Cauchy distributions also arise geometrically in optics through Huygens' principle.

The remaining exponents $1 < \alpha \leq 2$ can be constructed from a random mixture. Begin with $\varphi(\lambda) = \cos\lambda$, the characteristic function of a Rademacher random variable satisfying

$$\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{2}.$$

If $(c_n)_{n\geq 0}$ is a probability mass function, then

$$\varphi_c(\lambda) = \sum_{n=0}^{\infty} c_n \cos(\lambda)^n$$

is the characteristic function of $X_1 + \cdots + X_T$, where the X_j are independent copies of X, the nonnegative integer-valued random variable T is independent of them, and $\mathbb{P}(T = n) = c_n$. More generally, the following elementary fact will be useful.

Proposition 9.5 Let $(\varphi_n)_{n\geq 0}$ be characteristic functions and let $c_n \geq 0$ satisfy

$$\sum_{n=0}^{\infty} c_n = 1.$$

Then $\sum_{n=0}^{\infty} c_n\varphi_n$ is a characteristic function.

Proof. For every n, choose a random variable Y_n with characteristic function φ_n. Let T be independent of the collection $(Y_n)_{n\geq 0}$ and satisfy $\mathbb{P}(T = n) = c_n$. Conditioning on T gives

$$\mathbb{E}[e^{i\lambda Y_T}] = \sum_{n=0}^{\infty} c_n \mathbb{E}[e^{i\lambda Y_n}] = \sum_{n=0}^{\infty} c_n \varphi_n(\lambda).$$

□

Fix $0 < \alpha \leq 2$, set $\beta = \alpha/2$, and define

$$c_0 = 0 \qquad \text{and} \qquad c_n = (-1)^{n+1}\binom{\beta}{n}, \quad n \geq 1.$$

Because $0 < \beta \leq 1$, all these coefficients are nonnegative. The generalized binomial expansion yields

$$\sum_{n=1}^{\infty} c_n x^n = 1 - (1-x)^{\beta}.$$

Taking $x = 1$ shows that $\sum_{n=0}^{\infty} c_n = 1$. Proposition 9.5, applied with $\varphi_n(\lambda) = \cos(\lambda)^n$, now shows that

$$\varphi_c(\lambda) = 1 - (1 - \cos\lambda)^{\alpha/2}$$

is a characteristic function.

Since $1 - \cos\lambda = \lambda^2/2 + O(\lambda^4)$ as $\lambda \to 0$,

$$\varphi_c(\lambda) = 1 - \frac{|\lambda|^{\alpha}}{2^{\alpha/2}} + o(|\lambda|^{\alpha}).$$

Consequently, for each fixed $\lambda \in \mathbb{R}$,

$$\varphi_c\left(\frac{\lambda}{n^{1/\alpha}}\right)^n = \left(1 - \frac{|\lambda|^{\alpha}}{2^{\alpha/2} n} + o\left(\frac{1}{n}\right)\right)^n \longrightarrow \exp\left(-\frac{|\lambda|^{\alpha}}{2^{\alpha/2}}\right).$$

The left-hand side is the characteristic function of a normalized sum of independent copies of the random variable with characteristic function φ_c. The Lévy continuity theorem therefore proves that the limiting function is a characteristic function. Rescaling its random variable by $\sqrt{2}$ produces the characteristic function $e^{-|\lambda|^{\alpha}}$.

Theorem 9.6 The function $\lambda \mapsto e^{-|\lambda|^{\alpha}}$ is a characteristic function if and only if $0 < \alpha \leq 2$. For every such α, it is the characteristic function of a strictly symmetric α-stable law.

Proof. *Forward implication.* Proposition 9.2 rules out $\alpha > 2$, while $\alpha \le 0$ does not give a function continuous at the origin with value 1.

Reverse implication. The mixture construction above proves the assertion for every $0 < \alpha \le 2$. The scaling identity in Definition 9.1(1) then gives strict symmetric α-stability. $\square$

Thus normalized sums need not converge only to Gaussian distributions. With a normalization of order $n^{1/\alpha}$, heavy-tailed summands can instead converge to a non-Gaussian stable law.

Lecture 10 — Thursday, February 07, 2019

We next give a third proof of the central limit theorem. Unlike the proofs based on Stirling's formula and characteristic function expansions, this argument replaces the summands one at a time by Gaussian random variables. It also leads naturally to a central limit theorem for non-identically distributed variables.

Lemma 10.1 Let $f\colon \mathbb{R} \to \mathbb{C}$ be k times continuously differentiable. If $|f^{(k)}(x)| \le b$ for every $x \in \mathbb{R}$, then

$$\left| f(x+h) - \sum_{j=0}^{k-1} \frac{f^{(j)}(x)}{j!} h^j \right| \le \frac{b}{k!} |h|^k$$

for all $x, h \in \mathbb{R}$.

Proof. Taylor's theorem with integral remainder gives, when $h \ge 0$,

$$f(x+h) - \sum_{j=0}^{k-1} \frac{f^{(j)}(x)}{j!} h^j = \frac{1}{(k-1)!} \int_0^h (h-t)^{k-1} f^{(k)}(x+t)\,\mathrm{d}t.$$

Taking absolute values and integrating $b(h-t)^{k-1}$ gives $bh^k/k!$. When $h < 0$, the same argument applies after reversing the limits of integration. $\square$

If both f'' and $f^{(3)}$ are bounded, Lemma 10.1 gives two useful estimates for the second-order remainder

$$R_f(x,h) = f(x+h) - f(x) - f'(x)h - \frac{f''(x)}{2} h^2.$$

Indeed,

$$|R_f(x,h)| \le \frac{\|f'''\|_\infty}{6}|h|^3 \qquad \text{and} \qquad |R_f(x,h)| \le \|f''\|_\infty h^2.$$

Thus, with $C_f = \max\{\|f'''\|_\infty/6, \|f''\|_\infty\}$,

$$|R_f(x,h)| \le C_f\big(|h|^3 \wedge h^2\big). \tag{10.1}$$

Theorem 10.2 Suppose that $X_1, X_2, \dots$ are independent and identically distributed, with

$$\mathbb{E}[X_1] = 0, \qquad \mathbb{E}[X_1^2] = 1, \qquad \text{and} \qquad \mathbb{E}[|X_1|^3] < \infty.$$

Let N have the standard Gaussian distribution. If $f\colon \mathbb{R} \to \mathbb{C}$ is bounded and three times continuously differentiable, with $\|f'''\|_\infty \le b$, then

$$\left|\mathbb{E}\left[f\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)\right] - \mathbb{E}[f(N)]\right| \le \frac{b}{6\sqrt{n}}\left(\mathbb{E}[|X_1|^3] + 2\sqrt{\frac{2}{\pi}}\right).$$

Proof. Let $Z_1, \dots, Z_n$ be independent standard Gaussian random variables, independent of the X_j. Define

$$W_k = \frac{Z_1+\cdots+Z_{k-1}+X_{k+1}+\cdots+X_n}{\sqrt{n}}.$$

Replacing the variables one at a time and applying the triangle inequality gives

$$\begin{aligned}&\left|\mathbb{E}\left[f\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\right)\right]\right| \\ &\qquad \le \sum_{k=1}^n \left|\mathbb{E}\left[f\left(W_k + \frac{X_k}{\sqrt{n}}\right) - f\left(W_k + \frac{Z_k}{\sqrt{n}}\right)\right]\right|.\end{aligned}$$

The random variable W_k is independent of both X_k and Z_k. By Lemma 10.1,

$$f\left(W_k + \frac{X_k}{\sqrt{n}}\right) = f(W_k) + f'(W_k)\frac{X_k}{\sqrt{n}} + \frac{f''(W_k)}{2}\frac{X_k^2}{n} + R_{k,X},$$

where $|R_{k,X}| \le b|X_k|^3/(6n^{3/2})$. The analogous expansion with Z_k has a remainder bounded by $b|Z_k|^3/(6n^{3/2})$. Since X_k and Z_k have the same first two moments, independence causes the first three terms in the two expected expansions to cancel. Therefore

$$\left|\mathbb{E}\left[f\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)\right] - \mathbb{E}[f(N)]\right|$$

$$\leq \frac{b}{6n^{3/2}} \sum_{k=1}^{n} \left(\mathbb{E}[|X_k|^3] + \mathbb{E}[|Z_k|^3]\right)$$
$$= \frac{b}{6\sqrt{n}} \left(\mathbb{E}[|X_1|^3] + 2\sqrt{\frac{2}{\pi}}\right),$$

because $(Z_1 + \cdots + Z_n)/\sqrt{n}$ has the standard Gaussian distribution. □

Lecture 11 — Tuesday, February 12, 2019

To understand the non-identically distributed case, use the truncated remainder estimate (10.1). For centered independent variables $X_{n1}, \ldots, X_{nk_n}$, set

$$s_n^2 = \sum_{j=1}^{k_n} \mathbb{E}[X_{nj}^2].$$

Replace X_{nj} by an independent Gaussian variable Z_{nj} with the same variance. The corresponding remainder is bounded by a constant multiple of

$$\frac{|X_{nj}|^3}{s_n^3} \wedge \frac{X_{nj}^2}{s_n^2},$$

and the same estimate holds with Z_{nj} in place of X_{nj}. This observation leads to the following theorem.

Theorem 11.1 (Lindeberg–Feller central limit theorem) For every n, let $X_{n1}, \ldots, X_{nk_n}$ be independent random variables with mean zero. Suppose that $s_n^2 > 0$ and that, for every $\varepsilon > 0$,

$$\frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}| > \varepsilon s_n\}}\right] \longrightarrow 0. \tag{11.1}$$

Then

$$\frac{1}{s_n} \sum_{j=1}^{k_n} X_{nj} \implies \mathcal{N}(0,1).$$

Proof. Let Z_{nj} be independent centered Gaussian random variables, independent of the array, with

$$\mathbb{E}[Z_{nj}^2] = \mathbb{E}[X_{nj}^2].$$

For a bounded function f with bounded second and third derivatives, the replacement argument and (10.1) show that the difference between the expectations obtained from the X_{nj} and the Z_{nj} is at most

$$C_f \sum_{j=1}^{k_n} \mathbb{E}\left[\frac{|X_{nj}|^3}{s_n^3} \wedge \frac{X_{nj}^2}{s_n^2} + \frac{|Z_{nj}|^3}{s_n^3} \wedge \frac{Z_{nj}^2}{s_n^2} \right].$$

For any $\varepsilon > 0$, the contribution from the original variables is bounded by

$$\varepsilon \frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}[X_{nj}^2] + \frac{1}{s_n^2} \sum_{j=1}^{k_n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}| > \varepsilon s_n\}} \right] = \varepsilon + o(1).$$

Condition (11.1) also implies

$$\max_{1 \le j \le k_n} \frac{\mathbb{E}[X_{nj}^2]}{s_n^2} \longrightarrow 0.$$

Indeed, each variance ratio is at most ε^2 plus the left-hand side of (11.1). Writing $a_{nj}^2 = \mathbb{E}[Z_{nj}^2]/s_n^2$, the Gaussian contribution is at most

$$\sum_{j=1}^{k_n} \frac{\mathbb{E}[|Z_{nj}|^3]}{s_n^3} = 2\sqrt{\frac{2}{\pi}} \sum_{j=1}^{k_n} a_{nj}^3 \le 2\sqrt{\frac{2}{\pi}} \max_j a_{nj} \longrightarrow 0.$$

It follows that the replacement error tends to zero. Choosing $f(x) = e^{i\lambda x}$ shows that the characteristic function of the normalized row sum converges pointwise to the characteristic function of $\mathcal{N}(0,1)$. The conclusion follows from the Lévy continuity theorem. □

For identically distributed summands with finite variance, the truncation needed above follows directly from the dominated convergence theorem. More explicitly, for every $k > 0$,

$$X_j^2 \wedge \frac{|X_j|^3}{\sqrt{n}} \le X_j^2 \mathbb{1}_{\{|X_j| \ge k\}} + \frac{k^3}{\sqrt{n}},$$

and hence

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[X_j^2 \wedge \frac{|X_j|^3}{\sqrt{n}} \right] \le \mathbb{E}\left[X_1^2 \mathbb{1}_{\{|X_1| \ge k\}} \right] + \frac{k^3}{\sqrt{n}}.$$

First letting $n \to \infty$ and then $k \to \infty$ recovers the identically distributed central limit theorem without assuming a finite third moment.

The same estimate gives a convenient sufficient condition for a triangular array whose nth row contains exactly n terms.

Corollary 11.2 For each positive integer n, let $X_{n1}, \ldots, X_{nn}$ be independent random variables with mean zero, and suppose that

$$\sum_{j=1}^{n} \mathbb{E}[X_{nj}^2] = n.$$

Define

$$\overline{\varepsilon}_{n,k} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>k\}}\right].$$

If

$$\sup_{n\geq 1} \overline{\varepsilon}_{n,k} \longrightarrow 0 \quad \text{as } k \to \infty,$$

then

$$\frac{X_{n1} + \cdots + X_{nn}}{\sqrt{n}} \implies \mathcal{N}(0,1).$$

Proof. Fix $\varepsilon > 0$ and $k > 0$. Once n is large enough that $\varepsilon\sqrt{n} \geq k$,

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[X_{nj}^2 \mathbb{1}_{\{|X_{nj}|>\varepsilon\sqrt{n}\}}\right] \leq \overline{\varepsilon}_{n,k} \leq \sup_{m\geq 1} \overline{\varepsilon}_{m,k}.$$

The final expression can be made arbitrarily small by choosing k sufficiently large. Thus the Lindeberg condition (11.1) holds with $s_n = \sqrt{n}$, and Theorem 11.1 gives the conclusion. $\square$

2. Conditional Expectation and Martingales

Conditional Expectation

Conditional expectation formalizes the expected value of a random variable after some information has been revealed. A sub-σ-field represents the available information, so the expression

$$\mathbb{E}\left[\underbrace{X}_{\text{random variable}} \mid \underbrace{\mathcal{G}}_{\text{available information}}\right]$$

must itself be a random variable whose value depends only on $\mathcal{G}$.

The discrete case makes this interpretation concrete. Let X and Y be independent, with X uniformly distributed on $\{1,2,3\}$ and Y uniformly distributed on $\{1,2\}$. The information in $\sigma(X+Y)$ reveals the value of the sum but not the individual values of X and Y. Equivalently, it reveals which cell of the partition in Figure 13 contains the outcome.

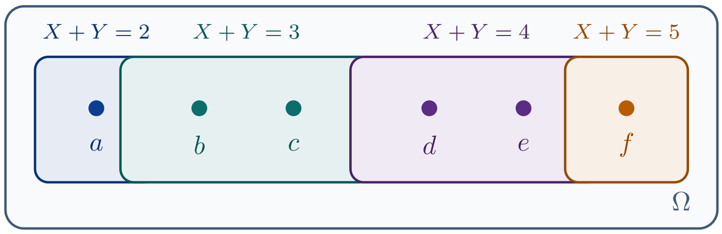

FIGURE 13

Label the six equally likely outcomes so that $a=(1,1)$, $b=(1,2)$, $c=(2,1)$, $d=(2,2)$, $e=(3,1)$, and $f=(3,2)$. The conditional expectation of Y is constant on each level set of $X+Y$. In fact,

$$Z = \mathbb{E}[Y \mid \sigma(X+Y)]$$

satisfies

$$Z(a)=1, \qquad Z(b)=Z(c)=Z(d)=Z(e)=\frac{3}{2}, \qquad \text{and} \qquad Z(f)=2.$$

For example, on the event $\{X+Y=3\}=\{b,c\}$, the two possible values of Y are 2 and 1, each with conditional probability $1/2$.

Example 11.3 At the two extremes of available information,

$$\mathbb{E}[Y \mid \{\varnothing,\Omega\}] = \mathbb{E}[Y] \qquad \text{and} \qquad \mathbb{E}[Y \mid 2^{\Omega}] = Y.$$

More generally, conditioning on Y itself gives $\mathbb{E}[Y \mid \sigma(Y)] = Y$.

For a finite partition, conditional expectation can be constructed by averaging over each cell. The following definition extends this idea to an arbitrary sub-σ-field.

Definition 11.4 Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, let $\mathcal{G}\subseteq\mathcal{F}$ be a sub-σ-field, and let X be integrable. An integrable random variable Z is a **conditional expectation of X given $\mathcal{G}$** if the following hold:

1. Z is $\mathcal{G}$-measurable.
2. For every $A \in \mathcal{G}$,
$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[Z\mathbb{1}_A].$$

When it exists, this random variable is denoted by $\mathbb{E}[X \mid \mathcal{G}]$.

Condition (1) says that the conditional expectation uses only the information in $\mathcal{G}$. Condition (2) says that it preserves the average of X on every event that can be distinguished using that information.

Proposition 11.5 If Z_1 and Z_2 both satisfy Definition 11.4, then $Z_1 = Z_2$ almost surely.

Proof. The event $A = \{Z_1 > Z_2\}$ belongs to $\mathcal{G}$. By Definition 11.4(2),
$$\mathbb{E}[(Z_1 - Z_2)\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A] - \mathbb{E}[X\mathbb{1}_A] = 0.$$
The integrand is nonnegative and is strictly positive on A, so $\mathbb{P}(A) = 0$. Interchanging Z_1 and Z_2 shows that $\mathbb{P}(Z_2 > Z_1) = 0$, proving the claim. □

Proposition 11.6 Let X and Y be integrable, and let $a, b \in \mathbb{R}$.

1. If $X \geq 0$ almost surely, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ almost surely.
2. Conditional expectation is linear:
$$\mathbb{E}[aX + bY \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}] \quad \text{almost surely.}$$
3. If X is independent of $\mathcal{G}$, then
$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X] \quad \text{almost surely.}$$

Proof. For property (1), set $Z = \mathbb{E}[X \mid \mathcal{G}]$. For every positive integer m, the event $A_m = \{Z \leq -1/m\}$ belongs to $\mathcal{G}$, and
$$0 \leq \mathbb{E}[X\mathbb{1}_{A_m}] = \mathbb{E}[Z\mathbb{1}_{A_m}] \leq -\frac{1}{m}\mathbb{P}(A_m).$$

Thus $\mathbb{P}(A_m) = 0$. Since $\{Z < 0\} = \bigcup_{m=1}^{\infty} A_m$, positivity follows.

For property (2), the random variable

$$Z = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}]$$

is integrable and $\mathcal{G}$-measurable. For every $A \in \mathcal{G}$,

$$\mathbb{E}[Z\mathbb{1}_A] = a\mathbb{E}[X\mathbb{1}_A] + b\mathbb{E}[Y\mathbb{1}_A] = \mathbb{E}[(aX + bY)\mathbb{1}_A].$$

Hence Z is a version of $\mathbb{E}[aX + bY \mid \mathcal{G}]$, and Proposition 11.5 gives the asserted equality.

For property (3), the constant random variable $Z = \mathbb{E}[X]$ is $\mathcal{G}$-measurable. Independence gives, for every $A \in \mathcal{G}$,

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[X]\mathbb{P}(A) = \mathbb{E}[Z\mathbb{1}_A].$$

The conclusion again follows from Proposition 11.5. □

Lemma 11.7 If $Z = \mathbb{E}[X \mid \mathcal{G}]$, then, for every bounded $\mathcal{G}$-measurable random variable Y,

$$\mathbb{E}[YX] = \mathbb{E}[YZ].$$

Proof. The proof follows the standard extension from indicators to bounded measurable functions:

1. If $Y = \mathbb{1}_A$ for some $A \in \mathcal{G}$, the identity is exactly Definition 11.4(2).
2. By linearity, the identity holds for every $\mathcal{G}$-measurable simple random variable.
3. If Y is bounded and nonnegative, choose $\mathcal{G}$-measurable simple random variables $Y_n \uparrow Y$. Since $|Y_nX| \leq \|Y\|_\infty |X|$ and $|Y_nZ| \leq \|Y\|_\infty |Z|$, the dominated convergence theorem extends the identity to Y.
4. For a general bounded Y, apply the preceding step to Y^+ and Y^-, and subtract the two identities.

□

Proposition 11.8 If Y is bounded and $\mathcal{G}$-measurable, then

$$\mathbb{E}[XY \mid \mathcal{G}] = Y\mathbb{E}[X \mid \mathcal{G}] \quad \text{almost surely.} \tag{11.2}$$

Proof. The right-hand side of (11.2) is integrable and $\mathcal{G}$-measurable. For every $A \in \mathcal{G}$, Lemma 11.7, applied to the bounded $\mathcal{G}$-measurable random variable $Y\mathbb{1}_A$, gives

$$\mathbb{E}[XY\mathbb{1}_A] = \mathbb{E}\left[Y\mathbb{1}_A\mathbb{E}[X \mid \mathcal{G}]\right].$$

The defining conditions and Proposition 11.5 prove the result. □

Lecture 12 — Thursday, February 14, 2019

Existence of Conditional Expectation and the Radon–Nikodym Theorem

It remains to prove that conditional expectations actually exist. The construction is an application of the Radon–Nikodym theorem.

Definition 12.1 Let μ and ν be measures on $(\Omega, \mathcal{F})$. The measure ν is **absolutely continuous with respect to μ**, written $\nu \ll \mu$, if, for every $A \in \mathcal{F}$,

$$\mu(A) = 0 \Longrightarrow \nu(A) = 0.$$

Theorem 12.2 (Radon–Nikodym theorem) Let μ and ν be σ-finite measures on $(\Omega, \mathcal{F})$, and suppose that $\nu \ll \mu$. Then there exists an $\mathcal{F}$-measurable function $f\colon \Omega \to [0, \infty]$ such that, for every $A \in \mathcal{F}$,

$$\nu(A) = \int_A f \,\mathrm{d}\mu.$$

The function f is unique up to μ-null sets and is denoted by $\mathrm{d}\nu/\,\mathrm{d}\mu$.

Theorem 12.3 Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-field. Then $\mathbb{E}[X \mid \mathcal{G}]$ exists. Together with Proposition 11.5, it is unique up to almost-sure equality.

Proof. First suppose that $X \geq 0$. Define a set function ν on $(\Omega, \mathcal{G})$ by

$$\nu(A) = \mathbb{E}[X \mathbb{1}_A] = \int_A X \, \mathrm{d}\mathbb{P}, \quad A \in \mathcal{G}.$$

This is the measure obtained by weighting $\mathbb{P}$ by X, rather than merely a restriction of $\mathbb{P}$. If $(A_j)_{j\geq 1}$ are pairwise disjoint members of $\mathcal{G}$, the monotone convergence theorem gives

$$\nu\left(\bigcup_{j=1}^{\infty} A_j\right) = \mathbb{E}\left[X \sum_{j=1}^{\infty} \mathbb{1}_{A_j}\right] = \sum_{j=1}^{\infty} \mathbb{E}[X \mathbb{1}_{A_j}] = \sum_{j=1}^{\infty} \nu(A_j).$$

Moreover, $\nu(\Omega) = \mathbb{E}[X] < \infty$, and $\nu \ll \mathbb{P}|_{\mathcal{G}}$. By the Radon–Nikodym theorem, there is a nonnegative $\mathcal{G}$-measurable function Z such that

$$\nu(A) = \int_A Z \, \mathrm{d}\mathbb{P} \quad \text{for every } A \in \mathcal{G}.$$

Taking $A = \Omega$ shows that Z is integrable, and the displayed identity is exactly Definition 11.4(2).

For a general integrable X, write $X = X^+ - X^-$. The first part supplies conditional expectations $Z^+ = \mathbb{E}[X^+ \mid \mathcal{G}]$ and $Z^- = \mathbb{E}[X^- \mid \mathcal{G}]$. Their difference $Z = Z^+ - Z^-$ is integrable and $\mathcal{G}$-measurable, and for every $A \in \mathcal{G}$,

$$\mathbb{E}[Z \mathbb{1}_A] = \mathbb{E}[X^+ \mathbb{1}_A] - \mathbb{E}[X^- \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A].$$

Thus Z is a conditional expectation of X given $\mathcal{G}$. $\square$

The Radon–Nikodym derivative therefore provides the general existence theorem. When X is square integrable, conditional expectation also has a geometric description as an orthogonal projection.

Theorem 12.4 Suppose that $X \in L^2(\mathcal{F})$. Then $\mathbb{E}[X \mid \mathcal{G}]$ is the orthogonal projection of X onto the closed subspace $L^2(\mathcal{G})$. Equivalently, it is the unique $Y \in L^2(\mathcal{G})$, up to almost-sure equality, that minimizes

$$\mathbb{E}[(X - Y)^2].$$

Figure 14 depicts this projection in the Hilbert space $L^2(\mathcal{F})$.

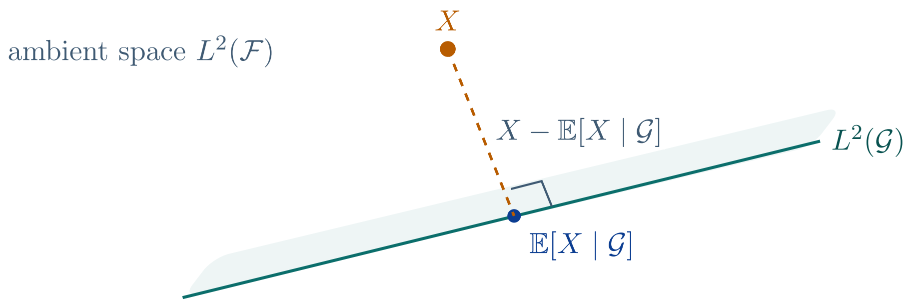

FIGURE 14

Proof. The space $L^2(\mathcal{F})$, with inner product $\langle U, V\rangle = \mathbb{E}[UV]$, is a Hilbert space, and $L^2(\mathcal{G})$ is a closed subspace. The Hilbert space projection theorem therefore gives a unique decomposition

$$X = Y + R, \qquad Y \in L^2(\mathcal{G}), \qquad \text{and} \qquad R \in L^2(\mathcal{G})^{\perp}.$$

For every $A \in \mathcal{G}$, the indicator $\mathbb{1}_A$ belongs to $L^2(\mathcal{G})$, so

$$\mathbb{E}[X\mathbb{1}_A] - \mathbb{E}[Y\mathbb{1}_A] = \mathbb{E}[R\mathbb{1}_A] = 0.$$

Thus Y satisfies both conditions in Definition 11.4, and Proposition 11.5 shows that $Y = \mathbb{E}[X \mid \mathcal{G}]$ almost surely.

If $V \in L^2(\mathcal{G})$, then $Y - V \in L^2(\mathcal{G})$ is orthogonal to $R = X - Y$. The Pythagorean identity gives

$$\mathbb{E}[(X-V)^2] = \mathbb{E}[(R+Y-V)^2] = \mathbb{E}[R^2] + \mathbb{E}[(Y-V)^2] \geq \mathbb{E}[R^2] = \mathbb{E}[(X-Y)^2].$$

Equality holds if and only if $V = Y$ almost surely. $\square$

Lecture 13 — Tuesday, February 19, 2019

Properties of Conditional Expectation

Theorem 13.1 (Conditional Jensen inequality) Let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be convex. If X and $\varphi(X)$ are integrable, then

$$\varphi(\mathbb{E}[X \mid \mathcal{G}]) \le \mathbb{E}[\varphi(X) \mid \mathcal{G}] \quad \text{almost surely.}$$

Proof. Figure 15 illustrates an affine minorant of φ.

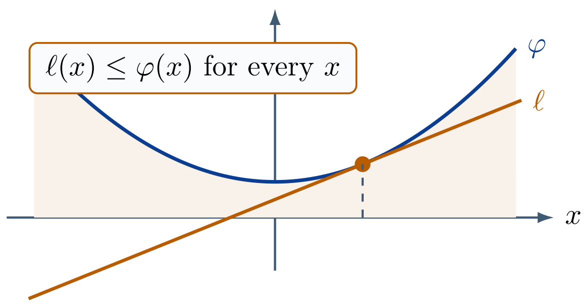

FIGURE 15

A finite convex function on $\mathbb{R}$ is the pointwise supremum of a countable collection of affine minorants $\ell_m(x) = a_m x + b_m$. For every m, positivity and linearity from Proposition 11.6 give

$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \ge \mathbb{E}[\ell_m(X) \mid \mathcal{G}] = \ell_m(\mathbb{E}[X \mid \mathcal{G}])\,.$$

Taking the supremum over m yields

$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \ge \sup_m \ell_m(\mathbb{E}[X \mid \mathcal{G}]) = \varphi(\mathbb{E}[X \mid \mathcal{G}])\,.$$

□

Remark 13.2 The same proof extends to convex functions $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ by representing φ as a supremum of affine minorants on $\mathbb{R}^d$.

Proposition 13.3 (Tower property) Suppose that $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$. For every integrable X, the following hold:

1.
$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{H}] \quad \text{almost surely.}$$

2.
$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}] \quad \text{almost surely.}$$

In particular, taking $\mathcal{H} = \{\varnothing, \Omega\}$ gives $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[X]$.

Proof. For (1), the random variable $\mathbb{E}[X \mid \mathcal{H}]$ is $\mathcal{H}$-measurable and hence $\mathcal{G}$-measurable. Conditioning a measurable random variable on $\mathcal{G}$ leaves it unchanged.

For (2), set $Z = \mathbb{E}[X \mid \mathcal{G}]$. For every $A \in \mathcal{H} \subseteq \mathcal{G}$,

$$\mathbb{E}[Z\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A].$$

Thus $\mathbb{E}[Z \mid \mathcal{H}]$ and $\mathbb{E}[X \mid \mathcal{H}]$ satisfy the same defining identities. Proposition 11.5 proves their almost-sure equality. $\square$

Proposition 13.4 Let $1 \le p < \infty$, and suppose that $X \in L^p$. Then

$$\|\mathbb{E}[X \mid \mathcal{G}]\|_p \le \|X\|_p.$$

For $p = \infty$, the same conclusion holds with the essential supremum norm.

Proof. For $1 \le p < \infty$, apply Theorem 13.1 to the convex function $\varphi(x) = |x|^p$:

$$|\mathbb{E}[X \mid \mathcal{G}]|^p \le \mathbb{E}[|X|^p \mid \mathcal{G}].$$

Taking expectations and using Proposition 13.3 gives

$$\mathbb{E}\left[|\mathbb{E}[X \mid \mathcal{G}]|^p\right] \le \mathbb{E}[|X|^p].$$

Taking pth roots proves the claim. If $X \in L^\infty$, then $-\|X\|_\infty \le X \le \|X\|_\infty$ almost surely. Positivity and linearity preserve these inequalities under conditional expectation, which proves the L^∞ assertion. $\square$

Martingales

Definition 13.5 A **filtration** is an increasing sequence of σ-fields

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots .$$

A stochastic process $(X_n)_{n\geq 0}$ is **adapted** to $(\mathcal{F}_n)_{n\geq 0}$ if X_n is $\mathcal{F}_n$-measurable for every n. Thus $\mathcal{F}_n$ represents the information available by time n, and adaptedness rules out dependence on future information.

Definition 13.6 A process $(X_n)_{n\geq 0}$ is a **martingale with respect to $(\mathcal{F}_n)_{n\geq 0}$** if the following hold:

1. $\mathbb{E}[|X_n|] < \infty$ for every n.
2. $(X_n)_{n\geq 0}$ is adapted to $(\mathcal{F}_n)_{n\geq 0}$.
3. the "fair game condition"

   $$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n \quad \text{almost surely}$$

   holds for every n.

Definition 13.7 An integrable adapted process $(X_n)_{n\geq 0}$ is a **supermartingale** if

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n \quad \text{almost surely}$$

for every n. It is a **submartingale** if the reverse inequality holds.

Remark 13.8 A submartingale has conditionally nonnegative expected increments, while a supermartingale has conditionally nonpositive expected increments. The terminology refers to the inequality imposed on conditional expectations, not to the visual direction suggested by the prefixes.

Proposition 13.9 Let $Y_1, Y_2, \ldots$ be independent and identically distributed random vectors, each uniformly distributed on the ball $B(0,r) \subseteq \mathbb{R}^d$, and set

$$X_n = \sum_{j=1}^{n} Y_j \qquad \text{and} \qquad \mathcal{F}_n = \sigma(Y_1, \ldots, Y_n).$$

If $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ is subharmonic and $\mathbb{E}[|\varphi(X_n)|] < \infty$ for every n, then $(\varphi(X_n))_{n\geq 0}$ is a submartingale.

Proof. The mean value inequality for a subharmonic function says that its value at the center of a ball is at most its average over that ball. Since Y_{n+1} is independent of $\mathcal{F}_n$,

$$\mathbb{E}[\varphi(X_{n+1}) \mid \mathcal{F}_n] = \frac{1}{|B(0,r)|} \int_{B(0,r)} \varphi(X_n + y)\,\mathrm{d}y \geq \varphi(X_n).$$

Adaptedness and integrability complete the verification. □

Proposition 13.10 If $(X_n)_{n\geq 0}$ is a martingale and φ is convex, with $\varphi(X_n)$ integrable for every n, then $(\varphi(X_n))_{n\geq 0}$ is a submartingale.

Proof. The conditional Jensen inequality gives

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

□

Lemma 13.11 If $(X_n)_{n\geq 0}$ is a martingale and $m \geq n$, then

$$\mathbb{E}[X_m \mid \mathcal{F}_n] = X_n \quad \text{almost surely.}$$

Proof. Proceed by induction on $m-n$. The case $m = n$ follows from $\mathcal{F}_n$-measurability. If $m > n$, the tower property and the martingale condition give

$$\mathbb{E}[X_m \mid \mathcal{F}_n] = \mathbb{E}[\mathbb{E}[X_m \mid \mathcal{F}_{m-1}] \mid \mathcal{F}_n] = \mathbb{E}[X_{m-1} \mid \mathcal{F}_n] = X_n,$$

where the last equality is the induction hypothesis. □

Examples of Martingales

1. **Random walks.** Let $Y_1, Y_2, \ldots$ be independent and identically distributed integrable random variables with mean zero. Define $X_n = Y_1 + \cdots + Y_n$ and $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n + \mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = X_n.$$

2. **Multiplicative models.** Let $Y_1, Y_2, \ldots$ be independent and identically distributed with $\mathbb{E}[Y_n] = 1$. The variable Y_n may be interpreted as the one-period growth factor of an asset. Let X_0 be integrable and independent of the sequence, and define

$$X_n = X_0 Y_1 \cdots Y_n \qquad \text{and} \qquad \mathcal{F}_n = \sigma(X_0, Y_1, \ldots, Y_n).$$

   Then $(X_n)_{n\geq 0}$ is a martingale because

$$\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n \mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = X_n \mathbb{E}[Y_{n+1}] = X_n.$$

3. **A quadratic random walk martingale.** Let $Y_1, Y_2, \ldots$ be independent and identically distributed with mean zero and variance one. Define $S_n = Y_1 + \cdots + Y_n$, $X_n = S_n^2 - n$, and $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. Then

$$\begin{aligned}\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[(S_n + Y_{n+1})^2 - (n+1) \mid \mathcal{F}_n] \\ &= S_n^2 + 2S_n \mathbb{E}[Y_{n+1}] + \mathbb{E}[Y_{n+1}^2] - n - 1 \\ &= S_n^2 - n = X_n.\end{aligned}$$

4. **Branching processes.** Galton–Watson processes were introduced to model the survival of family names and now serve as a fundamental model of population growth. Let μ be a probability distribution on $\{0, 1, 2, \ldots\}$, and let $\xi_{n,j}$, for $n \geq 0$ and $j \geq 1$, be independent random variables with distribution μ. The variable $\xi_{n,j}$ is the number of offspring produced by the jth individual in generation n.

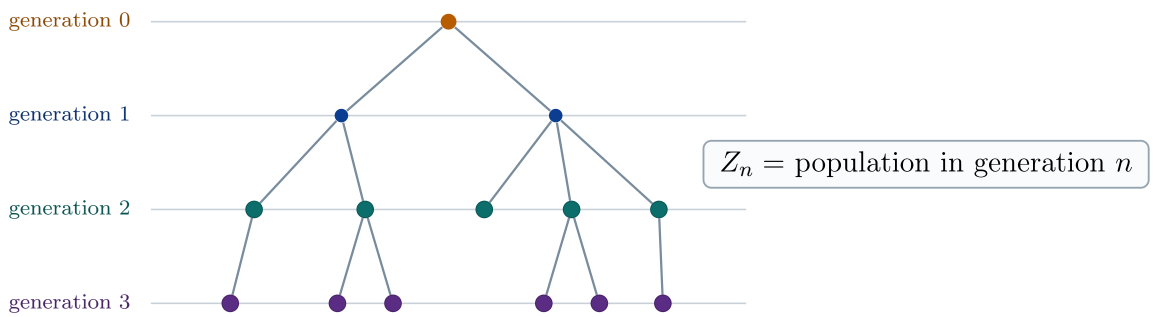

FIGURE 16

Starting from $Z_0 = 1$, define the population recursively by

$$Z_{n+1} = \sum_{j=1}^{Z_n} \xi_{n,j}.$$

Figure 16 shows one possible family tree. Let $m = \mathbb{E}[\xi_{0,1}] \in (0, \infty)$, called the Malthusian parameter, and let

$$\mathcal{F}_n = \sigma(\xi_{k,j} : 0 \le k < n,\ j \ge 1).$$

Then Z_n is $\mathcal{F}_n$-measurable, and independence gives

$$\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] = \sum_{j=1}^{\infty} \mathbb{1}_{\{Z_n \ge j\}} \mathbb{E}[\xi_{n,j} \mid \mathcal{F}_n] = mZ_n.$$

Therefore $M_n = Z_n/m^n$ is a nonnegative martingale.

Theorem 13.12 (Nonnegative martingale convergence theorem) If $(X_n)_{n\ge0}$ is a nonnegative martingale, then there is a finite random variable X_∞ such that

$$X_n \longrightarrow X_\infty \quad \text{almost surely.}$$

The proof rests on the **upcrossing inequality**, which we will prove later: a path that fails to converge must cross some fixed interval infinitely many times, whereas a nonnegative martingale has only finitely many such upcrossings almost surely.

Corollary 13.13 For the Galton–Watson process above, suppose that $m = 1$ and $\mu \ne \delta_1$. Then $Z_n = 0$ eventually almost surely.

Proof. When $m = 1$, the population process $(Z_n)_{n\geq 0}$ itself is a nonnegative martingale. By Theorem 13.12, it converges almost surely to a finite limit. Since every Z_n is integer valued, each convergent path is eventually constant.

Fix $k \geq 1$, and let q_k be the probability that the total number of offspring of k individuals is exactly k. The assumptions $\mathbb{E}[\xi_{0,1}] = 1$ and $\mu \neq \delta_1$ imply $q_k < 1$. Conditional on $Z_N = k$, independence between generations shows that the probability of remaining equal to k for all later generations is bounded by q_k^r for every positive integer r, and is therefore zero. Taking a countable union over N and $k \geq 1$ shows that the eventual constant cannot be positive. Hence it must be 0. □

Martingale Transforms and Stopping Times

Suppose that $(X_n)_{n\geq 0}$ records the cumulative winnings in a fair game. A betting strategy may choose the stake for the next round using all information currently available, but it may not depend on the outcome of that round. This is the predictability condition in the following definition.

Definition 13.14 Let $(X_n, \mathcal{F}_n)_{n\geq 0}$ be a martingale, and let $(H_n)_{n\geq 0}$ be an adapted process. The **martingale transform** of X by H is the process $H \cdot X$ defined by

$$(H \cdot X)_0 = 0 \qquad \text{and} \qquad (H \cdot X)_n = \sum_{j=0}^{n-1} H_j (X_{j+1} - X_j), \quad n \geq 1.$$

Proposition 13.15 If every H_n is bounded, then $H \cdot X$ is a martingale.

Proof. Write $M = H \cdot X$. Each M_n is integrable and $\mathcal{F}_n$-measurable. Moreover,

$$\begin{aligned} \mathbb{E}[M_{n+1} \mid \mathcal{F}_n] &= \mathbb{E}[M_n + H_n(X_{n+1} - X_n) \mid \mathcal{F}_n] \\ &= M_n + H_n\big(\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] - X_n\big) = M_n. \end{aligned}$$

□

Definition 13.16 A random variable T with values in $\{0, 1, 2, \dots\} \cup \{\infty\}$ is a **stopping time** with respect to $(\mathcal{F}_n)_{n\geq 0}$ if

$$\{T \leq n\} \in \mathcal{F}_n \quad \text{for every } n \geq 0.$$

Equivalently, $\{T = n\} \in \mathcal{F}_n$ for every $n \geq 0$.

Example 13.17 First hitting times, such as $T = \inf\{n : X_n \geq c\}$, are stopping times because we can determine by time n whether the level has been reached. By contrast, the time at which a process attains its ultimate maximum is generally not a stopping time: deciding that the current value is the ultimate maximum requires knowledge of the future. The same obstruction occurs when one tries to stop one step before an unpredictable event.

Theorem 13.18 (Stopped martingale theorem) If $(X_n, \mathcal{F}_n)_{n\geq 0}$ is a martingale and T is a stopping time, then the stopped process $(X_{T\wedge n})_{n\geq 0}$ is a martingale.

Proof. For $j \geq 0$, set $H_j = \mathbb{1}_{\{T>j\}}$. Since

$$\{T > j\} = \{T \leq j\}^{\mathsf{c}} \in \mathcal{F}_j,$$

the process H is adapted and bounded. The telescoping identity

$$X_{T\wedge n} = X_0 + \sum_{j=0}^{n-1} \mathbb{1}_{\{T>j\}}(X_{j+1} - X_j)$$

expresses the stopped process as $X_0 + H \bullet X$. Proposition 13.15 proves the claim. □

Corollary 13.19 (Optional stopping theorem) If $(X_n, \mathcal{F}_n)_{n\geq 0}$ is a martingale and $T \leq b$ almost surely for some finite b, then

$$\mathbb{E}[X_T] = \mathbb{E}[X_0].$$

Proof. By Theorem 13.18, the stopped process is a martingale. Since $X_{T\wedge b} = X_T$ almost surely,

$$\mathbb{E}[X_T] = \mathbb{E}[X_{T\wedge b}] = \mathbb{E}[X_0].$$

□

Lecture 14 — Thursday, February 21, 2019

Upcrossings quantify how often a sequence oscillates between two fixed levels. Figure 17 illustrates the successive entrance and exit times used below.

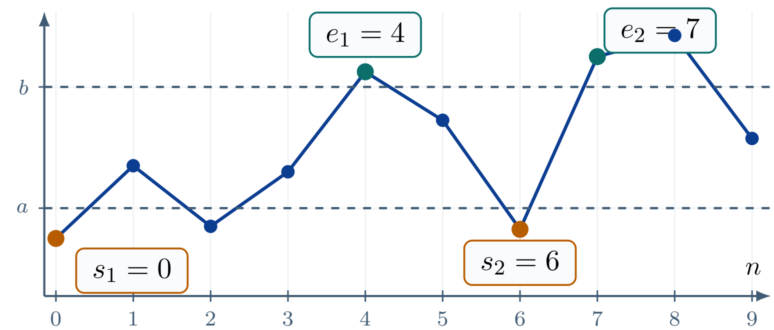

FIGURE 17

Definition 14.1 Let $(x_n)_{n\geq 0}$ be a real sequence, and fix $a < b$. With the convention $\inf \varnothing = \infty$, define

$$s_1 = \inf\{n \geq 0 : x_n \leq a\}, \qquad e_1 = \inf\{n > s_1 : x_n \geq b\},$$
$$s_k = \inf\{n > e_{k-1} : x_n \leq a\}, \qquad e_k = \inf\{n > s_k : x_n \geq b\}, \quad k \geq 2.$$

The pair (s_k, e_k) is the ***k*th upcrossing** of $[a, b]$. The number of completed upcrossings by time n is

$$U_n(a, b) = \max\{k \geq 0 : e_k \leq n\},$$

and the total number of upcrossings is

$$U(a, b) = \lim_{n\to\infty} U_n(a, b).$$

Proposition 14.2 For a real sequence $(x_n)_{n\geq 0}$, the following statements are equivalent:

1. The sequence has a limit in the extended real line $[-\infty, \infty]$.
2. For every pair $a < b$ of real numbers, $U(a, b) < \infty$.

3. For every pair $a < b$ of rational numbers, $U(a, b) < \infty$.

Proof. If (x_n) has an extended-real limit, it eventually remains on one side of at least one endpoint of every interval $[a, b]$, so every upcrossing count is finite. Thus property (1) implies property (2), and property (2) immediately implies property (3).

Conversely, if (x_n) has no extended-real limit, then

$$\liminf_{n\to\infty} x_n < \limsup_{n\to\infty} x_n.$$

Choose rational numbers $a < b$ strictly between these two quantities. The sequence falls below a and later rises above b infinitely often, so $U(a, b) = \infty$. This contradicts property (3). □

Proposition 14.3 (Doob's upcrossing inequality) Let $(X_n, \mathcal{F}_n)_{n\geq 0}$ be a nonnegative martingale, and let $0 < a < b$. Then

$$\mathbb{E}[U_n(a, b)] \leq \frac{a}{b-a} \quad \text{for every } n \geq 0.$$

Consequently,

$$\mathbb{E}[U(a, b)] \leq \frac{a}{b-a}.$$

Proof. Apply Definition 14.1 pathwise to (X_n). Consider the strategy

$$H_j = \mathbb{1}_{\{s_k \leq j < e_k \text{ for some } k\}},$$

which holds one unit of the asset after a purchase at or below a and until the subsequent sale at or above b. The decision H_j depends only on $X_0, \ldots, X_j$, so it is $\mathcal{F}_j$-measurable.

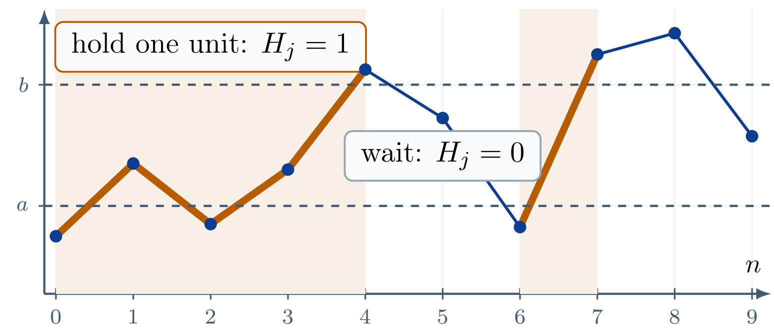

FIGURE 18

The transform $M = H \cdot X$ is a martingale by Proposition 13.15. Every completed transaction, illustrated in Figure 18, earns at least $b-a$. At time n, there may also be one unfinished transaction; because its purchase price is at most a and $X_n \geq 0$, its gain is at least $-a$. Therefore

$$M_n \geq (b-a)U_n(a,b) - a.$$

Taking expectations and using $\mathbb{E}[M_n] = \mathbb{E}[M_0] = 0$ gives

$$0 \geq (b-a)\mathbb{E}[U_n(a,b)] - a.$$

The first assertion follows. Since $U_n(a,b) \uparrow U(a,b)$, the monotone convergence theorem gives the second. □

Proof of Theorem 13.12. For every pair of positive rational numbers $a < b$, Proposition 14.3 gives $\mathbb{E}[U(a,b)] < \infty$, and hence $U(a,b) < \infty$ almost surely. Since there are only countably many such rational pairs, all these upcrossing counts are finite on one event of probability one. Proposition 14.2 then shows that X_n converges on this event to a limit $X_\infty \in [0,\infty]$.

Fatou's lemma gives

$$\mathbb{E}[X_\infty] \leq \liminf_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X_0] < \infty.$$

Therefore $X_\infty < \infty$ almost surely. □

Example 14.4 For the critical Galton–Watson process in Example (4), $\mathbb{E}[Z_n] = 1$ for every n, but Corollary 13.13 gives $Z_n \to 0$ almost surely. This example shows that almost-sure convergence need not imply convergence of expectations.

Proposition 14.5 Let $Y_1, Y_2, \dots$ be independent and identically distributed random variables such that

$$Y_1 \geq -b \quad \text{almost surely}, \qquad \mathbb{E}[Y_1] = 0, \qquad \text{and} \qquad \mathrm{Var}(Y_1) \in (0, \infty].$$

For $c > 0$, set $X_0 = c$ and $X_n = c + Y_1 + \cdots + Y_n$. Then

$$\mathbb{P}(X_n \leq 0 \text{ for some } n) = 1.$$

Proof. Let $T = \inf\{n \geq 0 : X_n \leq 0\}$. By Theorem 13.18, $(X_{T\wedge n})_{n\geq 0}$ is a martingale. The lower bound on the increments gives

$$X_{T\wedge n} \geq -b \quad \text{almost surely}.$$

Thus $(X_{T\wedge n} + b)_{n\geq 0}$ is a nonnegative martingale and converges almost surely by Theorem 13.12.

On the event $\{T = \infty\}$, the original random walk (X_n) therefore converges, which would force $Y_n = X_n - X_{n-1} \to 0$. Since Y_1 is nondegenerate, there is an $\varepsilon > 0$ such that $\mathbb{P}(|Y_1| > \varepsilon) > 0$. Independence and the second Borel–Cantelli lemma imply that $|Y_n| > \varepsilon$ occurs infinitely often almost surely. Hence $\mathbb{P}(T = \infty) = 0$. □

Lecture 15 — Thursday, February 28, 2019

The nonnegative martingale convergence theorem admits a useful extension: a supermartingale (X_n) converges almost surely to a finite limit whenever

$$\sup_{n\geq 0} \mathbb{E}[X_n^-] < \infty.$$

The same upcrossing method proves this stronger form after applying the appropriate supermartingale upcrossing inequality.

Martingales can exhibit several distinct types of long-run behavior:

1. A nonnegative martingale converges almost surely to a finite limit by Theorem 13.12.
2. Without a lower bound, a martingale may converge to $+\infty$ almost surely; the negative part

can then carry increasingly rare but increasingly large compensating values so that every finite-time expectation remains unchanged.

3. A simple symmetric random walk visits every integer infinitely often. In particular,

$$\limsup_{n\to\infty} X_n = \infty \qquad \text{and} \qquad \liminf_{n\to\infty} X_n = -\infty \quad \text{almost surely.}$$

Theorem 15.1 Let $(X_n, \mathcal{F}_n)_{n\geq 0}$ be a martingale whose increments satisfy $|X_{n+1} - X_n| \leq b$ almost surely for every n. Define

$$C = \{X_n \text{ converges to a finite limit}\},$$
$$O = \left\{\liminf_{n\to\infty} X_n = -\infty \text{ and } \limsup_{n\to\infty} X_n = \infty\right\}.$$

Then $\mathbb{P}(C \cup O) = 1$.

Proof. Subtracting X_0, we may assume that $X_0 = 0$. For $K > 0$, let

$$T_K = \inf\{n \geq 0 : X_n \leq -K\}.$$

The assumption that the increments are bounded gives $X_{T_K \wedge n} \geq -K - b$. The stopping construction and the possible overshoot below $-K$ are shown in Figure 19.

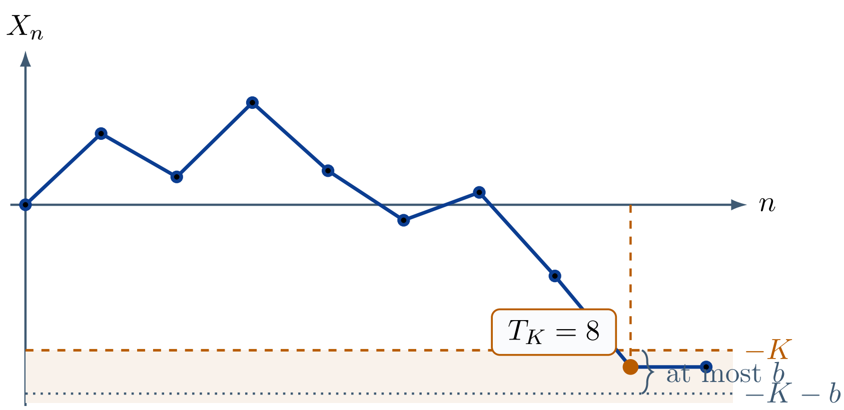

FIGURE 19

The shifted stopped process $(X_{T_K \wedge n} + K + b)_{n\geq 0}$ is a nonnegative martingale, so Theorem 13.12 shows that it converges almost surely. On $\{T_K = \infty\}$, it differs from X_n by a constant, so X_n

converges there. Taking the union over positive integers K, we find that X_n converges almost surely on

$$A_- := \left\{ \inf_{n \geq 0} X_n > -\infty \right\}.$$

Applying the same argument to $(-X_n)$ shows that X_n converges almost surely on

$$A_+ := \left\{ \sup_{n \geq 0} X_n < \infty \right\}.$$

Apart from a null set, $A_- \cup A_+ \subseteq C$. Hence, on C^{c}, both the infimum is $-\infty$ and the supremum is $+\infty$, which is exactly the event O. Thus $C^{\mathsf{c}} \subseteq O$ almost surely, and the conclusion follows. □

Corollary 15.2 If (X_n) is a martingale with $X_{n+1} - X_n \geq -b$ almost surely for every n, then X_n converges to a finite limit almost surely on the event

$$\left\{ \inf_{n \geq 0} X_n > -\infty \right\}.$$

Proof. The argument based on lower stopping in the proof of Theorem 15.1 uses only the lower bound on the increments, so it applies without change. □

Theorem 15.3 (Conditional Borel–Cantelli lemma) Let $(\mathcal{F}_n)_{n \geq 0}$ be a filtration, and let $A_n \in \mathcal{F}_n$ for every $n \geq 1$. Then, up to a null set,

$$\{A_n \text{ occurs infinitely often}\} = \left\{ \sum_{n=1}^{\infty} \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) = \infty \right\}.$$

Unlike the second classical Borel–Cantelli lemma, Theorem 15.3 does not assume independence. When the events are independent and $\mathcal{F}_n = \sigma(A_1, \dots, A_n)$, the conditional probabilities reduce to the constants $\mathbb{P}(A_n)$, recovering the classical result.

Lecture 16 — Tuesday, March 05, 2019

Stopping also raises a convergence question. If $T < \infty$ almost surely, then $X_{T\wedge n} \to X_T$ almost surely. To pass from

$$\mathbb{E}[X_{T\wedge n}] = \mathbb{E}[X_0]$$

to $\mathbb{E}[X_T] = \mathbb{E}[X_0]$, we need a condition that permits the interchange of limit and expectation. Uniform integrability provides precisely that condition. More generally, it characterizes when convergence in probability strengthens to convergence in L^1:

$$M_n \longrightarrow M \quad \text{in probability} \Longrightarrow \mathbb{E}[|M_n - M|] \longrightarrow 0.$$

Definition 16.1 A collection of integrable random variables $\{X_i\}_{i\in I}$ is **uniformly integrable** if, for every $\varepsilon > 0$, there exists $b > 0$ such that

$$\mathbb{E}\left[|X_i| \mathbb{1}_{\{|X_i|>b\}}\right] < \varepsilon$$

for every $i \in I$.

Remark 16.2 Every finite collection of integrable random variables is uniformly integrable. For an infinite collection, uniform integrability requires a single tail threshold that works for every member of the collection.

Theorem 16.3 Suppose that $X_n \to X$ in probability. The following statements are equivalent:

1. The sequence $(X_n)_{n\geq 1}$ is uniformly integrable.
2. The sequence converges to X in L^1:

$$\mathbb{E}[|X_n - X|] \longrightarrow 0.$$

3. All the variables are integrable and

$$\mathbb{E}[|X_n|] \longrightarrow \mathbb{E}[|X|].$$

Proof. (1) *implies* (2). The direct tail function

$$x \mapsto |x|\mathbb{1}_{\{|x|\geq b\}}$$

is discontinuous, with the two jumps displayed in Figure 20. It is more convenient to split each variable into a clipped part and a continuous tail remainder.

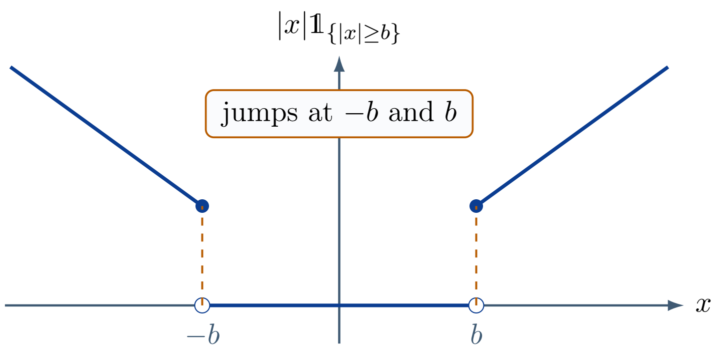

FIGURE 20

Define

$$\varphi(x) := \begin{cases} x+b, & x \leq -b, \\ x-b, & x \geq b, \\ 0, & \text{otherwise}, \end{cases}$$

which is continuous. Set $\eta(x) := x - \varphi(x)$. Then

$$\eta(x) = \begin{cases} -b, & x \leq -b, \\ b, & x \geq b, \\ x, & \text{otherwise}, \end{cases}$$

so η is bounded by b. The resulting clipping function is shown in Figure 21.

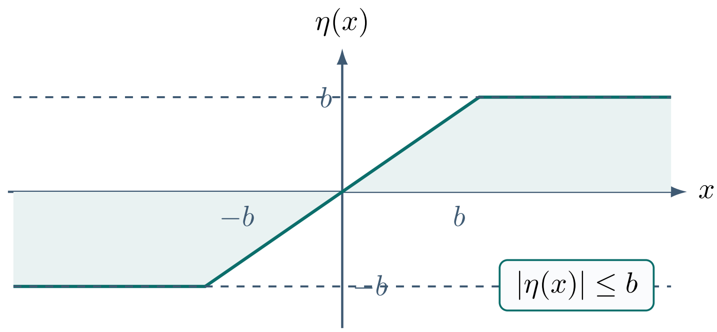

FIGURE 21

The sequence (X_n) is uniformly bounded in L^1, and convergence in probability therefore implies that X is integrable: extract an almost surely convergent subsequence and apply Fatou's lemma. The same subsequence argument gives

$$\mathbb{E}[|\varphi(X)|] \le \sup_n \mathbb{E}\left[|X_n| \mathbb{1}_{\{|X_n| \ge b\}}\right].$$

The triangle inequality gives

$$\mathbb{E}\left[|X_n - X|\right] \le \mathbb{E}\left[|\varphi(X_n)|\right] + \mathbb{E}\left[|\varphi(X)|\right] + \mathbb{E}\left[|\eta(X_n) - \eta(X)|\right].$$

Now $\varphi(X_n) = \pm(|X_n| - b)$ on $\{|X_n| \ge b\}$, so

$$|\varphi(X_n)| \le |X_n| \mathbb{1}_{\{|X_n| \ge b\}} \qquad \text{and} \qquad |\varphi(X)| \le |X| \mathbb{1}_{\{|X| \ge b\}}.$$

Given $\varepsilon > 0$, uniform integrability allows us to choose b so that the first two expectations are each less than $\varepsilon/3$, uniformly in n. Also, $\eta(X_n) \to \eta(X)$ in probability and $|\eta(X_n) - \eta(X)| \le 2b$. Boundedness upgrades convergence in probability to convergence in L^1, so

$$\mathbb{E}\left[|\eta(X_n) - \eta(X)|\right] \to 0.$$

Hence $\mathbb{E}[|X_n - X|] \to 0$.

(2) *implies* (3). The reverse triangle inequality gives

$$\left|\mathbb{E}\left[|X_n|\right] - \mathbb{E}\left[|X|\right]\right| \le \mathbb{E}\left[|X_n - X|\right] \to 0.$$

(3) *implies* (1). We can approximate the tail function by the continuous function

$$\psi_b(x) := \begin{cases} |x|, & |x| \geq b, \\ 0, & |x| \leq b-1, \\ \text{linear}, & x \in [b-1, b], \\ \text{linear}, & x \in [-b, -b+1]. \end{cases}$$

which is illustrated in Figure 22. Notice that ψ_b is continuous but unbounded, so convergence in probability alone does not justify passing its expectation to the limit. A subsequence argument avoids this difficulty.

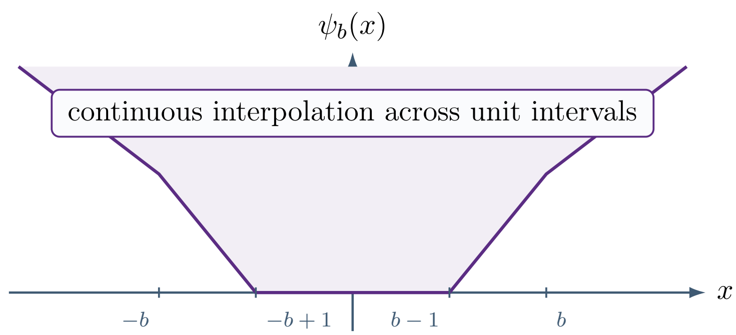

FIGURE 22

Take an arbitrary subsequence of (X_n). Since convergence in probability implies the subsequence principle, it has a further subsequence, denoted again by (X_n), that converges to X almost surely. Then $|X_n| \to |X|$ almost surely, and the hypothesis gives $\mathbb{E}[|X_n|] \to \mathbb{E}[|X|]$. The Scheffé theorem implies

$$\mathbb{E}\big[||X_n| - |X||\big] \longrightarrow 0.$$

Moreover,

$$|X_n - X| \leq ||X_n| - |X|| + 2|X|\mathbb{1}_{\{X_n X \leq 0\}}.$$

The second term converges almost surely to zero and is dominated by $2|X|$, so the dominated convergence theorem shows that this further subsequence converges to X in L^1. Since every subsequence has such a further subsequence, the entire sequence converges in L^1.

Finally, every L^1-convergent sequence is uniformly integrable. Indeed, the tail of X is small for a sufficiently large cutoff, $\mathbb{E}[|X_n - X|]$ is uniformly small for all sufficiently large n, and the remaining finite initial collection is uniformly integrable. This proves property (1). □

Lecture 17 — Thursday, March 07, 2019

Lemma 17.1 If $\{\mathcal{G}_i\}_{i\in I}$ is a collection of sub-σ-fields and X is integrable, then

$$\{\mathbb{E}[X \mid \mathcal{G}_i]\}_{i\in I}$$

is uniformly integrable.

Proof. Let $Y_i = \mathbb{E}[X \mid \mathcal{G}_i]$. By the conditional Jensen inequality,

$$|Y_i| \leq \mathbb{E}[|X| \mid \mathcal{G}_i].$$

With $A_i = \{\mathbb{E}[|X| \mid \mathcal{G}_i] > b\}$, it follows that

$$\mathbb{E}\left[|Y_i|\mathbb{1}_{\{|Y_i|>b\}}\right] \leq \mathbb{E}[|X|\mathbb{1}_{A_i}].$$

Markov's inequality gives

$$\mathbb{P}(A_i) \leq \frac{\mathbb{E}[|X|]}{b}.$$

Since X is integrable, absolute continuity of the integral gives

$$\sup_{i\in I} \mathbb{E}[|X|\mathbb{1}_{A_i}] \longrightarrow 0 \quad \text{as } b \to \infty,$$

which proves uniform integrability. $\square$

Lemma 17.2 If $X_n \to X$ in L^1, then, for every $A \in \mathcal{F}$,

$$\mathbb{E}[X_n \mathbb{1}_A] \longrightarrow \mathbb{E}[X\mathbb{1}_A].$$

Proof.

$$|\mathbb{E}\left[(X_n - X)\mathbb{1}_A\right]| \leq \mathbb{E}\left[|X_n - X|\mathbb{1}_A\right] \leq \mathbb{E}\left[|X_n - X|\right] \to 0.$$

$\square$

Theorem 17.3 For a martingale $(X_n, \mathcal{F}_n)_{n\geq 0}$, the following statements are equivalent:

1. The collection $\{X_n : n \geq 0\}$ is uniformly integrable.
2. There is an integrable X such that $X_n \to X$ almost surely and in L^1.
3. There is an integrable X such that $X_n \to X$ in L^1.
4. There is an integrable X such that

$$X_n = \mathbb{E}[X \mid \mathcal{F}_n] \quad \text{almost surely for every } n.$$

(In general, a martingale given by $X_n = \mathbb{E}[X \mid \mathcal{F}_n]$ is called a *Doob martingale*.)

Proof. (1) *implies* (2). Uniform integrability implies $\sup_n \mathbb{E}[|X_n|] < \infty$, and hence $\sup_n \mathbb{E}[X_n^-] < \infty$. The supermartingale extension of the convergence theorem therefore gives an almost-sure finite limit X. Theorem 16.3 then upgrades this convergence to convergence in L^1.

(2) *implies* (3) This is immediate.

(3) *implies* (4). Assume that $X_m \to X$ in L^1. Fix n and $A \in \mathcal{F}_n$. For $m \geq n$, the multistep martingale property gives

$$\mathbb{E}[X_m \mathbb{1}_A] = \mathbb{E}[X_n \mathbb{1}_A].$$

Letting $m \to \infty$ and applying Lemma 17.2 gives

$$\mathbb{E}[X_n \mathbb{1}_A] = \mathbb{E}[X \mathbb{1}_A].$$

Because X_n is $\mathcal{F}_n$-measurable, this is precisely the defining property of $\mathbb{E}[X \mid \mathcal{F}_n]$.

(4) *implies* (1). This follows from Lemma 17.1 applied to the collection $(\mathcal{F}_n)_{n\geq 0}$. □

The limit of a martingale generated by successive conditional expectations can be identified explicitly.

Theorem 17.4 (Lévy's upward theorem) Let $(\mathcal{F}_n)_{n\geq 0}$ be a filtration, let X be integrable, and define

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{n\geq 0} \mathcal{F}_n\right).$$

Then

$$\mathbb{E}[X \mid \mathcal{F}_n] \longrightarrow \mathbb{E}[X \mid \mathcal{F}_\infty]$$

almost surely and in L^1.

Proof. Let $M_n = \mathbb{E}[X \mid \mathcal{F}_n]$. Lemma 17.1 and Theorem 17.3 show that $M_n \to Y$ almost surely and in L^1 for some integrable Y. The limit Y is $\mathcal{F}_\infty$-measurable. If $A \in \bigcup_n \mathcal{F}_n$, then $A \in \mathcal{F}_N$ for some N, and, for $n \geq N$,

$$\mathbb{E}[M_n \mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A].$$

Passing to the limit in L^1 gives $\mathbb{E}[Y\mathbb{1}_A] = \mathbb{E}[X\mathbb{1}_A]$. The collection $\bigcup_n \mathcal{F}_n$ is a π-system, so Dynkin's π-λ theorem extends this identity to every $A \in \mathcal{F}_\infty$. Consequently,

$$Y = \mathbb{E}[X \mid \mathcal{F}_\infty] \quad \text{almost surely.}$$

□

Corollary 17.5 (Lévy's zero–one law) If $A \in \mathcal{F}_\infty$, then

$$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] \longrightarrow \mathbb{1}_A$$

almost surely and in L^1.

Proof. Apply Theorem 17.4 with $X = \mathbb{1}_A$. Since $\mathbb{1}_A$ is $\mathcal{F}_\infty$-measurable,

$$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_\infty] = \mathbb{1}_A.$$

□

Recall the Kolmogorov zero–one law from STA2111H. The proof was rather involved back then; however, with the power of martingales, it is now a simple corollary of the Lévy zero–one law:

Corollary 17.6 (Kolmogorov's zero–one law) Let $X_1, X_2, \ldots$ be independent random variables, and let

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots)$$

be their tail σ-field. If $A \in \mathcal{T}$, then $\mathbb{P}(A) \in \{0, 1\}$.

Proof. Set $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Since $A \in \mathcal{T}$, the event A is independent of $\mathcal{F}_n$ for every n, and hence

$$\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}_n] = \mathbb{P}(A) \quad \text{almost surely.}$$

On the other hand, $A \in \mathcal{F}_\infty = \sigma(X_1, X_2, \ldots)$, so Corollary 17.5 shows that the left-hand side converges almost surely to $\mathbb{1}_A$. Therefore $\mathbb{1}_A = \mathbb{P}(A)$ almost surely. An indicator can be almost surely constant only with value 0 or 1. $\square$

Lecture 18 — Tuesday, March 12, 2019

3. Martingale Applications and Inequalities

Applications of Martingales

Recall the first Borel–Cantelli lemma, which states that for any sequence of events $(A_n)_{n \geq 1}$,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \Longrightarrow \mathbb{P}(A_n \text{ occurs infinitely often}) = 0.$$

Indeed, if

$$N = \sum_{n=1}^{\infty} \mathbb{1}_{A_n}$$

is the number of events that occur, then

$$\mathbb{E}[N] = \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$$

implies $N < \infty$ almost surely.

The second Borel–Cantelli lemma supplies the converse under independence:

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \Longrightarrow \mathbb{P}(A_n \text{ occurs infinitely often}) = 1.$$

Independence is essential for this classical converse. For example, if $A_n = A_1$ for every n, then $N = \infty$ on A_1 and $N = 0$ on A_1^{c}, regardless of whether $\sum_n \mathbb{P}(A_n)$ diverges.

The conditional Borel–Cantelli lemma replaces independence by conditional probabilities. Its proof is a direct application of the dichotomy of Theorem 15.1:

Proof of Theorem 15.3. Set

$$S_n = \sum_{i=1}^{n} \mathbb{1}_{A_i}, \qquad B_n = \sum_{i=1}^{n} \mathbb{P}(A_i \mid \mathcal{F}_{i-1}), \qquad \text{and} \qquad M_n = S_n - B_n.$$

The process $(M_n, \mathcal{F}_n)_{n \geq 0}$ is a martingale because

$$\mathbb{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] = \mathbb{E}[\mathbb{1}_{A_n} \mid \mathcal{F}_{n-1}] - \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) = 0.$$

Moreover, $|M_n - M_{n-1}| \leq 1$. By Theorem 15.1, almost every path either has a finite limit or oscillates between $-\infty$ and $+\infty$.

If $M_n = S_n - B_n$ has a finite limit, the nondecreasing sequences (S_n) and (B_n) must either both have finite limits or both diverge to $+\infty$. If M_n oscillates, then $M_n \leq S_n$ implies $S_n \to \infty$, while $-M_n \leq B_n$ implies $B_n \to \infty$. Thus $S_n \to \infty$ if and only if $B_n \to \infty$, almost surely. Since $S_n \to \infty$ precisely when the events A_n occur infinitely often, the theorem follows. $\square$

Our second application is the Pólya urn. Begin with $R_0 = r$ red balls and $G_0 = g$ green balls, where r, g are positive integers. Fix a positive integer c. At each step, draw one ball uniformly from the urn, replace it, and add c new balls of the same color. If R_n and G_n are the numbers of red and green balls after n draws, then

$$R_n + G_n = r + g + cn.$$

Example 18.1 If $r = 6$, $g = 8$, and $c = 2$, a green draw followed by a red draw changes the composition as follows:

$$(R_0, G_0) = (6, 8), \qquad (R_1, G_1) = (6, 10), \qquad \text{and} \qquad (R_2, G_2) = (8, 10).$$

Proposition 18.2 Both colors are drawn infinitely often almost surely. In particular,

$$R_n \longrightarrow \infty \qquad \text{and} \qquad G_n \longrightarrow \infty \quad \text{almost surely.}$$

Proof. Let A_n be the event that the nth draw is green, and let $\mathcal{F}_n$ contain the first n draws. Since $G_{n-1} \geq g$,

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n \mid \mathcal{F}_{n-1}) = \sum_{n=1}^{\infty} \frac{G_{n-1}}{r + g + c(n-1)} \geq \sum_{n=1}^{\infty} \frac{g}{r + g + c(n-1)} = \infty.$$

The conditional Borel–Cantelli lemma shows that green is drawn infinitely often. The same argument with red in place of green completes the proof. □

Theorem 18.3 The green proportion

$$X_n = \frac{G_n}{R_n + G_n} = \frac{G_n}{r + g + cn}$$

is a bounded martingale and hence converges almost surely to a random variable $L \in [0, 1]$. Moreover,

$$L \sim \text{Beta}\left(\frac{g}{c}, \frac{r}{c}\right).$$

Proof. Let $\mathbb{1}_n^G$ and $\mathbb{1}_n^R$ be the indicators that the nth draw is green and red, respectively, and set

$$D_n = (R_n + G_n)(R_{n-1} + G_{n-1}).$$

Since $G_n = G_{n-1} + c\mathbb{1}_n^G$ and $R_n = R_{n-1} + c\mathbb{1}_n^R$, the original increment calculation becomes

$$X_n - X_{n-1} = \frac{G_n R_{n-1} - G_{n-1} R_n}{D_n} = \frac{c}{D_n}\left(R_{n-1}\mathbb{1}_n^G - G_{n-1}\mathbb{1}_n^R\right).$$

Because $D_n = (r+g+cn)(r+g+c(n-1))$ is deterministic and

$$\mathbb{E}[\mathbb{1}_n^G \mid \mathcal{F}_{n-1}] = \frac{G_{n-1}}{R_{n-1}+G_{n-1}} \qquad \text{and} \qquad \mathbb{E}[\mathbb{1}_n^R \mid \mathcal{F}_{n-1}] = \frac{R_{n-1}}{R_{n-1}+G_{n-1}},$$

we obtain

$$\mathbb{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] = \frac{c}{D_n}\left(R_{n-1}\frac{G_{n-1}}{R_{n-1}+G_{n-1}} - G_{n-1}\frac{R_{n-1}}{R_{n-1}+G_{n-1}}\right) = 0.$$

Thus (X_n) is a martingale. Since $0 \le X_n \le 1$, Theorem 13.12 gives an almost-sure limit $L \in [0,1]$.

To identify the limit law, set $a = g/c$ and $b = r/c$, and write

$$(q)_m = q(q+1)\cdots(q+m-1)$$

for the rising factorial. The probability of observing m green draws followed by ℓ red draws is

$$\begin{aligned} &\frac{g}{g+r}\frac{g+c}{g+r+c}\cdots\frac{g+c(m-1)}{g+r+c(m-1)} \\ &\qquad \times \frac{r}{g+r+cm}\frac{r+c}{g+r+c(m+1)}\cdots\frac{r+c(\ell-1)}{g+r+c(m+\ell-1)} = \frac{(a)_m(b)_\ell}{(a+b)_{m+\ell}}. \end{aligned}$$

This probability depends only on the numbers of green and red draws, not on their order. Consequently, if K_n is the number of green draws among the first n draws, then

$$\mathbb{P}(K_n = m) = \binom{n}{m}\frac{(a)_m(b)_{n-m}}{(a+b)_n}, \quad 0 \le m \le n. \tag{18.1}$$

For $c = 1$ and integer r, g, this is the factorial expression

$$\frac{n!}{m!(n-m)!}\frac{(g+m-1)!}{(g-1)!}\frac{(r+n-m-1)!}{(r-1)!}\frac{(g+r-1)!}{(g+r+n-1)!},$$

which is the **beta-binomial** probability mass function.

If $m/n \to x \in (0,1)$, the gamma function ratio asymptotic

$$\frac{\Gamma(t+q)}{\Gamma(t+1)} \sim t^{q-1} \quad \text{as } t \to \infty$$

applied to (18.1) gives

$$n\mathbb{P}(K_n = m) \longrightarrow \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}.$$

The corresponding Riemann sums show that K_n/n converges in distribution to $\mathrm{Beta}(a,b)$. Since

$$X_n = \frac{g + cK_n}{r + g + cn} \qquad \text{and} \qquad \left| X_n - \frac{K_n}{n} \right| = O\left(\frac{1}{n}\right),$$

the same distributional limit holds for X_n. But $X_n \to L$ almost surely, so the law of L is $\mathrm{Beta}(a,b) = \mathrm{Beta}(g/c, r/c)$. $\square$

Remark 18.4 When $c = 1$, the law $\mathrm{Beta}(g,r)$ also has an interpretation through order statistics. If $r + g - 1$ independent random variables are uniformly distributed on $[0,1]$, then their rth largest value has the same distribution as L.

Doob's Inequality

Doob's maximal inequality is the process-level analogue of Markov's inequality. For a nonnegative random variable W and $\lambda > 0$, Markov's inequality follows from

$$\mathbb{P}(W \geq \lambda) = \mathbb{E}[\mathbb{1}_{\{W \geq \lambda\}}] \leq \frac{1}{\lambda}\mathbb{E}[W\mathbb{1}_{\{W \geq \lambda\}}] \leq \frac{1}{\lambda}\mathbb{E}[W].$$

Doob's inequality controls the maximum of an entire submartingale by its terminal value.

Theorem 18.5 (Doob's L^1 maximal inequality) If $(Z_k, \mathcal{F}_k)_{k=0}^n$ is a nonnegative submartingale and $\lambda > 0$, then

$$\mathbb{P}\left(\max_{0 \leq k \leq n} Z_k \geq \lambda\right) \leq \frac{1}{\lambda}\mathbb{E}\left[Z_n \mathbb{1}_{\{\max_{0 \leq k \leq n} Z_k \geq \lambda\}}\right] \leq \frac{1}{\lambda}\mathbb{E}[Z_n].$$

Proof. Let

$$\tau = \inf\{0 \leq k \leq n : Z_k \geq \lambda\},$$

with $\tau = \infty$ if the set is empty. For every $j \leq n$, the event $\{\tau = j\}$ belongs to $\mathcal{F}_j$. The submartingale property gives

$$\mathbb{E}[Z_n \mathbb{1}_{\{\tau = j\}}] = \mathbb{E}[\mathbb{E}[Z_n \mid \mathcal{F}_j]\mathbb{1}_{\{\tau = j\}}] \geq \mathbb{E}[Z_j \mathbb{1}_{\{\tau = j\}}].$$

Therefore

$$\lambda \mathbb{P}\left(\max_{0 \leq k \leq n} Z_k \geq \lambda\right) = \lambda \sum_{j=0}^{n} \mathbb{P}(\tau = j)$$

$$
\begin{aligned}
&\le \sum_{j=0}^{n} \mathbb{E}[Z_j \mathbb{1}_{\{\tau=j\}}] \le \sum_{j=0}^{n} \mathbb{E}[Z_n \mathbb{1}_{\{\tau=j\}}] \\
&= \mathbb{E}\left[Z_n \mathbb{1}_{\{\max_{0\le k\le n} Z_k \ge \lambda\}}\right] \le \mathbb{E}[Z_n].
\end{aligned}
$$

□

Example 18.6 Let $S_n = \xi_1 + \cdots + \xi_n$, where the ξ_i are independent, centered, and square integrable. Then (S_n) is a martingale, and (S_n^2) is a nonnegative submartingale. Theorem 18.5 gives

$$
\mathbb{P}\left(\max_{k\le n} S_k^2 \ge \lambda\right) \le \frac{\mathbb{E}[S_n^2]}{\lambda} = \frac{\sum_{i=1}^n \mathrm{Var}(\xi_i)}{\lambda}.
$$

Equivalently, for $\alpha > 0$,

$$
\mathbb{P}\left(\max_{k\le n} |S_k| \ge \alpha\right) \le \frac{\sum_{i=1}^n \mathrm{Var}(\xi_i)}{\alpha^2}.
$$

This is Kolmogorov's maximal inequality!

For every $\theta > 0$ for which the relevant moment-generating functions are finite, $(e^{\theta S_n})$ is also a nonnegative submartingale. Hence

$$
\mathbb{P}\left(\max_{k\le n} S_k > c\right) \le e^{-\theta c}\mathbb{E}[e^{\theta S_n}] = e^{-\theta c}\prod_{i=1}^n \mathbb{E}[e^{\theta \xi_i}].
$$

If the increments are independent standard Gaussian random variables, the right-hand side is

$$
\exp\left(-\theta c + \frac{n\theta^2}{2}\right).
$$

Optimizing at $\theta = c/n$ yields the one-sided **Chernoff bound**

$$
\mathbb{P}\left(\max_{k\le n} S_k > c\right) \le e^{-c^2/(2n)}.
$$

Applying the same estimate to $(-S_k)$ and using a union bound gives

$$
\mathbb{P}\left(\max_{k\le n} |S_k| > c\right) \le 2e^{-c^2/(2n)}.
$$

Theorem 18.7 (Law of the iterated logarithm) Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed with mean zero and variance one, and set $S_n = \xi_1 + \cdots + \xi_n$. Then

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \qquad \text{and} \qquad \liminf_{n\to\infty} \frac{S_n}{\sqrt{n\log\log n}} = -\sqrt{2}$$

almost surely. In particular, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{n\log\log n}} < \sqrt{2} + \varepsilon\right) = 1.$$

The theorem also implies

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} = \infty \qquad \text{and} \qquad \liminf_{n\to\infty} \frac{S_n}{\sqrt{n}} = -\infty$$

almost surely, while $S_n/n^\alpha \to 0$ almost surely for every $\alpha > 1/2$.

Lecture 19 — Tuesday, March 19, 2019

The L^1 maximal inequality has an L^p counterpart for every $p > 1$.

Theorem 19.1 (Doob's L^p maximal inequality) Let $(X_k, \mathcal{F}_k)_{k=0}^n$ be a submartingale, let $p > 1$, and define

$$\overline{X}_n = \max_{0\le k\le n} X_k^+.$$

Then

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

Proof. Fix $b > 0$. The layer-cake formula and Theorem 18.5 give

$$\begin{aligned}
\mathbb{E}[(\overline{X}_n \wedge b)^p] &= \int_0^b p y^{p-1} \mathbb{P}(\overline{X}_n \ge y)\,\mathrm{d}y \\
&\le p \int_0^b y^{p-2} \mathbb{E}[X_n^+ \mathbb{1}_{\{\overline{X}_n \ge y\}}]\,\mathrm{d}y \\
&= \frac{p}{p-1} \mathbb{E}\left[X_n^+ (\overline{X}_n \wedge b)^{p-1}\right].
\end{aligned}$$

Hölder's inequality now yields

$$\mathbb{E}[(\overline{X}_n \wedge b)^p] \le \frac{p}{p-1} \|X_n^+\|_p \mathbb{E}[(\overline{X}_n \wedge b)^p]^{(p-1)/p}.$$

If the expectation on the left is zero, the desired bound is immediate. Otherwise, division by its $(p-1)/p$-power gives

$$\mathbb{E}[(\overline{X}_n \wedge b)^p] \le \left(\frac{p}{p-1}\right)^p \|X_n^+\|_p^p.$$

Letting $b \to \infty$ and applying the monotone convergence theorem proves the result. ☐

Several important consequences follow.

Proposition 19.2 Let $(X_n, \mathcal{F}_n)_{n\ge 0}$ be a martingale.

1. For every $p > 1$,

$$\mathbb{E}\left[\max_{0\le k\le n} |X_k|^p\right] \le \left(\frac{p}{p-1}\right)^p \mathbb{E}[|X_n|^p].$$

Proof. The process $(|X_n|)$ is a nonnegative submartingale by the conditional Jensen inequality. Apply Theorem 19.1 to this process. ☐

2. There is no analogous inequality with a universal constant when $p = 1$. To see this, let (S_n) be a simple symmetric random walk with $S_0 = 1$, and set

$$T := \inf\{n \ge 0 : S_n = 0\} \qquad \text{and} \qquad M_n := S_{n\wedge T}.$$

Then (M_n) is a nonnegative martingale and $\mathbb{E}[|M_n|] = 1$ for every n. Set $\overline{M}_n = \max_{k\le n} M_k$. The gambler's ruin calculation gives, for $m \ge 1$,

$$\mathbb{P}(\overline{M}_T \ge m) = \frac{1}{m}.$$

If Y is nonnegative and integer valued, then

$$Y = \sum_{j=1}^{\infty} \mathbb{1}_{\{Y\ge j\}} \qquad \text{and} \qquad \mathbb{E}[Y] = \sum_{j=1}^{\infty} \mathbb{P}(Y \ge j).$$

Applying this identity to $\overline{M}_T$ yields

$$\mathbb{E}[\overline{M}_T] = \sum_{m=1}^{\infty} \frac{1}{m} = \infty.$$

Since $\overline{M}_n \uparrow \overline{M}_T$ almost surely, the monotone convergence theorem gives

$$\mathbb{E}[\overline{M}_n] \longrightarrow \infty,$$

even though $\mathbb{E}[|M_n|] = 1$ for every n.

3. If $p > 1$ and $\sup_n \|X_n\|_p < \infty$, then there is an $X_\infty \in L^p$ such that $X_n \to X_\infty$ almost surely and in L^p.

Proof. Let $B = \sup_n \|X_n\|_p$. Since $\sup_n \mathbb{E}[|X_n|] \le B$, the martingale convergence theorem gives an almost-sure finite limit X_∞. Set $\overline{X}_n = \max_{k \le n} |X_k|$ and $\overline{X}_\infty = \sup_{k \ge 0} |X_k|$. By property (1),

$$\mathbb{E}[\overline{X}_n^p] \le \left(\frac{p}{p-1}\right)^p B^p.$$

The monotone convergence theorem gives

$$\mathbb{E}[\overline{X}_\infty^p] \le \left(\frac{p}{p-1}\right)^p B^p < \infty.$$

Since $|X_n - X_\infty|^p \le 2^p \overline{X}_\infty^p$, the dominated convergence theorem implies

$$\mathbb{E}[|X_n - X_\infty|^p] \longrightarrow 0.$$

□

Theorem 19.3 (Wald's first identity) Let $Y_1, Y_2, \ldots$ be independent and identically distributed integrable random variables, set $X_n = Y_1 + \cdots + Y_n$, and let $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$. If $\mu = \mathbb{E}[Y_1]$ and T is an $(\mathcal{F}_n)$-stopping time with $\mathbb{E}[T] < \infty$, then

$$\mathbb{E}[X_T] = \mu \mathbb{E}[T].$$

Proof. The process $M_n = X_n - \mu n$ is a martingale. The bounded optional stopping theorem applied to $T \wedge n$ gives

$$\mathbb{E}[X_{T\wedge n}] = \mu\mathbb{E}[T \wedge n].$$

The monotone convergence theorem gives $\mathbb{E}[T \wedge n] \to \mathbb{E}[T]$. In addition,

$$\mathbb{E}[|X_T - X_{T\wedge n}|] \le \sum_{j=n+1}^{\infty} \mathbb{E}[|Y_j|\mathbb{1}_{\{T\ge j\}}] = \mathbb{E}[|Y_1|] \sum_{j=n+1}^{\infty} \mathbb{P}(T \ge j) \longrightarrow 0.$$

Here $\{T \ge j\} \in \mathcal{F}_{j-1}$, so it is independent of Y_j, and the tail sum tends to zero because

$$\mathbb{E}[T] = \sum_{j=1}^{\infty} \mathbb{P}(T \ge j) < \infty.$$

Thus $X_{T\wedge n} \to X_T$ in L^1, and passing to the limit proves the identity. □

Theorem 19.4 (Wald's second identity) Under the setup of Theorem 19.3, suppose in addition that $\mathbb{E}[Y_1] = 0$ and $\sigma^2 = \mathbb{E}[Y_1^2] < \infty$. Then

$$\mathbb{E}[X_T^2] = \sigma^2\mathbb{E}[T].$$

Proof. The process

$$M_n = X_n^2 - n\sigma^2$$

is a martingale. Bounded optional stopping at $T \wedge n$ gives

$$\mathbb{E}[X_{T\wedge n}^2] = \sigma^2\mathbb{E}[T \wedge n].$$

The stopped sums converge in L^2. Indeed, orthogonality of the centered independent increments gives

$$\mathbb{E}[(X_T - X_{T\wedge n})^2] = \sum_{j=n+1}^{\infty} \mathbb{E}[Y_j^2\mathbb{1}_{\{T\ge j\}}] = \sigma^2 \sum_{j=n+1}^{\infty} \mathbb{P}(T \ge j) \longrightarrow 0.$$

Therefore $\mathbb{E}[X_{T\wedge n}^2] \to \mathbb{E}[X_T^2]$, while the monotone convergence theorem gives $\mathbb{E}[T \wedge n] \to \mathbb{E}[T]$. Passing to the limit yields

$$\mathbb{E}[X_T^2] = \sigma^2\mathbb{E}[T].$$

□

Lecture 20 — Thursday, March 21, 2019

Martingales and Compounding Errors

Independent centered errors accumulate on the scale $\sqrt{n}$, rather than n. Indeed, if $X_1, \ldots, X_n$ are independent and $\mathrm{Var}(X_i) \le 1$, then

$$\sqrt{\mathrm{Var}(X_1 + \cdots + X_n)} = \sqrt{\sum_{i=1}^{n} \mathrm{Var}(X_i)} \le \sqrt{n}.$$

Martingale theory extends this phenomenon to nonlinear functions of independent inputs.

Theorem 20.1 Let $X_1, \ldots, X_n$ be independent random variables with $\mathrm{Var}(X_i) \le 1$. Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is square integrable under their joint law and is 1-Lipschitz in each coordinate:

$$|f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x_k', \ldots, x_n)| \le |x_k - x_k'|.$$

Then

$$\mathrm{Var}(f(X_1, \ldots, X_n)) \le n.$$

Proof. We first record the one-dimensional estimate. If U' is an independent copy of U, then

$$\mathbb{E}[(U - U')^2] = 2\,\mathrm{Var}(U).$$

Consequently, if h is 1-Lipschitz, then

$$\mathrm{Var}(h(U)) = \frac{1}{2}\mathbb{E}[(h(U) - h(U'))^2] \le \frac{1}{2}\mathbb{E}[(U - U')^2] = \mathrm{Var}(U).$$

Set $Y = f(X_1, \ldots, X_n)$, define $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$, and let

$$M_k = \mathbb{E}[Y \mid \mathcal{F}_k], \quad 0 \le k \le n.$$

This is the **exposure martingale**: it reveals one input at a time. Conditional on $\mathcal{F}_{k-1}$, the random variable M_k is a 1-Lipschitz function of X_k, because averaging over the unrevealed coordinates preserves the Lipschitz constant. The conditional version of the one-dimensional estimate therefore gives

$$\mathbb{E}[(M_k - M_{k-1})^2 \mid \mathcal{F}_{k-1}] = \mathrm{Var}(M_k \mid \mathcal{F}_{k-1}) \leq \mathrm{Var}(X_k) \leq 1.$$

The martingale increments are orthogonal, as shown in Lemma 20.2 below. Hence

$$\mathrm{Var}(Y) = \mathbb{E}[(M_n - M_0)^2] = \sum_{k=1}^{n} \mathbb{E}[(M_k - M_{k-1})^2] \leq n.$$

□

Lemma 20.2 If $(M_k)_{k\geq 0}$ is an L^2-martingale, then, for $i \neq j$,

$$\mathbb{E}\left[(M_i - M_{i-1})(M_j - M_{j-1})\right] = 0.$$

Proof. By symmetry, it suffices to suppose that $i < j$. Then

$$\mathbb{E}\left[(M_i - M_{i-1})(M_j - M_{j-1})\right] = \mathbb{E}\left[(M_i - M_{i-1})\,\mathbb{E}\left[M_j - M_{j-1} \mid \mathcal{F}_{j-1}\right]\right] = 0.$$

□

Corollary 20.3 Let $(M_k)_{k\geq 0}$ be an L^2-martingale. Then

$$\mathbb{E}\left[(M_n - M_0)^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(M_i - M_{i-1})^2\right].$$

Proof. Expand

$$M_n - M_0 = \sum_{i=1}^{n} (M_i - M_{i-1})$$

and use orthogonality of the increments. □

Lecture 21 — Tuesday, March 26, 2019

Theorem 21.1 (Azuma–Hoeffding inequality) Let $(X_k, \mathcal{F}_k)_{k=0}^n$ be a martingale, and suppose that deterministic constants $c_1, \dots, c_n$ satisfy

$$|X_k - X_{k-1}| \le c_k \quad \text{almost surely for every } 1 \le k \le n.$$

Then, for every $t > 0$,

$$\mathbb{P}(X_n - X_0 \ge t) \le \exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right).$$

Consequently,

$$\mathbb{P}(|X_n - X_0| \ge t) \le 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right).$$

The Gaussian exponent has the correct order in t and in the total squared increment bound.

Proof. Set $Y_k = X_k - X_{k-1}$. Conditional on $\mathcal{F}_{k-1}$, we have $\mathbb{E}[Y_k \mid \mathcal{F}_{k-1}] = 0$ and $|Y_k| \le c_k$. Convexity of the exponential function implies that, for every $y \in [-c_k, c_k]$,

$$e^{\lambda y} \le \frac{c_k + y}{2c_k} e^{\lambda c_k} + \frac{c_k - y}{2c_k} e^{-\lambda c_k}.$$

Taking conditional expectations gives

$$\mathbb{E}[e^{\lambda Y_k} \mid \mathcal{F}_{k-1}] \le \frac{e^{\lambda c_k} + e^{-\lambda c_k}}{2} = \cosh(\lambda c_k) \le e^{\lambda^2 c_k^2/2}.$$

The final inequality follows term by term from the power series, because $(2m)! \ge 2^m m!$.

Iterated conditioning now yields

$$\mathbb{E}[e^{\lambda(X_n - X_0)}] = \mathbb{E}\left[e^{\lambda(X_{n-1} - X_0)}\mathbb{E}[e^{\lambda Y_n} \mid \mathcal{F}_{n-1}]\right] \le e^{\lambda^2 c_n^2/2}\mathbb{E}[e^{\lambda(X_{n-1} - X_0)}] \le \exp\left(\frac{\lambda^2}{2}\sum_{k=1}^n c_k^2\right).$$

Markov's inequality therefore gives, for every $\lambda > 0$,

$$\mathbb{P}(X_n - X_0 \geq t) \leq \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{k=1}^{n} c_k^2\right).$$

The exponent is minimized at $\lambda = t/\sum_{k=1}^n c_k^2$, which proves the one-sided bound. Applying it to the martingale $(-X_k)$ and adding the two tail probabilities proves the two-sided bound. □

Example 21.2 Place k balls independently and uniformly into n bins. Let Z_j be the location of the jth ball, and let $Y = f(Z_1, \dots, Z_k)$ be the number of occupied bins. With the exposure martingale

$$X_i = \mathbb{E}[Y \mid Z_1, \dots, Z_i],$$

changing one ball location changes Y by at most one, so $|X_i - X_{i-1}| \leq 1$. Theorem 21.1 gives

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) \leq 2e^{-t^2/(2k)}.$$

To compute the center of concentration, let I_j indicate that bin j is occupied. Since a fixed bin remains empty with probability $(1 - 1/n)^k$,

$$\mathbb{E}[Y] = \sum_{j=1}^{n} \mathbb{E}[I_j] = n\left(1 - \left(1 - \frac{1}{n}\right)^k\right).$$

Example 21.3 In the **Erdős–Rényi random graph** $G(n, p)$, each possible edge between n vertices is present independently with probability p. A proper coloring assigns colors to the vertices so that adjacent vertices receive different colors. The **chromatic number** $\chi(G)$ is the smallest number of colors in a proper coloring. A proper three-coloring of a sample graph is shown in Figure 23.

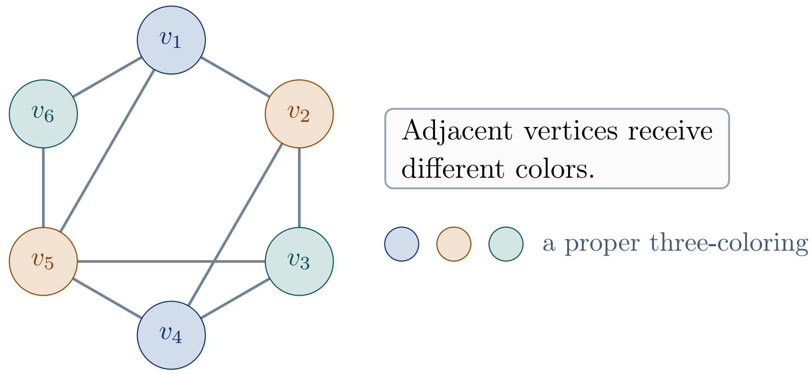

FIGURE 23

Changing any collection of edges incident to one fixed vertex changes the chromatic number by at most one. Indeed, after removing that vertex, the two graphs agree; reinserting it requires at most one additional color. Label the vertices $v_1, \dots, v_n$, and partition the possible edges into blocks

$$E_k = \{\{v_k, v_j\} : k < j \leq n\}, \quad 1 \leq k \leq n-1.$$

Figure 24 illustrates these blocks for five vertices.

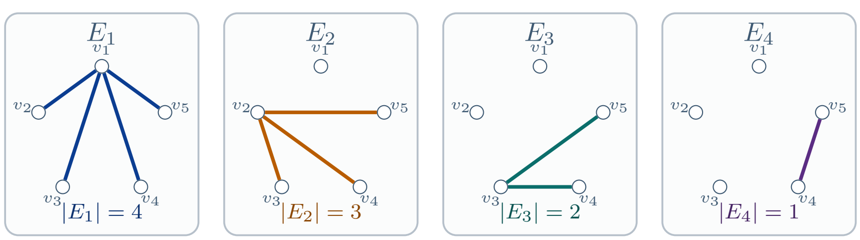

FIGURE 24

Let $Y_e = \mathbb{1}_{\{e \text{ is present}\}}$, so $\chi = f((Y_e)_e)$. Define

$$\mathcal{F}_k = \sigma(Y_e : e \in E_1 \cup \cdots \cup E_k) \qquad \text{and} \qquad X_k = \mathbb{E}[\chi \mid \mathcal{F}_k].$$

Since changing one vertex exposure block changes χ by at most one, $|X_k - X_{k-1}| \leq 1$. The Azuma–Hoeffding inequality yields

$$\mathbb{P}(|\chi - \mathbb{E}[\chi]| \geq t) \leq 2e^{-t^2/(2n)}.$$

Corollary 21.4 (McDiarmid's inequality) Let $X_1, \dots, X_n$ be independent random variables, where X_k takes values in a set S_k. Suppose that $f : S_1 \times \cdots \times S_n \to \mathbb{R}$ satisfies the coordinatewise bounded differences condition

$$\left| f(s_1, \dots, s_{k-1}, s_k, s_{k+1}, \dots, s_n) - f(s_1, \dots, s_{k-1}, s_k', s_{k+1}, \dots, s_n) \right| \le c_k \tag{21.1}$$

whenever the two inputs differ only in coordinate k. If $Y = f(X_1, \dots, X_n)$, then, for every $t > 0$,

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \ge t) \le 2 \exp\left(-\frac{t^2}{2 \sum_{k=1}^n c_k^2} \right).$$

Proof. Let

$$\mathcal{F}_k = \sigma(X_1, \dots, X_k) \qquad \text{and} \qquad M_k = \mathbb{E}[Y \mid \mathcal{F}_k], \quad 0 \le k \le n.$$

This is the exposure martingale associated with Y, and its endpoints are $M_0 = \mathbb{E}[Y]$ and $M_n = Y$. Fix $k \in \{1, \dots, n\}$ and condition on $X_1, \dots, X_{k-1}$. For $s \in S_k$, set

$$g_k(s) = \mathbb{E}\left[f(X_1, \dots, X_{k-1}, s, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_{k-1} \right].$$

By (21.1), $|g_k(s) - g_k(s')| \le c_k$ for all $s, s' \in S_k$. Moreover,

$$M_k = g_k(X_k) \qquad \text{and} \qquad M_{k-1} = \mathbb{E}[g_k(X_k) \mid \mathcal{F}_{k-1}].$$

Thus M_k, conditionally on $\mathcal{F}_{k-1}$, takes values in an interval of length at most c_k. In particular, $|M_k - M_{k-1}| \le c_k$ almost surely. Applying the Azuma–Hoeffding inequality to $M_n - M_0 = Y - \mathbb{E}[Y]$ gives the claimed two-sided estimate. □

4. Brownian Motion and Continuous Martingales

Brownian Motion

Brownian motion arises as the universal diffusive scaling limit of random walks. Let

$$S_n = X_1 + \cdots + X_n,$$

where $X_1, X_2, \dots$ are independent and identically distributed random variables with mean 0 and variance 1. A typical linearly interpolated random walk path is illustrated in Figure 25.

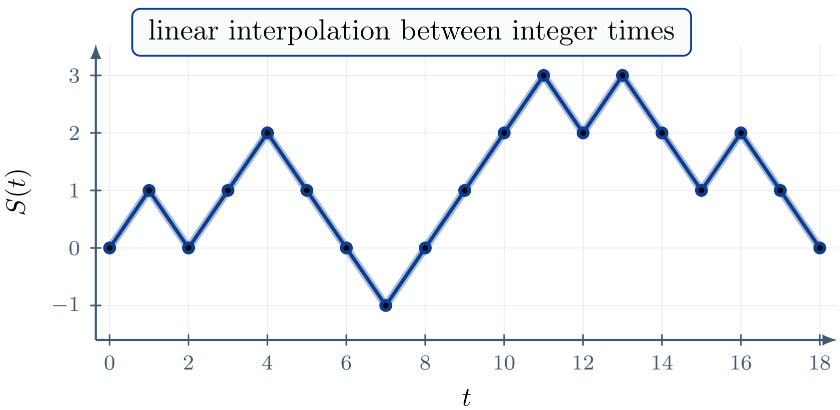

FIGURE 25

The ordinary central limit theorem describes only the endpoint:

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

Questions involving the whole path require a stronger limit theorem. For example, we may ask for the limits of

$$\frac{1}{\sqrt{n}} \max_{0\le i\le n} S_i \qquad \text{or} \qquad \frac{1}{\sqrt{n}} \max_{0\le i\le n} \left(S_i + \frac{i}{\sqrt{n}}\right).$$

The second expression adds a deterministic drift before rescaling. The limiting laws of these path functionals are universal: apart from the mean and variance normalization, they do not depend on the distribution of the increments.

To formulate the path-level result, set $S_0 = 0$ and extend S_n to a continuous function $S : [0,\infty) \to \mathbb{R}$ by linear interpolation:

$$S(t) = S_k + (t-k)(S_{k+1} - S_k), \quad k \le t \le k+1.$$

Define the diffusively rescaled process

$$B_n(t) = \frac{S(nt)}{\sqrt{n}}, \quad t \ge 0.$$

Let

$$\mathcal{C} = C([0,\infty),\mathbb{R})$$

and equip $\mathcal{C}$ with the topology of uniform convergence on compact sets. This topology is induced, for example, by the metric

$$d(f,g) = \sum_{m=1}^{\infty} 2^{-m} \left(\sup_{0\le t\le m} |f(t)-g(t)| \wedge 1 \right).$$

A sequence f_n converges to f in this topology precisely when it converges uniformly on every compact time interval. Thus each B_n is a random variable with values in the measurable space $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$.

Theorem 21.5 (Donsker's invariance principle) The random elements B_n converge in distribution in $\mathcal{C}$ to a Brownian motion B:

$$B_n \xrightarrow{d} B.$$

Convergence in distribution in $\mathcal{C}$ means that

$$\mathbb{E}[\varphi(B_n)] \longrightarrow \mathbb{E}[\varphi(B)]$$

for every bounded continuous function $\varphi : \mathcal{C} \to \mathbb{R}$. For example, the map

$$f \longmapsto \max_{0\le t\le 1} f(t)$$

is continuous, although it is not bounded. Its truncations

$$\varphi_M(f) = \left(\max_{0\le t\le 1} f(t) \right) \wedge M \vee (-M), \quad M > 0,$$

are bounded and continuous. The continuous mapping theorem therefore turns Theorem 21.5 into limit theorems for maxima and many other continuous path functionals.

Definition 21.6 (Brownian motion) A stochastic process $B = (B(t))_{t\ge 0}$ is a **standard Brownian motion** if its sample paths are continuous almost surely and it has the following properties:

1. $B(0) = 0$ almost surely.
2. If $0 \le t_0 < t_1 < \cdots < t_m$, then the increments

$$B(t_1) - B(t_0), \ldots, B(t_m) - B(t_{m-1})$$

are independent.

3. If $0 \le s < t$, then
$$B(t) - B(s) \sim \mathcal{N}(0, t-s).$$

Theorem 21.7 Standard Brownian motion exists, and its law on $\mathcal{C}$ is unique.

The uniqueness assertion follows because (1)–(3) determine every finite-dimensional distribution, while the Borel σ-algebra on $\mathcal{C}$ is generated by coordinate evaluations. Existence requires the construction of a process with continuous sample paths. Norbert Wiener was the first to give a rigorous construction, but it was notoriously complicated. Paul Lévy later developed a simpler approach. That approach, developed below, first defines consistent values at dyadic times and then proves that the successive piecewise-linear approximations converge uniformly. Figure 26 illustrates the first refinement step.

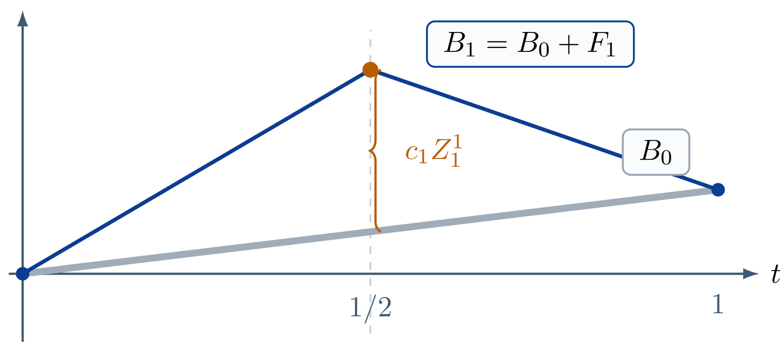

FIGURE 26

Lecture 22 — Thursday, March 28, 2019

Lévy's Construction of Brownian Motion

Lemma 22.1 If X and Y are independent $\mathcal{N}(0,\sigma^2)$ random variables, then $X+Y$ and $X-Y$ are independent, and each has distribution $\mathcal{N}(0, 2\sigma^2)$.

Proof. The vector $(X+Y, X-Y)$ is jointly Gaussian. Its two coordinates have variance $2\sigma^2$, and their covariance is

$$\mathrm{Cov}(X+Y, X-Y) = \mathrm{Var}(X) - \mathrm{Var}(Y) = 0.$$

Uncorrelated jointly Gaussian random variables are independent, which proves the claim. $\square$

Lemma 22.2 If $Z \sim \mathcal{N}(0,1)$, then, for every $x > 0$,

$$\mathbb{P}(Z > x) \le \frac{1}{\sqrt{2\pi}\,x} e^{-x^2/2} \qquad \text{and} \qquad \mathbb{P}(|Z| > x) \le \frac{1}{x} e^{-x^2/2}.$$

Proof. Since $y/x \ge 1$ when $y \ge x$,

$$\mathbb{P}(Z > x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2}\,\mathrm{d}y \le \frac{1}{\sqrt{2\pi}\,x} \int_x^\infty y e^{-y^2/2}\,\mathrm{d}y = \frac{1}{\sqrt{2\pi}\,x} e^{-x^2/2}.$$

Symmetry gives the two-sided estimate, because $2/\sqrt{2\pi} \le 1$. $\square$

It is enough initially to construct Brownian motion on $[0,1]$. Independent copies on successive unit intervals can then be joined at their endpoints to produce the process on $[0,\infty)$.

For $n \ge 0$, define the dyadic grid

$$D_n = \left\{ \frac{k}{2^n} : 0 \le k \le 2^n \right\} \qquad \text{and} \qquad D_\infty = \bigcup_{n=0}^\infty D_n.$$

Let

$$Z_* \qquad \text{and} \qquad \left\{ Z_n^k : n \ge 1,\ 1 \le k \le 2^{n-1} \right\}$$

be independent standard Gaussian random variables. Set $F_0(t) = tZ_*$. For $n \ge 1$, let

$$c_n = 2^{-(n+1)/2}$$

and define F_n to be the continuous, piecewise-linear function that vanishes on D_{n-1} and satisfies

$$F_n\left(\frac{2k-1}{2^n}\right) = c_n Z_n^k, \quad 1 \le k \le 2^{n-1}.$$

Equivalently, F_n is linear on every interval between consecutive points of D_n. Figure 27 shows F_0

and the single triangular correction F_1.

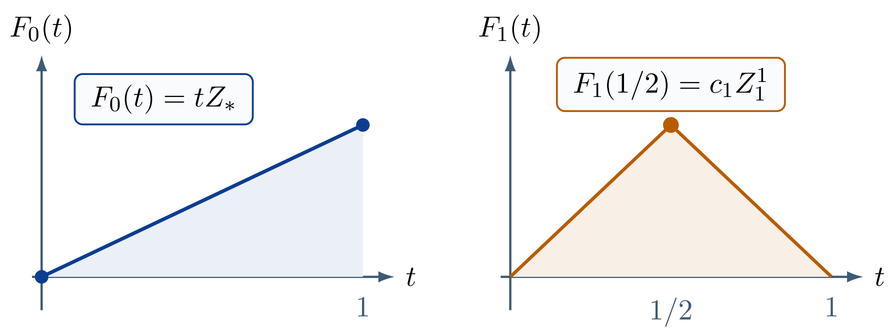

FIGURE 27

Define

$$B_n(t) = \sum_{j=0}^{n} F_j(t), \quad 0 \le t \le 1.$$

The following consistency statement contains the essential finite-dimensional part of the construction.

Lemma 22.3 For all nonnegative integers m, n, the approximations satisfy the following:

1. $B_n(t) = B_m(t)$ for every $t \in D_{\min\{m,n\}}$.
2. The increments
$$B_n\left(\frac{k}{2^n}\right) - B_n\left(\frac{k-1}{2^n}\right), \quad 1 \le k \le 2^n,$$
are independent $\mathcal{N}(0, 2^{-n})$ random variables.

Proof. For (1), because F_j vanishes on D_{j-1}, adding the jth correction never changes any previously assigned dyadic value.

For (2), we prove the result by induction on n. At level 0, the only increment is Z_*, which has distribution $\mathcal{N}(0, 1)$. Suppose the conclusion holds at level $n-1$, and write

$$X_k^{n-1} = B_{n-1}\left(\frac{k}{2^{n-1}}\right) - B_{n-1}\left(\frac{k-1}{2^{n-1}}\right).$$

The two level-n increments obtained by bisecting the kth parent interval are

$$X_{2k-1}^n = \frac{1}{2} X_k^{n-1} + c_n Z_n^k, \tag{22.1}$$

$$X_{2k}^n = \frac{1}{2} X_k^{n-1} - c_n Z_n^k. \tag{22.2}$$

Both are centered Gaussian random variables, and

$$\mathrm{Var}(X_{2k-1}^n) = \frac{1}{4} 2^{-(n-1)} + c_n^2 = 2^{-(n+1)} + 2^{-(n+1)} = 2^{-n}.$$

The same calculation applies to X_{2k}^n, while

$$\mathrm{Cov}(X_{2k-1}^n, X_{2k}^n) = \frac{1}{4} 2^{-(n-1)} - c_n^2 = 0.$$

Hence the two child increments are independent by Lemma 22.1. Distinct pairs depend on independent parent increments and independent Gaussian corrections, so all 2^n increments are independent. □

Lecture 23 — Tuesday, April 02, 2019

It remains to prove that the series of corrections converges uniformly. With $\|f\|_\infty = \sup_{0 \le t \le 1} |f(t)|$, the shape of the triangular functions gives

$$\|F_n\|_\infty = c_n \max_{1 \le k \le 2^{n-1}} |Z_n^k|.$$

Set

$$m_n = 2c_n \sqrt{n} = 2\sqrt{n}\, 2^{-(n+1)/2}.$$

The union bound and Lemma 22.2 imply

$$\mathbb{P}(\|F_n\|_\infty > m_n) \le 2^{n-1} \mathbb{P}(|Z| > 2\sqrt{n}) \le \frac{2^{n-1}}{2\sqrt{n}} e^{-2n} \le \frac{1}{2} \left(\frac{2}{e^2} \right)^n.$$

The final series is summable. The Borel–Cantelli lemma therefore shows that $\|F_n\|_\infty \le m_n$ eventually, almost surely. Since

$$\sum_{n=1}^\infty m_n = \sqrt{2} \sum_{n=1}^\infty \sqrt{n}\, 2^{-n/2} < \infty,$$

we obtain

$$\sum_{n=0}^{\infty} \|F_n\|_\infty < \infty \quad \text{almost surely.}$$

Consequently, $B_n = \sum_{j=0}^{n} F_j$ converges uniformly to a continuous function

$$B = \sum_{j=0}^{\infty} F_j.$$

For each $t \in D_\infty$, the sequence $B_n(t)$ eventually stabilizes by (1). It follows from (2) that B has independent Gaussian increments of the required variances at dyadic times. For arbitrary $0 \le t_0 < \cdots < t_m \le 1$, approximate every t_j by dyadic times and use continuity. The corresponding increment vectors converge almost surely, so their characteristic functions converge; the limiting increments remain independent and have distributions $\mathcal{N}(0, t_j - t_{j-1})$. Thus B satisfies Definition 21.6 on $[0,1]$, completing the existence part of Theorem 21.7.

Theorem 23.1 There is an almost surely finite random variable C such that

$$|B(s+t) - B(s)| \le C\sqrt{t \log\left(\frac{2}{t}\right)}$$

whenever $0 < t \le 1$ and $s, s+t \in [0,1]$.

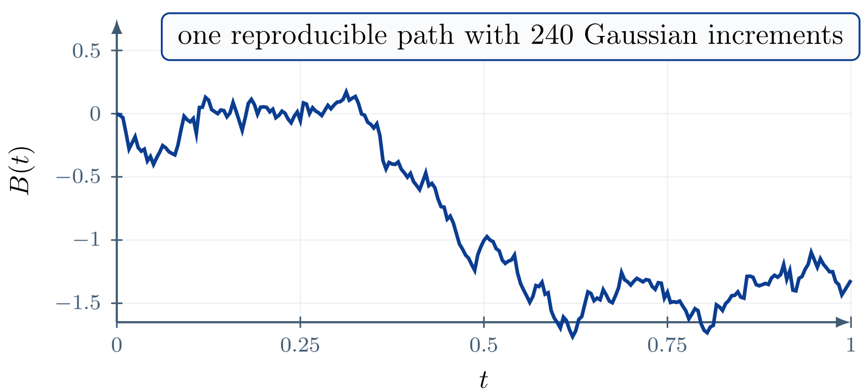

FIGURE 28

Figure 28 suggests the irregularity quantified by Theorem 23.1. The order $\sqrt{t\log(1/t)}$ is sharp for the uniform modulus of continuity.

Proof. The Borel–Cantelli argument in the construction can be strengthened to give an almost surely finite C_0 such that, for every $n \ge 1$,

$$\|F_n\|_\infty \le C_0\sqrt{n}\,2^{-n/2}.$$

Because F_n is piecewise linear on intervals of length 2^{-n}, its derivative exists except at finitely many dyadic points and satisfies

$$\|F_n'\|_\infty \le C_0\sqrt{n}\,2^{n/2}.$$

Fix a nonnegative integer ℓ. Splitting the series at level ℓ gives

$$|B(s+t)-B(s)| \le t\sum_{n=0}^{\ell}\|F_n'\|_\infty + 2\sum_{n=\ell+1}^{\infty}\|F_n\|_\infty \le C_1\left(t\sqrt{\ell+1}\,2^{\ell/2} + \sqrt{\ell+1}\,2^{-\ell/2}\right).$$

Choose ℓ so that $2^{-\ell-1} < t \le 2^{-\ell}$. Both terms on the right are then bounded by a constant multiple of $\sqrt{t\log(2/t)}$. Absorbing the finitely many low-level terms and deterministic constants into a new random variable C proves the result. □

Proposition 23.2 The random constant in Theorem 23.1 may be chosen so that there are deterministic constants $c_1, c_2 > 0$ for which

$$\mathbb{P}(C > x) \le c_1 e^{-c_2 x^2}, \quad x \ge 1.$$

Proof. Let

$$C_0 = \sup_{n\ge 1}\frac{\|F_n\|_\infty}{2c_n\sqrt{n}}.$$

For $x \ge 1$, a union bound and Lemma 22.2 give

$$\mathbb{P}(C_0 > x) \le \sum_{n=1}^{\infty} 2^{n-1}\mathbb{P}(|Z| > 2x\sqrt{n}) \le \sum_{n=1}^{\infty}\frac{2^{n-1}}{2x\sqrt{n}}e^{-2x^2 n} \le c_1' e^{-c_2' x^2}.$$

The constant C constructed in the proof of Theorem 23.1 is bounded by a deterministic affine function of $1 + C_0 + |Z_*|$. The standard Gaussian tail bound therefore yields the stated estimate after the constants are adjusted. □

Theorem 23.3 Almost surely, a Brownian path is not monotone on any nondegenerate interval.

Proof. Fix rational numbers $a, b \geq 0$ with $a < b$. If B is nondecreasing on $[a, b]$, then all n increments over the equal subdivision of this interval are nonnegative. These increments are independent centered Gaussian random variables, so

$$\mathbb{P}\big(B \text{ is nondecreasing on } [a, b]\big) \leq 2^{-n}$$

for every n. The probability is therefore zero. The same argument applies to nonincreasing paths. Every nondegenerate interval contains an interval with rational endpoints, so a countable union completes the proof. $\square$

Theorem 23.4 Almost surely,

$$\limsup_{t\to\infty} B(t) = +\infty \qquad \text{and} \qquad \liminf_{t\to\infty} B(t) = -\infty.$$

Proof. At integer times,

$$B(n) = \sum_{k=1}^{n} \big(B(k) - B(k-1)\big),$$

where the summands are independent standard Gaussian random variables. The law of the iterated logarithm gives

$$\limsup_{n\to\infty} \frac{B(n)}{\sqrt{2n \log\log n}} = 1 \qquad \text{and} \qquad \liminf_{n\to\infty} \frac{B(n)}{\sqrt{2n \log\log n}} = -1$$

almost surely. The conclusion at continuous times follows because the integers form a subsequence of $[0, \infty)$. $\square$

Theorem 23.5 For every $a > 0$, the rescaled process

$$\widetilde{B}(t) = \frac{1}{\sqrt{a}} B(at), \quad t \geq 0,$$

is a standard Brownian motion.

Proof. The process starts at zero and has continuous paths. Its increments over disjoint time intervals are independent, and, for $0 \leq s < t$,

$$\widetilde{B}(t) - \widetilde{B}(s) = \frac{B(at) - B(as)}{\sqrt{a}} \sim \mathcal{N}(0, t - s).$$

Thus all conditions in Definition 21.6 hold. □

Theorem 23.6 For every $s \geq 0$, the process

$$\widetilde{B}(t) = B(s+t) - B(s), \quad t \geq 0,$$

is a standard Brownian motion.

Proof. Continuity and the initial value are immediate. The independent increments property of B is unchanged by a deterministic time shift, and

$$\widetilde{B}(t) - \widetilde{B}(r) = B(s+t) - B(s+r) \sim \mathcal{N}(0, t-r)$$

whenever $0 \leq r < t$. □

Theorem 23.7 Define

$$R(0) = 0 \qquad \text{and} \qquad R(t) = tB(1/t) \quad \text{for } t > 0.$$

Then R is a standard Brownian motion.

Proof. First, $B(u)/u \to 0$ almost surely as $u \to \infty$. To see this, apply the Doob L^2 maximal inequality on increasingly fine rational grids and then use path continuity. For every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{0 \leq u \leq 2^{n+1}} |B(u)| > \varepsilon 2^n\right) \leq \frac{4\mathbb{E}[B(2^{n+1})^2]}{\varepsilon^2 2^{2n}} = \frac{8}{\varepsilon^2 2^n}.$$

The series of these probabilities is finite. The Borel–Cantelli lemma, first for rational $\varepsilon > 0$, therefore gives $B(u)/u \to 0$. Hence $R(t) \to 0$ as $t \downarrow 0$, so R has continuous paths.

The process R is centered and Gaussian. For $s, t > 0$,

$$\mathbb{E}[R(s)R(t)] = st \min\left\{\frac{1}{s}, \frac{1}{t}\right\} = \min\{s, t\}.$$

Thus R has the same finite-dimensional distributions as a standard Brownian motion. Together with continuity, this proves the assertion. □

Corollary 23.8 Almost surely,

$$\limsup_{t\downarrow 0} \frac{B(t)}{t} = +\infty \qquad \text{and} \qquad \liminf_{t\downarrow 0} \frac{B(t)}{t} = -\infty.$$

Proof. By Theorem 23.7, $R(t) = tB(1/t)$ is a Brownian motion. Consequently,

$$\frac{R(t)}{t} = B(1/t).$$

The conclusion follows from Theorem 23.4 by letting $t \downarrow 0$. $\square$

Theorem 23.9 Almost surely, a Brownian path is differentiable at no time $t \geq 0$.

Proof. Corollary 23.8 proves nondifferentiability at zero, and Theorem 23.6 gives the same conclusion at every fixed deterministic time. To obtain the simultaneous assertion for all times, an argument on a countable grid is needed.

If a continuous function is differentiable at some $t \in [0, N]$, then there is an integer $m \geq 1$ such that

$$|B(t+h) - B(t)| \leq m|h| \quad \text{whenever } |h| \leq \frac{1}{m}$$

and $t + h \geq 0$. Put $\delta_k = 2^{-k}$, and choose j with $j\delta_k \leq t < (j+1)\delta_k$. For all sufficiently large k, each of the three consecutive grid increments

$$B((j+r+1)\delta_k) - B((j+r)\delta_k), \quad r = 0, 1, 2,$$

then has absolute value at most $6m\delta_k$. Each such increment has distribution $\mathcal{N}(0, \delta_k)$, and the three are independent. Since

$$\mathbb{P}(|\mathcal{N}(0,\delta_k)| \leq 6m\delta_k) = \mathbb{P}(|Z| \leq 6m\sqrt{\delta_k}) \leq C_m\sqrt{\delta_k},$$

a union bound over the at most $N/\delta_k + 2$ possible locations gives probability at most $C_{m,N}\sqrt{\delta_k}$. This is summable in k. The Borel–Cantelli lemma rules out the displayed three-increment event for all sufficiently large k. Taking the countable union over m, N proves that no differentiability point exists. $\square$

Proposition 23.10 For all $s, t \geq 0$,

$$\mathbb{E}[B(s)B(t)] = \operatorname{Cov}(B(s), B(t)) = \min\{s, t\}.$$

Proof. Suppose $s \leq t$. Since $B(t) - B(s)$ is independent of $B(s)$ and has mean zero,

$$\mathbb{E}[B(s)B(t)] = \mathbb{E}\big[B(s)\big(B(s) + B(t) - B(s)\big)\big] = \mathbb{E}[B(s)^2] = s.$$

Symmetry in s and t proves the formula. ☐

Definition 23.11 The process

$$U(t) = e^{-t/2} B(e^t), \quad t \in \mathbb{R},$$

is called an **Ornstein–Uhlenbeck process**.

Proposition 23.12 The Ornstein–Uhlenbeck process U is **stationary**. Equivalently, for every fixed $a \in \mathbb{R}$, the processes $(U(a+t))_{t \in \mathbb{R}}$ and $(U(t))_{t \in \mathbb{R}}$ have the same law.

Proof. The process is centered and Gaussian. By Proposition 23.10, for $s \leq t$,

$$\operatorname{Cov}(U(s), U(t)) = e^{-(s+t)/2} \min\{e^s, e^t\} = e^{-(t-s)/2}.$$

Hence, for arbitrary s, t,

$$\operatorname{Cov}(U(s), U(t)) = e^{-|t-s|/2}.$$

This covariance depends only on the time difference. A centered Gaussian process is determined by its covariance function, so every time shift has the same finite-dimensional distributions. ☐

Corollary 23.13 For every $s, h \in \mathbb{R}$,

$$\operatorname{Cov}(U(s), U(s+h)) = e^{-|h|/2}.$$

Proof. This is the covariance formula obtained in the proof of Proposition 23.12. ☐

Brownian Bridges

Definition 23.14 If B is a standard Brownian motion, then

$$R(t) = B(t) - tB(1), \quad 0 \le t \le 1,$$

is called a **standard Brownian bridge**. Its two endpoints are fixed: $R(0) = R(1) = 0$.

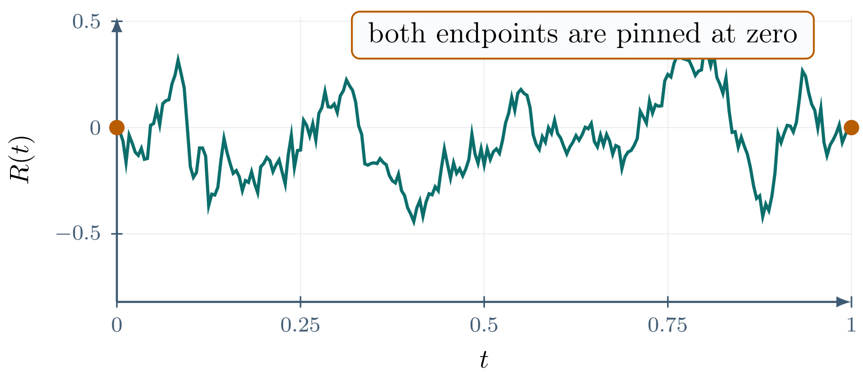

FIGURE 29

The pinned endpoints of a typical path are shown in Figure 29. The bridge is a centered Gaussian process and has the following basic properties.

Proposition 23.15 For $0 \le s, t \le 1$, the following hold:

1. $$\mathbb{E}[R(s)R(t)] = \min\{s,t\} - st.$$
2. $$\mathrm{Var}(R(t)) = t(1-t).$$
3. The time-reversed process $\widetilde{R}(t) = R(1-t)$ is again a standard Brownian bridge.

Proof. Proposition 23.10 gives

$$\mathbb{E}[R(s)R(t)] = \mathbb{E}[(B(s) - sB(1))(B(t) - tB(1))] = \min\{s,t\} - st - ts + st = \min\{s,t\} - st,$$

which proves (1). Setting $s = t$ proves (2). For the last assertion, $\widetilde{R}$ is a centered Gaussian process with endpoints zero and

$$\mathbb{E}[\widetilde{R}(s)\widetilde{R}(t)] = \min\{1-s, 1-t\} - (1-s)(1-t) = \min\{s,t\} - st.$$

Hence $\widetilde{R}$ and R have the same finite-dimensional distributions, proving (3). □

Moreover,

$$\mathrm{Cov}(R(t), B(1)) = \mathrm{Cov}(B(t) - tB(1), B(1)) = t - t = 0.$$

Because the pair is jointly Gaussian, the entire bridge process is independent of $B(1)$. This makes precise the interpretation of R as Brownian motion with its endpoint pinned at zero.

Lecture 24 — Thursday, April 04, 2019

Continuous Martingales

Definition 24.1 Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration. An adapted stochastic process $(M_t)_{t\geq 0}$ is a **continuous martingale** if its sample paths are continuous almost surely, $M_t \in L^1$ for every $t \geq 0$, and

$$\mathbb{E}[M_t \mid \mathcal{F}_s] = M_s$$

whenever $0 \leq s \leq t$.

Example 24.2 With respect to the natural Brownian filtration $\mathcal{F}_t = \sigma(B(u) : 0 \leq u \leq t)$, both

$$M_t = B(t) \qquad \text{and} \qquad N_t = B(t)^2 - t$$

are continuous martingales. Indeed, if $0 \leq s \leq t$, then $B(t) - B(s)$ is independent of $\mathcal{F}_s$, has mean zero, and has variance $t - s$. Consequently,

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

Proposition 24.3 Let

$$T = \inf\{t \geq 0 : B(t) \in \{-a, b\}\}, \quad a, b > 0.$$

Then

$$\mathbb{P}(B(T) = -a) = \frac{b}{a+b}, \qquad \mathbb{P}(B(T) = b) = \frac{a}{a+b}, \qquad \text{and} \qquad \mathbb{E}[T] = ab.$$

Moreover,

$$\mathbb{E}[T^2] \leq 4\mathbb{E}[B(T)^4].$$

Proof. Let $T_n = T \wedge n$. The stopped process $B(T_n)$ remains in $[-a, b]$. Optional stopping applied first to $B(t)$ and then to $B(t)^2 - t$, followed by bounded and monotone convergence, gives

$$\mathbb{E}[B(T)] = 0 \qquad \text{and} \qquad \mathbb{E}[T] = \mathbb{E}[B(T)^2].$$

If $p = \mathbb{P}(B(T) = b)$, the first identity yields $bp - a(1-p) = 0$, so $p = a/(a+b)$. The other exit probability follows, and substitution into the second identity gives

$$\mathbb{E}[T] = a^2 \frac{b}{a+b} + b^2 \frac{a}{a+b} = ab.$$

To control the second moment, observe that

$$H_t = B(t)^4 - 6tB(t)^2 + 3t^2$$

is a martingale. This follows by expanding $B(t) = B(s) + (B(t) - B(s))$ and using the first four moments of a centered Gaussian increment. Optional stopping at T_n gives

$$3\mathbb{E}[T_n^2] = 6\mathbb{E}[T_n B(T_n)^2] - \mathbb{E}[B(T_n)^4].$$

By the Cauchy–Schwarz inequality,

$$3\mathbb{E}[T_n^2] \leq 6\sqrt{\mathbb{E}[T_n^2]\mathbb{E}[B(T_n)^4]},$$

and hence

$$\mathbb{E}[T_n^2] \leq 4\mathbb{E}[B(T_n)^4].$$

Letting $n \to \infty$, using monotone convergence on the left and bounded convergence on the right, proves the final assertion. $\square$

This proposition is the continuous-time analogue of the gambler's ruin calculation. It is also the basic ingredient in the following embedding result.

Lemma 24.4 Every integrable, mean-zero probability distribution on $\mathbb{R}$ is a mixture of mean-zero distributions supported on at most two points.

Proof. Let μ be the distribution in question and set

$$m = \int_{(0,\infty)} v\,\mu(\mathrm{d}v) = \int_{(-\infty,0)} (-u)\,\mu(\mathrm{d}u).$$

If $m = 0$, then μ is concentrated at zero. Suppose $m > 0$. On pairs $u < 0 < v$, define the finite measure

$$\nu(\mathrm{d}u, \mathrm{d}v) = \frac{v-u}{m}\,\mu(\mathrm{d}u)\mu(\mathrm{d}v).$$

Its total mass is $1 - \mu(\{0\})$; add an atom of mass $\mu(\{0\})$ at $(0,0)$. Given $u < 0 < v$, let $Q_{u,v}$ be the mean-zero two-point distribution defined by

$$Q_{u,v}(\{u\}) = \frac{v}{v-u} \qquad \text{and} \qquad Q_{u,v}(\{v\}) = \frac{-u}{v-u}.$$

For a Borel set $A \subset (-\infty, 0)$, its mass under the mixture is

$$\int_{A\times(0,\infty)} \frac{v}{v-u}\,\nu(\mathrm{d}u,\mathrm{d}v) = \frac{1}{m}\int_A \mu(\mathrm{d}u)\int_{(0,\infty)} v\,\mu(\mathrm{d}v) = \mu(A).$$

The analogous calculation on $(0,\infty)$, together with the atom at zero, shows that the mixture has distribution μ. □

Theorem 24.5 (Skorohod embedding theorem) Let X be an integrable random variable with $\mathbb{E}[X] = 0$. On a probability space carrying a Brownian motion, possibly enlarged by independent initial randomness, there is a stopping time T such that the following hold:

1. $B(T)$ has the same distribution as X.
2. If $\mathbb{E}[X^2] < \infty$, then
$$\mathbb{E}[T] = \mathbb{E}[B(T)^2] = \mathbb{E}[X^2].$$
3. If $\mathbb{E}[X^4] < \infty$, then
$$\mathbb{E}[T^2] \le 4\mathbb{E}[X^4] < \infty.$$

Proof. Sample a pair (U, V), independently of B, according to the mixing distribution in Lemma 24.4; thus either $U = V = 0$ or $U < 0 < V$. Make this pair measurable at time zero and define

$$T = \inf\{t \geq 0 : B(t) \in \{U, V\}\}.$$

On the event $U = V = 0$, take $T = 0$. Otherwise, Theorem 23.4 ensures that $T < \infty$ almost surely. Conditional on $U = u < 0 < v = V$, Proposition 24.3 gives

$$\mathbb{P}(B(T) = u \mid U = u, V = v) = \frac{v}{v-u} \qquad \text{and} \qquad \mathbb{P}(B(T) = v \mid U = u, V = v) = \frac{-u}{v-u}.$$

Thus the conditional law is $Q_{u,v}$, and the mixture identity proves (1).

The same proposition yields

$$\mathbb{E}[T \mid U, V] = -UV = \mathbb{E}[B(T)^2 \mid U, V],$$

so integration proves (2). Its fourth-moment estimate gives

$$\mathbb{E}[T^2 \mid U, V] \leq 4\mathbb{E}[B(T)^4 \mid U, V].$$

Integrating once more proves (3). □

The strong Markov property permits the embedding to be restarted after each stopping time. Figure 30 illustrates the successive embedded increments.

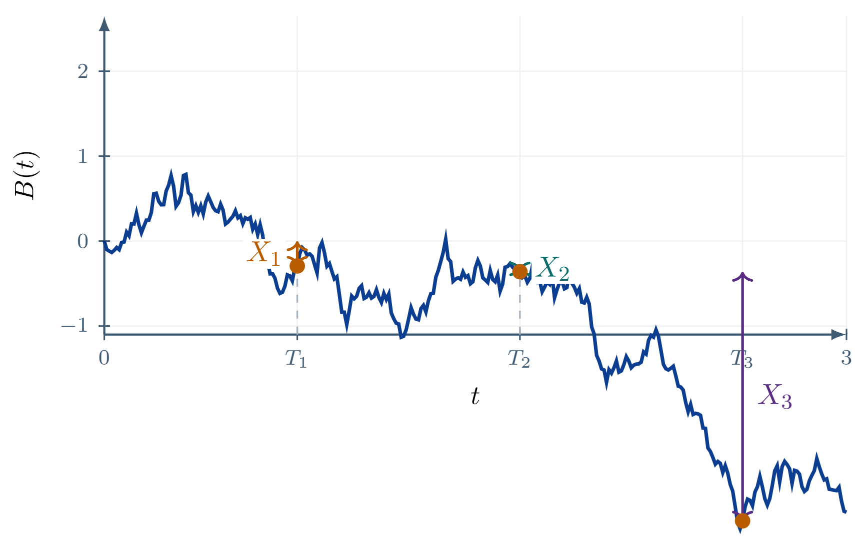

FIGURE 30

Theorem 24.6 (Central limit theorem (via Brownian embedding!)) Let $X_1, X_2, \dots$ be independent and identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$. Then

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Proof. Iterating Theorem 24.5 with the strong Markov property, define stopping times

$$0 = T_0^{(n)} \le T_1^{(n)} \le \cdots \le T_n^{(n)}$$

such that the spatial increments are independent and satisfy

$$B(T_i^{(n)}) - B(T_{i-1}^{(n)}) \stackrel{d}{=} \frac{X_i}{\sqrt{n}}.$$

Brownian scaling shows that the duration increments have the same law as τ_i/n, where $\tau_1, \tau_2, \dots$ are independent copies of an integrable embedding time for X and $\mathbb{E}[\tau_i] = \sigma^2$. Consequently,

$$T_n^{(n)} \stackrel{d}{=} \frac{1}{n} \sum_{i=1}^{n} \tau_i \xrightarrow{\mathbb{P}} \sigma^2$$

by the weak law of large numbers. Brownian path continuity then implies

$$B(T_n^{(n)}) - B(\sigma^2) \xrightarrow{\mathbb{P}} 0.$$

On the other hand,

$$B(T_n^{(n)}) \stackrel{d}{=} \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

Since $B(\sigma^2) \sim \mathcal{N}(0, \sigma^2)$, the desired convergence follows. $\square$

Appendix: Lecture and Tutorial Index