Lectures 12–15 Martingales

Conditional expectation · stopping and convergence · the martingale CLT · concentration

Reading guide. Section 1 is a reminder of L^2 and orthogonal projection, and defines conditional expectation. Sections 1–2 are Lecture 12; Section 3 is Lecture 13; Section 4 is Lecture 14; Section 5 is Lecture 15. Everything is done in L^2. Facts taken from outside the course are in boxes marked QUOTED. Examples marked WORKED IN CLASS are done on the board. Exercises are typed A (routine), B (a worked example or short proof) and C (longer; AI assistance encouraged); the homework for a lecture is the exercise block closing its section.

1 Reminder: L^2, projection, conditional expectation

1.1 L^2 and orthogonal projection

L^2 is the space of real random variables X with $\mathbb{E}X^2 < \infty$, with inner product and norm

$$\langle X, Z\rangle = \mathbb{E}[XZ], \qquad \|X\|_2 = \big(\mathbb{E}X^2\big)^{1/2}.$$

For a finite family of random variables $Y = (Y_1, \dots, Y_n)$ (with values in any $\mathbb{R}^m$) put

$$L^2(Y) = \big\{\, f(Y) \;:\; f : \mathbb{R}^m \to \mathbb{R},\ \ \mathbb{E}f(Y)^2 < \infty \,\big\},$$

the functions of Y with finite second moment. It is a linear subspace of L^2 containing the constants, and $L^2(Y) \subset L^2(Y, Y')$ for any further family Y'.

QUOTED · Hilbert space

L^2 is complete: a sequence that is Cauchy in $\|\cdot\|_2$ converges in $\|\cdot\|_2$. Each $L^2(Y)$ is a closed subspace. For a closed subspace $V \subset L^2$ and $X \in L^2$ there is a unique $P_V X \in V$ with $\|X - P_V X\|_2 = \min_{W \in V} \|X - W\|_2$, characterized by $\langle X - P_V X, W\rangle = 0$ for all $W \in V$. The map P_V is linear with $\|P_V X\|_2 \le \|X\|_2$; $\|X\|_2^2 = \|P_V X\|_2^2 + \|X - P_V X\|_2^2$; and if $V \subset W$ are closed subspaces then $P_V P_W = P_W P_V = P_V$.

1.2 Conditional expectation

Definition 1.1 (conditional expectation)**.** For $X \in L^2$ and a finite family Y, $\mathbb{E}[X \mid Y] := P_{L^2(Y)} X$ is the orthogonal projection of X onto $L^2(Y)$.

This is the definition of Lectures 1–2: $\mathbb{E}[X \mid Y]$ is the minimizer of $\mathbb{E}(X - f(Y))^2$ over all functions f with $f(Y) \in L^2$, i.e. the best mean-square prediction of X by any function of Y. The best *linear* prediction is the projection onto the smaller subspace $\operatorname{span}(1, Y_1, \dots, Y_n)$; the two differ in general (Example 1.7) and coincide for Gaussian vectors (Theorem 1.6).

Lemma 1.2 (characterization)**.** *Let $Z \in L^2(Y)$. Then $Z = \mathbb{E}[X \mid Y]$ if and only if $\mathbb{E}[(X - Z)f(Y)] = 0$ for every $f(Y) \in L^2(Y)$, and it suffices to check bounded f.*

Proof. The first statement is the characterization in the box. For the second, truncate: $f_M = \max(-M, \min(f, M))$ satisfies $f_M(Y) \to f(Y)$ in L^2, and $\mathbb{E}[(X - Z)f_M(Y)] \to \mathbb{E}[(X - Z)f(Y)]$ by Cauchy–Schwarz. □

Proposition 1.3 (rules)**.** *Let $X, X' \in L^2$, $a, b \in \mathbb{R}$, and let Y, Y' be finite families.*

(i) Linearity. $\mathbb{E}[aX + bX' \mid Y] = a\mathbb{E}[X \mid Y] + b\mathbb{E}[X' \mid Y]$.

(ii) Averaging. $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}X$.

(iii) Taking out what is known. *If $g(Y) \in L^2(Y)$ and $g(Y)X \in L^2$, then $g(Y)\mathbb{E}[X \mid Y] \in L^2$ and $\mathbb{E}[g(Y)X \mid Y] = g(Y)\mathbb{E}[X \mid Y]$; in particular $\mathbb{E}[g(Y) \mid Y] = g(Y)$.*

(iv) Tower. $\mathbb{E}\big[\mathbb{E}[X \mid Y, Y'] \,\big|\, Y\big] = \mathbb{E}[X \mid Y] = \mathbb{E}\big[\mathbb{E}[X \mid Y] \,\big|\, Y, Y'\big]$.

(v) Independence. *If X is independent of Y then $\mathbb{E}[X \mid Y] = \mathbb{E}X$.*

(vi) Positivity. *If $X \geq 0$ then $\mathbb{E}[X \mid Y] \geq 0$; hence $X \leq X'$ implies $\mathbb{E}[X \mid Y] \leq \mathbb{E}[X' \mid Y]$, and $|\mathbb{E}[X \mid Y]| \leq \mathbb{E}[|X| \mid Y]$.*

(vii) Pythagoras. $\mathbb{E}X^2 = \mathbb{E}\big(\mathbb{E}[X \mid Y]\big)^2 + \mathbb{E}\big(X - \mathbb{E}[X \mid Y]\big)^2$.

(viii) Contraction. $\mathbb{E}\big(\mathbb{E}[X \mid Y]\big)^2 \leq \mathbb{E}X^2$, *and* $\mathbb{E}\big(\mathbb{E}[X \mid Y] - \mathbb{E}[X' \mid Y]\big)^2 \leq \mathbb{E}(X - X')^2$.

Proof. (i), (iv), (vii), (viii) are properties of orthogonal projections listed in the box, (iv) because $L^2(Y) \subset L^2(Y, Y')$. Write $Z = \mathbb{E}[X \mid Y]$ and use Lemma 1.2. (ii) Test against $f \equiv 1$. (iii) For bounded g: $g(Y)Z \in L^2(Y)$ and $\mathbb{E}[(g(Y)X - g(Y)Z)f(Y)] = \mathbb{E}[(X - Z)(gf)(Y)] = 0$ for bounded f. For general g truncate, $g_M = \max(-M, \min(g, M))$. Then $g_M(Y)X \to g(Y)X$ in L^2, so $\mathbb{E}[g_M(Y)X \mid Y] \to \mathbb{E}[g(Y)X \mid Y]$ in L^2 by (viii); and $\mathbb{E}[g_M(Y)X \mid Y] = g_M(Y)Z$ with $\|g_M(Y)Z\|_2 \leq \|g_M(Y)X\|_2 \leq \|g(Y)X\|_2$ for every M, so letting $M \to \infty$, $g(Y)Z \in L^2$ and $g_M(Y)Z \to g(Y)Z$ in L^2. The two limits agree. (v) $\mathbb{E}[(X - \mathbb{E}X)f(Y)] = \mathbb{E}[X - \mathbb{E}X]\,\mathbb{E}f(Y) = 0$. (vi) Test against $f(Y) = \mathbf{1}\{Z < 0\}$: $\mathbb{E}[Z\mathbf{1}\{Z < 0\}] = \mathbb{E}[X\mathbf{1}\{Z < 0\}] \geq 0$, so $\mathbb{P}(Z < 0) = 0$; the rest follows by linearity applied to $X' - X \geq 0$ and to $\pm X \leq |X|$. □

Definition 1.4 (conditional variance)**.** $\operatorname{Var}(X \mid Y) := \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2 \,\big|\, Y\big]$. By (ii) and (vii), $\operatorname{Var} X = \mathbb{E} \operatorname{Var}(X \mid Y) + \operatorname{Var} \mathbb{E}[X \mid Y]$.

Lemma 1.5 (freezing)**.** *Let $U \in L^2(Y)$ take values in $\mathbb{R}^p$, let Z be a random variable in $\mathbb{R}^q$ independent of Y, and let $h : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ satisfy $h(U, Z) \in L^2$. Put $\bar{h}(u) = \mathbb{E}\, h(u, Z)$. Then*

$$\mathbb{E}\big[h(U, Z) \mid Y\big] = \bar{h}(U).$$

Proof. $\bar{h}(U) \in L^2(Y)$. For bounded f, independence of Z from (U, Y) allows integrating over Z first: $\mathbb{E}[h(U, Z)f(Y)] = \mathbb{E}\big[\bar{h}(U)f(Y)\big]$. Apply Lemma 1.2. □

Theorem 1.6 (Gaussian conditioning, Lectures 1–2)**.** *Let (X, Y) be jointly Gaussian in $\mathbb{R}^p \times \mathbb{R}^m$ with Σ_{YY} invertible, and put $A = \Sigma_{XY}\Sigma_{YY}^{-1}$. Then $X - AY$ is independent of Y, and*

$$\mathbb{E}[X \mid Y] = \mu_X + A(Y - \mu_Y), \qquad \operatorname{Var}(X \mid Y) = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$$

(componentwise in the first identity; the second is the conditional covariance matrix, which does not depend on Y). That is, for jointly Gaussian (X, Y) the best prediction among all functions of Y is the linear one, and the conditional law of X given Y is Gaussian with the displayed mean and covariance.

Example 1.7 (a non-Gaussian pair)**.** Let Y be uniform on $[-1, 1]$ and $X = Y^2$ (the failure case of Lecture 2). Then $\operatorname{Cov}(X, Y) = \mathbb{E}Y^3 = 0$, so the best linear prediction of X from Y is the constant $\mathbb{E}X = \frac{1}{3}$, with residual variance $\operatorname{Var} Y^2 = \frac{4}{45}$; while $\mathbb{E}[X \mid Y] = Y^2$ by rule (iii), with residual zero.

2 Martingales

2.1 Filtrations

Fix a sequence of random variables $Y_1, Y_2, \dots$ (each with values in some $\mathbb{R}^m$). For $n \geq 0$ we define

$$\mathbb{E}[X \mid \mathcal{F}_n] := \mathbb{E}[X \mid Y_1, \dots, Y_n], \qquad L^2(\mathcal{F}_n) := L^2(Y_1, \dots, Y_n),$$

with $\mathbb{E}[X \mid \mathcal{F}_0] := \mathbb{E}X$ and $L^2(\mathcal{F}_0)$ the constants. We abbreviate $\mathbb{E}_n[X] := \mathbb{E}[X \mid \mathcal{F}_n]$. The symbol $\mathcal{F}_n$ stands for "the first n variables"; the sequence $(\mathcal{F}_n)_{n\geq 0}$ is the *filtration* generated by (Y_n). By the tower rule,

$$\mathbb{E}_m \mathbb{E}_n = \mathbb{E}_n \mathbb{E}_m = \mathbb{E}_m \qquad (m \leq n).$$

Definition 2.1. A sequence $(X_n)_{n\geq 0}$ is *adapted* if $X_n \in L^2(\mathcal{F}_n)$ for every n, and *predictable* if $X_n \in L^2(\mathcal{F}_{n-1})$ for every $n \geq 1$.

2.2 Martingales and orthogonal increments

Definition 2.2 (martingale)**.** Let $(\mathcal{F}_n)$ be the filtration generated by a sequence (Y_n). A sequence $(M_n)_{n\geq 0}$ in L^2 is a *martingale* with respect to $(\mathcal{F}_n)$ if it is adapted and

$$\mathbb{E}_n[M_{n+1}] = M_n \qquad \text{for all } n \geq 0;$$

a *submartingale* if it is adapted and $\mathbb{E}_n[M_{n+1}] \geq M_n$; a *supermartingale* if it is adapted and $\mathbb{E}_n[M_{n+1}] \leq M_n$. The *differences* of M are $\xi_n = M_n - M_{n-1}$, $n \geq 1$; the martingale property reads $\mathbb{E}_{n-1}[\xi_n] = 0$.

The filtration is part of the data. Every sequence (M_n) generates its own filtration (take $Y_n = M_n$), and a martingale with respect to a filtration $(\mathcal{F}_n)$ is also a martingale with respect to its own (adaptedness gives $L^2(M_0, \dots, M_n) \subset L^2(\mathcal{F}_n)$; apply the tower rule), but not conversely. Below, "martingale" means martingale with respect to the filtration of Section 2.1 unless stated otherwise.

By the tower rule, $\mathbb{E}_m[M_n] = M_m$ for $m \leq n$, and $\mathbb{E}M_n = \mathbb{E}M_0$. Conversely, if $(\xi_n)_{n\geq 1}$ is adapted with $\mathbb{E}_{n-1}\xi_n = 0$, then $M_n = M_0 + \sum_{k\leq n} \xi_k$ is a martingale for any constant M_0.

Lemma 2.3 (orthogonal increments)**.** *Let M be a martingale with differences ξ_k. If $W \in L^2(\mathcal{F}_{n-1})$ then $\mathbb{E}[W\xi_n] = 0$. Consequently $\mathbb{E}[\xi_m \xi_n] = 0$ for $m \neq n$, and*

$$\mathbb{E}M_n^2 = \mathbb{E}M_0^2 + \sum_{k=1}^{n} \mathbb{E}\xi_k^2, \qquad \mathbb{E}(M_n - M_m)^2 = \sum_{k=m+1}^{n} \mathbb{E}\xi_k^2 \quad (m \leq n).$$

Proof. $\mathbb{E}_{n-1}\xi_n = 0$ says that the projection of ξ_n onto $L^2(\mathcal{F}_{n-1})$ is 0, i.e. $\xi_n \perp L^2(\mathcal{F}_{n-1})$ (Lemma 1.2 with $Z = 0$). Since M_0 and ξ_m, $m < n$, lie in $L^2(\mathcal{F}_{n-1})$, the terms of $M_n = M_0 + \xi_1 + \cdots + \xi_n$ are pairwise orthogonal, and the identities are Pythagoras. □

2.3 Examples

Example 2.4. (a) *Sums of independent variables.* WORKED IN CLASS Let $Y_1, Y_2, \dots$ be independent and centered with variances $s_k^2 = \mathbb{E}Y_k^2$, and let $S_n = \sum_{k\leq n} Y_k$. Then

$$S_n, \qquad S_n^2 - \sum_{k\leq n} s_k^2, \qquad \frac{e^{\lambda S_n}}{\prod_{k\leq n} \mathbb{E}e^{\lambda Y_k}} \quad (\lambda \in \mathbb{R},\ \mathbb{E}e^{2\lambda Y_k} < \infty)$$

are martingales. Indeed, by rules (iii) and (v), $\mathbb{E}_n S_{n+1} = S_n + \mathbb{E}Y_{n+1} = S_n$; $\mathbb{E}_n[S_{n+1}^2] = S_n^2 + 2S_n\mathbb{E}Y_{n+1} + \mathbb{E}Y_{n+1}^2 = S_n^2 + s_{n+1}^2$; and $\mathbb{E}_n[e^{\lambda S_{n+1}}] = e^{\lambda S_n}\,\mathbb{E}e^{\lambda Y_{n+1}}$. For the simple random walk ($Y_k = \pm 1$ with probability $\frac{1}{2}$) these are S_n, $S_n^2 - n$ and $e^{\lambda S_n}/(\cosh\lambda)^n$.

(b) *Exponential martingale, i.i.d. case.* If the Y_k are i.i.d. with $\psi(\lambda) = \log \mathbb{E}e^{\lambda Y_1}$, the third martingale in (a) is $M_n = \exp(\lambda S_n - n\psi(\lambda))$.

(c) *Doob martingale.* For $f(Y_1, \dots, Y_N) \in L^2$, the sequence $M_n = \mathbb{E}_n[f(Y_1, \dots, Y_N)]$, $0 \le n \le N$, is a martingale by the tower rule, with $M_0 = \mathbb{E}f$, $M_N = f$. Hence $f - \mathbb{E}f = \sum_{k=1}^N \xi_k$ and, by Lemma 2.3, $\operatorname{Var} f = \sum_{k \le N} \mathbb{E}\xi_k^2$.

(d) *Stochastic gradient descent.* Let $\ell(\theta; z)$ be a loss, $\theta \in \mathbb{R}^d$, with population loss $L(\theta) = \mathbb{E}\,\ell(\theta; z)$, and assume $\mathbb{E}\|\nabla\ell(\theta; z)\|^2 < \infty$ for every θ. Let $z_1, z_2, \dots$ be i.i.d. samples, $\mathcal{F}_n = (z_1, \dots, z_n)$, let $\eta_0, \eta_1, \dots$ be deterministic step sizes, and let

$$\theta_{n+1} = \theta_n - \eta_n \nabla\ell(\theta_n; z_{n+1}), \qquad n \ge 0, \quad \theta_0 \text{ fixed.}$$

Then θ_n is a function of $z_1, \dots, z_n$, and

$$\theta_{n+1} = \theta_0 - \sum_{k=0}^{n} \eta_k \nabla L(\theta_k) - M_{n+1}, \qquad M_{n+1} := \sum_{k=0}^{n} \eta_k \big(\nabla\ell(\theta_k; z_{k+1}) - \nabla L(\theta_k)\big), \quad M_0 := 0.$$

$(M_n)_{n \ge 0}$ is a martingale with respect to $(\mathcal{F}_n)$, coordinatewise: it is adapted, and its differences $\xi_{k+1} = \eta_k(\nabla\ell(\theta_k; z_{k+1}) - \nabla L(\theta_k))$ satisfy $\mathbb{E}_k\xi_{k+1} = 0$ by the freezing lemma with $U = \theta_k \in L^2(\mathcal{F}_k)$ and $Z = z_{k+1}$ independent of $\mathcal{F}_k$, which gives $\mathbb{E}_k[\nabla\ell(\theta_k; z_{k+1})] = \nabla L(\theta_k)$. So the SGD iterate is θ_0, minus the accumulated population-gradient steps, minus a martingale.

Example 2.5 (likelihood ratios)**.** WORKED IN CLASS Let $Y_1, Y_2, \dots$ be i.i.d. with density p, and let q be another density with $\int q^2/p < \infty$. Then

$$M_n = \prod_{k=1}^{n} \frac{q(Y_k)}{p(Y_k)}$$

is a martingale: $\mathbb{E}_n M_{n+1} = M_n\,\mathbb{E}\big[q(Y_{n+1})/p(Y_{n+1})\big] = M_n \int q = M_n$. It is nonnegative with $\mathbb{E}M_n = 1$; $M_n \to 0$ in probability if $q \ne p$ (Exercise 2.10), so a martingale can converge to 0 while its expectation stays 1.

Example 2.6 (Pólya's urn)**.** WORKED IN CLASS An urn holds r red and b black balls. At each step a ball is drawn uniformly at random and returned with one more ball of the same colour. Let R_n be the number of red balls after n draws, $Y_n = \mathbf{1}\{\text{draw } n \text{ is red}\}$, and $F_n = R_n/(r + b + n)$. Then $\mathbb{E}_n[R_{n+1}] = R_n + F_n$, so

$$\mathbb{E}_n[F_{n+1}] = \frac{R_n + F_n}{r + b + n + 1} = \frac{F_n(r + b + n) + F_n}{r + b + n + 1} = F_n :$$

F_n is a martingale with values in $[0, 1]$. Its differences are $\xi_{n+1} = (Y_{n+1} - F_n)/(r + b + n + 1)$, with $\mathbb{E}_n\xi_{n+1}^2 = F_n(1 - F_n)/(r + b + n + 1)^2 \le \frac{1}{4}(r + b + n + 1)^{-2}$, so $\sum_n \mathbb{E}\xi_n^2 < \infty$. Section 4 shows this forces F_n to converge, and identifies the (random) limit: Example 4.4.

Exercises for Lecture 12

Exercise 2.7. TYPE A Derive rules (iii) and (iv) of Proposition 1.3 directly from the box in Section 1.1, i.e. from $P_V P_W = P_V$ for $V \subset W$ and from $g(Y)Z \in L^2(Y)$ when $Z \in L^2(Y)$, g bounded.

Exercise 2.8. type A Let (X, Y) be jointly Gaussian, X scalar, with Σ_{YY} invertible. Using Theorem 1.6 and the rules, compute $\mathbb{E}[X^2 \mid Y]$ and $\mathbb{E}[e^X \mid Y]$.

Exercise 2.9. type A Let $Y_1, \dots, Y_N$ be independent with means μ_k. Write out the Doob martingale of $f = \sum_{k \le N} Y_k$ and of $f = \prod_{k \le N} Y_k$, and its differences in each case.

Exercise 2.10. type B In Example 2.5, show $\mathbb{E}\sqrt{M_n} = \rho^n$ with $\rho = \sum \sqrt{pq} < 1$ when $q \neq p$, and deduce $\mathbb{P}(M_n \ge \varepsilon) \to 0$ for every $\varepsilon > 0$. Explain why this does not contradict $\mathbb{E}M_n = 1$ or the L^2 convergence theorem of Section 4.

Exercise 2.11 (Efron–Stein)**.** type B Let $Y_1, \dots, Y_N$ be independent, $f(Y_1, \dots, Y_N) \in L^2$, and let Y_k' be an independent copy of Y_k, independent of everything else; write $f^{(k)}$ for f with Y_k replaced by Y_k'. Let ξ_k be the differences of the Doob martingale. (a) Use the freezing lemma to show $\mathbb{E}_{k-1} f = \mathbb{E}_k f^{(k)}$, hence $\xi_k = \mathbb{E}_k[f - f^{(k)}]$. (b) Use rule (viii) to show $\mathbb{E}\xi_k^2 \le \mathbb{E}(f - f^{(k)})^2$. (c) Conclude $\operatorname{Var} f \le \frac{1}{2} \sum_{k \le N} \mathbb{E}(f - f^{(k)})^2$, with equality when $f = \sum_k g_k(Y_k)$.

Exercise 2.12 (SGD on a quadratic)**.** type C Let $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d., centered, with variance σ^2, and consider SGD for $L(\theta) = \frac{1}{2}\mathbb{E}(\theta - \varepsilon)^2$ with step sizes $\eta_k \in (0, 1]$:

$$\theta_{k+1} = \theta_k - \eta_k(\theta_k - \varepsilon_{k+1}) = (1 - \eta_k)\theta_k + \eta_k \varepsilon_{k+1}.$$

(a) Write $\theta_{n+1} = \theta_0 - \sum_{k \le n} \eta_k L'(\theta_k) - M_{n+1}$ as in Example 2.4(d), identifying M_{n+1} explicitly. (b) Derive $\mathbb{E}\theta_{n+1}^2 = (1-\eta_n)^2 \mathbb{E}\theta_n^2 + \eta_n^2 \sigma^2$ and solve it. (c) Show $\mathbb{E}\theta_n^2 \to 0$ if $\eta_n \to 0$ and $\sum \eta_n = \infty$; show that for constant $\eta_n \equiv \eta$, $\mathbb{E}\theta_n^2 \to \eta\sigma^2/(2-\eta)$. (d) Write $\theta_n = \theta_0 \prod_{k<n}(1-\eta_k) + \sum_{k<n} w_{n,k}\varepsilon_{k+1}$ with explicit weights $w_{n,k}$. Which condition on (η_k) makes $\sum_k \eta_k \varepsilon_{k+1}$ converge in L^2? (Theorem 4.2 below.)

3 Stopping

Throughout, M is a martingale with differences ξ_k with respect to the filtration of Section 2.1.

3.1 The Doob decomposition and quadratic variation

Proposition 3.1 (Doob decomposition)**.** *Let $(X_n)_{n \ge 0}$ be adapted. There is exactly one way to write $X_n = X_0 + M_n + A_n$ with M a martingale, A predictable, and $M_0 = A_0 = 0$, namely*

$$A_n = \sum_{k=1}^{n} \big(\mathbb{E}_{k-1} X_k - X_{k-1}\big), \qquad M_n = \sum_{k=1}^{n} \big(X_k - \mathbb{E}_{k-1} X_k\big).$$

X is a submartingale if and only if A is nondecreasing.

Proof. The displayed A is predictable, the displayed M has $\mathbb{E}_{k-1}[X_k - \mathbb{E}_{k-1}X_k] = 0$, and they sum to $X_n - X_0$. If $X_0 + M' + A'$ is another such decomposition, $M - M' = A' - A$ is a predictable martingale, so $M_n - M_n' = \mathbb{E}_{n-1}[M_n - M_n'] = M_{n-1} - M_{n-1}' = \cdots = 0$. The last claim reads off $A_n - A_{n-1} = \mathbb{E}_{n-1}X_n - X_{n-1}$. □

Definition 3.2 (conditional variances, quadratic variation)**.** $\sigma_k^2 := \mathbb{E}_{k-1}[\xi_k^2]$ and $\langle M \rangle_n := \sum_{k=1}^n \sigma_k^2$.

Lemma 3.3. *M_n^2 is a submartingale, $M_n^2 - \langle M \rangle_n$ is a martingale, and $\mathbb{E}\langle M \rangle_n = \mathbb{E}M_n^2 - \mathbb{E}M_0^2$. Equivalently, $\langle M \rangle$ is the predictable part of the Doob decomposition of M^2.*

Proof. $\mathbb{E}_{n-1}[M_n^2] - M_{n-1}^2 = \mathbb{E}_{n-1}[(M_{n-1} + \xi_n)^2] - M_{n-1}^2 = 2M_{n-1}\mathbb{E}_{n-1}\xi_n + \mathbb{E}_{n-1}\xi_n^2 = \sigma_n^2 \ge 0$. □

3.2 Martingale transforms

Definition 3.4. Let $H = (H_k)_{k\geq 1}$ be predictable with $\mathbb{E}\big[H_k^2\sigma_k^2\big] < \infty$ for every k (automatic if H is bounded). The *transform* of M by H is

$$(H \cdot M)_n := \sum_{k=1}^{n} H_k\, \xi_k, \qquad (H \cdot M)_0 := 0.$$

H_k is a stake chosen at time $k-1$, placed on the k-th increment. The hypothesis says the stakes are square-integrable against the conditional variances: $\mathbb{E}[H_k^2\xi_k^2] = \mathbb{E}[H_k^2\sigma_k^2]$ (apply rule (iii) to $\min(H_k^2, C)$ and ξ_k^2, then let $C \to \infty$), so $H_k\xi_k \in L^2$.

Proposition 3.5. *$H \cdot M$ is a martingale with differences $H_k\xi_k$, and*

$$\mathbb{E}\,(H \cdot M)_n^2 = \sum_{k\leq n} \mathbb{E}\big[H_k^2\sigma_k^2\big].$$

If H is bounded then moreover $\langle H \cdot M\rangle_n = \sum_{k\leq n} H_k^2\sigma_k^2$. If X is a submartingale (supermartingale) and $H \geq 0$ is predictable with $H_k(X_k - X_{k-1}) \in L^2$, then $(H \cdot X)_n := \sum_{k\leq n} H_k(X_k - X_{k-1})$ is a submartingale (supermartingale).

Proof. $H_k\xi_k \in L^2(\mathcal{F}_k)$, and by rule (iii), $\mathbb{E}_{k-1}[H_k\xi_k] = H_k\mathbb{E}_{k-1}\xi_k = 0$; the second-moment identity is Lemma 2.3 together with $\mathbb{E}[H_k^2\xi_k^2] = \mathbb{E}[H_k^2\sigma_k^2]$. For bounded H, $(H_k\xi_k)^2 \in L^2$ and rule (iii) gives $\mathbb{E}_{k-1}[H_k^2\xi_k^2] = H_k^2\sigma_k^2$. For a submartingale and $H \geq 0$, $\mathbb{E}_{k-1}[H_k(X_k - X_{k-1})] = H_k\mathbb{E}_{k-1}[X_k - X_{k-1}] \geq 0$. □

No predictable betting strategy changes the mean of a martingale: $\mathbb{E}(H \cdot M)_n = 0$ for all n (Exercise 3.19).

3.3 Stopping times and optional stopping

Definition 3.6 (stopping time)**.** A random variable τ with values in $\{0, 1, 2, \dots\} \cup \{\infty\}$ is a *stopping time* if $\mathbf{1}\{\tau \leq n\} \in L^2(\mathcal{F}_n)$ for every n, i.e. whether $\tau \leq n$ is a function of $Y_1, \dots, Y_n$.

The first entrance time $\min\{n : X_n \in B\}$ of an adapted sequence into a set B is a stopping time; so is $\tau \wedge N = \min(\tau, N)$ for a constant N; the last visit to B is not. For a stopping time τ the *stopped sequence* is $(M_{n\wedge\tau})_{n\geq 0}$.

Proposition 3.7 (stopped martingales)**.** *If τ is a stopping time then $M_{n\wedge\tau} = M_0 + (H \cdot M)_n$ with $H_k = \mathbf{1}\{\tau \geq k\}$. In particular $(M_{n\wedge\tau})_{n\geq 0}$ is a martingale (sub-, super-, if M is), with differences $\mathbf{1}\{\tau \geq n\}\xi_n$ and $\langle M_{\cdot\wedge\tau}\rangle_n = \langle M\rangle_{n\wedge\tau}$.*

Proof. $M_{n\wedge\tau} - M_0 = \sum_{k\leq n} \mathbf{1}\{\tau \geq k\}\xi_k$, and $\mathbf{1}\{\tau \geq k\} = 1 - \mathbf{1}\{\tau \leq k-1\} \in L^2(\mathcal{F}_{k-1})$ is bounded. Apply Proposition 3.5. □

Corollary 3.8 (optional stopping, bounded case)**.** *Let $\tau \leq N$ be a stopping time bounded by a constant. If M is a martingale then $\mathbb{E}M_\tau = \mathbb{E}M_0$. If X is a submartingale then $\mathbb{E}X_0 \leq \mathbb{E}X_\tau \leq \mathbb{E}X_N$; for a supermartingale the inequalities reverse.*

Proof. $M_\tau = M_{N\wedge\tau}$ and the stopped martingale has constant mean, so $\mathbb{E}M_\tau = \mathbb{E}M_{0\wedge\tau} = \mathbb{E}M_0$; the same gives $\mathbb{E}X_\tau \geq \mathbb{E}X_0$ for a submartingale. For the upper bound, $X_N - X_\tau = (H \cdot X)_N$ with $H_k = \mathbf{1}\{\tau \leq k-1\} \geq 0$, which has nonnegative mean by Proposition 3.5. □

Boundedness of τ cannot be dropped: for the simple random walk and $\tau = \min\{n : S_n = 1\}$, $\mathbb{E}S_\tau = 1 \neq 0 = \mathbb{E}S_0$. For unbounded τ one applies the corollary to $\tau \wedge n$ and passes to the limit, as in the examples below.

Proposition 3.9 (quadratic variation at a stopping time)**.** *For every stopping time τ and every n,*

$$\mathbb{E}M_{\tau\wedge n}^2 = \mathbb{E}M_0^2 + \mathbb{E}\langle M\rangle_{\tau\wedge n}.$$

Proof. $M^2 - \langle M\rangle$ is a martingale (Lemma 3.3); apply Corollary 3.8 at the bounded time $\tau \wedge n$. □

Example 3.10 (Wald's identities)**.** Let $Y_1, Y_2, \ldots$ be i.i.d. with mean μ and variance σ^2, $S_n = \sum_{k\leq n} Y_k$, and τ a stopping time.

(a) $\mathbb{E}S_{\tau\wedge n} = \mu\,\mathbb{E}(\tau \wedge n)$, by Corollary 3.8 for the martingale $S_n - n\mu$.

(b) $\mathbb{E}(S_{\tau\wedge n} - \mu(\tau\wedge n))^2 = \sigma^2\,\mathbb{E}(\tau\wedge n)$, by Proposition 3.9 for the same martingale, whose quadratic variation is $n\sigma^2$.

(c) If $\mathbb{E}\tau < \infty$ then $\mathbb{E}S_\tau = \mu\,\mathbb{E}\tau$.

Proof of (c). First, $\mathbb{P}(\tau = \infty) \leq \mathbb{P}(\tau > m) \leq \mathbb{E}\tau/m$ for every m, so $\tau < \infty$ with probability one and S_τ is defined. The plan is to let $n \to \infty$ in (a).

The right side of (a). $\mathbb{E}(\tau\wedge n) = \sum_{m=0}^{n-1} \mathbb{P}(\tau > m)$, which increases to $\sum_{m\geq 0} \mathbb{P}(\tau > m) = \mathbb{E}\tau$.

The left side of (a). Write $D_n = S_\tau - S_{\tau\wedge n}$. On $\{\tau \leq n\}$, $D_n = 0$; on $\{\tau > n\}$, $D_n = Y_{n+1} + \cdots + Y_\tau$. In both cases

$$D_n = \sum_{k>n} \mathbf{1}\{\tau \geq k\}\, Y_k, \qquad \text{so} \qquad |D_n| \leq \sum_{k>n} \mathbf{1}\{\tau \geq k\}\, |Y_k|.$$

The event $\{\tau \geq k\} = \{\tau \leq k-1\}^c$ is decided by $Y_1, \ldots, Y_{k-1}$, hence independent of Y_k, so $\mathbb{E}[\mathbf{1}\{\tau \geq k\}|Y_k|] = \mathbb{P}(\tau \geq k)\,\mathbb{E}|Y_1|$. Summing the nonnegative terms,

$$\mathbb{E}|D_n| \leq \mathbb{E}|Y_1| \sum_{k>n} \mathbb{P}(\tau \geq k) = \mathbb{E}|Y_1|\,\mathbb{E}(\tau - n)_+,$$

and $\mathbb{E}(\tau - n)_+ = \sum_{m\geq n} \mathbb{P}(\tau > m)$ is the tail of a convergent series, so $\mathbb{E}|D_n| \to 0$. (With $n = 0$ the same bound gives $\mathbb{E}|S_\tau| \leq \mathbb{E}|Y_1|\,\mathbb{E}\tau < \infty$.)

Conclusion. $\big|\mathbb{E}S_\tau - \mu\mathbb{E}\tau\big| \leq \big|\mathbb{E}S_\tau - \mathbb{E}S_{\tau\wedge n}\big| + |\mu|\,\big|\mathbb{E}(\tau\wedge n) - \mathbb{E}\tau\big| \leq \mathbb{E}|D_n| + |\mu|\,\mathbb{E}(\tau - n)_+ \to 0.$

3.4 Two random walks

Example 3.11 (gambler's ruin)**.** WORKED IN CLASS Let S be the simple symmetric random walk, a, b positive integers, $\tau = \min\{n : S_n \in \{-b, a\}\}$, and $c = \max(a, b)$.

τ is finite. By Proposition 3.9 with $\langle S\rangle_n = n$, $\mathbb{E}(\tau\wedge n) = \mathbb{E}S_{\tau\wedge n}^2 \leq c^2$. Hence $\mathbb{P}(\tau > n) = \mathbb{P}(\tau\wedge n = n) \leq \mathbb{E}(\tau\wedge n)/n \leq c^2/n \to 0$, and $\mathbb{E}\tau = \lim_n \mathbb{E}(\tau\wedge n) \leq c^2$.

Exit probabilities. Since $\tau \wedge n$ is a bounded stopping time, Corollary 3.8 gives $\mathbb{E}S_{\tau\wedge n} = 0$ for every n. Split according to whether the walk has exited by time n:

$$S_{\tau\wedge n} = S_\tau\,\mathbf{1}\{\tau \le n\} + S_n\,\mathbf{1}\{\tau > n\}.$$

On $\{\tau > n\}$ the walk is still strictly between $-b$ and a, so $|S_n| \le c$ there and $|\mathbb{E}[S_n\mathbf{1}\{\tau > n\}]| \le c\,\mathbb{P}(\tau > n) \to 0$. Taking expectations,

$$\mathbb{E}\big[S_\tau\mathbf{1}\{\tau \le n\}\big] = -\mathbb{E}\big[S_n\mathbf{1}\{\tau > n\}\big] \longrightarrow 0.$$

On the other hand $S_\tau \in \{a, -b\}$, so $\mathbb{E}[S_\tau\mathbf{1}\{\tau \le n\}] = a\,\mathbb{P}(S_\tau = a, \tau \le n) - b\,\mathbb{P}(S_\tau = -b, \tau \le n)$, and as $n \to \infty$ the events $\{S_\tau = a, \tau \le n\}$ increase to $\{S_\tau = a\}$ because $\tau < \infty$ with probability one. With $P = \mathbb{P}(S_\tau = a)$ and $\mathbb{P}(S_\tau = -b) = 1 - P$ we get $aP - b(1-P) = 0$:

$$\mathbb{P}(S_\tau = a) = \frac{b}{a+b}, \qquad \mathbb{P}(S_\tau = -b) = \frac{a}{a+b}.$$

Expected exit time. $\mathbb{E}(\tau\wedge n) = \mathbb{E}S^2_{\tau\wedge n} \to \mathbb{E}S^2_\tau = a^2\frac{b}{a+b} + b^2\frac{a}{a+b} = ab$, the left side increasing to $\mathbb{E}\tau$ and the right side converging as above. So $\mathbb{E}\tau = ab$.

Example 3.12 (biased random walk)**.** WORKED IN CLASS Let $Y_k = +1$ with probability p and -1 with probability $q = 1-p$, $p \ne q$, and let τ, a, b be as above. Put $\rho = q/p$. Since $\mathbb{E}\rho^{Y_1} = p\rho + q/\rho = q + p = 1$, the sequence ρ^{S_n} is a martingale (Example 2.4(a) with $e^\lambda = \rho$).

τ is finite. Any $a+b$ consecutive steps $+1$ force an exit, so $\mathbb{P}(\tau > m(a+b)) \le (1-p^{a+b})^m$, and $\mathbb{E}\tau < \infty$.

Exit probabilities. Corollary 3.8 gives $\mathbb{E}\rho^{S_{\tau\wedge n}} = 1$. Split as in Example 3.11: $\rho^{S_{\tau\wedge n}} = \rho^{S_\tau}\mathbf{1}\{\tau \le n\} + \rho^{S_n}\mathbf{1}\{\tau > n\}$; on $\{\tau > n\}$, $\rho^{S_n} \le \max(\rho^a, \rho^{-b})$, so the second term has expectation at most $\max(\rho^a, \rho^{-b})\mathbb{P}(\tau > n) \to 0$, while $\mathbb{E}[\rho^{S_\tau}\mathbf{1}\{\tau \le n\}] \to \rho^a P + \rho^{-b}(1-P)$ with $P = \mathbb{P}(S_\tau = a)$. Hence $1 = P\rho^a + (1-P)\rho^{-b}$:

$$\mathbb{P}(S_\tau = a) = \frac{\rho^b - 1}{\rho^{a+b} - 1}.$$

As $\rho \to 1$ this tends to $b/(a+b)$.

Expected exit time. By Wald's identity (Example 3.10(c)), $\mathbb{E}\tau = \dfrac{\mathbb{E}S_\tau}{p-q} = \dfrac{aP - b(1-P)}{p-q}$.

3.5 Upcrossings

Definition 3.13 (upcrossings)**.** Fix $a < b$. Define stopping times $\sigma_1 = \min\{k \ge 0 : M_k \le a\}$, $\tau_1 = \min\{k > \sigma_1 : M_k \ge b\}$, and inductively $\sigma_{j+1} = \min\{k > \tau_j : M_k \le a\}$, $\tau_{j+1} = \min\{k > \sigma_{j+1} : M_k \ge b\}$ (each $\min\emptyset = \infty$). The number of *upcrossings* of $[a,b]$ completed by time n is

$$U_n[a,b] := \max\{j : \tau_j \le n\}.$$

Theorem 3.14 (Doob's upcrossing inequality)**.** *For every martingale M, every $a < b$ and every n,*

$$(b-a)\,\mathbb{E}\,U_n[a,b] \;\le\; \mathbb{E}(M_n - a)^+ - \mathbb{E}(M_0 - a)^+ \;\le\; \mathbb{E}|M_n| + |a|.$$

Proof. Let $X_n = (M_n - a)^+$. It is a submartingale: $X_n \ge M_n - a$ and $X_n \ge 0$ give, by rule (vi), $\mathbb{E}_{n-1}X_n \ge \max(M_{n-1} - a, 0) = X_{n-1}$. Let

$$H_k = \mathbf{1}\{\sigma_j < k \le \tau_j \text{ for some } j \ge 1\},$$

the strategy that holds one unit from the moment the path is at or below a until the moment it is at or above b. H_k is decided by $M_0, \dots, M_{k-1}$, so H is predictable, and $0 \le H \le 1$. On a completed upcrossing, X rises from $X_{\sigma_j} = 0$ to $X_{\tau_j} \ge b - a$; on an uncompleted final attempt, from 0 to $X_n \ge 0$. Hence

$$(H \cdot X)_n \;\ge\; (b-a)\, U_n[a,b].$$

On the other hand $X_n - X_0 = (H \cdot X)_n + ((1-H) \cdot X)_n$ and $\mathbb{E}((1-H) \cdot X)_n \ge 0$ by Proposition 3.5, since $1 - H \ge 0$ and X is a submartingale. So $(b-a)\mathbb{E}U_n \le \mathbb{E}(H \cdot X)_n \le \mathbb{E}X_n - \mathbb{E}X_0$. Finally $(M_n - a)^+ \le |M_n| + |a|$. □

Corollary 3.15. *If $\sup_n \mathbb{E}|M_n| < \infty$ (for instance if $\sup_n \mathbb{E}M_n^2 < \infty$), then for every $a < b$ the total number of upcrossings $U_\infty[a,b] = \lim_n U_n[a,b]$ satisfies*

$$\mathbb{E}\, U_\infty[a,b] \le \frac{\sup_n \mathbb{E}|M_n| + |a|}{b-a} < \infty, \qquad \text{so} \qquad \mathbb{P}\big(U_\infty[a,b] = \infty\big) = 0.$$

Remark 3.16. A sequence that crosses every interval $[a,b]$ with rational endpoints only finitely often cannot oscillate, so it has a limit in $[-\infty, \infty]$. Corollary 3.15 is therefore the statement that a martingale bounded in L^1 converges with probability one; this is Doob's martingale convergence theorem, and the argument just sketched is its standard proof. We do not pursue it: Section 4 proves convergence in L^2 directly, which is what the rest of the course uses.

3.6 Ville's inequality

Proposition 3.17 (Ville)**.** *Let Z be a nonnegative supermartingale. Then for every $u > 0$,*

$$\mathbb{P}\Big(\sup_{n \ge 0} Z_n \ge u\Big) \le \frac{\mathbb{E}Z_0}{u}.$$

Proof. Let $\tau = \min\{n : Z_n \ge u\}$. By Corollary 3.8 for the supermartingale Z at the bounded time $\tau \wedge n$, and nonnegativity,

$$\mathbb{E}Z_0 \;\ge\; \mathbb{E}Z_{\tau\wedge n} \;\ge\; \mathbb{E}\big[Z_{\tau\wedge n}\mathbf{1}\{\tau \le n\}\big] \;\ge\; u\,\mathbb{P}(\tau \le n).$$

Let $n \to \infty$. □

A nonnegative supermartingale with $Z_0 = 1$ exceeds u at some time with probability at most $1/u$, whatever the stopping rule. For the likelihood ratio of Example 2.5 under p this is the validity of a sequential test at every stopping time; Section 5 applies it to exponential supermartingales.

Exercises for Lecture 13

Exercise 3.18. TYPE A Compute $\langle M \rangle_n$ for: the simple random walk; the exponential martingale of Example 2.4(a); the Doob martingale of $f = \sum_{k \le N} Y_k$ with independent Y_k; the likelihood ratio of Example 2.5; the transform $H \cdot S$ of the simple random walk by $H_k = S_{k-1}$ (check the hypothesis of Definition 3.4, compute $\mathbb{E}(H \cdot S)_n^2$, and identify $(H \cdot S)_n$ in closed form).

Exercise 3.19 (the doubling strategy)**.** TYPE B Let S be the simple random walk and $\tau = \min\{k : Y_k = +1\}$. The doubling strategy stakes $H_k = 2^{k-1}\mathbf{1}\{\tau \ge k\}$ on the k-th step, for $k \le N$. (a) Show H is predictable and $(H \cdot S)_N = 1$ on $\{\tau \le N\}$ and $= -(2^N - 1)$ on $\{\tau > N\}$. (b) Verify directly that $\mathbb{E}(H \cdot S)_N = 0$, and compute $\langle H \cdot S \rangle_N$. (c) Explain why letting $N \to \infty$ does not produce a strategy with positive expected gain.

Exercise 3.20. type B Let S be the simple symmetric random walk and $\tau_1 = \min\{n : S_n = 1\}$. (a) Use Example 3.11 with $a = 1$ and $b \to \infty$ to show $\mathbb{P}(\tau_1 < \infty) = 1$. (b) Show $\mathbb{E}\tau_1 = \infty$ (otherwise Wald's identity gives $1 = \mathbb{E}S_{\tau_1} = 0$). (c) Apply Corollary 3.8 to the martingale $e^{\lambda S_n}/(\cosh\lambda)^n$ at $\tau_1 \wedge n$ and let $n \to \infty$ to compute $\mathbb{E}\big[(\cosh\lambda)^{-\tau_1}\big]$ for $\lambda > 0$.

Exercise 3.21 (upcrossings of a random walk)**.** type B Let S be the simple symmetric random walk. (a) Show that $\mathbb{E}U_n[0,1] \to \infty$ (every return to 0 followed by a step up is an upcrossing), and compare with the bound of Theorem 3.14, which grows like $\mathbb{E}|S_n| \asymp \sqrt{n}$. (b) For $M_n = \sum_{k\le n} \varepsilon_k/k$ with i.i.d. signs, show $\mathbb{E}U_\infty[a,b] < \infty$ for every $a < b$, and give the explicit bound.

Exercise 3.22 (branching processes)**.** type C Let $Z_0 = 1$ and $Z_{n+1} = \sum_{i=1}^{Z_n} \zeta_{n+1,i}$, where the $\zeta_{n,i}$ are i.i.d. nonnegative integers with mean m and variance s^2, independent of the past. (a) Use the freezing lemma to show $\mathbb{E}_n Z_{n+1} = mZ_n$, so $W_n = Z_n/m^n$ is a martingale. (b) Use Definition 1.4 to show $\operatorname{Var} Z_{n+1} = s^2 m^n + m^2 \operatorname{Var} Z_n$, hence $\operatorname{Var} W_{n+1} = \operatorname{Var} W_n + s^2/m^{n+2}$. (c) For $m > 1$ conclude $\sup_n \mathbb{E}W_n^2 < \infty$; by Theorem 4.2 (Lecture 14), $W_n \to W_\infty$ in L^2 with $\mathbb{E}W_\infty = 1$ and $\operatorname{Var} W_\infty = s^2/(m(m-1))$. (d) Let $\tau = \min\{n : Z_n = 0\}$ be the extinction time. Show $\{\tau \le n\}$ is decided by $Z_0, \dots, Z_n$, and that for $m < 1$, $\mathbb{P}(\tau > n) \le \mathbb{E}Z_n = m^n$. (e) Simulate for Poisson offspring with $m = 1.5$ and plot W_n along several runs.

4 Limit theorems

Throughout, M is a martingale with differences ξ_k and conditional variances σ_k^2.

4.1 The maximal inequality

Proposition 4.1 (Doob's maximal inequality)**.** *For every n and $u > 0$,*

$$\mathbb{P}\Big(\max_{k\le n} |M_k| \ge u\Big) \le \frac{\mathbb{E}M_n^2}{u^2}, \qquad \text{and} \qquad \mathbb{E}\max_{k\le n} M_k^2 \le 4\,\mathbb{E}M_n^2.$$

Proof. Let $\tau = \min\{k \le n : |M_k| \ge u\}$, with $\tau = n$ if no such k exists; $\tau \le n$ is a stopping time. M^2 is a submartingale (Lemma 3.3), so by Corollary 3.8

$$\mathbb{E}M_n^2 \ge \mathbb{E}M_\tau^2 \ge \mathbb{E}\big[M_\tau^2 \mathbf{1}\{\max_{k\le n}|M_k| \ge u\}\big] \ge u^2\,\mathbb{P}\Big(\max_{k\le n}|M_k| \ge u\Big).$$

The second inequality is Exercise 4.13. □

4.2 Convergence in L^2

Theorem 4.2 (L^2 convergence)**.** *Suppose $\sum_{k\ge1}\mathbb{E}\xi_k^2 < \infty$, equivalently $\sup_n \mathbb{E}M_n^2 < \infty$. Then:*

(i) there is $M_\infty \in L^2$ with $\mathbb{E}(M_n - M_\infty)^2 = \sum_{k>n}\mathbb{E}\xi_k^2 \to 0$, and $\mathbb{E}_n M_\infty = M_n$ for every n;

(ii) for every n and $u > 0$, $\mathbb{P}\Big(\sup_{k\ge n}|M_k - M_n| \ge u\Big) \le \dfrac{1}{u^2}\displaystyle\sum_{k>n}\mathbb{E}\xi_k^2$.

Proof. (i) $\mathbb{E}(M_n - M_m)^2 = \sum_{k=n+1}^m \mathbb{E}\xi_k^2$ for $n < m$ (Lemma 2.3), so (M_n) is Cauchy in L^2 and converges to some M_∞ by completeness, with $\mathbb{E}(M_n - M_\infty)^2 = \lim_m \mathbb{E}(M_n - M_m)^2$. Since $\mathbb{E}_n M_m =$

M_n for $m \ge n$ and $\mathbb{E}_n$ is a contraction (rule (viii)), $\mathbb{E}_n M_\infty = \lim_m \mathbb{E}_n M_m = M_n$. (ii) For fixed n, $(M_k - M_n)_{k \ge n}$ is a martingale with differences ξ_k, $k > n$. Proposition 4.1 with horizon m gives $\mathbb{P}(\max_{n \le k \le m} |M_k - M_n| \ge u) \le u^{-2} \sum_{k=n+1}^{m} \mathbb{E}\xi_k^2$; let $m \to \infty$. □

Statement (ii) controls the entire future of the path from time n on; it is the form in which convergence will be used.

Example 4.3 (weighted sums)**.** Let $\varepsilon_1, \varepsilon_2, \dots$ be martingale differences with $\mathbb{E}_{k-1}\varepsilon_k^2 \le 1$ and let a_k be predictable with $\sum_k \mathbb{E}a_k^2 < \infty$. Then $M_n = \sum_{k \le n} a_k \varepsilon_k$ is the transform of $\sum_{k \le n} \varepsilon_k$ by a (Definition 3.4), with $\mathbb{E}M_n^2 = \sum_{k \le n} \mathbb{E}[a_k^2 \mathbb{E}_{k-1}\varepsilon_k^2] \le \sum_{k \le n} \mathbb{E}a_k^2$, and Theorem 4.2 applies. In particular, for deterministic η_k,

$$\sum_k \eta_k \varepsilon_k \ \text{ converges in } L^2 \ \text{ whenever } \ \sum_k \eta_k^2 < \infty.$$

This is the Robbins–Monro condition on the step sizes of SGD (Lecture 16).

Example 4.4 (Bayesian learning of a coin)**.** WORKED IN CLASS Let θ be a random variable with values in $[0,1]$ (the prior), and given θ let $Y_1, Y_2, \dots$ be i.i.d. with $\mathbb{P}(Y_k = 1 \mid \theta) = \theta$, $\mathbb{P}(Y_k = 0 \mid \theta) = 1 - \theta$. The *posterior mean* after n observations is

$$M_n := \mathbb{E}_n[\theta] = \mathbb{E}[\theta \mid Y_1, \dots, Y_n], \qquad M_0 = \mathbb{E}\theta.$$

It is a martingale: adapted by definition, and $\mathbb{E}_n M_{n+1} = \mathbb{E}_n \mathbb{E}_{n+1}\theta = \mathbb{E}_n \theta = M_n$ by the tower rule. It is also bounded, $0 \le M_n \le 1$, so $\sup_n \mathbb{E}M_n^2 \le 1$ and Theorem 4.2 gives $M_n \to M_\infty$ in L^2.

The limit is θ, with a rate. M_n is the best mean-square prediction of θ by a function of $Y_1, \dots, Y_n$ (Definition 1.1), and the empirical frequency $\bar{Y}_n = \frac{1}{n}\sum_{k \le n} Y_k$ is one such function, so

$$\mathbb{E}(\theta - M_n)^2 \le \mathbb{E}(\theta - \bar{Y}_n)^2 = \mathbb{E}\big[\operatorname{Var}(\bar{Y}_n \mid \theta)\big] = \mathbb{E}\Big[\frac{\theta(1-\theta)}{n}\Big] \le \frac{1}{4n},$$

where the middle equality is the freezing lemma: given θ, $\bar{Y}_n$ is an average of n i.i.d. Bernoulli(θ) variables. Hence $M_n \to \theta$ in L^2, and $M_\infty = \theta$. Since $M_\infty = \theta$, Theorem 4.2(i) gives $\sum_{k>n} \mathbb{E}\xi_k^2 = \mathbb{E}(\theta - M_n)^2$, and Theorem 4.2(ii) reads

$$\mathbb{P}\Big(\sup_{k \ge n} |M_k - M_n| \ge u\Big) \le \frac{\mathbb{E}(\theta - M_n)^2}{u^2} \le \frac{1}{4nu^2} :$$

after n observations the posterior mean will never again move by more than u, with probability at least $1 - 1/(4nu^2)$, whatever the prior.

Pólya's urn is the uniform prior. For θ uniform on $[0,1]$ and $s = \sum_{k \le n} Y_k$, the Beta integral $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \int_0^1 \theta^s (1-\theta)^{n-s}\, d\theta = \frac{s!\,(n-s)!}{(n+1)!}$ gives $\mathbb{P}(Y_{n+1} = 1 \mid Y_1, \dots, Y_n) = M_n = \frac{s+1}{n+2}$: the sequence (Y_n) has the law of the colours drawn from Pólya's urn with $r = b = 1$, and M_n is the red fraction F_n of Example 2.6. The random limit F_∞ found there is the coin's bias θ, and the convergence $F_n \to F_\infty$ is Bayesian consistency.

4.3 The martingale central limit theorem

Fix k and martingale differences $\xi_1, \dots, \xi_k$ with $\mathbb{E}|\xi_i|^3 < \infty$, conditional variances $\sigma_i^2 = \mathbb{E}_{i-1}\xi_i^2$, and

$$S_j = \sum_{i \le j} \xi_i, \qquad V_j = \sum_{i \le j} \sigma_i^2 = \langle S \rangle_j, \qquad S_0 = V_0 = 0.$$

Normalization. A central limit theorem is scale-free, so we normalize the quadratic variation: the *variance budget* is 1, and the standing hypothesis is

$$V_k \le 1,$$

i.e. the martingale is observed while at most unit quadratic variation has accumulated. This costs nothing. A martingale with more variance is stopped when the budget is spent, at

$$\tau = \min\{n \ge 0 :\ V_{n+1} > 1\},$$

a stopping time because $V_{n+1} = V_n + \sigma_{n+1}^2 \in L^2(\mathcal{F}_n)$; the stopped martingale $S_{n\wedge\tau}$ has differences $\xi_n \mathbf{1}\{\tau \ge n\} = \xi_n \mathbf{1}\{V_n \le 1\}$, conditional variances $\sigma_n^2 \mathbf{1}\{V_n \le 1\}$, and quadratic variation $V_{n\wedge\tau} \le 1$ (Proposition 3.7). Corollary 4.6 below is proved this way.

The comparison object is the martingale with the same conditional variances and Gaussian increments. Let $g, g_1, g_2, \dots$ be i.i.d. $\mathcal{N}(0,1)$, independent of $Y_1, Y_2, \dots$, enlarge the filtration to $\mathcal{G}_n = (Y_1, \dots, Y_n, g_1, \dots, g_n)$, and put

$$G_n = \sum_{i \le n} \sigma_i \, g_i.$$

Since $\sigma_i \in L^2(\mathcal{F}_{i-1}) \subset L^2(\mathcal{G}_{i-1})$ and g_i is independent of $\mathcal{G}_{i-1}$, the freezing lemma gives $\mathbb{E}[\sigma_i g_i \mid \mathcal{G}_{i-1}] = 0$ and $\mathbb{E}[\sigma_i^2 g_i^2 \mid \mathcal{G}_{i-1}] = \sigma_i^2$: G is a martingale with respect to $(\mathcal{G}_n)$ with the same conditional variances and the same quadratic variation V_n as S, but with conditionally Gaussian steps. Conditionally on $Y_1, Y_2, \dots$, $G_k \sim \mathcal{N}(0, V_k)$, and $G_k + \sqrt{1 - V_k}\, g \sim \mathcal{N}(0,1)$ exactly: the Gaussian martingale, *completed* by an independent Gaussian carrying the unspent budget $1 - V_k$, is standard normal. The theorem says the same holds approximately for S.

Theorem 4.5 (martingale CLT, quantitative form)**.** *Suppose $V_k \le 1$. Then for every $f \in C^4(\mathbb{R})$ with f', f'', f''', f'''' bounded,*

$$\Big|\mathbb{E} f\big(S_k + \sqrt{1 - V_k}\, g\big) - \mathbb{E} f(g)\Big| \ \le\ \frac{\|f'''\|_\infty}{6} \sum_{i=1}^k \mathbb{E}|\xi_i|^3 \ +\ \frac{\|f''''\|_\infty}{8} \sum_{i=1}^k \mathbb{E}\, \sigma_i^4,$$

and consequently

$$\big|\mathbb{E} f(S_k) - \mathbb{E} f(g)\big| \ \le\ \frac{\|f'''\|_\infty}{6} \sum_{i=1}^k \mathbb{E}|\xi_i|^3 \ +\ \frac{\|f''''\|_\infty}{8} \sum_{i=1}^k \mathbb{E}\, \sigma_i^4 \ +\ \frac{\|f''\|_\infty}{2}\, \mathbb{E}\big(1 - V_k\big).$$

Since $\sum_i \sigma_i^4 \le V_k \max_i \sigma_i^2 \le \max_i \sigma_i^2$, the second term is at most $\frac{1}{8}\|f''''\|_\infty \mathbb{E} \max_i \sigma_i^2$. Nothing is assumed about independence.

Corollary 4.6 (martingale CLT)**.** *For each n let $\xi_{n,1}, \dots, \xi_{n,k_n}$ be martingale differences (for a filtration that may depend on n) with conditional variances $\sigma_{n,i}^2$ and $V_{n,k_n} = \sum_i \sigma_{n,i}^2$. If*

$$\sum_i \mathbb{E}|\xi_{n,i}|^3 \to 0, \qquad \mathbb{E} \max_i \sigma_{n,i}^2 \to 0, \qquad \mathbb{E}\big|V_{n,k_n} - 1\big| \to 0,$$

then $\sum_i \xi_{n,i} \Rightarrow \mathcal{N}(0,1)$.

Proof. Drop the index n and stop at τ as above: $S'_k = S_{k\wedge\tau}$ has $V'_k = V_{k\wedge\tau} \le 1$, third moments and conditional variances no larger than those of S, and by Lemma 2.3

$$\mathbb{E}(S_k - S'_k)^2 = \mathbb{E}\big(V_k - V'_k\big) \le \mathbb{E}\big(V_k - 1\big)_+ + \mathbb{E} \max_i \sigma_i^2 \to 0,$$

since the discarded variance $V_k - V_{k\wedge\tau}$ is at most $(V_k-1)+\sigma^2_{\tau+1}$ on $\{\tau<k\}$ and 0 otherwise. Hence $|\mathbb{E}f(S_k)-\mathbb{E}f(S'_k)| \le \|f'\|_\infty\sqrt{\mathbb{E}(S_k-S'_k)^2}\to 0$, while $\mathbb{E}(1-V'_k)\le \mathbb{E}|1-V_k|+\mathbb{E}(V_k-V'_k)\to 0$, so Theorem 4.5 applied to S' gives $\mathbb{E}f(S'_k)\to\mathbb{E}f(g)$. Smooth f with bounded derivatives of all orders determine convergence in distribution (Lectures 3–4). □

The third hypothesis cannot be dropped: if $\eta_1,\dots,\eta_k$ are i.i.d. signs and $\varepsilon\in\{0,\sqrt{2}\}$ is an independent fair coin flipped at time 0, then $\xi_i=\varepsilon\eta_i/\sqrt{k}$ are martingale differences with $\mathbb{E}V_k=1$ but $V_k=\varepsilon^2$, and S_k is a half–half mixture of δ_0 and $\mathcal{N}(0,2)$. The quadratic variation has to concentrate, not merely have the right mean.

Proof of the theorem

Lemma 4.7 (the heat semigroup)**.** *For f as in Theorem 4.5 put $\phi(x,v)=\mathbb{E}f(x+\sqrt{v}\,g)$, $x\in\mathbb{R}$, $v\ge 0$. Then:*

(i) $\phi(x,0)=f(x)$, $\phi(0,1)=\mathbb{E}f(g)$, *and* $\partial_x^j\phi(x,v)=\mathbb{E}f^{(j)}(x+\sqrt{v}g)$ *for* $j\le 4$, *so* $\|\partial_x^j\phi\|_\infty\le\|f^{(j)}\|_\infty$.

(ii) Semigroup. *If $g'\sim\mathcal{N}(0,1)$ is independent of g, then for $a,v\ge 0$, $\mathbb{E}\,\phi(x+\sqrt{a}\,g',v)=\phi(x,v+a)$. More generally, if $X,A,V\in L^2(\mathcal{G})$ for a family $\mathcal{G}$ independent of g', with $A,V\ge 0$, then $\mathbb{E}[\phi(X+\sqrt{A}\,g',V)\mid\mathcal{G}]=\phi(X,V+A)$; in particular $\mathbb{E}\,\phi(X,V)=\mathbb{E}f(X+\sqrt{V}\,g')$.*

(iii) Heat equation. *$\phi(x,\cdot)$ is twice continuously differentiable on $[0,\infty)$ (one-sided at 0), with $\partial_v\phi=\frac12\partial_{xx}\phi$ and $\partial_v^2\phi=\frac14\partial_x^4\phi$; hence $\|\partial_v\phi\|_\infty\le\frac12\|f''\|_\infty$ and $\|\partial_v^2\phi\|_\infty\le\frac14\|f''''\|_\infty$.*

Proof. (i) Differentiate under $\mathbb{E}$; the derivatives are bounded. (ii) $\sqrt{a}\,g'+\sqrt{v}\,g\sim\mathcal{N}(0,a+v)$ for independent g,g' (Lectures 1–2), so $\mathbb{E}f(x+\sqrt{a}g'+\sqrt{v}g)=\phi(x,v+a)$; the conditional version is the freezing lemma, and the last claim is the case $A=0$ read backwards. (iii) For $v>0$, apply Stein's identity $\mathbb{E}[g\,h(g)]=\mathbb{E}h'(g)$ (Lectures 3–4) to $h(y)=f'(x+\sqrt{v}\,y)$:

$$\partial_v\phi(x,v)=\mathbb{E}\Big[f'(x+\sqrt{v}g)\,\frac{g}{2\sqrt{v}}\Big]=\frac{1}{2\sqrt{v}}\mathbb{E}h'(g)=\frac12\mathbb{E}f''(x+\sqrt{v}g)=\tfrac12\partial_{xx}\phi(x,v).$$

The right side is continuous in $v\ge 0$, so $\phi(x,\cdot)$ is C^1 on $[0,\infty)$ with this derivative; the same computation for f'' gives $\partial_v\partial_{xx}\phi=\frac12\partial_x^4\phi$, hence $\partial_v^2\phi=\frac14\partial_x^4\phi$. □

The proof compares S and G through the processes

$$\Psi_n^S:=\phi(S_n,\,1-V_n),\qquad \Psi_n^G:=\phi(G_n,\,1-V_n),\qquad 0\le n\le k,$$

the expected value of f at the current position completed by an independent Gaussian carrying the unspent budget $1-V_n$ (this is where $V_k\le 1$ is used). Both start at $\Psi_0=\phi(0,1)=\mathbb{E}f(g)$, and by (ii) their terminal expectations are the completed martingales: $\mathbb{E}\Psi_k^S=\mathbb{E}f(S_k+\sqrt{1-V_k}\,g)$ and $\mathbb{E}\Psi_k^G=\mathbb{E}f(G_k+\sqrt{1-V_k}\,g)$.

Lemma 4.8. *Ψ^G is a martingale with respect to $(\mathcal{G}_n)$. In particular $\mathbb{E}f\big(G_k+\sqrt{1-V_k}\,g\big)=\mathbb{E}\Psi_k^G=\mathbb{E}\Psi_0^G=\mathbb{E}f(g)$.*

Proof. G_{n-1}, σ_n and $1-V_n=1-V_{n-1}-\sigma_n^2\ge 0$ are in $L^2(\mathcal{G}_{n-1})$ and g_n is independent of $\mathcal{G}_{n-1}$, so by the semigroup property (ii),

$$\mathbb{E}\big[\Psi_n^G\,\big|\,\mathcal{G}_{n-1}\big]=\mathbb{E}\big[\phi(G_{n-1}+\sigma_n g_n,\,1-V_n)\,\big|\,\mathcal{G}_{n-1}\big]=\phi\big(G_{n-1},\,1-V_n+\sigma_n^2\big)=\Psi_{n-1}^G.\qquad\square$$

For S the same identity holds only approximately, and the defect is what the theorem measures.

Lemma 4.9. *For $1 \le n \le k$,*

$$\big|\mathbb{E}_{n-1}\big[\Psi_n^S - \Psi_{n-1}^S\big]\big| \;\le\; \frac{\|f'''\|_\infty}{6}\,\mathbb{E}_{n-1}|\xi_n|^3 + \frac{\|f''''\|_\infty}{8}\,\sigma_n^4.$$

Proof. Write $x = S_{n-1}$, $\xi = \xi_n$, $\sigma^2 = \sigma_n^2$ and $w = 1 - V_n$, so $1 - V_{n-1} = w + \sigma^2$; all of x, σ^2, w are in $L^2(\mathcal{F}_{n-1})$ and $w \ge 0$. Taylor's theorem in x at (x,w) and in v at (x,w) give

$$\begin{aligned}
\phi(x+\xi, w) &= \phi(x,w) + \xi\,\partial_x\phi(x,w) + \tfrac12\xi^2\,\partial_{xx}\phi(x,w) + R_1, && |R_1| \le \tfrac16\|f'''\|_\infty|\xi|^3,\\
\phi(x, w+\sigma^2) &= \phi(x,w) + \sigma^2\,\partial_v\phi(x,w) + R_2, && |R_2| \le \tfrac12\sigma^4\|\partial_v^2\phi\|_\infty \le \tfrac18\|f''''\|_\infty\sigma^4.
\end{aligned}$$

Subtracting, and using the heat equation $\sigma^2\partial_v\phi(x,w) = \frac12\sigma^2\partial_{xx}\phi(x,w)$,

$$\Psi_n^S - \Psi_{n-1}^S = \phi(x+\xi,w) - \phi(x,w+\sigma^2) = \xi\,\partial_x\phi(x,w) + \tfrac12\big(\xi^2 - \sigma^2\big)\,\partial_{xx}\phi(x,w) + R_1 - R_2.$$

The coefficients $\partial_x\phi(x,w)$ and $\partial_{xx}\phi(x,w)$ are bounded and in $L^2(\mathcal{F}_{n-1})$, so by rule (iii) the martingale property $\mathbb{E}_{n-1}\xi = 0$ kills the first term and the definition of the conditional variance, $\mathbb{E}_{n-1}[\xi^2 - \sigma^2] = 0$, kills the second. What remains is $\mathbb{E}_{n-1}[R_1 - R_2]$. □

For Gaussian increments the two Taylor expansions are replaced by the exact semigroup identity, which is why Ψ^G is an exact martingale: Lemma 4.9 says that S is Gaussian "to second order" at every step, and the remainders are the cost of the third and fourth orders.

Proof of Theorem 4.5. Completed form. By Lemma 4.8 and the remarks before it,

$$\mathbb{E}f\big(S_k + \sqrt{1-V_k}\,g\big) - \mathbb{E}f(g) = \mathbb{E}\Psi_k^S - \mathbb{E}\Psi_k^G = \mathbb{E}\Psi_k^S - \mathbb{E}\Psi_0^S = \sum_{n=1}^{k}\mathbb{E}\big[\mathbb{E}_{n-1}(\Psi_n^S - \Psi_{n-1}^S)\big],$$

and Lemma 4.9 bounds the sum by $\frac{\|f'''\|_\infty}{6}\sum_n\mathbb{E}|\xi_n|^3 + \frac{\|f''''\|_\infty}{8}\sum_n\mathbb{E}\sigma_n^4$.

Removing the completion. $f(S_k) = \phi(S_k, 0)$ and $\mathbb{E}f(S_k + \sqrt{1-V_k}\,g) = \mathbb{E}\phi(S_k, 1-V_k)$, while by the mean value theorem in v and (iii), $|\phi(x,0) - \phi(x,v)| \le v\|\partial_v\phi\|_\infty \le \frac{v}{2}\|f''\|_\infty$. So the two expectations differ by at most $\frac{\|f''\|_\infty}{2}\mathbb{E}(1-V_k)$. □

Remark 4.10 (independent summands). If the ξ_i are independent with variances σ_i^2 summing to one, then $V_k = 1$, the completion is empty, and the bound reads $|\mathbb{E}f(S_k) - \mathbb{E}f(g)| \le \frac{\|f'''\|_\infty}{6}\sum\mathbb{E}|\xi_i|^3 + \frac{\|f''''\|_\infty}{8}\sum\sigma_i^4$. The hybrid $\sum_{j\le i}\xi_j + \sum_{j>i}g_j$ of Lectures 3–4 has become $S_i + \sqrt{1-V_i}\,g$: with random σ_j the Gaussian tail $\sum_{j>i}\sigma_j g_j$ is not independent of ξ_i (the later σ_j depend on ξ_i), so it is replaced by one Gaussian whose variance $1 - V_i$ is known at time $i-1$.

4.4 Examples for the CLT

Example 4.11 (i.i.d. summands). WORKED IN CLASS Let $X_1, \dots, X_n$ be i.i.d., centered, with unit variance and $\mathbb{E}|X_1|^3 < \infty$, and $\xi_i = X_i/\sqrt{n}$. Then $\sigma_i^2 = 1/n$, $V_n = 1$ (the completion is empty), $\sum\mathbb{E}|\xi_i|^3 = \mathbb{E}|X_1|^3/\sqrt{n}$, $\sum\sigma_i^4 = 1/n$, and Theorem 4.5 gives $|\mathbb{E}f(S_n) - \mathbb{E}f(g)| \le \frac{\|f'''\|_\infty\mathbb{E}|X_1|^3}{6\sqrt{n}} + \frac{\|f''''\|_\infty}{8n}$.

Example 4.12 (adaptively collected data). WORKED IN CLASS Let $\varepsilon_1, \varepsilon_2, \dots$ be i.i.d., centered, with unit variance and $\mathbb{E}|\varepsilon_1|^3 < \infty$, adapted to a filtration $(\mathcal{F}_i)$ with ε_i independent of $\mathcal{F}_{i-1}$, and let a_i be predictable with $|a_i| \le A$. Then $\xi_i = a_i\varepsilon_i$ are martingale differences with

$$\sigma_i^2 = a_i^2, \qquad \mathbb{E}_{i-1}|\xi_i|^3 = |a_i|^3\,\mathbb{E}|\varepsilon_1|^3, \qquad V_k = \sum_{i\le k} a_i^2.$$

Let $v_k \to \infty$ be deterministic with $\mathbb{E}|V_k/v_k - 1| \to 0$. The normalized differences $a_i\varepsilon_i/\sqrt{v_k}$ satisfy $\sum_i \mathbb{E}|\cdot|^3 \le A\,\mathbb{E}|\varepsilon_1|^3\,\mathbb{E}V_k/v_k^{3/2} \to 0$ and $\max_i \sigma_i^2/v_k \le A^2/v_k \to 0$, so by Corollary 4.6

$$\frac{1}{\sqrt{v_k}}\sum_{i\le k} a_i\varepsilon_i \;\Rightarrow\; \mathcal{N}(0,1).$$

This covers least squares with covariates a_i chosen as a function of past responses (bandits, active learning, sequential experiments): Exercises 4.16 and 4.17.

Exercises for Lecture 14

Exercise 4.13. TYPE B Let $M_n^* = \max_{k\le n}|M_k|$. (a) Show $|M|$ is a submartingale (rule (vi)). (b) Refine the proof of Proposition 4.1 to $u\,\mathbb{P}(M_n^* \ge u) \le \mathbb{E}\big[|M_n|\mathbf{1}\{M_n^* \ge u\}\big]$. (c) Use $\mathbb{E}Z^2 = \int_0^\infty 2u\mathbb{P}(Z \ge u)\,du$ and Cauchy–Schwarz to conclude $\mathbb{E}(M_n^*)^2 \le 4\mathbb{E}M_n^2$.

Exercise 4.14. TYPE A Let ε_k be i.i.d. signs. Show that $\sum_k \varepsilon_k/k$ converges in L^2 and that $\sum_k \varepsilon_k/\sqrt{k}$ does not (compute $\mathbb{E}(M_m - M_n)^2$). What does Theorem 4.2(ii) say about the first series after n terms?

Exercise 4.15. TYPE A Verify directly that $\phi(x,v) = \mathbb{E}f(x + \sqrt{v}g)$ solves $\partial_v\phi = \tfrac12\partial_{xx}\phi$ when $f(x) = e^{i\theta x}$, and when $f(x) = x^4$.

Exercise 4.16 (a two-armed bandit). TYPE B Two arms give rewards $\mu_1 + \varepsilon$ and $\mu_2 + \varepsilon$, with ε as in Example 4.12. At each round the player pulls the arm with the larger current empirical mean with probability $1-\delta$, and the other arm with probability $\delta \in (0, \tfrac12)$, using an independent coin. Let $N_1(n)$ be the number of pulls of arm 1 in the first n rounds and $\hat\mu_1(n)$ its empirical mean. (a) Show $N_1(n)$ dominates a $\mathrm{Bin}(n,\delta)$ variable, so $N_1(n) \to \infty$. (b) Write $N_1(n)(\hat\mu_1(n) - \mu_1) = \sum_{i\le n} a_i\varepsilon_i$ with $a_i \in \{0,1\}$ predictable. (c) Assuming $N_1(n)/n \to p_1 \in (0,1)$ in L^2, show $\sqrt{N_1(n)}\,(\hat\mu_1(n) - \mu_1) \Rightarrow \mathcal{N}(0,1)$.

Exercise 4.17 (least squares for an autoregression). TYPE C Let $X_0 = 0$, $X_{k+1} = \rho X_k + \varepsilon_{k+1}$ with $|\rho| < 1$ and ε_k i.i.d. $\mathcal{N}(0,\sigma^2)$. The least squares estimate of ρ from $X_0, \dots, X_n$ satisfies

$$\hat\rho_n - \rho = \frac{\sum_{k<n} X_k\varepsilon_{k+1}}{\sum_{k<n} X_k^2}.$$

(a) Show the numerator M_n is a martingale with $\sigma_{k+1}^2 = \sigma^2 X_k^2$. (b) Show $\mathbb{E}X_k^2 \to \sigma^2/(1-\rho^2)$ and $\mathrm{Var}\big(\frac1n\sum_{k<n}X_k^2\big) = O(1/n)$ (Wick's theorem for the fourth moments; $\mathrm{Cov}(X_j^2, X_k^2)$ decays geometrically in $|j-k|$), so $\frac1n\sum_{k<n}X_k^2 \to \sigma^2/(1-\rho^2)$ in L^2. (c) Apply Corollary 4.6 to $M_n/\sqrt{n\sigma^4/(1-\rho^2)}$ and conclude $\sqrt{n}(\hat\rho_n - \rho) \Rightarrow \mathcal{N}(0, 1-\rho^2)$ (quote that dividing by a sequence converging in probability to a positive constant does not change a distributional limit). (d) Simulate for $\rho = 0.9$, $n = 2000$, 10^3 runs, and compare with the predicted Gaussian.

Exercise 4.18 (SGD on a quadratic: fluctuations). TYPE C In Exercise 2.12 take $\eta_k = \eta/k$ with $\eta > \tfrac12$ and $\mathbb{E}|\varepsilon_1|^3 < \infty$. (a) Show the weights satisfy $w_{n,k} = \frac{\eta}{k}\big(\frac{k}{n}\big)^\eta(1 + o(1))$ uniformly in $k \le n$. (b) Show $n\,\mathrm{Var}\,\theta_n \to \eta^2\sigma^2/(2\eta - 1)$. (c) Apply Theorem 4.5 (the summands are independent, so V_k is deterministic) to prove $\sqrt{n}\,\theta_n \Rightarrow \mathcal{N}\big(0, \eta^2\sigma^2/(2\eta-1)\big)$. What changes for $\eta < \tfrac12$?

5 Concentration

Throughout, M is a martingale with differences ξ_k and conditional variances σ_k^2.

5.1 Azuma–Hoeffding

Lemma 5.1 (Hoeffding's lemma)**.** *Let $\mathbb{E}\xi = 0$ and $a \le \xi \le a + c$ for constants a, c. Then $\mathbb{E}e^{\lambda\xi} \le e^{\lambda^2 c^2/8}$ for all $\lambda \in \mathbb{R}$. Conditionally: if $\mathbb{E}_{k-1}\xi = 0$ and $A \le \xi \le A + c$ with $A \in L^2(\mathcal{F}_{k-1})$ and c a constant, then $\mathbb{E}_{k-1}e^{\lambda\xi} \le e^{\lambda^2 c^2/8}$.*

Proof. We prove the case $|\xi| \le c$ with the constant $e^{\lambda^2 c^2/2}$; the sharp constant is Exercise 5.12. By convexity of $x \mapsto e^{\lambda x}$ on $[-c, c]$,

$$e^{\lambda\xi} \le \frac{c - \xi}{2c} e^{-\lambda c} + \frac{c + \xi}{2c} e^{\lambda c},$$

so $\mathbb{E}e^{\lambda\xi} \le \cosh(\lambda c) \le e^{\lambda^2 c^2/2}$, the last step by comparing Taylor series ($(2k)! \ge 2^k k!$). The conditional version is the same computation with $\mathbb{E}_{k-1}$ (rules (iii), (vi)). □

Theorem 5.2 (Azuma–Hoeffding)**.** *Suppose $A_k \le \xi_k \le A_k + c_k$ with A_k predictable and c_k constants. Then for every $u > 0$,*

$$\mathbb{P}\big(M_n - M_0 \ge u\big) \le \exp\Big(-\frac{2u^2}{\sum_{k\le n} c_k^2}\Big),$$

and the same bound holds for $\mathbb{P}(M_n - M_0 \le -u)$.

Proof. By rules (ii)–(iv) and Lemma 5.1,

$$\mathbb{E}e^{\lambda(M_n - M_0)} = \mathbb{E}\Big[e^{\lambda(M_{n-1} - M_0)}\, \mathbb{E}_{n-1} e^{\lambda\xi_n}\Big] \le e^{\lambda^2 c_n^2/8}\, \mathbb{E}e^{\lambda(M_{n-1} - M_0)} \le \cdots \le \exp\Big(\frac{\lambda^2}{8}\sum_{k\le n} c_k^2\Big).$$

Chernoff's bound $\mathbb{P}(M_n - M_0 \ge u) \le e^{-\lambda u}\mathbb{E}e^{\lambda(M_n - M_0)}$ with $\lambda = 4u/\sum c_k^2$ gives the claim; apply it to $-M$ for the lower tail. □

For the simple random walk, $\mathbb{P}(S_n \ge u) \le e^{-u^2/2n}$.

5.2 McDiarmid's inequality

Theorem 5.3 (bounded differences)**.** *Let $Y_1, \dots, Y_N$ be independent and let f be bounded with*

$$\big|f(y_1, \dots, y_N) - f(y_1, \dots, y_{k-1}, y_k', y_{k+1}, \dots, y_N)\big| \le c_k \qquad \text{for all arguments and all } k.$$

Then for every $u > 0$, $\mathbb{P}\big(f(Y) - \mathbb{E}f(Y) \ge u\big) \le \exp\big(-2u^2/\sum_{k\le N} c_k^2\big)$, and the same for the lower tail.

Proof. Let $M_k = \mathbb{E}_k f$ be the Doob martingale. By the freezing lemma, $M_k = g_k(Y_1, \dots, Y_k)$ with $g_k(y_1, \dots, y_k) = \mathbb{E}f(y_1, \dots, y_k, Y_{k+1}, \dots, Y_N)$, and $|g_k(\dots, y, \dots) - g_k(\dots, y', \dots)| \le c_k$ when the arguments differ only in the k-th place. Hence, given $Y_1, \dots, Y_{k-1}$, the difference $\xi_k = g_k(Y_1, \dots, Y_k) - g_{k-1}(Y_1, \dots, Y_{k-1})$ lies in $[A_k, A_k + c_k]$ with $A_k = \inf_y g_k(Y_1, \dots, Y_{k-1}, y) - g_{k-1}(Y_1, \dots, Y_{k-1})$ predictable. Apply Theorem 5.2 to $M_N - M_0 = f - \mathbb{E}f$. □

Example 5.4 (Hoeffding's inequality)**.** For $f = \frac{1}{n}\sum_{i\le n} Y_i$ with $Y_i \in [0,1]$, $c_i = 1/n$ and $\mathbb{P}(\bar{Y}_n - \mathbb{E}\bar{Y}_n \ge u) \le e^{-2nu^2}$.

Example 5.5 (the generalization gap)**.** WORKED IN CLASS Let $z_1, \dots, z_n$ be i.i.d., $\mathcal{H}$ a class of predictors, $\ell(h,z) \in [0,1]$ a loss, $L(h) = \mathbb{E}\ell(h,z)$, $\hat{L}_n(h) = \frac{1}{n}\sum_i \ell(h,z_i)$, and

$$Z = \sup_{h\in\mathcal{H}} \big(L(h) - \hat{L}_n(h)\big).$$

Changing one z_i changes each $\hat{L}_n(h)$, hence Z, by at most $1/n$. Theorem 5.3 gives, for any class $\mathcal{H}$,

$$\mathbb{P}\big(Z \ge \mathbb{E}Z + u\big) \le e^{-2nu^2}.$$

Proposition 5.6 (symmetrization)**.** *With $\varepsilon_1, \dots, \varepsilon_n$ independent uniform signs independent of the data,*

$$\mathbb{E}Z \le 2\,\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i\le n} \varepsilon_i\, \ell(h, z_i).$$

Proof. Let $z_1', \dots, z_n'$ be an independent copy of the data and $\mathbb{E}'$ the expectation over the copy, so $L(h) = \mathbb{E}'\hat{L}_n'(h)$. Then

$$\mathbb{E}Z = \mathbb{E}\sup_h \mathbb{E}'\big[\hat{L}_n'(h) - \hat{L}_n(h)\big] \le \mathbb{E}\sup_h \big[\hat{L}_n'(h) - \hat{L}_n(h)\big] = \mathbb{E}\sup_h \frac{1}{n}\sum_i \big(\ell(h,z_i') - \ell(h,z_i)\big).$$

The joint law of the family $\big(\ell(h,z_i') - \ell(h,z_i)\big)_{h,i}$ is unchanged if the i-th term is multiplied by ε_i, since z_i and z_i' are exchangeable. Hence the right side equals $\mathbb{E}\sup_h \frac{1}{n}\sum_i \varepsilon_i(\ell(h,z_i') - \ell(h,z_i)) \le 2\mathbb{E}\sup_h \frac{1}{n}\sum_i \varepsilon_i \ell(h,z_i)$. □

Corollary 5.7 (finite classes)**.** *If $|\mathcal{H}| = m$ then $\mathbb{E}Z \le 2\sqrt{2\log m/n}$, hence $\mathbb{P}\big(Z \ge 2\sqrt{2\log m/n} + u\big) \le e^{-2nu^2}$.*

Proof. Conditionally on the data, $X_h = \frac{1}{n}\sum_i \varepsilon_i \ell(h,z_i)$ satisfies $\mathbb{E}e^{\lambda X_h} = \prod_i \cosh(\lambda\ell(h,z_i)/n) \le e^{\lambda^2/2n}$. The proof of Proposition 2.2 of Lectures 9–11 uses only this bound, so $\mathbb{E}\max_h X_h \le \sqrt{2\log m/n}$. □

For infinite classes the right side of Proposition 5.6 is bounded by chaining, as in Dudley's inequality, with the Rademacher process in place of the Gaussian one: chaining bounds the mean of the supremum, McDiarmid its fluctuations.

5.3 Freedman's inequality

Lemma 5.8. *Let $\psi_c(\lambda) = (e^{\lambda c} - 1 - \lambda c)/c^2$, $\lambda > 0$. If $\mathbb{E}_{k-1}\xi = 0$, $\xi \le c$ and $\sigma^2 = \mathbb{E}_{k-1}\xi^2$, then*

$$\mathbb{E}_{k-1}e^{\lambda\xi} \le 1 + \psi_c(\lambda)\,\sigma^2 \le \exp\big(\psi_c(\lambda)\sigma^2\big).$$

Proof. $h(x) = (e^x - 1 - x)/x^2$ ($h(0) = \frac{1}{2}$) is increasing on $\mathbb{R}$ (Exercise 5.13), so $e^x \le 1 + x + x^2 h(\lambda c)$ for $x \le \lambda c$. With $x = \lambda\xi$ and $\lambda^2 h(\lambda c) = \psi_c(\lambda)$, $e^{\lambda\xi} \le 1 + \lambda\xi + \psi_c(\lambda)\xi^2$; take $\mathbb{E}_{k-1}$ (rule (vi)). □

Corollary 5.9. *Let $M_0 = 0$ and $\xi_k \le c$ for all k. For every $\lambda > 0$, $Z_n = \exp\big(\lambda M_n - \psi_c(\lambda)\langle M\rangle_n\big)$ is a nonnegative supermartingale with $Z_0 = 1$.*

Proof. $Z_n \le e^{\lambda nc}$ is bounded, and $\langle M\rangle_n$ is predictable, so $\mathbb{E}_{n-1}Z_n = Z_{n-1}e^{-\psi_c(\lambda)\sigma_n^2}\mathbb{E}_{n-1}e^{\lambda\xi_n} \le Z_{n-1}$ by Lemma 5.8. □

Theorem 5.10 (Freedman's inequality)**.** *Let $M_0 = 0$ and $\xi_k \le c$ for all k. Then for all $u, v > 0$,*

$$\mathbb{P}\Big(\exists\, n:\ M_n \ge u \ \text{ and } \ \langle M\rangle_n \le v\Big) \le \exp\Big(-\frac{u^2}{2(v+cu/3)}\Big).$$

Proof. On the event, $Z_n \ge \exp(\lambda u - \psi_c(\lambda)v)$ for the n in question, so by Ville's inequality (Proposition 3.17)

$$\mathbb{P}(\cdots) \le \mathbb{P}\Big(\sup_n Z_n \ge e^{\lambda u - \psi_c(\lambda)v}\Big) \le \exp\big(-\lambda u + \psi_c(\lambda)v\big).$$

For $0 < \lambda < 3/c$, $\psi_c(\lambda) \le \frac{\lambda^2}{2(1-\lambda c/3)}$ (Exercise 5.13). Take $\lambda = u/(v+cu/3) < 3/c$; then $1 - \lambda c/3 = v/(v+cu/3)$ and the exponent is

$$-\frac{u^2}{v+cu/3} + \frac{u^2}{(v+cu/3)^2}\cdot\frac{v+cu/3}{2} = -\frac{u^2}{2(v+cu/3)}.$$ □

Two properties distinguish Theorem 5.10 from Theorem 5.2: the bound depends on the realized quadratic variation $\langle M\rangle_n$ rather than on $\sum c_k^2$ (for $v \gg cu$ it is the Gaussian tail $e^{-u^2/2v}$; Bernstein's inequality for independent sums is the case of deterministic σ_k^2), and it holds simultaneously for all n.

Example 5.11 (state-dependent noise)**.** WORKED IN CLASS Let $M_n = \sum_{k\le n}\eta_k\xi_k$ with $|\xi_k| \le c$ and $\sigma_k^2 = s^2(\theta_{k-1})$ for some adapted sequence (θ_k) and a function $s^2 \le c^2$. Azuma gives $\mathbb{P}(M_n \ge u) \le \exp(-u^2/(2c^2\sum_{k\le n}\eta_k^2))$. Freedman gives, for every v,

$$\mathbb{P}\Big(\exists n:\ M_n \ge u,\ \sum_{k\le n}\eta_k^2 s^2(\theta_{k-1}) \le v\Big) \le \exp\Big(-\frac{u^2}{2(v+\eta_{\max}cu/3)}\Big),$$

where $\eta_{\max} = \max_k \eta_k$. When the path (θ_k) spends its time where s^2 is small, v can be taken much smaller than $c^2\sum\eta_k^2$. For SGD this is the regime where the gradient noise vanishes near the minimizer (Lecture 17).

QUOTED · Gaussian concentration

Let $g \sim \mathcal{N}(0, I_d)$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be L-Lipschitz for the Euclidean norm. Then $\mathbb{P}\big(f(g) - \mathbb{E}f(g) \ge u\big) \le e^{-u^2/2L^2}$ for all $u > 0$. This is the Gaussian analogue of Theorem 5.3, with L in place of $(\sum c_k^2)^{1/2}$ and no bounded-differences hypothesis; it can be proved by the interpolation of Lecture 11.

Exercises for Lecture 15

Exercise 5.12. TYPE A (a) Prove Lemma 5.1 with the constant $e^{\lambda^2c^2/8}$: write ξ as a convex combination of the endpoints a, $a + c$ to get $\mathbb{E}e^{\lambda\xi} \le e^{\varphi(\lambda c)}$ with $\varphi(t) = -pt + \log(1 - p + pe^t)$, $p = -a/c$, and show $\varphi(0) = \varphi'(0) = 0$, $\varphi'' \le \frac{1}{4}$. (b) For m fixed predictors and a loss in $[0,1]$, bound $\mathbb{P}(\max_{j\le m}|\hat{L}_n(h_j) - L(h_j)| \ge u)$ by a union bound over Example 5.4, and compare with Corollary 5.7.

Exercise 5.13. type A (a) Show $h(x) = (e^x - 1 - x)/x^2$ is increasing on $\mathbb{R}$. (Put $k(x) = x^3 h'(x) = (x-2)e^x + x + 2$; $k'(x) = (x-1)e^x + 1$ and $k''(x) = xe^x$, so $k' \geq 0$, k is increasing with $k(0) = 0$, and $h'(x) = k(x)/x^3 \geq 0$ on both half-lines.) (b) Show $\psi_c(\lambda) \leq \frac{\lambda^2}{2(1-\lambda c/3)}$ for $0 < \lambda < 3/c$, using $k! \geq 2 \cdot 3^{k-2}$.

Exercise 5.14 (linear predictors)**.** type B Let $x_1, \dots, x_n \in \mathbb{R}^d$ with $\|x_i\| \leq B$ and $\mathcal{H} = \{x \mapsto \langle w, x\rangle : \|w\| \leq 1\}$. Show

$$\mathbb{E} \sup_{\|w\| \leq 1} \frac{1}{n} \sum_i \varepsilon_i \langle w, x_i\rangle = \frac{1}{n} \mathbb{E} \Big\| \sum_i \varepsilon_i x_i \Big\| \leq \frac{1}{n} \Big(\sum_i \|x_i\|^2 \Big)^{1/2} \leq \frac{B}{\sqrt{n}},$$

and combine with Example 5.5 and Proposition 5.6 (quoting that composing with a 1-Lipschitz loss does not increase the Rademacher complexity) to obtain a generalization bound independent of d.

Exercise 5.15 (the Robbins–Monro noise)**.** type B Let $|\xi_k| \leq c$ and $\eta_k = \eta/k$. For $m \geq 2$ let $M_n = \sum_{k=m}^n \eta_k \xi_k$, $n \geq m$. (a) Show the steps are bounded by $\eta c/m$ and $\langle M \rangle_n \leq \eta^2 c^2/(m-1)$ for all n. (b) Deduce from Theorem 5.10

$$\mathbb{P}\Big(\exists\, n \geq m : \Big| \sum_{k=m}^n \eta_k \xi_k \Big| \geq u \Big) \leq 2 \exp\Big(- \frac{u^2}{2\big(\eta^2 c^2/(m-1) + \eta c u/(3m)\big)} \Big).$$

(c) Compare with Theorem 4.2(ii).

Exercise 5.16 (bounded differences, two examples)**.** type B (a) Let L_n be the length of the longest increasing subsequence of n i.i.d. uniform points in $[0,1]$. Show L_n has bounded differences 1, so $\mathbb{P}(|L_n - \mathbb{E}L_n| \geq u) \leq 2e^{-2u^2/n}$. (It is known that $\mathbb{E}L_n \sim 2\sqrt{n}$ and that the fluctuations are of order $n^{1/6}$.) (b) Let B_n be the minimal number of unit bins needed to pack n i.i.d. sizes in $[0,1]$. Show B_n has bounded differences 1, and that fluctuations of order $\sqrt{n}$ do occur (compare B_n with the sum of the sizes).

Exercise 5.17 (Freedman versus Azuma for SGD)**.** type C In Exercise 2.12 take $\eta_k = \eta/k$, $\eta > \frac{1}{2}$, and state-dependent noise $\varepsilon_{k+1} = s(\theta_k)\zeta_{k+1}$ with ζ_k i.i.d. signs and $s(\theta) = \min(|\theta|, 1)$. (a) Show that for fixed n the noise term $N_n = \sum_{k<n} w_{n,k} s(\theta_k)\zeta_{k+1}$ is the terminal value of a martingale with quadratic variation $\sum_{k<n} w_{n,k}^2 s(\theta_k)^2$. (b) Write the Azuma and Freedman bounds for $\mathbb{P}(|N_n| \geq u)$. (c) Show that on the event $\{|\theta_k| \leq \delta \text{ for } k \geq n/2\}$ Freedman's variance term is smaller than Azuma's by a factor of order δ^2, and set up a bootstrap argument for the rate of $\theta_n \to 0$. (d) Simulate and compare.

Exercise 5.18 (the operator norm)**.** type C Let M be an $N \times d$ matrix with independent entries. (a) If the entries lie in $[-1,1]$, show $\|M\|_{\mathrm{op}}$ has bounded differences 2 in each entry, so $\mathbb{P}(\|M\|_{\mathrm{op}} - \mathbb{E}\|M\|_{\mathrm{op}} \geq u) \leq e^{-u^2/(2Nd)}$. Compare the scale $\sqrt{Nd}$ with $\mathbb{E}\|M\|_{\mathrm{op}} \asymp \sqrt{N} + \sqrt{d}$ (Lectures 9–11), and check that grouping the entries into rows gives the same scale. (b) If the entries are i.i.d. $\mathcal{N}(0,1)$, show $M \mapsto \|M\|_{\mathrm{op}}$ is 1-Lipschitz for the Frobenius norm, and use the Gaussian concentration box with $\mathbb{E}\|M\|_{\mathrm{op}} \leq \sqrt{N} + \sqrt{d}$ (Lecture 11) to get $\mathbb{P}(\|M\|_{\mathrm{op}} \geq \sqrt{N} + \sqrt{d} + u) \leq e^{-u^2/2}$. (c) Which property of $\|\cdot\|_{\mathrm{op}}$ does (b) use that bounded differences cannot express? (Talagrand's convex distance inequality is the bounded-variables result that can.)