# Solutions to the exercises, Lectures 12–15

Martingales

Numbering follows the notes *Lectures 12–15: Martingales*. Rules (i)–(viii) are those of Proposition 1.3; $\mathbb{E}_n$ is conditional expectation given $Y_1, \dots, Y_n$.

## Lecture 12

### 2.7 tower and taking out what is known

Let $V = L^2(Y) \subset W = L^2(Y, Y')$. Rule (iv): $P_V P_W = P_V$ gives $\mathbb{E}[\mathbb{E}[X \mid Y, Y'] \mid Y] = \mathbb{E}[X \mid Y]$; and $P_W P_V = P_V$ (because $P_V X \in V \subset W$ is left fixed by $P_W$) gives $\mathbb{E}[\mathbb{E}[X \mid Y] \mid Y, Y'] = \mathbb{E}[X \mid Y]$. Rule (iii): let $Z = P_V X$ and $g$ bounded. Then $g(Y)Z \in V$, and for every $f(Y) \in V$, $\langle g(Y)X - g(Y)Z, f(Y)\rangle = \langle X - Z, (gf)(Y)\rangle = 0$ since $(gf)(Y) \in V$. By the characterization of the projection, $g(Y)Z = P_V(g(Y)X)$.

### 2.8 conditional moments of a Gaussian

By Theorem 1.6, $X = AY + b + R$ with $A = \Sigma_{XY}\Sigma_{YY}^{-1}$, $b = \mu_X - A\mu_Y$, and $R \sim \mathcal{N}(0, s^2)$ independent of $Y$, $s^2 = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$. Expanding and using rules (iii) and (v),

$$\mathbb{E}[X^2 \mid Y] = (AY + b)^2 + 2(AY + b)\mathbb{E}R + \mathbb{E}R^2 = (AY + b)^2 + s^2,$$
$$\mathbb{E}[e^X \mid Y] = e^{AY+b}\,\mathbb{E}e^R = e^{AY+b+s^2/2}.$$

### 2.9 Doob martingales of a sum and a product

Sum: by the freezing lemma, $M_n = \mathbb{E}_n \sum_k Y_k = \sum_{k\le n} Y_k + \sum_{k>n} \mu_k$, with differences $\xi_n = Y_n - \mu_n$. Product: independence gives $M_n = \prod_{k\le n} Y_k \prod_{k>n} \mu_k$ (freezing with $U = (Y_1, \dots, Y_n)$, $Z = (Y_{n+1}, \dots, Y_N)$, $\mathbb{E}\prod_{k>n} Y_k = \prod_{k>n} \mu_k$), hence $\xi_n = (Y_n - \mu_n)\prod_{k<n} Y_k \prod_{k>n} \mu_k$. (Both need $f \in L^2$; for the product $\mathbb{E}f^2 = \prod_k \mathbb{E}Y_k^2$.)

### 2.10 the likelihood ratio tends to zero

By independence, $\mathbb{E}\sqrt{M_n} = \prod_{k\le n} \mathbb{E}\sqrt{q(Y_k)/p(Y_k)} = \big(\int \sqrt{q/p}\,p\big)^n = \rho^n$ with $\rho = \int \sqrt{pq}$. By Cauchy–Schwarz $\rho \le (\int p)^{1/2}(\int q)^{1/2} = 1$, with equality only if $\sqrt{q}$ is a constant multiple of $\sqrt{p}$, i.e. $q = p$. Markov: $\mathbb{P}(M_n \ge \varepsilon) = \mathbb{P}(\sqrt{M_n} \ge \sqrt{\varepsilon}) \le \rho^n/\sqrt{\varepsilon} \to 0$. There is no contradiction: $\mathbb{E}M_n = 1$ is carried by rare large values of $M_n$. Theorem 4.2 does not apply because $\mathbb{E}M_n^2 = (\int q^2/p)^n$ and $\int q^2/p \ge (\int q)^2 = 1$ (Cauchy–Schwarz on $q = (q/\sqrt{p})\sqrt{p}$), with equality only if $q = p$; so $\sup_n \mathbb{E}M_n^2 = \infty$.

### 2.11 Efron–Stein

(a) Freezing with $U = (Y_1, \dots, Y_{k-1})$ and $Z = (Y_k, \dots, Y_N)$ gives $\mathbb{E}_{k-1}f = g_{k-1}(Y_1, \dots, Y_{k-1})$, $g_{k-1}(y) = \mathbb{E}f(y, Y_k, \dots, Y_N)$. Freezing with $U = (Y_1, \dots, Y_k)$ and $Z = (Y_k', Y_{k+1}, \dots, Y_N)$, which is independent of $U$ and has the same law as $(Y_k, \dots, Y_N)$, gives $\mathbb{E}_k f^{(k)} = g_{k-1}(Y_1, \dots, Y_{k-1})$ as well. Hence $\xi_k = \mathbb{E}_k f - \mathbb{E}_{k-1} f = \mathbb{E}_k[f - f^{(k)}]$. (b) Rule (viii): $\mathbb{E}\xi_k^2 = \mathbb{E}(\mathbb{E}_k[f - f^{(k)}])^2 \le \mathbb{E}(f - f^{(k)})^2$.

Summing over $k$ with Lemma 2.3 gives $\operatorname{Var} f \le \sum_k \mathbb{E}(f - f^{(k)})^2$. (c) The factor $\frac{1}{2}$. Taking out what is known, $\mathbb{E}\xi_k^2 = \mathbb{E}[\xi_k \mathbb{E}_k(f - f^{(k)})] = \mathbb{E}[\xi_k(f - f^{(k)})]$, and $\mathbb{E}[(\mathbb{E}_{k-1}f)(f - f^{(k)})] = 0$ because $\mathbb{E}[f - f^{(k)} \mid Y_1, \dots, Y_{k-1}] = 0$. So $\mathbb{E}\xi_k^2 = \mathbb{E}[G(f - f^{(k)})]$ with $G = g_k(Y_1, \dots, Y_k) = \mathbb{E}_k f$. Exchanging the roles of $Y_k$ and $Y_k'$ leaves the joint law unchanged and maps $f \leftrightarrow f^{(k)}$, $G \leftrightarrow G' = g_k(Y_1, \dots, Y_{k-1}, Y_k')$; averaging the two expressions, $\mathbb{E}\xi_k^2 = \frac{1}{2}\mathbb{E}[(G - G')(f - f^{(k)})]$. Now $G - G' = \mathbb{E}[f - f^{(k)} \mid Y_1, \dots, Y_k, Y_k']$ (freezing again), so by (viii) $\mathbb{E}(G - G')^2 \le \mathbb{E}(f - f^{(k)})^2$, and Cauchy–Schwarz gives $\mathbb{E}\xi_k^2 \le \frac{1}{2}\mathbb{E}(f - f^{(k)})^2$. For $f = \sum_k g_k(Y_k)$: $\xi_k = g_k(Y_k) - \mathbb{E}g_k(Y_k)$ and $\mathbb{E}(f - f^{(k)})^2 = 2 \operatorname{Var} g_k(Y_k)$, so both sides equal $\sum_k \operatorname{Var} g_k(Y_k)$.

### 2.12 SGD on a quadratic

(a) $L(\theta) = \frac{1}{2}(\theta^2 + \sigma^2)$, $L'(\theta) = \theta$, $\nabla\ell(\theta; \varepsilon) = \theta - \varepsilon$, so $M_{n+1} = \sum_{k \le n} \eta_k\big((\theta_k - \varepsilon_{k+1}) - \theta_k\big) = -\sum_{k \le n} \eta_k \varepsilon_{k+1}$ and $\theta_{n+1} = \theta_0 - \sum_{k \le n} \eta_k \theta_k - M_{n+1}$. (b) $\theta_{n+1}^2 = (1 - \eta_n)^2\theta_n^2 + 2(1 - \eta_n)\eta_n\theta_n\varepsilon_{n+1} + \eta_n^2\varepsilon_{n+1}^2$; the middle term has mean zero ($\theta_n$ is a function of $\varepsilon_1, \dots, \varepsilon_n$, independent of $\varepsilon_{n+1}$). Hence $a_{n+1} = (1 - \eta_n)^2 a_n + \eta_n^2\sigma^2$ with $a_n = \mathbb{E}\theta_n^2$, and

$$a_n = \theta_0^2 \prod_{k<n} (1 - \eta_k)^2 + \sigma^2 \sum_{k<n} \eta_k^2 \prod_{j=k+1}^{n-1} (1 - \eta_j)^2.$$

(c) Since $(1 - \eta)^2 \le 1 - \eta$ on $[0, 1]$, $a_{n+1} \le (1 - \eta_n)a_n + \eta_n^2\sigma^2$. Fix $\epsilon > 0$ and $n_0$ with $\eta_n\sigma^2 \le \epsilon$ for $n \ge n_0$; then $a_{n+1} - \epsilon \le (1 - \eta_n)(a_n - \epsilon)$, so $(a_n - \epsilon)_+ \le (a_{n_0} - \epsilon)_+ \prod_{k=n_0}^{n-1}(1 - \eta_k) \le (a_{n_0} - \epsilon)_+ e^{-\sum_{k=n_0}^{n-1} \eta_k} \to 0$. Thus $\limsup a_n \le \epsilon$ for every $\epsilon$. For constant $\eta$: $a_{n+1} - a^* = (1 - \eta)^2(a_n - a^*)$ with $a^* = \eta^2\sigma^2/(1 - (1 - \eta)^2) = \eta\sigma^2/(2 - \eta)$, so $a_n \to a^*$ geometrically. (d) Unrolling, $\theta_n = \theta_0 \prod_{k<n}(1 - \eta_k) + \sum_{k<n} w_{n,k}\varepsilon_{k+1}$ with $w_{n,k} = \eta_k \prod_{j=k+1}^{n-1}(1 - \eta_j)$. The undamped sum $\sum_k \eta_k\varepsilon_{k+1}$ has $\mathbb{E}(\cdot)^2 = \sigma^2 \sum \eta_k^2$, so it converges in $L^2$ if and only if $\sum_k \eta_k^2 < \infty$ (Theorem 4.2 for "if"; for "only if", the partial sums are not Cauchy).

## Lecture 13

### 3.18 quadratic variations

Simple random walk: $\sigma_k^2 = 1$, $\langle S\rangle_n = n$. Exponential martingale $M_n = e^{\lambda S_n}/(\cosh\lambda)^n$: $\xi_n = M_{n-1}(e^{\lambda Y_n}/\cosh\lambda - 1)$ and $\mathbb{E}(e^{\lambda Y}/\cosh\lambda - 1)^2 = \cosh 2\lambda/\cosh^2\lambda - 1 = \tanh^2\lambda$, so $\langle M\rangle_n = \tanh^2\lambda \sum_{k<n} M_k^2$. Doob martingale of $\sum_{k \le N} Y_k$: $\xi_k = Y_k - \mu_k$, $\langle M\rangle_n = \sum_{k \le n} \operatorname{Var} Y_k$. Likelihood ratio: $\xi_n = M_{n-1}(q(Y_n)/p(Y_n) - 1)$ and $\mathbb{E}(q/p(Y) - 1)^2 = \int q^2/p - 1 =: \chi^2(q\|p)$, so $\langle M\rangle_n = \chi^2(q\|p) \sum_{k<n} M_k^2$. Transform by $H_k = S_{k-1}$: $\mathbb{E}[H_k^2\sigma_k^2] = \mathbb{E}S_{k-1}^2 = k - 1 < \infty$, so Definition 3.4 applies; $\mathbb{E}(H \cdot S)_n^2 = \sum_{k \le n}(k-1) = n(n-1)/2$; and since $S_k^2 - S_{k-1}^2 = 2S_{k-1}Y_k + Y_k^2 = 2S_{k-1}Y_k + 1$, summing gives $(H \cdot S)_n = \sum_{k \le n} S_{k-1}Y_k = \frac{1}{2}(S_n^2 - n)$. (Check: $\mathbb{E}S_n^4 = 3n^2 - 2n$ gives $\mathbb{E}(S_n^2 - n)^2 = 2n^2 - 2n$.)

### 3.19 the doubling strategy

(a) $\{\tau \ge k\} = \{Y_1 = \cdots = Y_{k-1} = -1\}$ is decided by $Y_1, \dots, Y_{k-1}$, so $H$ is predictable (and bounded by $2^{N-1}$). On $\{\tau = j\}$, $j \le N$, the stakes $1, 2, \dots, 2^{j-2}$ are lost and $2^{j-1}$ is won: gain $2^{j-1} - (2^{j-1} - 1) = 1$. On $\{\tau > N\}$ all stakes are lost: $-(2^N - 1)$. (b) $\mathbb{E}(H \cdot S)_N = (1 - 2^{-N}) \cdot 1 - 2^{-N}(2^N - 1) = 0$, as Proposition 3.5 requires. $\langle H \cdot S\rangle_N = \sum_{k \le N} H_k^2 = \sum_{k \le N} 4^{k-1}\mathbf{1}\{\tau \ge k\}$, with expectation $\sum_{k \le N} 4^{k-1}2^{-(k-1)} = 2^N - 1$. (c) As $N \to \infty$, $(H \cdot S)_N \to 1$ (since $\tau < \infty$ with

probability one) while $\mathbb{E}(H \cdot S)_N = 0$ for every $N$: the limit is not the limit of the means. This is consistent with Theorem 4.2 because $\mathbb{E}(H \cdot S)_N^2 = 2^N - 1$ is unbounded, and with optional stopping because $\tau$ is unbounded and the stopped process is not bounded in $L^2$. The "sure win" requires unbounded capital; the loss of $2^N - 1$ with probability $2^{-N}$ has expectation $\approx 1$ for every $N$.

### 3.20 hitting time of 1

(a) $\mathbb{P}(\tau_1 < \infty) \geq \mathbb{P}(S \text{ hits } 1 \text{ before } -b) = b/(1+b)$ for every $b$; let $b \to \infty$. (b) If $\mathbb{E}\tau_1 < \infty$, Wald's identity (Example 3.10(c), $\mu = 0$) gives $\mathbb{E}S_{\tau_1} = 0$; but $S_{\tau_1} = 1$. (c) $M_n = e^{\lambda S_n}/(\cosh\lambda)^n$ is a martingale, so $\mathbb{E}M_{\tau_1 \wedge n} = 1$. Split: on $\{\tau_1 \leq n\}$, $M_{\tau_1} = e^{\lambda}(\cosh\lambda)^{-\tau_1}$; on $\{\tau_1 > n\}$, $S_n \leq 0$ so $0 \leq M_n \leq (\cosh\lambda)^{-n} \to 0$ uniformly ($\lambda > 0$). Hence $e^{\lambda}\mathbb{E}[(\cosh\lambda)^{-\tau_1}\mathbf{1}\{\tau_1 \leq n\}] \to 1$, and by monotone limits $\mathbb{E}[(\cosh\lambda)^{-\tau_1}] = e^{-\lambda}$. (With $s = 1/\cosh\lambda$: $\mathbb{E}s^{\tau_1} = (1 - \sqrt{1-s^2})/s$, the classical first-passage generating function.)

### 3.21 upcrossings of a random walk

(a) Between two consecutive times $k < k'$ with $S_k = 0$, $S_{k+1} = 1$ and $S_{k'} = 0$, $S_{k'+1} = 1$, the walk went from 1 down to 0 and back up to 1: a new upcrossing of $[0,1]$ is completed by time $k'+1$. Hence $U_n[0,1] \geq \#\{k < n : S_k = 0, S_{k+1} = 1\}$, whose expectation is $\frac{1}{2}\sum_{k<n}\mathbb{P}(S_k = 0)$. Since $\mathbb{P}(S_k = 0) = \binom{k}{k/2}2^{-k} \sim \sqrt{2/(\pi k)}$ for even $k$ (Stirling), this is $\sim \frac{1}{2}\sqrt{2n/\pi} = \sqrt{n/2\pi} \to \infty$. Theorem 3.14 gives $\mathbb{E}U_n[0,1] \leq \mathbb{E}S_n^+ \leq \mathbb{E}|S_n| \sim \sqrt{2n/\pi}$: the right order, within a factor 2. (b) $\mathbb{E}|M_n| \leq (\mathbb{E}M_n^2)^{1/2} \leq (\sum_k k^{-2})^{1/2} = \pi/\sqrt{6}$, so Corollary 3.15 gives $\mathbb{E}U_\infty[a,b] \leq (\pi/\sqrt{6} + |a|)/(b-a)$.

### 3.22 branching processes

(a) Freezing with $U = Z_n$ and $Z = (\zeta_{n+1,i})_{i \geq 1}$: $\mathbb{E}[\sum_{i \leq z}\zeta_{n+1,i}] = mz$, so $\mathbb{E}_n Z_{n+1} = mZ_n$ and $\mathbb{E}_n W_{n+1} = mZ_n/m^{n+1} = W_n$. (b) By the same freezing, $\operatorname{Var}(Z_{n+1} \mid Z_n) = s^2 Z_n$, so Definition 1.4 gives $\operatorname{Var} Z_{n+1} = \mathbb{E}[s^2 Z_n] + \operatorname{Var}(mZ_n) = s^2 m^n + m^2 \operatorname{Var} Z_n$. Dividing by $m^{2n+2}$: $\operatorname{Var} W_{n+1} = \operatorname{Var} W_n + s^2/m^{n+2}$. (c) $\operatorname{Var} W_n = s^2 \sum_{j=2}^{n+1} m^{-j} \leq s^2/(m(m-1))$ for $m > 1$, so $\sup_n \mathbb{E}W_n^2 < \infty$; Theorem 4.2 gives $W_n \to W_\infty$ in $L^2$, $\mathbb{E}W_\infty = \lim \mathbb{E}W_n = 1$, $\operatorname{Var} W_\infty = \lim \operatorname{Var} W_n = s^2/(m(m-1))$. (d) 0 is absorbing, so $\{\tau \leq n\} = \{Z_n = 0\}$, a function of $Z_n$. For $m < 1$, Markov's inequality: $\mathbb{P}(\tau > n) = \mathbb{P}(Z_n \geq 1) \leq \mathbb{E}Z_n = m^n$. (e) Poisson(1.5) offspring, 4000 runs, 25 generations: $\mathbb{E}W \approx 1.02$, $\operatorname{Var} W \approx 2.0$ (predicted $s^2/(m(m-1)) = 1.5/0.75 = 2$), extinction in 42% of runs; the surviving runs have $W_{25}$ spread over $(0, \sim 8)$.

## Lecture 14

### 4.13 Doob's $L^2$ maximal inequality

(a) $|M_{n-1}| = |\mathbb{E}_{n-1}M_n| \leq \mathbb{E}_{n-1}|M_n|$ by rule (vi); adaptedness and square-integrability are inherited. (b) Let $\tau = \min\{k \leq n : |M_k| \geq u\}$ ($\tau = n$ if none) and $A = \{M_n^* \geq u\}$. By Corollary 3.8 for the submartingale $|M|$, $\mathbb{E}|M_n| \geq \mathbb{E}|M_\tau| = \mathbb{E}[|M_\tau|\mathbf{1}_A] + \mathbb{E}[|M_\tau|\mathbf{1}_{A^c}]$. On $A^c$, $\tau = n$; on $A$, $|M_\tau| \geq u$. Hence $\mathbb{E}[|M_n|\mathbf{1}_A] \geq \mathbb{E}[|M_\tau|\mathbf{1}_A] \geq u\mathbb{P}(A)$. (c) $\mathbb{E}(M_n^*)^2 = \int_0^\infty 2u\mathbb{P}(M_n^* \geq u)\,du \leq \int_0^\infty 2\,\mathbb{E}\big[|M_n|\mathbf{1}\{M_n^* \geq u\}\big]du = 2\mathbb{E}\big[|M_n|M_n^*\big] \leq 2\sqrt{\mathbb{E}M_n^2}\sqrt{\mathbb{E}(M_n^*)^2}$, and $\mathbb{E}(M_n^*)^2 < \infty$ since $M_n^* \leq \sum_{k \leq n}|M_k|$. Divide.

### 4.14 random series

$\mathbb{E}(M_m - M_n)^2 = \sum_{n<k\le m} a_k^2$ for $M_n = \sum_{k\le n} a_k\varepsilon_k$. For $a_k = 1/k$ this is $\le 1/n$, so $(M_n)$ is Cauchy and converges in $L^2$ (Theorem 4.2). For $a_k = 1/\sqrt{k}$, $\mathbb{E}(M_{2n} - M_n)^2 = \sum_{n<k\le 2n} 1/k \ge \frac{1}{2}$ for all $n$: not Cauchy. For the first series, Theorem 4.2(ii) gives $\mathbb{P}(\sup_{k\ge n} |M_k - M_n| \ge u) \le u^{-2}\sum_{k>n} k^{-2} \le 1/(nu^2)$.

### 4.15 the heat equation, two checks

$f(x) = e^{i\theta x}$: $\phi(x,v) = e^{i\theta x}\mathbb{E}e^{i\theta\sqrt{v}g} = e^{i\theta x - \theta^2 v/2}$, so $\partial_v\phi = -\frac{\theta^2}{2}\phi$ and $\partial_{xx}\phi = -\theta^2\phi$. $f(x) = x^4$: $\phi(x,v) = \mathbb{E}(x + \sqrt{v}g)^4 = x^4 + 6vx^2 + 3v^2$ (odd moments vanish, $\mathbb{E}g^4 = 3$), so $\partial_v\phi = 6x^2 + 6v$ and $\frac{1}{2}\partial_{xx}\phi = 6x^2 + 6v$.

### 4.16 a two-armed bandit

Set-up: let $U_1, U_2, \dots$ be i.i.d. uniform on $[0,1]$, independent of the rewards, and implement the rule as: if arm 1 currently has the larger empirical mean, pull it iff $U_i \ge \delta$; otherwise pull it iff $U_i \ge 1-\delta$. Both have the required probabilities, and since $\delta < \frac{1}{2}$, $\{U_i \ge 1-\delta\} \subset \{\text{arm 1 pulled at round } i\}$ in either case. Take $\mathcal{F}_i$ to be generated by the rewards of rounds $1,\dots,i$ and $U_1,\dots,U_{i+1}$; then the pull indicator $a_i$ is in $L^2(\mathcal{F}_{i-1})$ and the reward noise $\varepsilon_i$ is independent of $\mathcal{F}_{i-1}$. (a) $N_1(n) = \sum_{i\le n} a_i \ge \sum_{i\le n} \mathbf{1}\{U_i \ge 1-\delta\} \sim \mathrm{Bin}(n,\delta)$, so $\mathbb{P}(N_1(n) \le n\delta/2) \to 0$ (Chebyshev). (b) The rewards of arm 1 are $\mu_1 + \varepsilon_i$ at the rounds with $a_i = 1$, so $N_1(n)\hat{\mu}_1(n) = \sum_{i\le n} a_i(\mu_1 + \varepsilon_i)$, i.e. $N_1(n)(\hat{\mu}_1(n) - \mu_1) = \sum_{i\le n} a_i\varepsilon_i$. (c) Example 4.12 with $v_n = p_1 n$ and $V_n = \sum a_i^2 = N_1(n)$: $\mathbb{E}|V_n/v_n - 1| \to 0$, so $\sum_{i\le n} a_i\varepsilon_i/\sqrt{p_1 n} \Rightarrow \mathcal{N}(0,1)$. Then $\sqrt{N_1(n)}(\hat{\mu}_1(n) - \mu_1) = \frac{\sum a_i\varepsilon_i}{\sqrt{p_1 n}} \cdot \sqrt{p_1 n/N_1(n)}$ and the second factor tends to 1 in probability.

### 4.17 least squares for an autoregression

(a) $X_k \in L^2(\mathcal{F}_k)$ and $\varepsilon_{k+1}$ is independent of $\mathcal{F}_k$, so $\mathbb{E}_k[X_k\varepsilon_{k+1}] = 0$ and $\mathbb{E}_k[X_k^2\varepsilon_{k+1}^2] = \sigma^2 X_k^2$; the transform hypothesis $\mathbb{E}[X_k^2\sigma^2] < \infty$ holds. (b) $X_k = \sum_{j\le k} \rho^{k-j}\varepsilon_j$, so $\mathbb{E}X_k^2 = \sigma^2(1-\rho^{2k})/(1-\rho^2) \to \sigma^2/(1-\rho^2)$ and $\mathrm{Cov}(X_j, X_k) = \rho^{|k-j|}\mathbb{E}X_{j\wedge k}^2 \le \rho^{|k-j|}\sigma^2/(1-\rho^2)$. By Wick's theorem, $\mathrm{Cov}(X_j^2, X_k^2) = 2\,\mathrm{Cov}(X_j, X_k)^2$, hence

$$\mathrm{Var}\Big(\sum_{k<n} X_k^2\Big) \le \frac{2\sigma^4}{(1-\rho^2)^2} \sum_{j,k<n} \rho^{2|j-k|} \le \frac{2\sigma^4}{(1-\rho^2)^2} \cdot \frac{1+\rho^2}{1-\rho^2}\, n,$$

so $\mathrm{Var}(\frac{1}{n}\sum_{k<n} X_k^2) = O(1/n)$, and $\frac{1}{n}\sum_{k<n} X_k^2 \to \sigma^2/(1-\rho^2)$ in $L^2$. (c) Put $c_n^2 = n\sigma^4/(1-\rho^2)$ and $\xi_{k+1} = X_k\varepsilon_{k+1}/c_n$. Then $V_n = \sigma^2\sum_{k<n} X_k^2/c_n^2 \to 1$ in $L^2$, hence in $L^1$; moreover

$$\sum_k \mathbb{E}|\xi_{k+1}|^3 = \frac{\mathbb{E}|\varepsilon_1|^3}{c_n^3} \sum_{k<n} \mathbb{E}|X_k|^3 = O(n \cdot n^{-3/2}) \to 0,$$
$$\mathbb{E}\max_k \sigma_{k+1}^2 \le \frac{\sigma^2}{c_n^2}\Big(\sum_{k<n} \mathbb{E}X_k^4\Big)^{1/2} = O(n^{-1} \cdot n^{1/2}) \to 0.$$

Corollary 4.6 gives $M_n/c_n \Rightarrow \mathcal{N}(0,1)$, and

$$\sqrt{n}(\hat{\rho}_n - \rho) = \frac{M_n}{c_n} \cdot \frac{c_n/\sqrt{n}}{\frac{1}{n}\sum_{k<n} X_k^2} \;\Rightarrow\; \mathcal{N}(0,1) \cdot \frac{\sigma^2/\sqrt{1-\rho^2}}{\sigma^2/(1-\rho^2)} = \mathcal{N}\big(0, 1-\rho^2\big).$$

(d) $\rho = 0.9$, $n = 2000$, $10^3$ runs: the sample of $\sqrt{n}(\hat{\rho}_n - \rho)$ has mean $-0.04$ and variance $0.189$ against the predicted $1 - \rho^2 = 0.19$.

### 4.18 SGD on a quadratic: fluctuations

Take $\frac12 < \eta \le 1$ (for $\eta > 1$ the finitely many factors $1-\eta/j < 0$ change nothing below). (a) $\log\prod_{j=k+1}^{n-1}(1-\eta/j) = -\eta\sum_{j=k+1}^{n-1}\frac1j + O(\sum_{j>k} j^{-2}) = -\eta\log\frac{n}{k} + O(1/k)$, so $w_{n,k} = \frac{\eta}{k}(\frac{k}{n})^\eta e^{O(1/k)}$: the ratio to $\frac{\eta}{k}(k/n)^\eta$ is bounded for all $k$ and tends to 1 as $k\to\infty$, uniformly in $n > k$. (b) $n\operatorname{Var}\theta_n = n\sigma^2\sum_{k<n} w_{n,k}^2$. The terms with $k \le K$ contribute $O(n^{1-2\eta}) \to 0$ for any fixed $K$; the rest is $(1+o(1))\,\eta^2\sigma^2 n^{1-2\eta}\sum_{k<n} k^{2\eta-2} \to \eta^2\sigma^2/(2\eta-1)$, since $\sum_{k<n} k^{2\eta-2} \sim n^{2\eta-1}/(2\eta-1)$ for $2\eta-2 > -1$. The initial term $\theta_0\prod_{k<n}(1-\eta_k) = O(\theta_0 n^{-\eta})$ is $o(n^{-1/2})$. (c) The summands $\sqrt{n}\,w_{n,k}\varepsilon_{k+1}/s_n$, $s_n^2 = n\operatorname{Var}\theta_n$, are independent, so $V$ is deterministic and equals 1. Third moments: $\sum_k(\sqrt{n}w_{n,k})^3 \asymp n^{3/2-3\eta}\sum_{k<n} k^{3\eta-3}$, which is $O(n^{-1/2})$ if $\eta > \frac23$, $O(n^{-1/2}\log n)$ if $\eta = \frac23$, and $O(n^{3/2-3\eta})$ if $\eta < \frac23$; all tend to 0 for $\eta > \frac12$. Largest variance: $\max_k nw_{n,k}^2 = O(n^{1-2\eta}) \to 0$. Theorem 4.5 gives $\sqrt{n}\theta_n/s_n \Rightarrow \mathcal{N}(0,1)$, i.e. $\sqrt{n}\theta_n \Rightarrow \mathcal{N}(0,\eta^2\sigma^2/(2\eta-1))$. Simulation ($\eta=\sigma=1$, $n=2000$, 2000 runs): sample variance 1.00. For $\eta < \frac12$, $\sum_k k^{2\eta-2}$ converges, so $\operatorname{Var}\theta_n \asymp n^{-2\eta} \gg 1/n$: the early steps dominate, the initial condition is not forgotten ($\theta_0 n^{-\eta}$ is of the same order), the largest-variance hypothesis fails, and $n^\eta\theta_n$ converges in $L^2$ to a non-Gaussian limit.

## Lecture 15

### 5.12 Hoeffding by symmetrization; a union bound

(a) Jensen: $\mathbb{E}e^{-\lambda\xi'} \ge e^{-\lambda\mathbb{E}\xi'} = 1$, so $\mathbb{E}e^{\lambda\xi} \le \mathbb{E}e^{\lambda\xi}\,\mathbb{E}e^{-\lambda\xi'} = \mathbb{E}e^{\lambda(\xi-\xi')}$. The pair $(\xi,\xi')$ has the same law as $(\xi',\xi)$, so $\xi-\xi'$ and $-(\xi-\xi')$ have the same law, hence so does $\varepsilon(\xi-\xi')$. Conditioning on $(\xi,\xi')$, $\mathbb{E}[e^{\lambda\varepsilon y}] = \cosh(\lambda y) \le e^{\lambda^2y^2/2}$ (compare Taylor series), and $|\xi-\xi'| \le c$. Sharpness: for $\xi = \pm c/2$, $\mathbb{E}e^{\lambda\xi} = \cosh(\lambda c/2) = 1 + \lambda^2c^2/8 + O(\lambda^4)$ agrees with $e^{\lambda^2c^2/8}$ to second order. (b) By Example 5.4 and its lower-tail version, $\mathbb{P}(|\hat L_n(h_j) - L(h_j)| \ge u) \le 2e^{-2nu^2}$ for each $j$, so the union bound gives $2me^{-2nu^2}$: with probability $1-\delta$ the maximum is at most $\sqrt{\log(2m/\delta)/(2n)}$. Corollary 5.7 gives $2\sqrt{2\log m/n} + \sqrt{\log(1/\delta)/(2n)}$ for the one-sided supremum: the same $\sqrt{\log m/n}$ scale with worse constants, but its proof (McDiarmid plus symmetrization) extends to infinite classes, where a union bound does not.

### 5.13 the function $\psi_c$

(a) With $k(x) = x^3h'(x) = (x-2)e^x + x + 2$: $k'(x) = (x-1)e^x + 1$, $k''(x) = xe^x$. So $k'$ decreases on $(-\infty,0]$ and increases on $[0,\infty)$ with $k'(0) = 0$, hence $k' \ge 0$ and $k$ is increasing with $k(0) = 0$. Thus $k \le 0$ on $x < 0$ and $k \ge 0$ on $x > 0$, and $h'(x) = k(x)/x^3 \ge 0$ on both half-lines; $h$ is continuous at 0 with $h(0) = \frac12$. (b) $e^x - 1 - x = \sum_{j\ge2} x^j/j!$ and $j! \ge 2\cdot3^{j-2}$, so for $0 < x < 3$, $e^x - 1 - x \le \frac{x^2}{2}\sum_{i\ge0}(x/3)^i = \frac{x^2}{2(1-x/3)}$. With $x = \lambda c$, $\psi_c(\lambda) \le \frac{\lambda^2}{2(1-\lambda c/3)}$.

### 5.14 linear predictors

For fixed signs, $\sup_{\|w\|\le1}\langle w,v\rangle = \|v\|$ with $v = \sum_i \varepsilon_i x_i$ (Cauchy–Schwarz, attained at $w = v/\|v\|$). Then $\mathbb{E}\|v\| \le (\mathbb{E}\|v\|^2)^{1/2} = (\sum_i \|x_i\|^2)^{1/2} \le B\sqrt{n}$, the cross terms vanishing because the signs are independent and centered. Dividing by $n$ gives $B/\sqrt{n}$. For a loss $\ell(h,z) = \varphi(h(x),y) \in [0,1]$ with $\varphi$ 1-Lipschitz in its first argument, the contraction property gives $\mathbb{E}\sup_h \frac1n\sum\varepsilon_i\ell(h,z_i) \le B/\sqrt{n}$, so by Proposition 5.6 and Example 5.5, with probability at least $1-\delta$,

$$\sup_{\|w\|\le1}\big(L(w) - \hat L_n(w)\big) \le \frac{2B}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

with no dependence on $d$.

### 5.15 the tail of the Robbins–Monro noise

(a) $|\eta_k\xi_k| \le \eta c/k \le \eta c/m$ for $k \ge m$, and $\langle M\rangle_n = \sum_{k=m}^n \eta_k^2 \mathbb{E}_{k-1}\xi_k^2 \le \eta^2c^2\sum_{k\ge m} k^{-2} \le \eta^2c^2\int_{m-1}^\infty x^{-2}dx = \eta^2c^2/(m-1)$. (b) Theorem 5.10 with step bound $\eta c/m$ and $v = \eta^2c^2/(m-1)$; the event $\{\langle M\rangle_n \le v\}$ always holds, so $\mathbb{P}(\exists n \ge m : M_n \ge u) \le \exp(-u^2/(2(v + \eta cu/(3m))))$; apply to $-M$ as well. (c) Theorem 4.2(ii) gives $\mathbb{P}(\sup_{n\ge m}|M_n| \ge u) \le u^{-2}\sum_{k\ge m}\eta_k^2\mathbb{E}\xi_k^2 \le \eta^2c^2/((m-1)u^2)$: the same scale $u \asymp \eta c/\sqrt{m}$, but a polynomial rather than exponential tail.

### 5.16 bounded differences, two examples

(a) Moving one point changes the longest increasing subsequence by at most 1: deleting a point from an increasing subsequence shortens it by at most one, and the new point can extend one by at most one. McDiarmid with $c_k = 1$ gives the stated bound. The true fluctuations are of order $n^{1/6}$ (Baik–Deift–Johansson): bounded differences see only that each point matters by at most 1, not that most points do not matter at all. (b) Changing one item changes the optimal number of bins by at most 1 (remove the old item from an optimal packing, put the new one in a fresh bin). $B_n \ge \sum_i s_i$ because each bin holds total size at most 1, and for uniform sizes it is known that $\mathbb{E}B_n = \mathbb{E}\sum_i s_i + O(\sqrt{n})$; since $\sum_i s_i$ has standard deviation $\sqrt{n/12}$, $B_n$ does fluctuate at scale $\sqrt{n}$, and McDiarmid's scale is correct.

### 5.17 Freedman versus Azuma for SGD

(a) As in 2.12(d), $\theta_n = \theta_0\prod_{k<n}(1-\eta_k) + N_n$ with $N_n = \sum_{k<n} w_{n,k}s(\theta_k)\zeta_{k+1}$. For fixed $n$, $m_j = \sum_{k<j} w_{n,k}s(\theta_k)\zeta_{k+1}$, $j \le n$, is the transform of the martingale $\sum_{k<j}\zeta_{k+1}$ by the bounded predictable $H_{k+1} = w_{n,k}s(\theta_k)$, so it is a martingale with $\langle m\rangle_j = \sum_{k<j} w_{n,k}^2 s(\theta_k)^2$, and $N_n = m_n$. (b) Azuma, with $|H_{k+1}\zeta_{k+1}| \le w_{n,k}$: $\mathbb{P}(|N_n| \ge u) \le 2\exp(-u^2/(2\sum_{k<n} w_{n,k}^2))$. Freedman, with steps bounded by $w^* = \max_k w_{n,k}$: $\mathbb{P}(|N_n| \ge u,\ \sum_{k<n} w_{n,k}^2 s(\theta_k)^2 \le v) \le 2\exp(-u^2/(2(v + w^*u/3)))$. (c) On $\{|\theta_k| \le \delta \text{ for } k \ge n_0\}$, $\sum_{k<n} w_{n,k}^2 s(\theta_k)^2 \le \delta^2\sum_{k\ge n_0} w_{n,k}^2 + \sum_{k<n_0} w_{n,k}^2$, and $\sum_{k<n_0} w_{n,k}^2 \asymp n^{-2\eta}n_0^{2\eta-1}$ is negligible against $\sum_{k<n} w_{n,k}^2 \asymp 1/n$ when $n_0 = o(n)$. So Freedman's variance term is $\approx \delta^2$ times Azuma's. The bootstrap is cleanest in $L^2$: since $s(\theta)^2 \le \theta^2$,

$$\mathbb{E}\theta_{k+1}^2 = (1-\eta_k)^2\mathbb{E}\theta_k^2 + \eta_k^2\mathbb{E}s(\theta_k)^2 \le \big(1 - 2\eta_k + 2\eta_k^2\big)\mathbb{E}\theta_k^2,$$

so $\mathbb{E}\theta_n^2 \le \theta_0^2\prod_{k<n}(1-2\eta_k+2\eta_k^2) \asymp n^{-2\eta}$: with noise that vanishes at the minimizer the rate is $n^{-\eta}$, faster than the $n^{-1/2}$ of Exercise 4.18 as soon as $\eta > \frac12$; Freedman upgrades this to a bound on the whole trajectory with high probability. (d) $\eta = 1$, $\theta_0 = 1$, 500 runs: $n^2\,\mathbb{E}\theta_n^2 \approx 4.2,\ 4.4,\ 5.1$ for $n = 200, 800, 3200$, consistent with $\mathbb{E}\theta_n^2 \asymp n^{-2}$ up to a slowly varying factor.

### 5.18 the operator norm

(a) If $M'$ differs from $M$ in one entry by $\Delta$, $|\|M'\|_{\mathrm{op}} - \|M\|_{\mathrm{op}}| \le \|M' - M\|_{\mathrm{op}} = |\Delta| \le 2$. McDiarmid with $Nd$ coordinates and $c_k = 2$: $\exp(-2u^2/(4Nd)) = e^{-u^2/(2Nd)}$. Grouping into $N$ rows in $[-1,1]^d$: changing a row changes the norm by at most $\|r_i' - r_i\| \le 2\sqrt{d}$, so $\sum c^2 = 4Nd$ again. The fluctuation scale $\sqrt{Nd}$ exceeds the mean $\asymp \sqrt{N} + \sqrt{d}$ when both are large, so the bound says nothing. (b) $|\|M'\|_{\mathrm{op}} - \|M\|_{\mathrm{op}}| \le \|M' - M\|_{\mathrm{op}} \le \|M' - M\|_F$, so the map is 1-Lipschitz for the Euclidean norm on $\mathbb{R}^{N\times d}$. The Gaussian box gives $\mathbb{P}(\|M\|_{\mathrm{op}} \ge \mathbb{E}\|M\|_{\mathrm{op}} + u) \le e^{-u^2/2}$, and $\mathbb{E}\|M\|_{\mathrm{op}} \le \sqrt{N} + \sqrt{d}$. (c) The Gaussian inequality uses that $\|\cdot\|_{\mathrm{op}}$ is Lipschitz for the *Euclidean* norm of the whole matrix: many simultaneous small changes move it by their $\ell^2$ size. Bounded

differences only record the worst change from one coordinate and add these up, an $\ell^1$-type quantity that is $\sqrt{Nd}$ times larger here. Talagrand's convex distance inequality recovers Gaussian-type concentration for convex Lipschitz functions of independent bounded variables, which $\|\cdot\|_{\mathrm{op}}$ is (a supremum of linear functionals $\langle v, Mu\rangle$).