Notes 20 : Azuma's inequality

Math 733-734: Theory of Probability *Lecturer: Sebastien Roch*

References: [Roc, Sections 3.2].

Recall:

THM 20.1 (Markov's inequality) *Let X be a non-negative random variable. Then, for all $b > 0$,*

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}. \tag{1}$$

DEF 20.2 (Moment-generating function) *The* moment-generating function *of X is the function*

$$M_X(s) = \mathbb{E}\left[e^{sX}\right],$$

defined for all $s \in \mathbb{R}$ where it is finite, which includes at least $s = 0$.

THM 20.3 (Chernoff-Cramér bound) *Assume X is a centered random variable such that $M_X(s) < +\infty$ for $s \in (-s_0, s_0)$ for some $s_0 > 0$. For any $\beta > 0$ and $s > 0$,*

$$\mathbb{P}[X \geq \beta] \leq \exp\left[-\left\{s\beta - \Psi_X(s)\right\}\right], \tag{2}$$

where

$$\Psi_X(s) = \log M_X(s),$$

is the cumulant-generating function of X.

EX 20.4 (Gaussian random variables) *Let $X \sim N(0, \nu)$ where $\nu > 0$ is the variance. We have*

$$M_X(s) \quad = \quad \exp\left(\frac{s^2\nu}{2}\right),$$

so that straightforward calculus gives for $\beta > 0$

$$\sup_{s>0}(s\beta - s^2\nu/2) = \frac{\beta^2}{2\nu}, \tag{3}$$

achieved at $s_\beta = \beta/\nu$. Plugging this into (2) *leads for $\beta > 0$ to the bound*

$$\mathbb{P}[X \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\nu}\right). \tag{4}$$

We say that a centered random variable X is *sub-Gaussian with variance factor* $\nu > 0$ if for all $s \in \mathbb{R}$

$$\Psi_X(s) \leq \frac{s^2\nu}{2},$$

which is denoted by $X \in \mathcal{G}(\nu)$. By the Chernoff-Cramér bound, it follows that

$$\mathbb{P}\left[X \geq \beta\right] \leq \exp\left(-\frac{\beta^2}{2\nu}\right). \tag{5}$$

THM 20.5 (Hoeffding's inequality) *Let $X_1, \ldots, X_n$ be independent random variables where, for each i, X_i takes values in $[a_i, b_i]$ with $-\infty < a_i \leq b_i < +\infty$. Let $S_n = \sum_{i \leq n} X_i$. For all $\beta > 0$,*

$$\mathbb{P}[S_n - \mathbb{E}S_n \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i \leq n}(b_i - a_i)^2}\right).$$

LEM 20.6 (Hoeffding's lemma) *Let X be a random variable taking values in $[a, b]$ for $-\infty < a \leq b < +\infty$. Then $X - \mathbb{E}X \in \mathcal{G}\left(\frac{1}{4}(b-a)^2\right)$.*

We will also need:

LEM 20.7 (Orthogonality of increments) *Let $\{M_n\}$ be a MG with $M_n \in \mathcal{L}^2$. Let $s \leq t \leq u \leq v$. Then,*

$$\langle M_t - M_s, M_v - M_u \rangle = 0.$$

1 Concentration for martingales

The Chernoff-Cramér method extends naturally to martingales. This observation leads to powerful new concentration inequalities that hold far beyond the case of sums of independent variables. In particular, it will allow us to prove one version of the *concentration phenomenon*, which can be stated informally as:

> If $X_1, \ldots, X_n$ are independent (or "weakly dependent") random variables, then the random variable $f(X_1, \ldots, X_n)$ is "close" to its mean $\mathbb{E}f(X_1, \ldots, X_n)$ provided that the function $f(x_1, \ldots, x_n)$ is not too "sensitive" to any of the coordinates x_i.

1.1 Azuma-Hoeffding inequality

The main result of this section is the following generalization of Hoeffding's inequality (THM 20.5).

THM 20.8 (Azuma-Hoeffding inequality) *Let $(Z_t)_{t\in\mathbb{Z}_+}$ be a martingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}_+}$. Assume that there are predictable processes (A_t) and (B_t) (i.e., $A_t, B_t \in \mathcal{F}_{t-1}$) and constants $0 < c_t < +\infty$ such that: for all $t \geq 1$, almost surely,*

$$A_t \leq Z_t - Z_{t-1} \leq B_t \qquad \text{and} \qquad B_t - A_t \leq c_t.$$

Then for all $\beta > 0$

$$\mathbb{P}[Z_t - Z_0 \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i\leq t} c_i^2}\right).$$

Applying this inequality to $(-Z_t)$ gives a tail bound in the other direction.

Proof:[Proof of THM 20.8] As in the Chernoff-Cramér method, we start by applying (the exponential version of) Markov's inequality (THM 20.1), for $s > 0$,

$$\mathbb{P}[Z_t - Z_0 \geq \beta] \leq \frac{\mathbb{E}\left[e^{s(Z_t - Z_0)}\right]}{e^{s\beta}} = \frac{\mathbb{E}\left[e^{s\sum_{r=1}^t (Z_r - Z_{r-1})}\right]}{e^{s\beta}}. \tag{6}$$

This time, however, the terms in the exponent are not independent. Instead, to exploit the martingale property, we condition on the filtration

$$\mathbb{E}\left[\mathbb{E}\left[e^{s\sum_{r=1}^t (Z_r - Z_{r-1})} \,\middle|\, \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[e^{s\sum_{r=1}^{t-1} (Z_r - Z_{r-1})}\, \mathbb{E}\left[e^{s(Z_t - Z_{t-1})} \,\middle|\, \mathcal{F}_{t-1}\right]\right].$$

The assumption in the statement implies that, conditioned on $\mathcal{F}_{t-1}$, the random variable $Z_t - Z_{t-1}$ lies in an interval of length c_t. Hence by Hoeffding's lemma (LEM 20.6), it holds almost surely that

$$\mathbb{E}\left[e^{s(Z_t - Z_{t-1})} \,\middle|\, \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{s^2 c_t^2/4}{2}\right) = \exp\left(\frac{c_t^2 s^2}{8}\right). \tag{7}$$

Arguing by induction, we obtain

$$\mathbb{E}\left[e^{s(Z_t - Z_0)}\right] \leq \exp\left(\frac{s^2 \sum_{r\leq t} c_r^2}{8}\right).$$

Put differently, we have proved that $Z_t - Z_0$ is sub-Gaussian with variance factor $\frac{1}{4}\sum_{r\leq t} c_r^2$. Choosing $s = \beta / \frac{1}{4}\sum_{r\leq t} c_r^2$ in (6) gives the result. ∎

1.2 Method of bounded differences

The power of the Azuma-Hoeffding inequality is that it produces tail inequalities for quantities other than sums of independent random variables. The setting is the following. Let $X_1, \dots, X_n$ be independent random variables where X_i is $\mathcal{X}_i$-valued for all i and let $X = (X_1, \dots, X_n)$. Assume that $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is a measurable function. Our goal is to characterize the concentration properties of $f(X)$ around its expectation in terms of its "discrete derivatives"

$$\begin{aligned} D_i f(x) := \sup_{y \in \mathcal{X}_i} & f(x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_n) \\ & - \inf_{y' \in \mathcal{X}_i} f(x_1, \dots, x_{i-1}, y', x_{i+1}, \dots, x_n), \end{aligned}$$

where $x = (x_1, \dots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. We think of $D_i f(x)$ as a measure of the "sensitivity" of f to its i-th coordinate.

To analyze the behavior of $f(X)$, the idea is to consider the Doob martingale

$$Z_i = \mathbb{E}[f(X) \,|\, \mathcal{F}_i], \tag{8}$$

where $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$, which is well-defined provided $\mathbb{E}|f(X)| < +\infty$. Note that $Z_n = \mathbb{E}[f(X) \,|\, \mathcal{F}_n] = f(X)$ and $Z_0 = \mathbb{E}[f(X)]$ so that we can write

$$f(X) - \mathbb{E}[f(X)] = \sum_{i=1}^{n} (Z_i - Z_{i-1}).$$

A clever observation relates the martingale differences to the discrete derivatives through the use of an independent copy of X. Let $X' = (X'_1, \dots, X'_n)$ be an independent copy of X and let

$$X^{(i)} = (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n).$$

Then

$$\begin{aligned} Z_i - Z_{i-1} &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X) \,|\, \mathcal{F}_{i-1}] \\ &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \,|\, \mathcal{F}_{i-1}] \\ &= \mathbb{E}[f(X) \,|\, \mathcal{F}_i] - \mathbb{E}[f(X^{(i)}) \,|\, \mathcal{F}_i] \\ &= \mathbb{E}[f(X) - f(X^{(i)}) \,|\, \mathcal{F}_i]. \end{aligned}$$

Note that we crucially used the independence of the X_ks in the second and third lines. But then, by Jensen's inequality,

$$|Z_i - Z_{i-1}| \leq \|D_i f\|_\infty. \tag{9}$$

By the orthogonality of increments of martingales in $\mathcal{L}^2$, we immediately obtain

$$\mathrm{Var}[f(X)] = \mathbb{E}[(Z_n - Z_0)^2] = \sum_{i=1}^{n} \mathbb{E}\left[(Z_i - Z_{i-1})^2\right] \leq \sum_{i=1}^{n} \|D_i f\|_\infty^2.$$

Moreover, by the Azuma-Hoeffding inequality (THM 20.8) and the fact that $Z_i - Z_{i-1} \in [-\|D_i f\|_\infty, \|D_i f\|_\infty]$,

$$\mathbb{P}[f(X) - \mathbb{E}[f(X)] \geq \beta] \leq \exp\left(-\frac{\beta^2}{2\sum_{i\leq n} \|D_i f\|_\infty^2}\right).$$

A more careful analysis, which we do not detail here (but see [Roc]), leads to better bounds:

THM 20.9 (Bounded differences inequality) *Let $X_1, \ldots, X_n$ be independent random variables where X_i is $\mathcal{X}_i$-valued for all i and let $X = (X_1, \ldots, X_n)$. Assume that $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is a measurable function with $\mathbb{E}[f(X)^2] < +\infty$. Then*

$$\mathrm{Var}[f(X)] \leq \frac{1}{4} \sum_{i=1}^{n} \mathbb{E}[D_i f(X)^2].$$

.

THM 20.10 (McDiarmid's inequality) *Let $X_1, \ldots, X_n$ be independent random variables where X_i is $\mathcal{X}_i$-valued for all i, and let $X = (X_1, \ldots, X_n)$. Assume $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is a measurable function such that $\|D_i f\|_\infty < +\infty$ for all i. Then for all $\beta > 0$*

$$\mathbb{P}[f(X) - \mathbb{E}f(X) \geq \beta] \leq \exp\left(-\frac{2\beta^2}{\sum_{i\leq n} \|D_i f\|_\infty^2}\right).$$

Terminology: Bounds on $\|D_i f\|_\infty$ are often phrased in terms of a Lipschitz condition under an appropriate metric. Recall that the *Hamming distance* is defined as

$$\rho(x, x') := \sum_{i=1}^{n} \mathbb{1}_{\{x_i \neq x_i'\}}.$$

DEF 20.11 (Lipschitz condition) *Let $0 < c < +\infty$. We say that the function $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is c-*Lipschitz *(with respect to the Hamming distance) if for all $x, x' \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$*

$$|f(x) - f(x')| \leq c\rho(x, x').$$

LEM 20.12 *If f is c-Lipschitz, then $\|D_i f\|_\infty \leq c$ for all i.*

1.3 Examples

In this section, we give a few examples.

EX 20.13 (Longest common subsequence) *Let $X_1, \dots, X_{2n}$ be independent uniform random variables in $\{-1, +1\}$. Let Z be the length of the longest common subsequence in $(X_1, \dots, X_n)$ and $(X_{n+1}, \dots, X_{2n})$, that is,*

$$\begin{aligned} Z = \max \big\{ k \,:& \exists 1 \le i_1 < i_2 < \cdots < i_k \le n \\ & \text{and } n+1 \le j_1 < j_2 < \cdots < j_k \le 2n \\ & \text{such that } X_{i_1} = X_{j_1}, X_{i_2} = X_{i_2}, \dots, X_{i_k} = X_{j_k} \big\}. \end{aligned}$$

Then, writing $Z = f(X_1, \dots, X_{2n})$, it follows that $\|D_i f\|_\infty \le 1$. Indeed, fix $\mathbf{x} = (x_1, \dots, x_{2n})$ and let $\mathbf{x}^{i,+}$ (respectively $\mathbf{x}^{i,-}$) be $\mathbf{x}$ where the i-th component is replaced with $+1$ (respectively -1). Assume w.l.o.g. that $f(\mathbf{x}^{i,-}) \le f(\mathbf{x}^{i,+})$. Then $|f(\mathbf{x}^{i,+}) - f(\mathbf{x}^{i,-})| \le 1$ because removing the i-th component (and its match) from a longest common subsequence when $x_i = +1$ (if present) decreases the length by 1. Since this is true for any $\mathbf{x}$, we have $\|D_i f\|_\infty \le 1$. Finally, by THM 20.9,

$$\mathrm{Var}[Z] \le \frac{1}{4} \sum_{i=1}^{2n} \|D_i f\|_\infty^2 \le \frac{n}{2}.$$

EX 20.14 (Balls and bins: empty bins) *Suppose we throw m balls into n bins independently, uniformly at random. The number of empty bins, $Z_{n,m}$, is centered at*

$$\mathbb{E} Z_{n,m} = n \left(1 - \frac{1}{n}\right)^m.$$

Writing $Z_{n,m}$ as the sum of indicators $\sum_{i=1}^n \mathbb{1}_{B_i}$, where B_i is the event that bin i is empty, is a natural first attempt at proving concentration around the mean. However there is a problem—the B_is are not independent. *Indeed, because there is a fixed number of bins, the event B_i intuitively makes the other such events less likely. Instead let X_j be the index of the bin in which ball j lands. The X_js are independent by construction and, moreover, $Z_{n,m} = f(X_1, \dots, X_m)$ where f is 1-Lipschitz. Indeed, moving a single ball changes the number of empty bins by at most 1 (if at all). Hence by the method of bounded differences*

$$\mathbb{P}\left[\left| Z_{n,m} - n\left(1 - \frac{1}{n}\right)^m \right| \ge b\sqrt{m} \right] \le 2e^{-2b^2}.$$

EX 20.15 (Concentration of measure on the hypercube) *For $A \subseteq \{0,1\}^n$ a subset of the hypercube and $r > 0$, we let*

$$A_r = \left\{ x \in \{0,1\}^n \,:\, \inf_{a \in A} \|x - a\|_1 \le r \right\},$$

be the points at ℓ^1 distance r from A. Fix $\varepsilon \in (0, 1/2)$ and assume that $|A| \geq \varepsilon 2^n$. Let λ_ε be such that $e^{-2\lambda_\varepsilon^2} = \varepsilon$. The following application of the method of bounded differences indicates that much of the uniform measure on the high-dimensional hypercube lies in a close neighborhood of any such "small" set A. This is an example of the concentration of measure phenomenon.

CLAIM 20.16

$$r > 2\lambda_\varepsilon \sqrt{n} \implies |A_r| \geq (1-\varepsilon)2^n.$$

Proof: *Let $X = (X_1, \ldots, X_n)$ be uniformly distributed in $\{0,1\}^n$. Note that the coordinates are in fact independent. The function $f(x) = \inf_{a \in A} \|x - a\|_1$ is* 1*-Lipschitz. Indeed changing one coordinate of x can only increase the ℓ^1 distance to the closest point to x by at most* 1 *(and vice versa, if changing a coordinate led to a decrease of more than* 1*, then reversing the change would lead to a contradiction). Hence McDiarmid's inequality (Theorem 20.10) gives*

$$\mathbb{P}\left[\mathbb{E}f(X) - f(X) \geq \beta\right] \leq \exp\left(-\frac{2\beta^2}{n}\right).$$

Choosing $\beta = \mathbb{E}f(X)$ and noting that $f(x) \leq 0$ if and only if $x \in A$ gives

$$\mathbb{P}[A] \leq \exp\left(-\frac{2(\mathbb{E}f(X))^2}{n}\right),$$

or, rearranging and using our assumption on A,

$$\mathbb{E}f(X) \leq \sqrt{\frac{1}{2} n \log \frac{1}{\mathbb{P}[A]}} \leq \sqrt{\frac{1}{2} n \log \frac{1}{\varepsilon}} = \lambda_\varepsilon \sqrt{n}.$$

By a second application of the method of bounded differences with $\beta = \lambda_\varepsilon \sqrt{n}$,

$$\mathbb{P}\left[f(X) \geq 2\lambda_\varepsilon \sqrt{n}\right] \leq \mathbb{P}\left[f(X) - \mathbb{E}f(X) \geq b\right] \leq \exp\left(-\frac{2\beta^2}{n}\right) = \varepsilon.$$

The result follows by observing that, with $r > 2\lambda_\varepsilon \sqrt{n}$,

$$\frac{|A_r|}{2^n} \geq \mathbb{P}\left[f(X) < 2\lambda_\varepsilon \sqrt{n}\right] \geq 1 - \varepsilon.$$

■

CLAIM 20.16 is striking for two reasons: 1) the radius $2\lambda_\varepsilon \sqrt{n}$ is much smaller than n, the diameter of $\{0,1\}^n$; and 2) it applies to any *A.*

2 Erdös-Rényi: exposure martingales and application to the chromatic number

Recall that an *undirected graph* (or *graph* for short) is a pair $G = (V, E)$ where V is the set of *vertices* (or *nodes* or *sites*) and

$$E \subseteq \{\{u, v\} \,:\, u, v \in V\},$$

is the set of *edges* (or *bonds*). If $e = \{u, v\} \in E$, we say that e is *incident* to u and v and that u and v are *adjacent*. A *subgraph* of $G = (V, E)$ is a graph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$.

DEF 20.17 (Erdös-Rényi graphs) *Let $V = [n]$ and $p \in [0, 1]$. The* Erdös-Rényi graph $G = (V, E)$ *on n vertices with density p is defined as follows: for each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability p independently of all other edges. We write $G \sim \mathbb{G}_{n,p}$ and we denote the corresponding measure by $\mathbb{P}_{n,p}$.*

Exposure martingales In the context of Erdös-Rényi graphs, a common way to apply the Azuma-Hoeffding inequality (THM 20.8) is to introduce a so-called *exposure martingale*. Let $G = (V, E) \sim \mathbb{G}_{n,p}$ and let F be any function on graphs such that $\mathbb{E}_{n,p}|F(G)| < +\infty$ for all n, p. Choose an arbitrary ordering of the vertices (i.e., think of V as $\{1, \ldots, n\}$) and, for $i = 1, \ldots, n$, denote by H_i the subgraph G restricted to its first i vertices: its vertex set is $V_i = \{1, \ldots, i\}$ and its edges are $E_i = \{\{u, v\} \in E \,:\, u, v \in V_i\}$. Then the filtration $\mathcal{H}_i = \sigma(H_1, \ldots, H_i)$, $i = 1, \ldots, n$, corresponds to exposing the vertices of G one at a time. The Doob martingale

$$Z_i = \mathbb{E}_{n,p}[F(G) \,|\, \mathcal{H}_i], \qquad i = 1, \ldots, n,$$

is known as a *vertex exposure martingale*. An alternative way to define the filtration is to consider instead the random variables $X_i = (\mathbb{1}_{\{\{i,j\} \in G\}} \,:\, 1 \leq j \leq i)$ for $i = 2, \ldots, n$. In words, X_i is a vector whose entries indicate the status (present or absent) of all potential edges incident to i and a vertex preceding it. Hence, $\mathcal{H}_i = \sigma(X_2, \ldots, X_i)$ for $i = 2, \ldots, n$ (and note that $\mathcal{H}_1$ is trivial as it corresponds to a graph with a single vertex and no edge). This representation has an important property: the X_is are *independent* as they pertain to disjoint subsets of edges. We are then in the setting of the method of bounded differences. Re-writing $F(G) = f(X_2, \ldots, X_n)$ for some function f, the vertex exposure martingale coincides with the martingale (8) used in that context.

Chromatic number As an example, consider the chromatic number $\chi(G)$, i.e., the smallest number of colors needed in a proper coloring of G (that is, an assignment of colors to the vertices such that any two adjacent vertices have different colors). Define

$$f_\chi(X_2, \ldots, X_n) := \chi(G).$$

We use the following combinatorial observation to bound $\|D_i f_\chi\|_\infty$.

LEM 20.18 *Altering the status (absent or present) of edges incident to a fixed vertex v changes the chromatic number by at most* 1.

Proof: Altering the status of edges incident to v increases the chromatic number by at most 1, since in the worst case one can simply use an extra color for v. On the other hand, if the chromatic number were to decrease by more than 1 after altering the status of edges incident to v, reversing the change and using the previous observation would produce a contradiction. ■

A fortiori, since X_i depends on a *subset* of the edges incident to node i, LEM 20.18 implies that f_χ is 1-Lipschitz. Hence, for all $0 < p < 1$ and n, by an immediate application of the McDiarmid's inequality (THM 20.10):

CLAIM 20.19

$$\mathbb{P}_{n,p}\left[|\chi(G) - \mathbb{E}_{n,p}[\chi(G)]| \geq b\sqrt{n-1}\right] \leq 2e^{-2b^2}.$$

Edge exposure can be defined in a manner similar to vertex exposure: reveal the edges one at a time in an arbitrary order. By LEM 20.18, the corresponding function is again 1-Lipschitz. Observe however that, for the chromatic number, edge exposure results in a much weaker bound as the $\Theta(n^2)$ random variables produce only a *linear in* n deviation for the same tail probability. (The reader may want to ponder the apparent paradox: using a larger number of independent variables seemingly leads to weaker concentration in this case.)

3 A maximal Azuma-Hoeffding inequality

By using Doob's subMG inequality instead of Markov's inequality in the proof of Azuma-Hoeffding, we obtain a maximal version, which is sometimes useful to avoid union bounds. We will not detail applications here, but see [Roc].

THM 20.20 (Maximal Azuma-Hoeffding inequality) *Let $(Z_t)_{t\in\mathbb{Z}_+}$ be a martingale with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}_+}$. Assume that there are predictable*

processes (A_t) and (B_t) (i.e., $A_t, B_t \in \mathcal{F}_{t-1}$) and constants $0 < c_t < +\infty$ such that: for all $t \geq 1$, almost surely,

$$A_t \leq Z_t - Z_{t-1} \leq B_t \qquad \textit{and} \qquad B_t - A_t \leq c_t.$$

Then for all $\beta > 0$

$$\mathbb{P}\left[\sup_{0\leq i\leq t}(Z_i - Z_0) \geq \beta\right] \leq \exp\left(-\frac{2\beta^2}{\sum_{i\leq t} c_i^2}\right).$$

We first give a proof of the subMG inequality using stopping times.

LEM 20.21 *Let (M_t) be a submartingale and $\tau \leq \sigma$ be stopping times. Then*

$$\mathbb{E}[M_{\tau\wedge t}] \leq \mathbb{E}[M_{\sigma\wedge t}].$$

Proof: Consider the predictable process

$$C_n^{(\tau)} = \mathbb{1}\{n \leq \tau\},$$

and similarly for σ. Note that by the assumption $\tau \leq \sigma$

$$C_n^{(\tau)} = \mathbb{1}\{n \leq \tau\} \leq \mathbb{1}\{n \leq \sigma\} = C_n^{(\sigma)}.$$

Observe that

$$\begin{aligned} M_{\sigma\wedge t} - M_{\tau\wedge t} &= (C^{(\sigma)} \bullet X)_n - (C^{(\tau)} \bullet X)_n \\ &= \sum_{1\leq i\leq t} (C_i^{(\sigma)} - C_i^{(\tau)})(X_i - X_{i-1}). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}\left[M_{\sigma\wedge t} - M_{\tau\wedge t}\right] &= \sum_{1\leq i\leq t} \mathbb{E}\left[(C_i^{(\sigma)} - C_i^{(\tau)})(M_i - M_{i-1})\right] \\ &= \sum_{1\leq i\leq t} \mathbb{E}\left[\mathbb{E}\left[(C_i^{(\sigma)} - C_i^{(\tau)})(M_i - M_{i-1}) \,|\, \mathcal{F}_{i-1}\right]\right] \\ &= \sum_{1\leq i\leq t} \mathbb{E}\left[(C_i^{(\sigma)} - C_i^{(\tau)})\mathbb{E}\left[M_i - M_{i-1} \,|\, \mathcal{F}_{i-1}\right]\right] \\ &\geq 0, \end{aligned}$$

where we used the subMG property. ∎

LEM 20.22 (Doob's submartingale inequality) *Let $\{M_t\}$ be a nonnegative submartingale. Then for $b > 0$*

$$\mathbb{P}\left[\sup_{0\le i\le t} M_i \ge b\right] \le \frac{\mathbb{E}[M_t]}{b}.$$

(Markov's inequality implies only $\sup_{0\le i\le t}\mathbb{P}[M_i \ge b] \le \frac{\mathbb{E}[M_t]}{b}$ from the fact that $\mathbb{E}[M_i] \le \mathbb{E}[M_t]$ for $i \le t$.)

Proof: Let $\tau = \inf\{i \ge 0 : M_i \ge b\}$. Let $\sigma = t$ in the lemma above. Then $\mathbb{E}[M_{\tau\wedge t}] \le \mathbb{E}[M_t]$. In addition,

$$\mathbb{E}[M_t] \ge \mathbb{E}[M_{\tau\wedge t}] \ge b\,\mathbb{P}[\tau \le t] = b\,\mathbb{P}\left[\sup_{0\le i\le t} M_i \ge b\right],$$

where we used the nonnegativity of $\{M_t\}$. ■

Before proving THM 20.20, we recall one last lemma:

LEM 20.23 *If $\{M_t\}$ is a martingale and ϕ is a convex function with $\mathbb{E}|\phi(M_t)| < +\infty$ for all t then $\{\phi(M_t)\}$ is a submartingale.*

Proof: By Jensen's inequality

$$\mathbb{E}[\phi(M_t)\,|\,\mathcal{F}_{t-1}] \ge \phi(\mathbb{E}[M_t\,|\,\mathcal{F}_{t-1}]) = \phi(M_{t-1}).$$

■

Proof:(Proof of THM 20.20) We simply note that, by Doob's subMG inequality, for $s > 0$

$$\begin{aligned}\mathbb{P}\left[\sup_{0\le i\le t}(Z_i - Z_0) \ge \beta\right] &= \mathbb{P}\left[\sup_{0\le i\le t} e^{s(Z_i - Z_0)} \ge e^{s\beta}\right] \\ &\le \frac{\mathbb{E}\left[e^{s(Z_t - Z_0)}\right]}{e^{s\beta}},\end{aligned}$$

where we used the fact that e^{sx} is increasing and convex for $s > 0$. The rest of the proof is unchanged. ■

References

[Roc] Sebastien Roch. Modern Discrete Probability: An Essential Toolkit. *Book in preparation*. http://www.math.wisc.edu/ roch/mdp/.