University of Washington Math 523A Lecture 2
Martingales: concentration I

Lecturer: Eyal Lubetzky

April 3, 2009

The Hoeffding-Azuma concentration inequality

Theorem 1 (Hoeffding '63)**.** *Let (X_t) be a martingale with respect to $(\mathcal{F}_t)$, and let $c_1, c_2, \ldots$ be real numbers such that $|X_t - X_{t-1}| \le c_t$ for all t. Then for any $a > 0$,*

$$\mathbb{P}(|X_n - X_0| \ge a) \le 2\exp\left(-\frac{a^2}{2\sum_{i=1}^n c_i^2}\right).$$

Proof. We will show a slightly stronger version of the above theorem: Instead of requiring that (X_t) is a martingale, we will assume that

$$\mathbb{E}[X_t \mid X_{t-1}] = X_{t-1} \text{ for all } t\,. \tag{0.1}$$

Let $Y_t \triangleq X_t - X_{t-1}$. By the above assumptions $\mathbb{E}[Y_t \mid X_{t-1}] = 0$ and $|Y_t| \le c_t$ for all t.

The following simple claim says that, if Z is a r.v. bounded by 1 and with mean 0, then the expected value of $\exp(\lambda Z)$ is maximized when splitting the mass equally between ± 1.

Claim 2. *Let Z be a random variable satisfying $|Z| \le 1$ and $\mathbb{E}Z = 0$. Then*

$$\mathbb{E}[e^{\lambda Z}] \le \cosh\lambda \le e^{\lambda^2/2} \quad \textit{for any } \lambda \in \mathbb{R}\,.$$

Proof of claim. As the function $g(x) = e^{\lambda x}$ is convex, the following holds for any $x \in [-1, 1]$:

$$e^{\lambda x} \le \frac{1-x}{2}e^{-\lambda} + \frac{1+x}{2}e^{-\lambda} = \cosh\lambda + x\sinh\lambda\,,$$

and therefore

$$\mathbb{E}[e^{\lambda Z}] \le \cosh\lambda + (\mathbb{E}Z)\sinh\lambda = \cosh\lambda\,.$$

The proof now follows from the fact that

$$\cosh\lambda = \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} \le \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2}\,. \qquad \blacksquare$$

Applying the above claim to Y_t/c_t given X_{t-1} (and then re-scaling by c_t), we now get

$$\mathbb{E}\left[\mathrm{e}^{\lambda Y_i} \mid X_{t-1}\right] \leq \mathrm{e}^{(\lambda c_t)^2/2} \text{ for all } \lambda > 0\,.$$

Let $\lambda > 0$ be specified later. Noticing that

$$\mathbb{E}\left[\mathrm{e}^{\lambda(X_n - X_0)}\right] = \mathbb{E}\left[\mathrm{e}^{\lambda(X_{n-1}-X_0)}\mathrm{e}^{\lambda Y_n}\right]\,,$$

it follows that

$$\begin{aligned}\mathbb{E}\left[\mathrm{e}^{\lambda(X_n - X_0)} \mid X_{n-1}\right] &= \mathbb{E}\left[\mathrm{e}^{\lambda(X_{n-1}-X_0)}\mathrm{e}^{\lambda Y_n} \mid X_{n-1}\right] = \mathrm{e}^{\lambda(X_{n-1}-X_0)}\mathbb{E}\left[\mathrm{e}^{\lambda Y_n} \mid X_{n-1}\right] \\ &\leq \mathrm{e}^{\lambda(X_{n-1}-X_0)}\mathrm{e}^{(\lambda c_n)^2/2}\,.\end{aligned}$$

Taking expectations,

$$\mathbb{E}\left[\mathrm{e}^{\lambda(X_n - X_0)}\right] \leq \mathbb{E}\left[\mathrm{e}^{\lambda(X_{n-1}-X_0)}\right]\mathrm{e}^{(\lambda c_n)^2/2}\,,$$

and by iterating we deduce that

$$\mathbb{E}\left[\mathrm{e}^{\lambda(X_n - X_0)}\right] \leq \exp\left[\tfrac{1}{2}\lambda^2 \sum_{i=1}^n c_i^2\right].$$

Applying Markov's inequality gives

$$\mathbb{P}(X_n - X_0 \geq a) = \mathbb{P}\left(\mathrm{e}^{\lambda(X_n - X_0)} \geq \mathrm{e}^{\lambda a}\right) \leq \exp\Big[-\lambda a + \tfrac{1}{2}\lambda^2 \sum_{i=1}^n c_i^2\Big]\,,$$

at which point a choice of $\lambda = a/\sum_{i=1}^n c_i^2$ concludes the proof (the bound on the event $X_0 - X_n \geq a$ follows from a similar argument). ■

Examples

1. Sums of i.i.d. variables

Consider Example 1 from the previous lesson: $Y_1, Y_2, \ldots$ are i.i.d. variables,

$$Y_1 \sim \begin{cases} 1 & p \\ -1 & q \triangleq 1-p \end{cases}$$

for some $\frac{1}{2} \leq p < 1$ fixed, and $S_n = \sum_{i=1}^n Y_i$.

- For $p = \frac{1}{2}$, applying Hoeffding's inequality to the martingale S_n (together with a choice of $c_t = 1$ for all t) gives the well known Chernoff bound:

$$\mathbb{P}\left(|S_n| \geq a\right) \leq 2\exp\left(-a^2/2n\right)\,.$$

- For $p > \frac{1}{2}$, setting $\mu = p - q$, and then applying Hoeffding's inequality to the martingale $X_n = S_n - n\mu$ (together with a choice of $c_t = 1 + \mu$ for all t) gives the bound:

$$\mathbb{P}\left(|S_n - n\mu| \geq a\sqrt{n}\right) \leq 2\exp\left(-a^2/2(1+\mu)\right)\,.$$

Remark. Note that if $N \sim \mathcal{N}(0,1)$ has a standard normal distribution, then for $a \gg 1$ we have

$$\mathbb{P}(N > a) = \frac{1}{\sqrt{2\pi}} \int_a^\infty \mathrm{e}^{-t^2/s} dt \approx \frac{1}{\sqrt{2\pi a}} \mathrm{e}^{-a^2/2} .$$

That is, Hoeffding's inequality implies that the tail of the martingale resembles that in a normal distributed random variable, and the exponent in this large-deviation bound is optimal.

2. Doob's martingale process

Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{E}Y^2 < \infty$. Let (X_t) be a sequence of r.v.'s adapted to a filtration $(\mathcal{F}_t) \subset \mathcal{F}$. The following sequence is *Doob's martingale process*:

$$Y_t \triangleq \mathbb{E}[Y \mid \mathcal{F}_t] .$$

To verify that this is a martingale, note that:

(i) Clearly $\mathbb{E}|Y_t| < \infty$ for all t, since the requirement $\mathbb{E}Y^2 < \infty$ implies that $\mathbb{E}Y_t^2 < \infty$.

(ii) For all t we have

$$\mathbb{E}\left[Y_{t+1} \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}[Y \mid \mathcal{F}_{t+1}] \mid \mathcal{F}_t\right] = \mathbb{E}\left[Y \mid \mathcal{F}_t\right] = Y_t ,$$

where the first inequality is by definition and the second one is due to the fact that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$.

To demonstrate how useful the above martingale can be, we need a few Graph Theoretic definitions. Let $G = (V, E)$ be a graph. That is, its vertex set is V (the vertices are unlabeled unless mentioned otherwise), and its set of edges is $E \subset \binom{V}{2}$ (undirected unless mentioned otherwise).

Definition 1. *An* independent set *of a graph $G = (V, E)$ is a set of vertices $S \subset V$ with no edges between them: For all $u, v \in S$ we have $(u, v) \notin E$. The maximum cardinality of an independent set of G is called the* independence number *of G, and denoted by $\alpha(G)$.*

Finding the independence number of a graph is a formidable problem in Computer Science (NP-hard to approximate even within a factor of $|V|^{1-\varepsilon}$). Nevertheless, the graph property of either having independent sets above a certain size, or not, is relatively intuitive. In that sense, the following parameter is even more difficult to understand:

Definition 2. *A* legal coloring *of a graph $G = (V, E)$ using k colors is a mapping $\varphi : V \to \{1, \ldots, k\}$ that maps neighboring vertices to distinct colors:*

$$\varphi(u) \neq \varphi(v) \text{ for any } (u, v) \in E .$$

The smallest integer k that admits a (legal) k-coloring is called the chromatic number *of G, and denoted by $\chi(G)$.*

Observe that a legal coloring of a graph $G = (V, E)$ by k colors is equivalent to finding k disjoint independent sets $S_1, \ldots, S_k$ that cover all the vertices: $V = \cup_i S_i$ (the set S_i is simply $\varphi^{-1}(i)$). Clearly, if the vertices of a graph can be covered by k independent sets, then they can also be covered by k *disjoint* independent sets, and so $\chi(G)$ is equivalent to the minimal number of independent sets needed to cover all vertices of G.

Definition 3. *The Erdős-Rényi random graph model $\mathcal{G}(n,p)$ is defined as follows: $G \sim \mathcal{G}(n,p)$ has the vertex set $\{1,\ldots,n\}$, and each edge (i,j) for $i \neq j \in V$ belongs to G with probability p, independent of the other edges.*

To study the concentration of $\chi(G)$ for $G \sim \mathcal{G}(n,p)$, we use Doob's martingale for graphs:

Definition 4 (Edge exposure martingale)**.** *Let f be a function on n-vertex graphs, set $m = \binom{n}{2}$ and let $e_1, e_2, \ldots, e_m$ be an arbitrary ordering of the edges of the complete graph on n vertices. For $G \sim \mathcal{G}(n,p)$ and $i \in \{1,\ldots,m\}$, let $A_i = \mathbf{1}_{\{e_i \in G\}}$, and define*

$$X_t \triangleq \mathbb{E}\left[f(G) \mid A_1, \ldots, A_t\right] \text{ for } t \in \{0,\ldots,m\}\,.$$

In other words, X_t is the expected value of f over all graphs $H \in \mathcal{G}(n,p)$ that agree with G on $\{e_1,\ldots,e_t\}$. According to this definition, our filtration exposes the edges of G sequentially. The variable X_t is the goal function f (playing the role of Y in the definition of Doob's martingale) averaged over all possible "continuations" of this sequence.

For a graph G and a subset S of its vertices, the *induced subgraph* of G on S, denoted by $G|_S$, is the restriction of the graph to S: its vertex set is S, and there is an edge between $u, v \in S$ iff this edge exists in G. Similar to the edge exposure martingale, one can define the following:

Definition 5 (Vertex exposure martingale)**.** *Let f be a function on n-vertex graphs, $m = n-1$ and $v_1, v_2, \ldots, v_n$ an arbitrary ordering of the vertices of the complete graph. For $G \sim \mathcal{G}(n,p)$, define*

$$X_t \triangleq \mathbb{E}\left[f(G) \mid G|_{\{v_1,\ldots,v_{t+1}\}}\right] \text{ for } t \in \{0,\ldots,m\}\,.$$

That is, in step i we reveal all the edges between the vertex v_{i+1} and its predecessors $\{v_1,\ldots,v_i\}$.

Note that in both of the above defined martingales, $X_0 = \mathbb{E}f(G)$, whereas the $X_m = f(G)$.

Theorem 3 (Shamir, Spencer '87)**.** *For any n, p, if $G \sim \mathcal{G}(n,p)$ then*

$$\mathbb{P}\left(|\chi(G) - \mathbb{E}[\chi(G)]| > \lambda\sqrt{n}\right) < 2\mathrm{e}^{-\lambda^2/2}\,.$$

Proof. Let (X_t) be the vertex exposure martingale. Crucially, $|X_t - X_{t-1}| \leq 1$ for all t, since:

- Fix some t, and note that X_{t-1} is an average of graphs H that agree with G on $\{v_1,\ldots,v_t\}$. We can therefore write:

$$X_{t-1} = \mathbb{E}\left[f(G) \mid G|_{\{v_1,\ldots,v_t\}}\right] = \mathbb{E}\left[\mathbb{E}\left[f(G) \mid G_{\{v_1,\ldots,v_{t+1}\}}\right] \mid G|_{\{v_1,\ldots,v_t\}}\right]\,.$$

- Let H_1 and H_2 be two possible such subgraphs on $\{v_1,\ldots,v_{t+1}\}$. Then we can map every extension of H_1 to the same extension of H_2 into a graph on n vertices (via a measure preserving map), and the only difference between the resulting graphs would be the edges between v_{t+1} and $\{v_1,\ldots,v_t\}$ (affecting at most 1 color). Therefore,

$$\left|\mathbb{E}\left[f(G) \mid G|_{\{v_1,\ldots,v_{t+1}\}} = H_1\right] - \mathbb{E}\left[f(G) \mid G|_{\{v_1,\ldots,v_{t+1}\}} = H_2\right]\right| \leq 1\,.$$

- As this holds for any H_1, H_2, we deduce that X_{t-1} is a weighted mean of values, whose pairwise differences are all at most 1. Thus, the distance of the mean X_{t-1} is also at most 1 from each of these values. In particular, this holds for $H = G|_{\{v_1,\ldots,v_{t+1}\}}$, realizing X_t.

The result now follows directly from Hoeffding's inequality. ■