Lecture 7-8: Martingales and Azuma's inequality CSE 525, Fall 2016
Instructor: James R. Lee

1 Martingales

We have seen that if $X = X_1 + \cdots + X_n$ is a sum of independent $\{0,1\}$ random variables, then X is tightly concentrated around its expected value $\mathbb{E}[X]$. The fact that the random variables were $\{0,1\}$-valued was not essential; similar concentration results hold if we simply assume that they are in some bounded range $[-L, L]$. One can also relax the independence assumption, as we will see next.

Consider a sequence of random variables $X_0, X_1, X_2, \ldots$. The sequence $\{X_i\}$ is called a *discrete-time martingale* if it holds that

$$\mathbb{E}[X_{i+1} \mid X_0, X_1, \ldots, X_i] = X_i$$

for every $i = 0, 1, 2, \ldots$. More generally, the sequence is $\{X_i\}$ is a martingale with respect to another sequence of random variables $\{Y_i\}$ if it for every i, it holds that

$$\mathbb{E}[X_{i+1} \mid Y_0, Y_1, \ldots, Y_i] = X_i\,.$$

Note that this is equivalent to $\mathbb{E}[X_{i+1} - X_i \mid Y_0, Y_1, \ldots, Y_i] = 0$. If one thinks of $\{Y_0, Y_1, \ldots, Y_i\}$ as all the "information" up to time i, then this says that the difference $X_{i+1} - X_i$ is unbiased conditioned on the past up to time i. Observe that for any i, we have

$$\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i \mid X_0, \ldots, X_{i-1}]] = \mathbb{E}[X_{i-1}] = \cdots = \mathbb{E}[X_0]\,.$$

Martingales form an extremely useful class of random processes that appear in a vast array of settings (e.g., finance, machine learning, information theory, statistical physics, etc.). The classic example is that of a Gambler whose bank roll is X_0. At each time, she chooses to play some game in the casino at some stakes. If we assume that every game is fair (that is, the expected utility from playing the game is 0), then the sequence $\{X_0, X_1, \ldots\}$ forms a martingale, where X_i is the amount of money she has at time i.

Remark 1.1. The correct level of generality at which to define martingales involves a filtration. Formally, this is an increasing sequence of σ-algebras on our measure space $(\Omega, \mu, \mathcal{F})$: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$. A sequence of random variables $\{X_i\}$ is a martingale with respect to the filtration $\{\mathcal{F}_i\}$ if $\mathbb{E}[X_{i+1} \mid \mathcal{F}_i] = X_i$ for every $i \geqslant 0$.

1.1 Doob martingales

One reason martingales are so powerful is that they model a situation where one gains progressively more information over time. Suppose that $\mathcal{U}$ is a set of objects, and $f : \mathcal{U} \to \mathbb{R}$. Let X be a random variable taking values in $\mathcal{U}$, and let $\{Y_i\}$ be another sequence of random variables. The associated *Doob martingale* is given by

$$X_i = \mathbb{E}[f(X) \mid Y_0, Y_1, \ldots, Y_i]\,.$$

In words, this is our "estimate" for the value of $f(X)$ given the information contained in $\{Y_0, \ldots, Y_i\}$. To see that this is always a martingale with respect to $\{Y_i\}$, observe that

$$\mathbb{E}[X_{i+1} \mid Y_0, \ldots, Y_i] = \mathbb{E}[\mathbb{E}[f(X) \mid Y_0, \ldots, Y_{i+1}] \mid Y_0, \ldots, Y_i] = \mathbb{E}[f(X) \mid Y_0, \ldots, Y_i] = X_i\,,$$

where we have used the tower rule of conditional expectations.

Balls in bins. Suppose we throw m balls into n bins one at a time. At step i, we place ball i in a uniformly random bin. Let $C_1, C_2, \ldots, C_m$ be the sequence of (random) choices, and let C denote the final configuration of the system, i.e. exactly which balls end up in which bins.

Now we can consider a functional like $f(C) = \#$ of empty bins. If $X_i = \mathbb{E}[f(C) \mid C_1, \ldots, C_i]$, then $\{X_i\}$ is a (Doob) martingale. It is straightforward to calculate that

$$\mathbb{E}[X_m] = \mathbb{E}[X_0] = \mathbb{E}[f(C)] = n \cdot \left(1 - \frac{1}{n}\right)^m .$$

Suppose we are interested the concentration of $X_m = f(C)$ around its mean value. Of course, we can write $X_m = Z_1 + \cdots Z_m$ where Z_i is the indicator of whether the ith been is empty after all the balls have been thrown. But note that the $\{Z_i\}$ variables are not independent—in particular, if I tell you that $Z_1 = 1$ (bin 1 is empty), it decreases slightly the likelihood that other bins are empty.

The vertex exposure filtration. Recall that $\mathcal{G}_{n,p}$ denotes the random graph model where an undirected graph on n vertices is chosen by including every edge independently with probability p. Suppose the vertices are numbered $\{1, 2, \ldots, n\}$ Let $G \sim \mathcal{G}_{n,p}$ and denote by G_i the induced subgraph on the vertices $\{1, \ldots, i\}$. G_0 denotes the empty graph.

Let $\chi(G)$ denote the chromatic number of G, and consider the Doob martingale

$$X_i = \mathbb{E}[\chi(G) \mid G_0, \ldots, G_i] .$$

If we wanted to understand concentration properties of $X_n = \chi(G)$, this seems even more daunting. The chromatic number is a very complicated parameter of a graph! Nevertheless, we will now see that martingale concentration inequalities allow us to achieve tight concentration using very limited information about a sequence of random variables.

2 The Hoeffding-Azuma inequality

Say that a martingale $\{X_i\}$ has L-bounded increments if

$$|X_{i+1} - X_i| \leqslant L$$

for all $i \geqslant 0$. (The preceding inequality is meant to hold with probability 1.)

Theorem 2.1. *For every $L > 0$, if $\{X_i\}$ is a martingale with L-bounded increments, then for every $\lambda > 0$ and $n \geqslant 0$, we have*

$$\mathbb{P}[X_n \geqslant X_0 + \lambda] \leqslant e^{-\frac{\lambda^2}{2L^2 n}}$$
$$\mathbb{P}[X_n \leqslant X_0 - \lambda] \leqslant e^{-\frac{\lambda^2}{2L^2 n}}$$

We will prove this in Section 3. It's useful to note the following special case of the theorem.

Corollary 2.2. *Suppose that $Z_1, Z_2, \ldots, Z_n$ are independent random variables taking values in the interval $[-L, L]$. Put $Z = Z_1 + \cdots + Z_n$ and $\mu = \mathbb{E}[Z]$. Then for every $\lambda > 0$, we have*

$$\mathbb{P}\left[Z \geqslant \mu + \lambda\right] \leqslant e^{-\lambda^2/(2L^2 n)}$$
$$\mathbb{P}\left[Z \geqslant \mu - \lambda\right] \leqslant e^{-\lambda^2/(2L^2 n)}$$

The Lispchitz condition. Recall the setting of Doob martingales, where $\mathcal{U}$ is a set. Suppose that we can describe every element $u \in \mathcal{U}$ by a sequence of values $u = (u_1, u_2, \ldots, u_n)$. (For instance, every configuration of m balls in n bins can be described by the sequence of which balls go into which bins.)

Say that f is *L-Lipschitz* if it holds that for every $i = 1, \ldots, n$ and for every two elements $u = (u_1, u_2, \ldots, u_i, \ldots, u_n) \in \mathcal{U}$ and $u' = (u_1, u_2, \ldots, u'_i, \ldots, u_n) \in \mathcal{U}$ that differ only in the ith coordinate, we have

$$|f(u) - f(u')| \leqslant L\,.$$

Let $Z = (Z_1, \ldots, Z_n)$ be a $\mathcal{U}$-valued random variable such that the random variables $\{Z_i\}$ are independent. We now confirm that the Doob martingale $X_i = \mathbb{E}[f(Z) \mid Z_1, \ldots, Z_i]$ has L-bounded increments.

Let Z'_{i+1} be an independent copy of Z_{i+1} conditioned on $Z_1, \ldots, Z_i$, and let $Z' = (Z_1, \ldots, Z_i, Z'_{i+1}, \ldots, Z_n)$. Then:

$$\begin{aligned}
|X_{i+1} - X_i| &= \left|\mathbb{E}[f(Z) \mid Z_1, \ldots, Z_{i+1}] - \mathbb{E}[f(Z) \mid Z_1, \ldots, Z_i]\right| \\
&= \left|\mathbb{E}\left[f(Z) - f(Z') \mid Z_1, \ldots, Z_{i+1}\right]\right| \\
&\leqslant \mathbb{E}\left[\left|f(Z) - f(Z')\right| \mid Z_1, \ldots, Z_{i+1}\right] \\
&\leqslant L\,,
\end{aligned}$$

where in the last step we have used the fact that the term inside the absolute value signs is always at most L by the L-Lipschitz property of f, and the fact that Z and Z' differ in at most one coordinate.

The number of empty bins. First let's apply this to balls and bins. Recall that for a sequence of choices $C_1, \ldots, C_m$ (where C_i is the bin that the ith ball is thrown into), we put $f(C_1, \ldots, C_m)$ to be the number of empty bins. Then clearly f is 1-Lipschitz: Changing the fate of ball i can only change the number of empty bins by 1. Therefore the corresponding martingale $X_i = \mathbb{E}[f(C_1, \ldots, C_m) \mid C_1, \ldots, C_i]$ has 1-bounded increments, and Azuma's inequality implies that

$$\mathbb{P}[X_n \geqslant X_0 + \lambda] \leqslant e^{-\frac{\lambda^2}{2m}}\,.$$

Recall that $X_0 = \mathbb{E}[X_n] = n(1 - \frac{1}{m})^n$. Consider the situation where $m = n$ and thus $X_0 \asymp \frac{n}{e}$. If we put $\lambda = C\sqrt{n}$, we see that with high probability we expect the number of empty bins to be in the interval $\frac{n}{e} \pm O(\sqrt{n})$.

The chromatic number. Similarly, consider the vertex exposure martingale. We have to be a little more careful here to describe a graph G by a sequence $(Z_1, \ldots, Z_n)$ of *independent* random variables. The key is to think about Z_i containing the information on edges from vertex i to the vertices $\{1, \ldots, i-1\}$ so that we have independence.

Since we can identify a graph G with the vector $(Z_1, \ldots, Z_n)$, we can think of the chromatic number as a function $\chi(Z_1, \ldots, Z_n)$. The function χ satisfies the 1-Lipschitz property because changing the edges adjacent to some vertex i can only change the chromatic number by 1. The chromatic number cannot increase by more than one because we could always color i a new color; it cannot decrease by more than one because if we could color the graph without vertex i with c colors, then we can color the whole graph with $c + 1$ colors.

So the martingale $X_i = \mathbb{E}[\chi(G) \mid Z_1, \ldots, Z_i] = \mathbb{E}[\chi(G) \mid G_1, \ldots, G_i]$ has 1-bounded increments and Azuma's inequality tells us that

$$\mathbb{P}[\chi(G) \geqslant \mathbb{E}[\chi(G)] + \lambda] \leqslant e^{-\frac{\lambda^2}{2n}}\,.$$

Even without having any idea how to compute $\mathbb{E}[\chi(G)]$, we are able to say something significant about its concentration properties. [By the way, if $G \sim \mathcal{G}_{n,1/2}$, then $\mathbb{E}[\chi(G)] = n/(2\log_2 n)$, so the concentration window here—which is $O(\sqrt{n})$—is again quite small with respect to the expectation. In the next lecture, we will see how a more clever use of Azuma's inequality can achieve even better concentration of $\chi(G)$.]

3 Proof of Azuma's inequality

We will actually prove the following generalization of Theorem 2.1.

Theorem 3.1. *Suppose that $\{X_i\}$ is a sequence of random variables satisfying the property that for every subset of distinct indices $i_1 < i_2 < \cdots < i_k$, we have*

$$\mathbb{E}\left[X_{i_1} X_{i_2} \cdots X_{i_k}\right] = 0\,.$$

Then for every $\lambda > 0$ and $n \geqslant 1$, it holds that

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geqslant \lambda\right] \leqslant \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^{n} \|X_i\|_\infty^2}\right).$$

Here, $\|X_i\|_\infty$ is the essential supremum of X_i, i.e. the least value L such that $|X_i| \leqslant L$ with probability one.

The reason Theorem 3.1 proves Theorem 2.1 is as follows: Suppose that $\{Z_i\}$ is a martingale with respect to the sequence of random variables $\{Y_i\}$, and let $X_i = Z_i - Z_{i-1}$. Consider distinct indices $i_1 < i_2 < \cdots < i_k$. Then:

$$\mathbb{E}[X_{i_1} \cdots X_{i_k}] = \mathbb{E}\left[X_{i_1} \cdots X_{i_{k-1}}\, \mathbb{E}\left[Z_{i_k} - Z_{i_k-1} \mid Y_0, \ldots, Y_{i_k-1}\right]\right] = 0\,,$$

where the final inequality follows from defining property of a martingale.

Proof of Theorem 3.1. Note that from our assumptions, we have that for any sequences of constants $\{a_i\}$ and $\{b_i\}$, we have

$$\mathbb{E}\left[\prod_{i=1}^{n}(a_i + b_i X_i)\right] = \prod_{i=1}^{n} a_i\,. \tag{3.1}$$

Also, observe that for any a, the functions $f(x) = e^{ax}$ is convex. Thus for $x \in [-1, 1]$, it lies below the line connecting e^{-a} to e^{a}. In other words, for $x \in [-1, 1]$,

$$e^{ax} \leqslant \frac{e^a + e^{-a}}{2} + x\frac{e^a - e^{-a}}{2} = \cosh(a) + x\sinh(a)\,.$$

Combining this with (3.1), we have for any t:

$$\mathbb{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] \leqslant \mathbb{E}\left[\prod_{i=1}^{n} \cosh(t\|X_i\|_\infty) + \frac{X_i}{\|X_i\|_\infty}\sinh(t\|X_i\|_\infty)\right] = \prod_{i=1}^{n} \cosh(t\|X_i\|_\infty) \leqslant e^{t^2\|X_i\|_\infty^2/2}\,,$$

where the final inequality follows from $\cosh(x) = \sum \frac{x^{2k}}{(2k)!} \leqslant \sum \frac{x^{2k}}{2^k k!} = e^{x^2/2}$.

Now we are in position to apply the method of Laplace transforms:

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i > \lambda\right] \leqslant \frac{\mathbb{E}[e^{t\sum_{i=1}^{n} X_i}]}{e^{t\lambda}} \leqslant e^{(t^2/2)\sum_{i=1}^{n}\|X_i\|_\infty^2 - t\lambda}\,.$$

Setting $t = \frac{\lambda}{\sum_{i=1}^{n}\|X_i\|_\infty^2}$ finishes the proof. □

4 Tighter concentration of the chromatic number

Previously, using the vertex exposure martingale we were able to prove reasonable concentration for $\chi(G)$ when $G \sim \mathcal{G}_{n,p}$. In what follows, we will put $p = n^{-\alpha}$ for some $\alpha > 0$. We will show that, surprisingly, if $\alpha > 5/6$, then with probability tending to one, $\chi(G)$ is concentrated on one of four values. In what follows, we will say that an event $\mathcal{E}_n$ (explicitly or implicitly indexed by n) holds "with high probability" if $\mathbb{P}(\mathcal{E}_n) \to 1$ as $n \to \infty$.

Lemma 4.1. *For any $c > 0$ and $\alpha > 5/6$, the following holds for $G \sim \mathcal{G}_{n,p}$: With high probability, every induced subgraph of size at most $c\sqrt{n}$ is 3-colorable.*

Proof sketch. Let S be the smallest subset of $V(G)$ that is not 3-colorable (if no such set exists, we are done). Then every $x \in S$ must have at least three neighbors in S, otherwise since $S \setminus \{x\}$ is 3-colorable, it would be the case that S is also 3-colorable. Thus the number of edges in the induced subgraph $G[S]$ is at least $3|S|/2$.

But it is unlikely that any set S with $|S| \leqslant c\sqrt{n}$ at least $3|S|/2$ edges inside it. To see this, let $t = |S|$, and we'll compute the probability for a fixed set S: It's at most

$$p^{3t/2}\binom{\binom{t}{2}}{3t/2} \leqslant p^{3t/2}O(t)^{3t/2}\,.$$

Now we take a union bound over all sets of size at most T:

$$\sum_{t \leqslant T} p^{3t/2}O(t)^{3t/2}\binom{n}{t} \leqslant O(pT)^{3T/2}O\left(\frac{n}{T}\right)^T\,.$$

The latter inequality holds as long as $T \ll n$. Now using $p = n^{-\alpha}$ and $T \leqslant c\sqrt{n}$, this is bounded by

$$O(n)^{(1/2-\alpha)3T/2}O(n)^{T/2}\,,$$

and the latter quantity is $o(1)$ as long as $3/2(1/2 - \alpha) < 1/2$, i.e. $\alpha > 5/6$. □

Theorem 4.2. *With high probability, $\chi(G)$ takes one of four different values.*

Proof. Fix a number $\varepsilon > 0$ that we will send to 0. Let $u = u(n, p, \varepsilon)$ be the smallest integer so that $\mathbb{P}[\chi(G) \leqslant u] > \varepsilon$. Observe that, by the choice of u, we have $\mathbb{P}[\chi(G) > u - 1] \geqslant 1 - \varepsilon$.

Let $Y = Y(G)$ be the minimal size of a set of vertices S such that $\chi(G \setminus S) \leqslant u$. Consider the vertex exposure martingale for $G \sim \mathcal{G}_{n,p}$. Note that Y is 1-Lipschitz with respect to the exposure process because we could always add the modified vertex S. Thus we can apply Azuma's inequality to the corresponding Doob martingale to conclude that

$$\mathbb{P}[Y \geqslant \mu + \lambda\sqrt{n}] \leqslant e^{-\lambda^2/2} \tag{4.1}$$

$$\mathbb{P}[Y \leqslant \mu - \lambda\sqrt{n}] \leqslant e^{-\lambda^2/2}\,, \tag{4.2}$$

where $\mu = \mathbb{E}[Y]$.

Let us choose λ so that $e^{-\lambda^2/2} = \varepsilon$. By the definition of u, we have $\mathbb{P}[Y = 0] > \varepsilon$. We conclude from (4.2) that $\mu \leqslant \lambda\sqrt{n}$. Now using (4.1), we see that $\mathbb{P}[Y \geqslant 2\lambda\sqrt{n}] \leqslant \varepsilon$.

By Lemma 4.1, we may assume that every subset of size at most $2\lambda\sqrt{n}$ is 3-colorable by throwing away an ε-fraction of graphs. Now observe that $Y < 2\lambda\sqrt{n}$ implies that G is $u + 3$ colorable since $G \setminus S$ is u-colorable and $|S| < 2\lambda\sqrt{n}$ so S can be colored with an additional 3 colors. We conclude that

$$\mathbb{P}[\chi(G) \in \{u, u + 1, u + 2, u + 3\}] \geqslant 1 - 3\varepsilon\,.$$

Sending $\varepsilon \to 0$ completes the proof. □