CS271 Randomness & Computation **Fall 2022**

Lecture 18: October 25

Instructor: Alistair Sinclair

Disclaimer: *These notes have not been subjected to the usual scrutiny accorded to formal publications. They may be distributed outside this class only with the permission of the Instructor.*

18.1 Martingales

The Chernoff/Hoeffding bounds for large deviations that we have been using up to now apply only to sums of *independent* random variables. In many contexts, however, such independence does not hold. In this lecture we study another setting in which large deviation bounds can be proved, namely martingales with bounded differences.

As a motivation, consider a fair game (i.e., the expected win/loss from each play of the game is zero). Suppose a gambler plays the game multiple times; neither his stakes, nor the outcome of the games need be independent, but each play is fair. Let Z_i denote the outcome of the i^{th} game and X_i the gambler's capital after game i. Fairness ensures that the expected capital after a game is the same as the capital before the game, i.e., $\mathrm{E}[X_i|Z_1...Z_{i-1}] = X_{i-1}$. A sequence X_i that has this property is called a martingale[1].

Definition 18.1 *Let $(Z_i)_{i=1}^n$ and $(X_i)_{i=1}^n$ be sequences of random variables on a common probability space such that $\mathrm{E}[X_i|Z_1..Z_{i-1}] = X_{i-1}$ for all i. (X_i) is called a* martingale *with respect to (Z_i). Moreover, the sequence $Y_i = X_i - X_{i-1}$, is called a* martingale difference sequence*. By definition, $\mathrm{E}[Y_i|Z_1..Z_{i-1}] = 0$ for all i.*

This definition can be generalized to abstract probability spaces as follows. Define a *filter* (or *filtration*) as an increasing sequence of σ-fields $\emptyset = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_n$ on some probability space. Let (X_i) be a sequence of random variables such that X_i is measurable with respect to $\mathcal{F}_i$. Then, (X_i) is a *martingale* with respect to $(\mathcal{F}_i)$ if $\mathrm{E}[X_i|\mathcal{F}_{i-1}] = X_{i-1}$ for all i. In what follows, we will usually identify the filter $\mathcal{F}_i$ with $(Z_1, \ldots, Z_i)$ for an underlying sequence of random variables (Z_i), as we did in the gambling example above. The formal interpretation is that $\mathcal{F}_i$ is the smallest σ-field with respect to which all of $Z_1, \ldots, Z_i$ are measurable.

18.2 The Doob Martingale

Martingales are ubiquitous; indeed, we can obtain a martingale from essentially *any* random variable as follows.

Claim 18.2 *Let A and (Z_i) be random variables on a common probability space. Then $X_i = \mathrm{E}[A|Z_1....Z_i]$ is a martingale (called the Doob martingale of A).*

[1] Historically, the term "martingale" referred to a popular gambling strategy: bet \$1 initially, if you lose bet \$2, then \$4, then \$8, and so on; stop after your first win. Assuming you have unlimited resources, you will win \$1 with probability 1.

Proof: Use the definition of X_i to get

$$\begin{aligned} \mathrm{E}[X_i|Z_1 \dots Z_{i-1}] &= \mathrm{E}[\mathrm{E}[A|Z_1 \dots Z_i]|Z_1 \dots Z_{i-1}] \\ &= \mathrm{E}[A|Z_1 \dots Z_{i-1}] = X_{i-1}. \end{aligned}$$

The second equality follows from the tower property of conditional expectations: if $\mathcal{F} \subseteq \mathcal{G}$ then $\mathrm{E}[\mathrm{E}[X|\mathcal{G}]|\mathcal{F}] = \mathrm{E}[X|\mathcal{F}]$. (The outer expectation simply "averages out" $\mathcal{G} - \mathcal{F}$.) ∎

Frequently in applications we will have $A = f(Z_1 \dots Z_n)$, i.e., A is determined by the random variables Z_i. In this case, $X_0 = \mathrm{E}[A]$ and $X_n = \mathrm{E}[A|Z_1 \dots Z_n] = A$. We can think of the martingale as revealing progressively more information about the random variable A. We begin with no information about A, and the value of the martingale is just the expectation $\mathrm{E}[A]$. At the end of the sequence we have specified all of the Z_i so we have complete information about A and the martingale has the (deterministic) value $A(Z_1, \dots, Z_n)$.

18.2.1 Examples

Coin tosses. A is the number of heads after N tosses, (Z_i) are the outcomes of the tosses, $X_i = \mathrm{E}[A|Z_1 \dots Z_i]$.

Note that in this case the martingale differences $Y_i = X_i - X_{i-1}$ are independent.

Balls & bins. m balls are thrown at random into n bins. For $1 \leq i \leq m$, let $Z_i \in \{1, \dots, n\}$ be the destination of the ith ball. Let $A(Z_1, \dots, Z_m)$ be the number of empty bins, and $X_i = \mathrm{E}[A|Z_1 \dots Z_i]$ the corresponding Doob martingale.

In this case the differences $Y_i = X_i - X_{i-1}$ are clearly *not* independent (because the position of the first $i-1$ balls certainly influences the expected change in the number of empty bins upon throwing the ith ball).

Random graphs: edge exposure martingale. In the $\mathcal{G}_{n,p}$ setting, let Z_i be an indicator of whether the i^{th} possible edge is present in the graph. Let $A = f(Z_1 \dots Z_{\binom{n}{2}})$ be any graph property (such as the size of a largest clique). Then $X_i = \mathrm{E}[A|Z_1 \dots Z_i]$ is a Doob martingale. Martingales defined with respect to this sequence (Z_i) are called "edge exposure" martingales.

Random graphs: vertex exposure martingale. Edge exposure reveals the random graph one edge at a time. Instead, we can reveal it one vertex at a time. Let $Z_i \in \{0,1\}^{n-i}$ be a vector of indicators of whether edges between vertex i and vertices $j > i$ are present. For any graph property $A = f(Z_1 \dots Z_n)$, the corresponding martingale $X_i = \mathrm{E}[A|Z_1 \dots Z_i]$ is called a "vertex exposure" martingale.

Max3SAT. Consider a random truth assignment to the variables of a 3SAT formula (as discussed in Lecture 6). Let Z_i be the assignment of the i^{th} variable. If $A(Z_1, \dots, Z_n)$ is the number of clauses satisfied, then $X_i = \mathrm{E}[A|Z_1 \dots Z_i]$ is a natural Doob martingale. Incidentally, it is precisely this martingale property that lies behind our derandomization (via the method of conditional probabilities) of the random assignment algorithm for Max3SAT that we saw in Lecture 6.

18.3 Azuma's inequality

We now prove an important concentration result for martingales (Theorem 18.3), known as Azuma's inequality. Azuma's inequality is usually attributed to Azuma [Az67] and Hoeffding [Ho63]. However, other versions appeared around the same time (notably one due to Steiger [St67]).

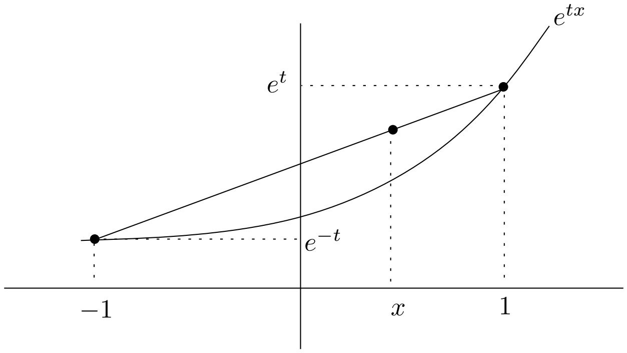

Figure 18.1: Convexity implies that $e^{tx} \leq \frac{1}{2}(1+x)e^t + \frac{1}{2}(1-x)e^{-t}$

Theorem 18.3 *Let (X_i) be a martingale with respect to the filter $(\mathcal{F}_i)$, and let $Y_i = X_i - X_{i-1}$ be the corresponding difference sequence. If the $c_i > 0$ are such that $|Y_i| \leq c_i$ for all i, then*

$$\left.\begin{array}{l}\Pr[X_n \geq X_0 + \lambda] \\ \Pr[X_n \leq X_0 - \lambda]\end{array}\right\} \quad \leq \quad \exp\left(-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}\right).$$

Note that Azuma's inequality provides a bound similar in form to the Chernoff bound, but without assuming independence. A key assumption is that the martingale differences Y_i are bounded in absolute value (though this can be relaxed somewhat as will become apparent in the proof). To recover Chernoff's bound, apply Azuma's inequality to the coin tossing martingale above with $c_i = 1$ for all i; and remember that $X_0 = \mathrm{E}[A]$ and $X_n = A$, where A is the number of heads (successes). This gives the bound $\exp(-\lambda^2/2n)$ for the probability of deviating more than λ either above or below the expectation, which is essentially the same as the simplest form of Chernoff bound we saw in Lecture 13 (Corollary 13.2).

In order to prove Azuma's inequality, we need a simple technical fact based on convexity of the exponential function.

Lemma 18.4 *Let Y be a random variable such that $Y \in [-1, +1]$ and $\mathrm{E}[Y] = 0$. Then for any $t \geq 0$, we have that $\mathrm{E}[e^{tY}] \leq e^{t^2/2}$.*

Proof: For any $x \in [-1, 1]$, $e^{tx} \leq \frac{1}{2}(1+x)e^t + \frac{1}{2}(1-x)e^{-t}$ by convexity (see Figure 18.1). Taking expectations,

$$\begin{aligned}
\mathrm{E}[e^{tY}] &\leq \frac{1}{2}e^t + \frac{1}{2}e^{-t} \quad \text{(since E}[Y] = 0) \\
&= \frac{1}{2}\left[(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots) + (1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \ldots)\right] \\
&= \left[1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \ldots\right] \\
&= \sum_{n=0}^{\infty} \frac{t^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!} = \sum_{n=0}^{\infty} \frac{(t^2/2)^n}{n!} = e^{\frac{t^2}{2}}.
\end{aligned}$$

■

Proof of Thm 18.3 (Azuma's Inequality): The proof follows a similar outline to that of the Chernoff bound in Lecture 13. We prove only the lower tail bound; the upper tail follows via a symmetrical argument.

First, for any $t > 0$ we have

$$\Pr[X_n - X_0 \geq \lambda] \quad = \quad \Pr[e^{t(X_n - X_0)} \geq e^{\lambda t}].$$

Applying Markov's inequality and writing $X_n = Y_n + X_{n-1}$,

$$\begin{aligned} \Pr[e^{t(X_n - X_0)} \geq e^{\lambda t}] &\leq e^{-\lambda t}\mathrm{E}[e^{t(X_n - X_0)}] && (18.1) \\ &= e^{-\lambda t}\mathrm{E}[e^{t(Y_n + X_{n-1} - X_0)}] \\ &= e^{-\lambda t}\mathrm{E}[\mathrm{E}[e^{t(Y_n + X_{n-1} - X_0)}|\mathcal{F}_{n-1}]]. && (18.2) \end{aligned}$$

To compute the inner expectation, factor out $E^{t(X_{n-1} - X_0)}$, which is constant given $\mathcal{F}_{n-1}$, and then apply Lemma 18.4 to the random variable $\frac{Y_n}{c_n}$ which has mean zero and takes values in $[-1, 1]$:

$$\begin{aligned} \mathrm{E}[e^{t(Y_n + X_{n-1} - X_0)}|\mathcal{F}_{n-1}] &= e^{t(X_{n-1} - X_0)}\mathrm{E}[e^{tY_n}|\mathcal{F}_{n-1}] \\ &\leq e^{t(X_{n-1} - X_0)}e^{t^2 c_n^2/2}. \end{aligned}$$

Substituting this result back into (18.2), we get

$$\Pr[X_n - X_0 \geq \lambda] \leq e^{-\lambda t}e^{t^2 c_n^2/2}\mathrm{E}[e^{t(X_{n-1} - X_0)}].$$

We can now handle the term $\mathrm{E}[e^{t(X_{n-1} - X_0)}]$ inductively in the same fashion as above to give

$$\Pr[X_n - X_0 \geq \lambda] \leq e^{t^2 \sum_{i=1}^n c_i^2/2 - \lambda t}.$$

Finally, since the above holds for any $t > 0$, we optimize our choice of t by taking $t = \frac{\lambda}{\sum c_i^2}$, which gives

$$\Pr[X_n - X_0 \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2\sum_i c_i^2}\right).$$

This completes the proof. ■

Exercise: When the range of Y_i is not symmetrical about 0, say $Y_i \in [a_i, b_i]$, one can still use the above by taking $c_i = \max\{|a_i|, |b_i|\}$. However, a better bound can be achieved as follows. First, prove an asymmetrical version of the technical lemma: When $\mathrm{E}[Y] = 0$ and $Y \in [a, b]$, $\mathrm{E}[e^{tY}] \leq e^{\frac{t^2}{8}(b-a)^2}$ for any $t > 0$. Then use this to prove the following version of Azuma's inequality, where $Y_i \in [a_i, b_i]$: $\Pr[X_n - X_0 \geq \lambda] \leq \exp(-\frac{2\lambda^2}{\sum(b_i - a_i)^2})$. For further variations on Azuma's inequality, see [McD98].

18.4 Applications of Azuma's Inequality

18.4.1 Gambling

In this setting, Z_i is the outcome of i-th game (which can depend on $Z_1, \ldots, Z_{i-1}$) and X_i is the capital at time i. Assuming the gambler doesn't quit and has unlimited capital, Azuma's inequality gives

$$\Pr[|X_n - C| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2nM^2}\right),$$

where C is the initial capital and the stakes and winnings are bounded by M (so $|X_i - X_{i-1}| \leq M$).

18.4.2 Coin Tossing

Here, Z_i is the outcome of the i-th coin toss and $X = f(Z_1, \dots, Z_n)$ is the number of heads after n tosses.

$|X_i - X_{i-1}| = |\mathrm{E}[f|Z_1, \dots, Z_i] - \mathrm{E}[f|Z_1, \dots, Z_{i-1}]| \le 1$, since the number of heads can't change by more than 1 after any coin toss. Therefore, by Azuma's inequality

$$\Pr[|X - \mathrm{E}[X]| \ge \lambda] \le 2\exp\left(-\frac{\lambda^2}{2n}\right),$$

which implies small probability of deviations of size $\omega(\sqrt{n})$. Note that this bound is essentially the same as Chernoff-Hoeffding.

Definition 18.5 *$f(Z_1, \dots, Z_n)$ is c-*Lipschitz *if changing the value of any one coordinate of f causes f to change by at most $\pm c$.*

Claim 18.6 *If f is c-Lipschitz and Z_i is independent of $Z_{i+1}, \dots, Z_n$ conditioned on $Z_1, \dots, Z_{i-1}$, then the Doob martingale X_i of f with respect to Z_i satisfies $|X_i - X_{i-1}| \le c$.*

Proof: Let $\hat{Z}_i$ be a random variable with the same distribution as Z_i conditioned on $Z_1, \dots, Z_{i-1}$, but independent of $Z_i, Z_{i+1}, \dots, Z_n$. Then

$$\begin{aligned} X_{i-1} &= \mathrm{E}[f(Z_1, \dots, Z_i, \dots, Z_n)|Z_1, \dots, Z_{i-1}] \\ &= \mathrm{E}[f(Z_1, \dots, \hat{Z}_i, \dots, Z_n)|Z_1, \dots, Z_{i-1}] \\ &= \mathrm{E}[f(Z_1, \dots, \hat{Z}_i, \dots, Z_n)|Z_1, \dots, Z_{i-1}, Z_i] \end{aligned}$$

Therefore, by the c-Lipschitz assumption:

$$|X_{i-1} - X_i| = |\mathrm{E}[f(Z_1, \dots, \hat{Z}_i, \dots, Z_n) - f(Z_1, \dots, Z_i, \dots, Z_n)|Z_1, \dots, Z_i]| \le c.$$

■

Exercise: Show that the (mild) independence assumption in Claim 18.6 is necessary.

18.4.3 Balls and Bins

We look at m balls and n bins. As usual, we are randomly throwing each ball into a bin. Here Z_i is the bin selected by the i-th ball and $X = f(Z_1, \dots, Z_m)$ is the number of empty bins.

Since each ball cannot change the number of empty bins by more than one, f is 1-Lipschitz:

$$\Pr[|X - \mathrm{E}[X]| \ge \lambda] \le 2\exp\left(-\frac{\lambda^2}{2m}\right).$$

This bound is useful whenever $\lambda \gg \sqrt{m}$. Note that this process can't be readily analyzed using Chernoff-Hoeffding bounds because the increments Y_i are *not* independent.

Incidentally, we also know that $\mathrm{E}[X] = n\left(1 - \frac{1}{n}\right)^m \sim ne^{-\frac{m}{n}}$ for $m = o(n^2)$, but we didn't use this fact (see the next example).

18.4.4 The chromatic number of a random graph $\mathcal{G}_{n,\frac{1}{2}}$

The *chromatic number* $\chi(G)$ of a graph G is the minimum number of colors required to color its vertices so that no two adjacent vertices receive the same color. Equivalently, since the set of vertices with a given color must form an independent set, $\chi(G)$ is also the size of a minimal partition of the vertices of G into independent sets. We are interested in a high probability estimate for $\chi(G)$, where G is drawn according to the distribution $\mathcal{G}_{n,\frac{1}{2}}$.

Let X denote the chromatic number of a random graph. Interestingly, it is much easier to obtain a large deviation bound on X than to compute its expectation. Recall that the *vertex exposure martingale* is a Doob martingale based on the random process that reveals the vertices of G one at a time. More precisely, we define a sequence of random variables $Z_1, \ldots, Z_n$, where Z_i encodes the edges between vertex i and vertices $i+1, \ldots, n$. For any graph G, the sequence $Z_1(G), \ldots, Z_n(G)$ uniquely determines G, so there is a function f such that $X = f(Z_1, \ldots, Z_n)$.

We observe that the function f is 1-Lipschitz: If we modify Z_i by adding edges incident to i, we can always obtain a proper coloring by choosing a new color for i; this increases the chromatic number by at most one. For similar reasons, removing edges incident to i cannot decrease the chromatic number by more than one. Applying Azuma's inequality to the Doob martingale of f immediately yields the following result:

Theorem 18.7 (Shamir and Spencer [SS87]) *Let X be the chromatic number of $G \in \mathcal{G}_{n,\frac{1}{2}}$. Then*

$$\Pr\left[|X - \mathrm{E}[X]| \geq \lambda\right] \leq 2\exp\left(-\frac{\lambda^2}{2n}\right).$$

Proof: Use the vertex exposure martingale and consider the random variable $X(Z_1, Z_2, \ldots, Z_n)$. Then X is 1-Lipschitz. The result follows from Azuma's inequality with $c_i = 1$. ■

Thus we see that deviations of size $\omega(\sqrt{n})$ are unlikely. Note that we proved this theorem without any knowledge of $\mathrm{E}[X]$! In the next lecture, we will use a more sophisticated martingale argument to compute $\mathrm{E}[X]$.

Note also that cliques (sets in which any two vertices are adjacent) are complementary to independent sets (no two vertices are adjacent). Since $\mathcal{G}_{n,\frac{1}{2}}$ and its complement have the same distribution finding the largest independent set is the same problem as finding the largest clique in this class of random graphs. As we already know from Lecture 7, the largest clique has size $\sim 2\log_2 n$ a.s. Since the set of vertices colored with any particular color must be an independent set, this implies that the chromatic number is a.s. at least $\frac{n}{2\log_2 n}(1 + o(1))$. Since $\frac{n}{2\log_2 n} \gg \sqrt{n}$, Theorem 18.7 gives tight concentration of the chromatic number.

References

[Az67] K. Azuma, "Weighted sums of certain dependent random variables," *Tokohu Mathematical Journal* **19** (1967), pp. 357–367.

[Ho63] W. Hoeffding, "Probability for sums of bounded random variables," *Journal of the American Statistical Association* **58** (1963), pp. 13–30.

[McD98] C. McDiarmid, "Concentration," in *Probabilistic Methods for Algorithmic Discrete Mathematics*, 1998, pp. 195-248.

[SS87] E. Shamir and J. Spencer, "Sharp concentration of the chromatic number on random graphs $\mathcal{G}_{n,p}$," *Combinatorica* **7** (1987), pp. 121–129.

[St67] W. Steiger, "Some Kolmogoroff-type inequalities for bounded random variables," *Biometrika* **54** (1967), pp. 641–647.