Lecture 19, 10/25/2020: Introduction to Martingales

1 Introduction and Background

We saw **Chernoff-Hoeffding bounds** for sum of independent random variables and applications. Last lecture we saw concentration bounds also hold for **negatively correlated random variables** and saw an application. However, there are other settings where we don't have independence or negative correlation and still concentration holds.

Martingales provide a **powerful framework** for such bounds and are also somewhat natural for algorithms. We saw an example of a martingale type process in the last lecture on rounding a fractional solution for the **Max K-Cover problem**.

1.1 Background on Conditional Probability

Suppose Ω is a probability space and $X : \Omega \to \mathbb{R}$ is a real-valued random variable.

If $A \subseteq \Omega$ is an event, then $E[X|A]$ is a real value, which, in the setting when X is discrete, is defined as:

$$E[X|A] = \frac{1}{Pr[A]} \sum_{\omega \in A} Pr[\omega] X(\omega).$$

In the continuous setting it is:

$$\frac{1}{Pr[A]} \int_A f(x) dx$$

where f is the probability density function (or similar function depending on the context).

Given random variables X and Y on Ω, the random variable $E[X|Y]$ is defined as follows. If $Z = E[X|Y]$, then $Z(\omega) = E[X|Y = Y(\omega)]$ (this is in the discrete setting). The meaning is the following: Y partitions Ω into parts where Y is constant in each part. $Y = b$ for some fixed b constitutes a part which can be alternatively thought as an event $A_b = \{\omega | Y(\omega) = b\}$. $E[X|Y]$ assigns a value to each ω equal to $E[X|A_b]$.
Claim: $E[E[X|Y]] = E[X]$.
Proof: Exercise.

When we write $E[X|Y_1, Y_2, ..., Y_n]$ where $Y_1, ..., Y_n$ are several random variables, then we are looking at the partition induced by $Y_1, ..., Y_n$ taking on different values in a joint way. The events correspond to:

$$A_{b_1, b_2, \ldots, b_n} = \{\omega \in \Omega | Y_i(\omega) = b_i\}.$$

Tower Property (Law of Total Expectation):
Lemma: $E[E[X|Y]|Z] = E[X|Z]$ (Assuming Z is a "coarser" partition of Ω than Y, or more formally, $\sigma(Z) \subseteq \sigma(Y)$).
Proof: Exercise.

Technically, a more formal way to describe this is via σ-algebras and filtrations but that requires more background.

2 Martingales

Defn: A sequence of random variables $X_0, X_1, X_2, \ldots$ is a **martingale sequence with respect to another s** $Y_0, Y_1, \ldots$ if for $n \geq 0$:

1. X_n is determined by $Y_0, \ldots, Y_n$.
2. $E[X_{n+1}|Y_0, \ldots, Y_n] = X_n$.

A sequence $X_0, X_1, \ldots$ is a **martingale** if it is a martingale w.r.t. itself. That is:

1. $E[|X_n|] < \infty$ and
2. $E[X_{n+1}|X_0, \ldots, X_n] = X_n$.

Example: Suppose a gambler starts with a random amount X_0 of money on day 0 and goes to the casino every day and plays some slot machine which is fair. Let Y_i be the winnings on day i (≤ 0 if he/she loses). Let X_i be the total amount that player has at the end of the i-th game. Because each game is **fair**, $E[Y_{i+1}] = 0$.

$$X_{n+1} = X_n + Y_{n+1}$$

$$E[X_{n+1}|Y_1, \ldots, Y_n] = X_n + E[Y_{n+1}|Y_1, \ldots, Y_n]$$

Since Y_{n+1} is independent of $Y_1, \ldots, Y_n$ and $E[Y_{n+1}] = 0$:

$$E[X_{n+1}|Y_1, \ldots, Y_n] = X_n + E[Y_{n+1}] = X_n + 0 = X_n.$$

The main thing that martingales allow one to capture is that the choice of how much to bet and which slot machine to bet on can be arbitrarily dependent on all the information/choices up to the previous step.

2.1 Doob Martingale

Martingales allow one to capture a particular type of phenomenon where we are interested in a function $f : U \to \mathbb{R}$ for some object X, and we have a random variable X that takes values in U. In other words, X is a **random object** chosen from U according to some process.

This process can be defined by a sequence of random variables $Y_0, Y_1, \dots, Y_n$. And X is determined by $Y_0, \dots, Y_n$. $Y_0, \dots, Y_i$ reveal partial information about X.

Let us define

$$X_i = E[f(X)|Y_0, \dots, Y_i].$$

We will assume that $E[|f(X)|] < \infty$.

Claim: $X_0, \dots, X_n$ is a martingale sequence with respect to $Y_0, \dots, Y_n$.

Proof:

$$E[X_{i+1}|Y_0, \dots, Y_i] = E[E[f(X)|Y_0, \dots, Y_{i+1}]|Y_0, \dots, Y_i]$$

(by definition of X_{i+1}) Using the **Tower Rule** ($E[E[A|B]|C] = E[A|C]$ if $\sigma(C) \subseteq \sigma(B)$):

$$\begin{aligned} E[E[f(X)|Y_0, \dots, Y_{i+1}]|Y_0, \dots, Y_i] &= E[f(X)|Y_0, \dots, Y_i] \\ &= X_i \end{aligned}$$

by definition.

Remark: In some settings it is easier to view $f(X)$ as another random variable Z and define $X_i = E[Z|Y_1, \dots, Y_i]$.

Example: Empty bins in balls and bins. We throw m balls into n bins independently. We think of this as a **sequential process** where we place ball i in the i-th step and Y_i is the random choice of the i-th ball (i.e., which bin it falls into). Let X be the # of empty bins after all m balls are thrown. Then:

$$X_i = E[X|Y_1, \dots, Y_i]$$

$$X_0 = E[X] = n\left(1 - \frac{1}{n}\right)^m$$

Example: Chromatic number of Random Graph $G(n, p)$. Let $G(n, p)$ be a graph on n vertices where each potential edge in the graph is chosen to be added independently with probability p. Let X be the **chromatic number** of the random graph. (Technically we should write $X_{n,p}$ to show dependence on n and p.)

We can define a **Doob martingale** called the **edge exposure martingale** where $Y_1, Y_2, \dots$ correspond to the binary random variables for picking the edges in some fixed order.

$$X_0 = E[\chi(G_{n,p})]$$

and $X_i = E[X|Y_1, \dots, Y_i]$.

We can define another **Doob martingale** called the **vertex exposure martingale** where $V_1, V_2, \dots, V_n$ is an ordering of vertices and Y_i reveals information about the edges of vertex i to vertices 1 to $i - 1$ in the random process.

3 Azuma-Hoeffding Inequality

Recall the additive change Hoeffding inequality. Let $X = \sum_{i=1}^{n} X_i$ where:

1. X_i are independent.
2. $X_i \in [a_i, b_i]$.

Then:

$$Pr[X - E[X] \geq \lambda] \leq e^{-\frac{\lambda^2}{2\sum_i (b_i - a_i)^2}}$$

and $Pr[X - E[X] \leq -\lambda] \leq e^{-\frac{\lambda^2}{2\sum_i (b_i - a_i)^2}}$.

A simpler form is when $[a_i, b_i] = [-c_i, c_i]$, in which case we have:

$$Pr[X - E[X] > \lambda] \leq e^{\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}}$$

Similarly for the lower tail.

Azuma-Hoeffding bound extends this to the martingale setting.
Theorem: Let $X_0, X_1, X_2, \ldots$ be a martingale sequence where

$$|X_i - X_{i-1}| \leq c_i \quad \forall i \geq 1.$$

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

The theorem also holds for the lower tail: $Pr[X_n - X_0 \leq -\lambda] \leq e^{\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}}$.

3.1 Proof of Azuma-Hoeffding

We need an auxiliary lemma.
Lemma (Hoeffding's Lemma): Let X be a random variable in $[-1, 1]$ with $E[X] = 0$. Then $E[e^{aX}] \leq e^{a^2/2}$.
Proof: The function e^{ax} is convex on $[-1, 1]$. For any $X \in [-1, 1]$, we can write $X = \frac{1+X}{2}(+1) + \frac{1-X}{2}(-1)$. Hence by convexity (Jensen's inequality):

$$e^{aX} \leq \frac{1+X}{2} e^{a} + \frac{1-X}{2} e^{-a} = \frac{e^a + e^{-a}}{2} + \frac{e^a - e^{-a}}{2} X.$$

If $X \in [-1, 1]$ with $E[X] = 0$, taking expectation on both sides:

$$E[e^{aX}] \leq \frac{e^a + e^{-a}}{2} + \frac{e^a - e^{-a}}{2} E[X] = \frac{e^a + e^{-a}}{2}.$$

By Taylor expansion:

$$\frac{e^a + e^{-a}}{2} = 1 + \frac{a^2}{2!} + \frac{a^4}{4!} + \cdots \leq 1 + \frac{a^2}{2} + \frac{(a^2/2)^2}{2!} + \cdots = e^{a^2/2}.$$

Corollary: If $E[X] = 0$ and $X \in [-c, c]$, then $E[e^{aX}] \leq e^{\frac{a^2 c^2}{2}}$.
Proof: Consider $X' = \frac{X}{c}$ and apply the previous Lemma.

3.1.1 Main Proof Steps

Let $t > 0$ be a parameter to be chosen later. By Markov's inequality:

$$Pr[X_n - X_0 \geq \lambda] \leq Pr[e^{t(X_n - X_0)} \geq e^{t\lambda}] \leq \frac{E[e^{t(X_n - X_0)}]}{e^{t\lambda}}.$$

So it boils down to estimating $E[e^{t(X_n - X_0)}]$.

Consider the differences $Z_i = X_i - X_{i-1}$. Recall that $|Z_i| \leq c_i$ and:

$$E[Z_i|X_0, \dots, X_{i-1}] = E[X_i - X_{i-1}|X_0, \dots, X_{i-1}]$$
$$= E[X_i|X_0, \dots, X_{i-1}] - X_{i-1} = X_{i-1} - X_{i-1} = 0.$$

Consider $E[e^{tZ_i}|X_0, \dots, X_{i-1}]$. By the Corollary with $a = t$ and $c = c_i$:

$$E[e^{tZ_i}|X_0, \dots, X_{i-1}] \leq e^{\frac{t^2 c_i^2}{2}}.$$

Now:

$$E[e^{t(X_n - X_0)}] = E[e^{t(Z_n + Z_{n-1} + \dots + Z_1)}] = E[e^{t(Z_{n-1} + \dots + Z_1)} \cdot e^{tZ_n}]$$

Using the property $E[AB] = E[A \cdot E[B|\text{information about } A]]$ and the fact that $e^{t(Z_{n-1} + \dots + Z_1)}$ is determined by $X_0, \dots, X_{n-1}$:

$$E[e^{t(X_n - X_0)}] = E[e^{t(Z_{n-1} + \dots + Z_1)} \cdot E[e^{tZ_n}|X_0, \dots, X_{n-1}]]$$

Applying the bound:

$$\leq E[e^{t(Z_{n-1} + \dots + Z_1)}]e^{\frac{t^2 c_n^2}{2}}$$

Repeating this recursively:

$$\leq E[e^{t(Z_{n-2} + \dots + Z_1)}] \cdot e^{\frac{t^2 c_{n-1}^2}{2}} \cdot e^{\frac{t^2 c_n^2}{2}}$$
$$\leq e^{(t^2 \sum_{i=1}^n c_i^2)/2}$$

Thus:

$$E[e^{t(X_n - X_0)}] \leq e^{\frac{\sum_{i=1}^n c_i^2}{2} \cdot t^2}.$$

Putting together with Markov's inequality:

$$Pr[X_n - X_0 \geq \lambda] \leq \frac{E[e^{t(X_n - X_0)}]}{e^{t\lambda}} \leq \frac{e^{t^2 \frac{\sum c_i^2}{2}}}{e^{t\lambda}} = e^{t^2 \left(\frac{\sum c_i^2}{2}\right) - t\lambda}$$

Choosing $t = \frac{\lambda}{\sum c_i^2}$ to minimize the exponent:

$$t^2 \left(\frac{\sum c_i^2}{2}\right) - t\lambda = \frac{\lambda^2}{(\sum c_i^2)^2} \frac{\sum c_i^2}{2} - \frac{\lambda^2}{\sum c_i^2} = \frac{\lambda^2}{2 \sum c_i^2} - \frac{\lambda^2}{\sum c_i^2} = -\frac{\lambda^2}{2 \sum c_i^2}.$$

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

Corollary: If $c_i \leq C \quad \forall i$ and $\lambda = C \cdot \alpha\sqrt{n}$, then:

$$Pr[|X_n - X_0| \geq C \cdot \alpha\sqrt{n}] \leq 2e^{-\frac{\alpha^2}{2}}.$$

(The 2 comes from bounding both tails).

4 McDiarmid's Inequality and Application

Chernoff-Hoeffding bound suggests that if $X = X_1 + X_2 + \cdots + X_n$ where $X_1, \ldots, X_n$ are independent and in a bounded range, then X has concentration.

Now consider the setting where we have an **arbitrary function** $f : U_1 \times U_2 \times \cdots \times U_n \to \mathbb{R}$. In other words f is a function of n "variables," each of which has domain U_i.

Defn: $f : U_1 \times U_2 \times \cdots \times U_n \to \mathbb{R}$ is c**-Lipschitz** for some $c > 0$ if:

$$|f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)| \leq c$$

for any $y, z \in U_i$, and any fixed x_j. In other words, changing one coordinate does not change the value of the function by more than c in absolute value.

A more refined definition is that f is $(c_1, \ldots, c_n)$**-Lipschitz** if:

$$|f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n)| \leq c_i$$

$\forall i$, $\forall y, z \in U_i$, and $\forall x_j$.

Theorem (McDiarmid's Inequality): Suppose f is $(c_1, \ldots, c_n)$-Lipschitz. Let $X_1, X_2, \ldots, X_n$ be **independent random variables** where $X_i \in U_i$. Then:

$$Pr[|f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)]| \geq \lambda] \leq 2e^{\frac{-\lambda^2}{2\sum_i c_i^2}}$$

Note that **independence is required**. In some settings it can be relaxed.

4.1 Proof Sketch (Reduction to Azuma's Inequality)

We reduce this to the Azuma's inequality via the **Doob martingale**. The main thing to note is where **independence** is used.

Let $\overline{X}$ denote $(X_1, \ldots, X_n)$ and $Z_i = E[f(\overline{X})|X_1, \ldots, X_i]$. $Z_0 = E[f(\overline{X})]$.

Recall that $Z_0, \ldots, Z_n$ is a **Doob martingale** sequence. We wish to apply Azuma's inequality to $Z_0, \ldots, Z_n$. For this we need to bound $Z_i - Z_{i-1}$.

$$Z_i - Z_{i-1} = E[f(\overline{X})|X_1, \ldots, X_i] - E[f(\overline{X})|X_1, \ldots, X_{i-1}].$$

Since $X_1, \ldots, X_n$ are independent, $X_{i+1}, \ldots, X_n$ are independent of $X_1, \ldots, X_i$.

It can be shown (using c_i-Lipschitz property and independence):

$$|Z_i - Z_{i-1}| \leq \sup_{u_1, \ldots, u_{i-1}} |E[f(u_1, \ldots, u_{i-1}, a, X_{i+1}, \ldots, X_n)] - E[f(u_1, \ldots, u_{i-1}, b, X_{i+1}, \ldots, X_n)]| \leq c_i$$

where $a, b \in U_i$. (The intermediate step involves sup over $X_1 = u_1, \ldots, X_{i-1} = u_{i-1}$ of the difference of conditional expectations, which is bounded by c_i when applying the expectation over $X_{i+1}, \ldots, X_n$). We used c_i-Lipschitzness on f and **independence** of $X_{i+1}, \ldots, X_n$ from $X_1, \ldots, X_i$.

Now we can apply Azuma's inequality to $Z_0, \ldots, Z_n$ and conclude that

$$Pr[Z_n - Z_0 > \lambda] \leq e^{-\frac{\lambda^2}{2\sum c_i^2}}.$$

We have $Z_0 = E[f(\overline{X})]$ and $Z_n = f(\overline{X})$.

4.2 Application: Balls and Bins

We throw m balls into n bins independently. Let $X_1, \ldots, X_m$ be the random choices of the m balls. $X_i \in [n] \quad \forall i \in [m]$. Define $f(X_1, X_2, \ldots, X_m)$ to be the # of empty bins. $E[f(X_1, \ldots, X_m)]$ is easy to calculate and is equal to $n\left(1-\frac{1}{n}\right)^m$.

We claim f is **1-Lipschitz** (specifically, $c_i = 1$ for all i). This is easy to verify. Changing the assignment of one ball can change the number of empty bins by **at most 1**.

Thus we get concentration (using $\sum c_i^2 = \sum_{i=1}^m 1^2 = m$):

$$Pr[|f(X_1, \ldots, X_m) - n\left(1-\frac{1}{n}\right)^m| > \lambda] \leq 2e^{\frac{-\lambda^2}{2m}}.$$

Hence if $\lambda = \alpha\sqrt{m}$, the probability is $\leq 2e^{-\frac{\alpha^2}{2}}$.

4.3 Application: Chromatic number of Random Graphs

Consider Random graph $G(n,p)$. Let $\chi(G)$ be the **chromatic number**. For instance, it is known that $E[\chi(G(n,\frac{1}{2}))] \approx \frac{n}{2\log_2 n}$. What about concentration?

We consider the **vertex exposure martingale**. Fix an ordering of vertices and let Y_i for vertex i denote the random vector of edges to vertices 1 to $i-1$. Let $f(Y_1, \ldots, Y_n)$ be the chromatic number of $G(n,p)$.

We claim f is **1-Lipschitz**. Changing the edges incident to a **single vertex** can change the chromatic number by **at most 1**.

Thus the **chromatic number** of $G(n,p)$ is **concentrated around its mean**.