Statistics 212: Lecture 6 (Feb. 11, 2025)

Concentration of Martingales

Instructor: Mark Sellke

Scribe: Nicholas Barnfield and Zimeng Li

1 Naive Concentration

Setup:

- Let (X_k) be a martingale with X_0 deterministic.
- An assumption that will be frequently used in this lecture:

$$|X_k - X_{k-1}| \le c_k, \text{for all } k = 1, \ldots, n \qquad (*)$$

- **Informal takeaway**: $|X_n - \mathbb{E}X_n| \le O(\sqrt{\sum c_k^2})$ with high probability. $\qquad (**)$

We then first take a look at the order of $\mathrm{Var}(X_n)$ under this setting. We claim that if (X_k) is a martingale that satisfy $(*)$, then

$$\mathrm{Var}(X_n) \le \sum_{k=1}^{n} c_k^2$$

Proof of this claim:
For martingale (X_k), we have that

$$\mathrm{Var}(X_n) = \mathbb{E}[X_n^2] - X_0^2 = \sum_{k=1}^{n} \mathbb{E}[X_k^2 - X_{k-1}^2] \tag{1}$$

$$\mathrm{Var}(X_k|\mathscr{F}_{k-1}) = \mathbb{E}[X_k^2|\mathscr{F}_{k-1}] - [\mathbb{E}[X_k|\mathscr{F}_{k-1}]]^2 = \mathbb{E}[X_k^2|\mathscr{F}_{k-1}] - X_{k-1}^2 \tag{2}$$

Notice that 2 leads to

$$\mathbb{E}[\mathrm{Var}(X_k|\mathscr{F}_{k-1})] = \mathbb{E}X_k^2 - \mathbb{E}X_{k-1}^2 \tag{3}$$

Combining 1 and 3, we can derive that

$$\mathrm{Var}(X_n) = \sum_{k=1}^{n} \mathbb{E}[\mathrm{Var}(X_k|\mathscr{F}_{k-1})] \tag{4}$$

Assuming $(*)$, we have

$$\mathrm{Var}(X_k|\mathscr{F}_{k-1}) = \mathbb{E}[(X_k - X_{k-1})^2|\mathscr{F}_{k-1}] \le c_k^2$$

which leads to the claimed result.

1.1 Examples

Example (Family of functions that satisfy the setup). Let $F:\mathbb{R}^n \to \mathbb{R}$ be a bounded function and $Y_1,\dots,Y_n$ be independent random variables. Suppose

$$|F(y_1,\dots,y_n) - F(y_1,\dots,y_{k-1},y_k',y_{k+1},\dots,y_n)| \le 1. \qquad (+)$$

for all k and define the martingale (X_k) by

$$X_k = \mathbb{E}[F(Y_1,\dots,Y_n)|Y_1,\dots,Y_k].$$

Example (A more specific one under this family). Consider a random graph G of vertices $v_1,\dots,v_n$, where edges e_{ij} appear independently with probability p_{ij}. Define the chromatic number as

$$\chi(G) = \min \text{ \# of colors needed to color vertices such that neighbors are different colors.}$$

and set $F(G) = \chi(G)$.

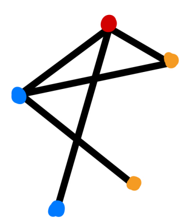

Figure 1: Example of G with $n = 5$ vertices and $\chi(G) = 3$

We are interested in giving an upper bound for $|\chi(G) - \mathbb{E}[\chi(G)]|$ using takeaway information. To achieve this, we make the following attempts by using different sets of $\{y_i\}$:

Try 1: Set $y_1 \dots, y_{\binom{n}{2}} \in \{0,1\}$ based on presence of each edge in G. Then $(+)$ holds since if we add 1 edge, we can always create a new color for either vertex. Hence,

$$|\chi(G) - \mathbb{E}[\chi(G)]| \le O(n) \qquad \text{w.h.p.}$$

Try 2: Reveal one vertex at a time. Add $v_1,\dots,v_n$ and reveal edges $e_{1,k},\dots,e_{k-1,k}$ when v_k is revealed. Let $y_1 = \varnothing, y_2 = \{e_{1,2}\}, y_3 = \{e_{1,3}, e_{2,3}\},\dots$, i.e. y_i is the set of edges revealed with the i-th revealed vertex. Again, $y_1,\dots,y_n$ are independent and $(+)$ holds, (this is because each v_k can get a new special color and won't affect all the previous edges) and so we have

$$|\chi(G) - \mathbb{E}[\chi(G)]| \le O(\sqrt{n}) \qquad \text{w.h.p.}$$

2 Two Improvements

In this part, we hope to improve on the concentration rate of $\mathrm{Var}(X_n)$ and we consider two ways:

(a) Efron-Stein variance bound.

(b) Subgaussian concentration (Azuma-Hoeffding inequality).

2.1 Efron-Stein Inequality

Theorem 2.1 (Efron-Stein Inequality)**.** *Let $Y_1,\dots,Y_n$ be independent random variables and $F:\mathbb{R}^n \to \mathbb{R}$ be a function, then*

$$\mathrm{Var}(F(y_1,\dots,y_n)) \le \sum_{k=1}^{n} \mathbb{E}[\mathrm{Var}(F(y_1,\dots,y_n)|y_{-k})]$$

where $y_{-k} = (y_1,\dots,y_{k-1},y_{k+1},\dots,y_n)$.

Proof. Denote $F = F(y_1,\dots,y_n)$ for notational simplicity. And let $X_k = \mathbb{E}[F(Y_1,\dots,Y_n)|Y_1,\dots,Y_k]$. Then

$$\mathrm{Var}(F) = \mathrm{Var}(\mathbb{E}(F|\mathscr{F}_n)) + \mathbb{E}[\mathrm{Var}(F|\mathscr{F}_n] \le \mathrm{Var}(\mathbb{E}(F|\mathscr{F}_n)) = \mathrm{Var}(X_n)$$

We've proved that $\mathrm{Var}(X_n) = \sum_{k=1}^{n} \mathbb{E}[\mathrm{Var}(X_k|\mathscr{F}_{k-1})]$ holds for any martingale, thus if we can show that

$$\mathbb{E}[\mathrm{Var}(X_k|y_1,\dots,y_{k-1})] \stackrel{?}{\le} \mathbb{E}[\mathrm{Var}(F|y_{-k})], \forall k \in [n] \tag{5}$$

then the proof is finished.
Notice that for any random variable A, $\mathbb{E}[A|\mathscr{G}] = \underset{B \text{ is } \mathscr{G}\text{-measurable}}{\operatorname{argmin}} \mathbb{E}[A-B]^2$ and the minimal value is $\mathbb{E}[\mathrm{Var}\, A|\mathscr{G}]$.
Thus for 5,

$$LHS = \min_{G_k: G_k \text{ is } \sigma(y_1,\dots,y_{k-1}) \text{ measurable}} \mathbb{E}[(X_k - G_k)^2]$$

$$RHS = \min_{G: G \text{ is } \sigma(y_{-1}) \text{ measurable}} \mathbb{E}[(F-G)^2]$$

i.e. both the left and right sides of 5 are minimal values of two functions, denoted as f and g respectively, then if we can show for any y, we can always find some x s.t. $f(x) \le g(y)$, then 5 is proved.
Thus it remains to show that for any given G that is $\sigma(y_{-k})$ measurable, we can find G_k $\sigma(y_1,\dots,y_{k-1})$ measurable s.t. $\mathbb{E}[(X_k - G_k)^2] \le \mathbb{E}[(F-G)^2]$. $\quad (***)$
Let

$$G_k = \mathbb{E}[G|y_1,\dots,y_{k-1}] = E[G|y_1,\dots,y_k]$$

where the second equality comes from that G is $\sigma(y_{-k})$ measurable and thus G is independent of y_k.
So we have $X_k - G_k = \mathbb{E}[F-G|y_1,\dots,y_k]$, then by Jensen's inequality, we have

$$\mathbb{E}[(X_k - G_k)^2] \le \mathbb{E}[(F-G)^2]$$

i.e. we've find G_k that satisfy $(***)$, which finishes the proof.

□

Example (Jack-Knife). We can (over)-estimate $\mathrm{Var}(F)$ by sampling iid copies y_i' of each y_i. One computes

$$\frac{1}{2}\sum_{k=1}^{n}[F(y_1,\dots,y_n) - F(y_1,\dots,y_k',\dots,y_n)]^2 \qquad \text{(Recall: } \mathrm{Var}(X) = \mathbb{E}[(X-X')^2]/2 \quad X, X' \text{ iid copies).}$$

Example (ES can be way off). Set $y_i \sim \mathrm{Unif}\{-1,+1\}$ and $F(y_1,\dots,y_n) = y_1 \cdots y_n$. Then, $\mathrm{Var}(F) = 1$ but $\sum_{k=1}^{n} \mathrm{Var}(F|y_{-k}) = n$.

Example (ES improvement). (First-passage percolation)

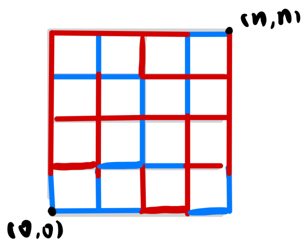

Figure 2: Example of grid

Assign each edge $\{1,2\}$ iid uniform. Set $F = $ min weight path from $(0,0) \to (n,n)$ and take each y_i to be a single edge. Then, (+) holds and so one gets the naive bound

$$|F - \mathbb{E}[F]| \le O(n) \qquad \text{w.h.p.} \qquad \Longrightarrow \mathrm{Var}(F) \le O(n^2).$$

But, using ES one obtains $\mathrm{Var}(F) \le O(n)$. The reasoning is that $\mathrm{Var}(F|y_{-k}) \lesssim \mathbb{P}(e_k \text{ is in all shortest paths})$, because switching the weight of e_k only changes F (and if so, by at most 1) if e_k is in all shortest paths in at least one of the examples. Then:

$$\sum_{\text{edges}} \mathbb{P}(e_k \text{ is in all shortest paths}) \le \mathbb{E}[\text{length of path}] \le O(n).$$

The moral is that although each edge *might* participate in the shortest path, only a few do on average, and Efron–Stein can take advantage.

2.2 Azuma-Hoeffding Inequality

Theorem 2.2 (Azuma-Hoeffding Inequality)**.** *Assume* $(*)$. *Then,* $\forall t \ge 0$,

$$\mathbb{P}(|X_n - X_0| \ge t) \le 2\exp\left(\frac{-t^2}{2\sum c_k^2}\right).$$

Lemma 2.3. *If* $u \in [-1,1]$ *has mean 0, then* $\mathbb{E}[e^{\lambda u}] \le e^{\lambda^2/2}$.

Proof. Note the fact that

$$e^{\lambda u} \le \frac{1}{2}(e^{\lambda} + e^{-\lambda}) + \frac{u}{2}(e^{\lambda} + e^{-\lambda}).$$

From this, we get

$$\mathbb{E}[e^{\lambda u}] \le \frac{1}{2}(e^{\lambda} + e^{-\lambda}) \le e^{\lambda^2/2}$$

where the first inequality is due to $\mathbb{E}\, u = 0$ and the second can be seen from a Taylor expansion. i.e. by Taylor expansion,

$$e^{\lambda^2/2} = 1 + \frac{\lambda^2}{2} + \cdots + \frac{\lambda^{2k}}{2^k(k!)} + \cdots$$

$$\frac{1}{2}(e^{\lambda} + e^{-\lambda}) = 1 + \frac{\lambda^2}{2} + \cdots + \frac{\lambda^{2k}}{(2k)!} + \cdots$$

and we have $\frac{(2k)!}{2^k k!} = 1 \times 3 \times 5 \times \cdots \times (2k-1) \ge 1$. ☐

Proof. (Azuma-Hoeffding) By induction, we will show that $\mathbb{E}[e^{\lambda X_m}] \le e^{\frac{1}{2}\lambda^2 \sum_{k=1}^m c_k^2}$. Observe that

$$\begin{aligned}\mathbb{E}[e^{\lambda(X_m - X_{m-1})} e^{\lambda X_{m-1}} | \mathscr{F}_{m-1}] &= e^{\lambda X_{m-1}} \mathbb{E}[e^{\lambda(X_m - X_{m-1})} \mathscr{F}_{m-1}] \qquad \text{(by martingale property)} \\ &\le e^{\lambda X_{m-1}} e^{\lambda^2 c_m^2 / 2}\end{aligned}$$

which completes the induction. (Here we apply the lemma to $\lambda(X_m - X_{m-1}) \in [-\lambda c_m, \lambda c_m]$.) Now, setting $\lambda = t / \sum_{k=1}^n c_k^2$, Markov's inequality yields

$$\mathbb{P}(X_n \ge t) \le \mathbb{E}[e^{\lambda X_n}] e^{-\lambda t} \le e^{t^2 / (2 \sum_k c_k^2)} e^{-t^2 / \sum_k c_k^2} = e^{-t^2 / (2 \sum_k c_k^2)}.$$

By symmetry, we obtain the same bound for $\mathbb{P}(X_n \le t)$ which completes the proof. □