[CS3958: Lecture 2] Concentration(cont'd), Matingale

Instructor: Chihao Zhang, Scribed by Yulin Wang

September 25, 2022

1 Concentration(cont'd)

1.1 Threshold Behavior of Random Graphs

The second moment method often refers to the use of variance (and hence Chebyshev's inequality) to analyze certain random structures. We demonstrate the method to analyze the *threshold behavior* of Erdős-Rényi random graphs. The notation $G(n,p)$ specifies a distribution over all simple undirected graphs with n vertices, where each of the $\binom{n}{2}$ possible edges appears with probability p independently. Therefore, the expected number of edges in the graph is $\binom{n}{2}p$ and the expected degree of each vertex is $(n-1)p$.

For certain graph properties, random graphs establish the so-called "threshold behavior". That is, in the model $G(n,p)$ it is often the case that there is a threshold function r such that:

$f(n) \ll g(n)$ iff $f = o_n(g)$, $f(n) \gg g(n)$ iff $g(n) \ll f(n)$.

- when $p \ll r(n)$, almost no graph satisfies the desired property;
- when $p \gg r(n)$, almost every graph has the desired property.

Formally, we have

Definition 1 (Threshold function) *Given a graph property P, a function* $r : \mathbb{N} \to [0,1]$ *is called a* threshold function *if:*

(a) if $p(n) \ll r(n), \mathbf{Pr}_{G\sim G(n,p(n))}[G \text{ satisfies } P] \to 0$ *when* $n \to \infty$*;*

(b) if $p(n) \gg r(n), \mathbf{Pr}_{G\sim G(n,p(n))}[G \text{ satisfies } P] \to 1$ *when* $n \to \infty$*;*

Next we will show that the property P = "G contains a 4-clique" has the threshold function $n^{-\frac{2}{3}}$.

A clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent, i.e., an induced complete subgraph.

Theorem 2 *The property "G contains a 4-clique" has a threshold function* $n^{-\frac{2}{3}}$.

Proof. First we verify (a) in Definition 1. For every $S \in \binom{[n]}{4}$, let X_S be the indicator of whether S is a clique, i.e.

For a vertex set S, we use $G[S]$ to denote the subgraph of G induced by S, i.e., $G[S] = \left(S, E(G) \cap \binom{S}{2}\right)$.

$$X_s = \begin{cases} 1, & \text{if } G[S] \text{ is a clique,} \\ 0, & \text{otherwise.} \end{cases}$$

Let $X = \sum_{S\in\binom{[n]}{4}} X_S$. Then X is the total number of 4-cliques in G. So G satisfies P iff $X > 0$. By the linearty of expectation, we have

$$\mathbf{E}[X] = \sum_{S\in\binom{[n]}{4}} \mathbf{E}[X_S] = \binom{n}{4}p^6 \approx \frac{n^4p^6}{24}.$$

Therefore, $\mathbf{E}[X] = o(1)$ when $p \ll n^{\frac{-2}{3}}$. Since X is a non-negative random variable, it follows by Markov inequality that $\mathbf{Pr}[X \geq 1] \leq o(1)$.

However, we could not use the same argument to prove (b), because in general, large expectation of a random variable does not imply large values with high probability. It is possible that almost all graphs contains no 4-clique but a small fraction of graphs contain a large number of 4-cliques, so that the expectation overall is large. Therefore, we have to consider the variance. First notice that

$$\mathbf{Pr}[X = 0] \leq \mathbf{Pr}[|X - \mathbf{E}[X]| \geq \mathbf{E}[X]] \leq \frac{\mathbf{Var}[X]}{(\mathbf{E}[X])^2},$$

where we apply Chebyshev's inequality to obtain the last inequality. Now we only need bound $\mathbf{Var}[X]$.

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}\left[\left(\sum_S X_S\right)^2\right] - \left(\mathbf{E}\left[\sum_S X_S\right]\right)^2 \\ &= \sum_{S \neq T} \mathbf{E}[X_S X_T] + \sum_S \mathbf{E}\left[X_S^2\right] - \sum_{S \neq T} \mathbf{E}[X_S]\,\mathbf{E}[X_T] - \sum_S \mathbf{E}[X_S]^2 \\ &= \underbrace{\sum_{|S \cap T| = 2} (\mathbf{E}[X_S X_T] - \mathbf{E}[X_S]\,\mathbf{E}[X_T])}_{A} + \underbrace{\sum_{|S \cap T| = 3} (\mathbf{E}[X_S X_T] - \mathbf{E}[X_S]\,\mathbf{E}[X_T])}_{B} \\ &\quad + \underbrace{\sum_S \left(\mathbf{E}\left[X_S^2\right] - \mathbf{E}[X_S]^2\right)}_{C}.\end{aligned}$$

When $|S \cap T| = 2$, there are 11 potential edges in $S \cup T$. Therefore, the probability that both S, T induce 4-cliques is p^{11}. We have

When $|S \cap T| = 0$ or 1, X_S and X_T are independent, so $\mathbf{E}[X_S X_T] = \mathbf{E}[X_S]\,\mathbf{E}[X_T]$.

$$A \leq \sum_{|S \cap T| = 2} \mathbf{E}[X_S X_T] = \binom{n}{2}\binom{n-2}{2}\binom{n-4}{2} p^{11} \approx n^6 p^{11}.$$

Similarly, for $|S \cap T| = 3$, the probability that both S an T induce 4-cliques is p^9, so it holds that

$$B \leq \sum_{|S \cap T| = 3} \mathbf{E}[X_S X_T] = \binom{n}{3}\binom{n-3}{1}\binom{n-4}{1} p^9 \approx n^5 p^9.$$

We also have $C \leq \sum_S \mathbf{E}[X_S] \leq n^4 p^6$. To sum up, since $p \gg n^{-\frac{2}{3}}$, we have

$$\mathbf{Var}[X] \leq n^6 p^{11} + n^5 p^9 + n^4 p^6 = o(\mathbf{E}[X]^2).$$

Finally, we get

$$\mathbf{Pr}[X = 0] \leq \frac{\mathbf{Var}[X]}{\mathbf{E}[X]^2} = o(1).$$

□

Recall that $\mathbf{E}[X]^2$ is $\Theta(n^8 p^{12})$. Intuitively, $\frac{n^6 p^{11} + n^5 p^9 + n^4 p^6}{n^8 p^{12}} = n^{-2} p^{-1} + n^{-3} p^{-3} + n^{-4} p^{-6} \ll n^{-4/3} + n^{-1} + 1$ when $p \gg n^{-\frac{2}{3}}$.

It is a common skill to use linearity and independence to simplify the estimation of expectations or variances.

1.2 Hoeffding's Inequality

Recall that the convenient form of the Chernoff bound is: for any $0 < \delta < 1$,

$$\mathbf{Pr}\left[X \ge (1+\delta)\mu\right] \le \exp\left\{\left(-\frac{\delta^2}{3}\mu\right)\right\};$$
$$\mathbf{Pr}\left[X \le (1-\delta)\mu\right] \le \exp\left\{\left(-\frac{\delta^2}{2}\mu\right)\right\}.$$

Example 1 (Tossing coins) *Given a coin which show "head" with probability p, we want to give an estimate $\hat{p}$ of the value p such that with high probability (say 99%), $\hat{p} \in [(1-\varepsilon)p, (1+\varepsilon)p]$. Assume we toss the coin T times. Let X denote the total number of heads, and $X_i \sim \mathtt{Ber}(p)$ be the indicator of whether the i-th toss gives a head. Let $\hat{p} = \frac{X}{T}$ be the estimate of p. Then by Chernoff bound, we have*

$$\mathbf{Pr}\left[|\hat{p}-p| \ge \varepsilon p\right] = \mathbf{Pr}\left[|X - pT| \ge \varepsilon pT\right] \le 2\exp\left\{\left(-\frac{\varepsilon^2}{3}\cdot pT\right)\right\} \le 0.01.$$

So it suffices to choose $T \ge \frac{3\log 200}{\varepsilon^2 p} = O\left(\frac{1}{\varepsilon^2}\right)$.

One of annoying restrictions of Chernoff bound is that each X_i needs to be a Bernoulli random variable. Hoeffding's inequality generalizes Chernoff bound by allowing X_i to follow any distribution, provided its value is almost surely bounded.

Theorem 3 (Hoeffding's inequality) *Let $X_1, \dots, X_n$ be independent random variables where each $X_i \in [a_i, b_i]$ for certain $a_i \le b_i$ with probability 1. Assume $\mathbf{E}[X_i] = p_i$ for every $1 \le i \le n$. Let $X = \sum_{i=1}^n X_i$ and $\mu \triangleq \mathbf{E}[X] = \sum_{i=1}^n p_i$, then*

$$\mathbf{Pr}\left[|X-\mu| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

for all $t \ge 0$.

Proof. You can see the proof in the notes. □

It is instructive to compare Hoeffding and Chernoff when X_i's are independent Bernoulli variables. Formally, let $X_1, \dots, X_n$ be i.i.d. random variables where $X_i \sim \mathtt{Ber}(p)$ for all $i = 1, \dots, n$. Set $X = \sum_{i=1}^n X_i$ and denote $\mathbf{E}[X] = np$ by μ. For $t = \delta\mu$, by Hoeffding's inequality, we have

$$\mathbf{Pr}\left[|X-\mu| \ge t\right] \le 2\exp\left(-2\delta^2 p^2 n\right).$$

By Chernoff Bound, we have

$$\mathbf{Pr}\left[|X-\mu| \ge t\right] \le 2\exp\left(-\frac{1}{3}\delta^2 pn\right).$$

Comparing the exponent, it is easy to see that for $p > 1/6$, Hoeffding's inequality is tighter up to a certain constant factor. However, for smaller

p, Chernoff bound is significantly better than Hoeffding's inequality, as its dependency to p is linear.

The following simple example demonstrates the difference. Suppose we have a box of N balls. Among them pN are red and $(1-p)N$ are blue. We draw a random ball from this box, record its color and put it back. The problem is in how many rounds we are sure about the value $\hat{p}$ (which is the percentage of red balls we record) we guess is within the range $(1 \pm 0.01)p$. The rounds required is $\Omega(1/p)$ if we apply Chernoff bound, and $\Omega(1/p^2)$ if we apply Hoeffding's inequality.

Example 2 (Meal delivery) *During the quarantine of our campus, the professors deliver meals for students using their private cars or trikes. Then a practical problem is how to estimate the amount of meals on a trike conveniently*[1]*.*

[1] See the news.

Assume we need to deliver $n > 200$ packed meals and we do not know the exact number n. Let $X_1, \dots, X_n$ be a sequence of independent and identically distributed random variables, representing the weight of each meal. For any i, $\mu = \mathbf{E}[X_i] = 300$, and $X_i \in [250, 350]$. We measure the total weight of n meals as $X = \sum_{i=1}^n X_i$, then we can give an estimate of n by $\hat{n} = \frac{X}{\mu}$. If we bound its error by a constant δ, then by Hoeffding's inequality, we have

$$\begin{aligned}\mathbf{Pr}[|\hat{n} - n| \geq \delta n] &= \mathbf{Pr}[|X - \mu n| \geq \delta \mu n] \\ &\leq 2\exp\left\{-\frac{2\delta^2\mu^2 n^2}{\sum_{i=1}^n (350-250)^2}\right\}.\end{aligned}$$

It follows that $\mathbf{Pr}[\hat{n} \in [0.95n, 1.05n]] \geq 99.97\%$ ($\delta = 0.05$).

2 Concentration on Martingales

In this section, we relax another restriction of Chernoff bound: the variables need to be mutually independent. If you need to review probability theory, see the notes.

In this note, we use the notation $\overline{X_{i,j}}$ to denote the sequence $X_i, \dots, X_j$ and $\overline{X_i}$ to denote the sequence $X_1, \dots, X_i$.

2.1 Martingales

The notion of martingale is used to describe fair games.

Example 3 (Fair games) *Consider a gambler who wins X_t dollars in the t-th round of a sequence of bets. If in each round, the game is fair, then $\mathbf{E}[X_t] = 0$ regardless of the history. The variables $\{X_t\}$ are not necessarily mutually independent, but if we use $Z_t = \sum_{i=0}^t X_t$ to denote the amount of money he won after t-th round, then clearly for every t, it holds that*

Proposition 4

$$\mathbf{E}[Z_{t+1} \mid X_0, \dots, X_t] = Z_t. \tag{1}$$

Proof. *Since Z_t is $\sigma(X_0, \dots, X_t)$-measurable, we have*

$$\mathbf{E}\left[Z_{t+1} \,\middle|\, \overline{X_{0,t}}\right] = \mathbf{E}\left[Z_t + X_{t+1} \,\middle|\, \overline{X_{0,t}}\right] = Z_t + \mathbf{E}\left[X_{t+1} \,\middle|\, \overline{X_{0,t}}\right] = Z_t$$

□

Taking expectation on the both sides of eq. (1), we have

$$\mathbf{E}[Z_{t+1}] = \mathbf{E}[Z_t] = \cdots = \mathbf{E}[Z_0] = Z_0.$$

We use the property to define *martingales*, i.e., martingales are those random processes satisfying Proposition 4.

Definition 5 *In a probability space $(\Omega, \mathcal{F}, \mathbf{Pr})$, a sequence of finite variables $\{Z_n\}_{n\geq 0}$ is a martingale if*

$$\forall n \geq 1, \mathbf{E}[Z_n \mid Z_1, \dots, Z_{n-1}] = Z_{n-1}.$$

Sometimes, we say $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t another sequence $\{X_n\}_{n\geq 0}$ if

$$\forall n \geq 1, \mathbf{E}[Z_n \mid X_1, \dots, X_{n-1}] = Z_{n-1}.$$

More formally, if for every $i \geq 1$, there exists a σ-algebra $\mathcal{F}_i$ satisfying $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$ and Z_i is $\mathcal{F}_i$-measurable, then we call $\{Z_n\}_{n\geq 0}$ a martingale if

Recall that $\mathbf{E}[Z \mid X] = \mathbf{E}[Z \mid \sigma(X)]$.

$$\forall n \geq 1, \mathbf{E}[Z_n \mid \mathcal{F}_{n-1}] = Z_{n-1}.$$

Here the sequence $\{\mathcal{F}_n\}_{n\geq 0}$ is called a filtration.

Similarly, we say $\{Z_n\}_{n\geq 0}$ a supermartingale *if*

$$\forall n \geq 1, \mathbf{E}[Z_n \mid \mathcal{F}_{n-1}] \leq Z_{n-1},$$

and a submartingale *if*

$$\forall n \geq 1, \mathbf{E}[Z_n \mid \mathcal{F}_{n-1}] \geq Z_{n-1}.$$

If $\{Z_n\}_{n\geq 0}$ is a martingale w.r.t. $\{X_n\}_{n\geq 0}$, then the following property is immediate.

Proposition 6 *For any $n \geq 1$, $\mathbf{E}[Z_n] = \mathbf{E}[Z_0]$.*

Example 4 (1-dim random walk) *Consider the random walk on $\mathbb{Z}$. One starts at 0 and in each round he toss a fair coin to determine the direction of moving distance 1. If we use $X_t \in \{-1, 1\}$ to denote the movement at time t. Let $Z_t = \sum_{i=1}^{t} X_t$ to denote the position at time t, then $Z_0 = 0$. It is obvious that $X_1, X_2 \dots$ are mutually independent random variables with $E[X_t] = 0$. Then $\{Z_t\}_{t\geq 1}$ is a martingale w.r.t. $\{X_t\}_{t\geq 1}$ since*

$$\mathbf{E}\left[Z_t \,\middle|\, \overline{X}_{t-1}\right] = \mathbf{E}\left[Z_{t-1} + X_t \,\middle|\, \overline{X}_{t-1}\right] = Z_{t-1} + \mathbf{E}\left[X_t \,\middle|\, \overline{X}_{t-1}\right] = Z_{t-1}.$$

Example 5 (The product of independent random variables) *Assume that* $X_1, \ldots, X_n$ *are* n *independent random variables with* $\mathbf{E}[X_i] = 1$. *Let* $P_k = \prod_{i=1}^{k} X_i$. *Then* $\{P_i\}_{i\geq 1}$ *is a martingale w.r.t.* $\{X_i\}_{i\geq 1}$ *since*

$$\mathbf{E}\left[P_i \,\middle|\, \overline{X}_{i-1}\right] = \mathbf{E}\left[P_{i-1} \cdot X_i \,\middle|\, \overline{X}_{i-1}\right] = P_{i-1} \cdot \mathbf{E}\left[X_i \,\middle|\, \overline{X}_{i-1}\right] = P_{i-1}.$$

Example 6 (Pólya's urn) *Initially, there are only one white and one black balls in the urn. In each round, we pick a ball uniformly at random from the urn. And then we return the picked ball and add an additional ball with the same color into the urn.*

Let X_n *denote the number of black balls in the urn after* n*-th round. Define* $Z_n \triangleq \frac{X_n}{n}$ *as the ratio of black balls after* n*-th round. Then* $\{Z_n\}_{n\geq 2}$ *is a martingal w.r.t.* $\{X_n\}_{n\geq 2}$ *since*

$$\begin{aligned}\mathbf{E}\left[Z_{n+1} \,\middle|\, \overline{X_{2,n}}\right] &= \frac{1}{n+1}\mathbf{E}\left[X_{n+1} \,\middle|\, \overline{X_{2,n}}\right] = \frac{1}{n+1}\mathbf{E}\left[Z_n(X_n+1) + (1-Z_n)Xn\right] \\ &= \frac{Z_n + X_n}{n+1} = \frac{X_n}{n} = Z_n.\end{aligned}$$

Example 7 (Doob's martingale) *An important family of martingales is the* Doob Sequence. *Let* $X_1, \ldots, X_n$ *be a sequence of (unnecessarily independent) random variables and* $f(\overline{X}_n) = f(X_1, \ldots, X_n) \in \mathbb{R}$ *be a function. For* $i \geq 0$, *we define*

$$Z_i = \mathbf{E}\left[f(\overline{X}_n) \,\middle|\, \overline{X}_i\right].$$

In particular, we have $Z_0 = \mathbf{E}\left[f(\overline{X}_n)\right]$ *and* $Z_n = f(\overline{X}_n)$. *In other words,* Z_n *is the value of the function given the input* $\overline{X}_n$ *and* Z_0 *is the average of the function value without any knowledge about the input. The sequence* $\{Z_i\}_{i\geq 0}$ *can be viewed as an sequence estimation of the function value with more and more information is revealed.*

Proposition *7* $\{Z_n\}_{n\geq 0}$ *is a martingale w.r.t.* $\{X_n\}_{n\geq 0}$.

Proof.

$$\mathbf{E}\left[Z_i \,\middle|\, \overline{X}_{i-1}\right] = \mathbf{E}\left[\mathbf{E}[f(\overline{X}_n) \,|\, \overline{X}_i] \,\middle|\, \overline{X}_{i-1}\right] = \mathbf{E}\left[f(\overline{X}_n) \,\middle|\, \overline{X}_{i-1}\right] = Z_{i-1}.$$

□

2.2 Azuma-Hoeffding's Inequality

With the knowledge of martingales, we are able to generalize Hoeffding's inequality:

Theorem 8 (Azuma-Hoeffding inequality) *Suppose we have a series of random variables* $\{X_n\}_{n\geq 1}$, *which satisfy* $X_i \in [a_i, b_i]$. *Without loss of generality, we assume* $\boldsymbol{E}(X_i) = 0$. *Otherwise, we can replace* X_i *with* $X_i - \boldsymbol{E}(X_i)$. *Let*

$S_k = \sum_{i=1}^{k} X_i$. *If* $\{S_n\}_{n\geq 0}$ *where* $S_k = \sum_{i=0}^{k} X_i$ *is a martingale w.r.t.* $\{X_n\}_{n\geq 0}$ *with* $X_i \in [a_i, b_i]$ *with probability* 1*, then*

$$\mathbf{Pr}\left[|S_n - S_0| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Proof. The proof is quite similar to our proof of Hoeffding inequality. You can see the proof in the notes. □