Mathematical Toolkit **Autumn 2019**

Lecture 15: November 19, 2019

Lecturer: Madhur Tulsiani

1 The Probabilistic Method: Independent Sets

Let us consider one more application of the *Probabilistic Method*, which is a powerful tool show the existence of objects with certain properties without necessarily constructing them. In the previous lecture we used probabilistic reasoning to show that there exists an assignment to a 3-SAT formula with m clauses satisfying $7m/8$ clauses, and then also gave an algorithm to find such an assignment. We will now use the method to show the existence of large independent sets in graphs.

Consider a graph $G = (V, E)$. A set $S \subseteq V$ is said to be an independent set if no edge lies completely within the set S. That is, $\forall e = \{i, j\}$, either $i \notin S$ or $j \notin S$. We are interested in finding a large independent set.

Let $N(i)$ denote the set of all neighbors of i i.e., $N(i) = \{j \mid \{i, j\}\} \in E$ and let $\deg(i) = |N(i)|$. Let us first consider a weaker statement which can be proved without any probabilistic reasoning at all.

Proposition 1.1 *Let $G = (V, E)$ be a graph with n vertices and let d be such that $deg(i) \leq d$ for all $i \in [n]$. Then there exists an independent set S of size $|S| \geq \frac{n}{d+1}$.*

Proof: Start with $S = \emptyset$ and consider the vertices of the graph in the order $1, \ldots, n$. When considering vertex i, if none of the neighbors of i (vertices in $N(i)$) are already included in S, then include i in S. At any step in this process, including a vertex in S removes at most d vertices from being included later. Since at the end, we finish processing all the n vertices, we must have $|S| \geq \frac{n}{d+1}$. ∎

The above bound is good in some cases, but the degrees of vertices in the graph might vary a lot and in particular asking for a uniform bound d which holds for all vertices might be too lossy (consider a "star" graph with one vertex connected to $n-1$ others, and no other edges). The following result gives a much better bound.

Theorem 1.2 *Let $G = (V, E)$ be a graph with n vertices. Then there exists an independent set S such that*

$$|S| \;\geq\; \sum_{i=1}^{n} \frac{1}{deg(i)+1} \;\geq\; \frac{n}{\max_i\{deg(i)\}+1}\,.$$

The main trick in such kind of problems is to set up the right kind of probabilistic experiment, the analysis is usually quite easy. In this question, we can't do everything independently unlike in some previous questions. Suppose that we do - and hence pursue the following idea: Put each v_i in S with probability p. We can't guarantee that we would not pick up both the endpoints of an edge to keep in S. However, this idea can also be made to work and is very useful in some settings. For now, we will prove the theorem using the observation that we can run the greedy algorithm starting with a *random* ordering of the vertices, instead of the fixed ordering $1, \ldots, n$. If we have an example where we have a single high-degree vertex surrounded by low-degree vertices, then in a random ordering we are much more likely to process one of the low-degree neighbors first (which are all good for the analysis).

Proof: Pick a random permutation π of the vertices $\{1, 2, \ldots n\}$. We define the set S as the set of all vertices which appear before all their neighbors in the ordering given by the permutation π.

$$S = \{i \mid \pi(i) < \pi(j) \ \ \forall j \in N(i)\} \ .$$

This is clearly an independent set since if $i \in S$, then for all $j \in N(i)$, we have $\pi(j) > \pi(i)$ and hence $j \notin S$. We now analyze the size of this independent set. We have $|S| = \sum_i X_i$, where

$$X_i = \begin{cases} 1 & \text{if } i \in S \\ 0 & \text{otherwise} \end{cases}$$

Thus, $\mathbb{E}[|S|] = \sum_i \mathbb{E}[X_i]$. To compute $\mathbb{E}[X_i]$, we notice that a random permutation of $[n]$ also induces a random ordering of the set $\{i\} \cup N(i)$. The probability that i appears before any of its neighbors in the ordering is $1/(\deg(i) + 1)$. Thus,

$$\mathbb{E}[X_i] = \frac{1}{\deg(i) + 1} \, ,$$

which gives

$$\mathbb{E}[|S|] \ = \ \sum_{i=1}^{n} \frac{1}{\deg(i) + 1} \, ,$$

and hence there must exist an independent set S with the above size. ∎

2 Inequalities

We will develop some inequalities which let us bound the probability of a random variable taking a value very far from its expectation.

2.1 Markov's Inequality

This is the most basic inequality we will use. This is useful if the only thing we know about a random variable is its expectation. It will also be useful to derive other inequalities later.

Lemma 2.1 (Markov's Inequality) *Let Z be non-negative variable. Then,*

$$\mathbb{P}[Z \geq t] \;\leq\; \frac{\mathbb{E}[Z]}{t}\,. \tag{1}$$

Proof: We start by considering the event $E \equiv \{Z \geq t\}$. We can then write,

$$\mathbb{E}[Z] \;=\; \mathbb{P}[E] \cdot \mathbb{E}[Z \mid E] + \mathbb{P}[E^{\mathsf{c}}] \cdot \mathbb{E}[Z \mid E^{\mathsf{c}}]\,.$$

Using non-negativity of Z, we get

$$\mathbb{E}[Z] \;\geq\; \mathbb{P}[E] \cdot \mathbb{E}[Z \mid E] \;\geq\; \mathbb{P}[E] \cdot t \;=\; \mathbb{P}[Z \geq t] \cdot t\,,$$

which completes the proof. ∎

2.2 Chebyshev's Inequality

The variance of a random variable X is defined as

$$\mathsf{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$$

Also, for two random variables X and Y, we define the covariance as

$$\mathsf{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]\,.$$

Lemma 2.2 (Chebyshev's inequality) *Let Z be a random variable and let $\mu = \mathbb{E}[Z]$. Then,*

$$\mathbb{P}[|Z - \mu| \geq t] \;\leq\; \frac{\mathsf{Var}[Z]}{t^2} \;=\; \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{t^2}\,. \tag{2}$$

Proof: Consider the non-negative random variable $(Z - \mu)^2$. Applying Markov's inequality we have

$$\mathbb{P}[|Z - \mu| \geq t] \;=\; \mathbb{P}\left[(Z-\mu)^2 \geq t^2\right] \;\leq\; \frac{\mathbb{E}\left[(Z-\mu)^2\right]}{t^2}\,.$$

∎

3 Coin tosses revisited

An unbiased coin is tossed n times. Probability that head shows up in each toss is $\frac{1}{2}$. Let Z be a random variable for the number of heads that have showed up after n tosses. We also have random variables X for i^{th} coin toss, where $X_i = 1$ if head shows up in i^{th} toss and 0 otherwise.

So we have

$$Z = \sum_{i=1}^{n} X_i \quad \text{and} \quad \mathbb{E}[Z] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{2}.$$

Let us now compare the kind of bounds we get using Markov's and Chebyshev's inequalities.

3.1 Application of Markov's inequality

Using Markov's inequality we have,

$$\mathbb{P}\left[Z \geq \frac{3n}{4}\right] \leq \frac{\mathbb{E}[Z]}{(3n/4)} \;\Rightarrow\; \mathbb{P}\left[Z \geq \frac{3n}{4}\right] \leq \frac{2}{3} \;\Rightarrow\; \mathbb{P}\left[Z - \frac{n}{2} \geq \frac{n}{4}\right] \leq \frac{2}{3}.$$

3.2 Application of Chebyshev's inequality

We want to show that Chebyshev's inequality gives a stronger bound on probability. For this we need to calculate the variance of Z. We do this calculation below in a way that applies in many other situations as well. We have

$$\mathsf{Var}[Z] \;=\; \mathbb{E}\left[Z^2\right] - (\mathbb{E}[Z])^2.$$

We observe that

$$\mathbb{E}\left[Z^2\right] \;=\; \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] \;=\; \mathbb{E}\left[\sum_{i,j} X_i X_j\right] \;=\; \sum_{i,j} \mathbb{E}\left[X_i X_j\right].$$

Similarly,

$$(\mathbb{E}[Z])^2 \;=\; \left(\mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)\right]\right)^2 \;=\; \sum_{i,j} \mathbb{E}[X_i]\,\mathbb{E}\left[X_j\right]$$

So we have

$$\begin{aligned}\mathsf{Var}\left[Z\right] &= \sum_{i,j}\mathbb{E}\left[X_iX_j\right] - \sum_{i,j}\mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \\ &= \sum_{i}(\mathbb{E}\left[X_i^2\right] - (\mathbb{E}\left[X_i\right])^2) + \sum_{i\neq j}(\mathbb{E}\left[X_i, X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]) \\ &= \sum_{i}\mathsf{Var}\left[X_i\right] + \sum_{i\neq j}\mathsf{Cov}\left[X_i, X_j\right],\end{aligned}$$

where $\mathsf{Cov}\left[X_i, X_j\right]$ denotes $\mathbb{E}\left[X_i \cdot X_j\right] - \mathbb{E}\left[X_i\right] \cdot \mathbb{E}\left[X_j\right]$. Since the coin tosses are independent, we have $\mathbb{E}\left[X_iX_j\right] = \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right]$ and hence $\mathsf{Cov}\left[X_i, X_j\right] = 0$. This yields,

$$\mathsf{Var}\left[Z\right] \;=\; \sum_{i}\mathsf{Var}\left[X_i\right] \quad \text{for independent random variables } X_i\,. \tag{3}$$

Also $\mathsf{Var}\left[X_i\right] = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}\left[X_i\right])^2 = p - p^2$, where $p = \mathbb{P}\left[X_i = 1\right]$. Here $p = \frac{1}{2}$, so $\mathsf{Var}\left[X_i\right] = \frac{1}{4}$ and hence, $\mathsf{Var}\left[Z\right] = \frac{n}{4}$. Applying Chebyshev's inequality we have,

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq t\right] \;\leq\; \frac{n}{4t^2}\,.$$

Setting $t = n/4$ and $t = \sqrt{n}$, gives the following bounds

$$\mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq \frac{n}{4}\right] \;\leq\; \frac{4}{n} \quad \text{and} \quad \mathbb{P}\left[\left|Z - \frac{n}{2}\right| \geq \sqrt{n}\right] \;\leq\; \frac{1}{4}$$

Thus, Chebyshev's inequality gives a much stronger bound on a deviation of $n/4$ from the mean, and can also bound the probability of deviations as small as $\sqrt{n}$. In particular, it gives a non-trivial bound whenever the deviation is larger than $\sqrt{\mathsf{Var}\left[Z\right]}$, a quantity which is referred to as the *standard deviation* of the random variable Z.

4 Threshold Phenomena in Random Graphs

We consider a model of Random Graphs by Erdős and Rényi [ER60]. To generate a random graph with n vertices, for every pair of vertices $\{i, j\}$, we put an edge independently with probability p. This model is denoted by $\mathcal{G}_{n,p}$.

Let G be a random $\mathcal{G}_{n,p}$ graph and let H be any fixed graph (on some constant number of vertices independent of n). We will be interested in understanding the probability that G contains a copy of H. We start by computing this when H is K_4, the clique on 4 vertices.

Definition 4.1 *We define* k-clique *to be a fully connected graph with* k *vertices.*

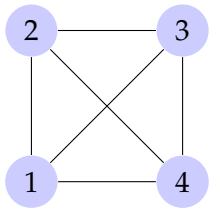

Figure 1: 4-Clique

As a convention, we will count a permutation of a copy of K_4 as the *same* copy. We define the random variable

$$Z \;=\; \text{number of copies of } K_4 \text{ in } G \;=\; \sum_C X_C\,,$$

where C ranges over all subsets of V of size 4 and the random variable X_C is defined as

$$X_C = \begin{cases} 1 & \text{if all pair of vertices in the set } C \text{ have an edge in between them} \\ 0 & \text{otherwise} \end{cases}.$$

We have $\mathbb{E}[X_C] = p^6$, since the probability of connecting all 4 vertices (using 6 edges) in the 4-tuple is p^6. So we have the expectation of Z :

$$\mathbb{E}[Z] \;=\; \sum_C \mathbb{E}[X_C] \;=\; \binom{n}{4} \cdot p^6$$

We observe that

$$\mathbb{E}[Z] \to 0 \;\text{ when } p \ll n^{-2/3} \qquad \text{and} \qquad \mathbb{E}[Z] \to \infty \;\text{ when } p \gg n^{-2/3}\,.$$

Here, by $p \ll n^{-2/3}$, we mean that $\lim_{n\to\infty}(p/n^{-2/3}) = 0$ and $p \gg n^{-2/3}$ is defined similarly. We will prove that there is in fact a threshold phenomenon in the probability that G contains a copy of K_4. When $p \ll n^{-2/3}$, the probability that a random graph G generated according to model $\mathcal{G}_{n,p}$ contains a copy of K_4, goes to 0 as $n \to \infty$. On the other hand, when $p \gg n^{-2/3}$, this probability tends to 1.

Theorem 4.2 *Let G be generated randomly according to the model $\mathcal{G}_{n,p}$ graph. We have that:*

- *If $p \ll n^{-2/3}$, then $\mathbb{P}[G \text{ contains a copy of } K_4] \to 0$ as $n \to \infty$.*
- *If $p \gg n^{-2/3}$, then $\mathbb{P}[G \text{ contains a copy of } K_4] \to 1$ as $n \to \infty$.*

Proof: As above, we define the random variable Z,

$$Z \;=\; \text{number of copies of } K_4 \text{ in } G \;=\; \sum_C X_C\,.$$

The case when $p \ll n^{-2/3}$ can be easily handled by Markov's inequality. We get that,

$$\mathbb{P}[Z > 0] \;=\; \mathbb{P}[Z \;\geq\; 1] \;\leq\; \frac{\mathbb{E}[Z]}{1}.$$

Since $\mathbb{E}[Z] \to 0$ as $n \to \infty$ when $p \ll n^{-2/3}$, we get that $\mathbb{P}[G \text{ contains a copy of } K_4] \to 0$. When $p \gg n^{-2/3}$, we want to show that $\mathbb{P}[Z > 0] \to 1$, i.e., $\mathbb{P}[Z = 0] \to 0$. We use Chebyshev's inequality to prove this. We first compute the variance of Z.

$$\mathsf{Var}[Z] \;=\; \mathsf{Var}\left[\sum_C X_C\right] \;=\; \sum_C \mathsf{Var}[X_C] + \sum_{C \neq D} \mathsf{Cov}[X_C, X_D]$$

Since $\mathbb{E}[X_C] = p^6$, we have $\mathsf{Var}[X_C] = p^6 - p^{12}$. Also, for two distinct sets C and D, we consider four different cases depending on the number of vertices they share.

- **Case 1:** $|C \cap D| = 0$. Since no vertex is shared, X_C and X_D are independent and hence $\mathsf{Cov}[X_C, X_D] = 0$.
- **Case 2:** $|C \cap D| = 1$. Since the variables X_C and X_D depend on *pairs* of vertices in the sets C and D, and the two sets do not share any pair, we still have $\mathsf{Cov}[X_C, X_D] = 0$.
- **Case 3:** $|C \cap D| = 2$. Since C and D share a pair of vertices, there are 11 pairs which must all have edges between them in G, for both X_C and X_D to be 1. Thus, we have $\mathbb{E}[X_C X_D] = p^{11}$ and

$$\mathsf{Cov}[X_C, X_D] \;=\; \mathbb{E}[X_C X_D] - \mathbb{E}[X_C] \cdot \mathbb{E}[X_D] \;=\; p^{11} - p^{12}.$$

- **Case 4:** $|C \cap D| = 3$. in this case C and D share 3 pairs and thus there are 9 distinct pairs of vertices which must all have edges between them for both X_C and X_D to be 1. Thus,

$$\mathsf{Cov}[X_C, X_D] \;=\; \mathbb{E}[X_C X_D] - \mathbb{E}[X_C] \cdot \mathbb{E}[X_D] \;=\; p^9 - p^{12}.$$

Also, there are $\binom{n}{6} \cdot \binom{6}{4}$ pairs C and D such that $|C \cap D| = 2$, and $\binom{n}{5} \cdot \binom{5}{4}$ pairs such that $|C \cap D| = 3$. Combining the above cases we have,

$$\begin{aligned}
\mathsf{Var}[Z] \;&=\; \sum_C \mathsf{Var}[X_C] + \sum_{C \neq D} \mathsf{Cov}[X_C, X_D] \\
&=\; \binom{n}{4} \cdot p^6(1 - p^6) + \binom{n}{6} \cdot \binom{6}{4} \cdot (p^{11} - p^{12}) + \binom{n}{5} \cdot \binom{5}{4} \cdot (p^9 - p^{12}) \\
&=\; O(n^4 p^6) + O(n^6 p^{11}) + O(n^5 p^9).
\end{aligned}$$

Applying Chebyshev's inequality gives

$$\begin{aligned}\mathbb{P}[Z=0] \;\leq\; \mathbb{P}[|Z-\mathbb{E}[Z]| \geq \mathbb{E}[Z]] \;&\leq\; \frac{\mathsf{Var}[Z]}{(\mathbb{E}[Z])^2} \\ &= \frac{1}{\binom{n}{4}^2 \cdot p^{12}} \cdot \left(O(n^4p^6) + O(n^6p^{11}) + O(n^5p^9)\right) \\ &= O\left(\frac{1}{n^4p^6}\right) + O\left(\frac{1}{n^2p}\right) + O\left(\frac{1}{n^3p^3}\right).\end{aligned}$$

For $p \gg n^{-2/3}$, all the terms on the right tend to 0 as $n \to \infty$. Hence, $\mathbb{P}[Z=0] \to 0$ as $n \to \infty$. ■

The above analysis can be extended to any graph H of a fixed size. Let Z_H be the number of copies of H in a random graph G generated according to $G_{n,p}$. Using the previous analysis, we have $\mathbb{E}[Z_H] = \Theta\left(n^{|V(H)|} \cdot p^{|E(H)|}\right)$. Hence, $\mathbb{E}[Z] \to 0$ when $p \ll n^{-|V(H)|/|E(H)|}$ and $\mathbb{E}[Z] \to \infty$ when $p \gg n^{-|V(H)|/|E(H)|}$. Thus, it might be tempting to conclude that $p = n^{-|V(H)|/|E(H)|}$ is the threshold probability for finding a copy of H. However, con-

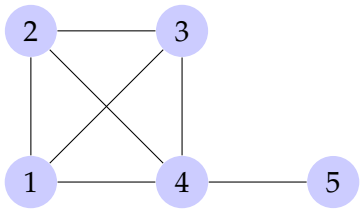

Figure 2: Subgraph H containing K_4

sider the graph in Figure 2. For this graph, we have $|V(H)|/|E(H)| = 5/7$. But for p such that $p \gg n^{-5/7}$ and $p \ll n^{-2/3}$, a random G is extremely unlikely to contain a copy of K_4 and thus also extremely unlikely to contain a copy of H. For an arbitrary graph H, it was shown by Bollobás [Bol81] that the threshold probability is $n^{-\lambda}$, where

$$\lambda \;=\; \min_{H' \subseteq H} \frac{|V(H')|}{|E(H')|}.$$

References

[Bol81] Béla Bollobás, *Threshold functions for small subgraphs*, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 90, Cambridge Univ Press, 1981, pp. 197–206. 8

[ER60] Paul Erdős and A Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci **5** (1960), 17–61. 5