Lecture 4: Probabilistic Inequalities, Random Graphs, $K_{2,2}$ in Bipartite Graphs

Combinatorial Methods (Winter 2023)
University of Toronto
Swastik Kopparty
Scribe: Aaron Ma, Anatoly Zavyalov

1 Basic Probabilistic Inequalities

Definition 1. *Let Ω be a finite set, which we call a* ***probability space****. A* ***probability measure*** *be a function $\mu \colon \Omega \to \mathbb{R}_{\geq 0}$ satisfying*

$$\sum_{w \in \Omega} \mu(w) = 1$$

Definition 2. *An S-**valued random variable** on a probability space Ω is a function $f \colon \Omega \to S$. We may sometimes write "r.v." to mean "random variable".*

1.1 A toy problem

Suppose we have n fair coins X_i each of which is 0 or 1 with equal probability $\frac{1}{2}$, suppose further that each coin toss is independent. Give, with proof, an upper bound on the probability that

$$\sum_{i=1}^{n} X_i \geq \frac{3n}{4}.$$

Definition 3. *If X is an $\mathbb{R}$-valued r.v., we can define the* ***expectation*** *of X by*

$$\mathbb{E}[X] := \sum_{x \in \mathbb{R}} \Pr[X = x] \cdot x$$

It is easy to check the following:

Claim 4 (Linearity of Expectation)**.** *If X and Y are $\mathbb{R}$-valued random variables and $\alpha \in \mathbb{R}$, then*

$$\mathbb{E}[\alpha X + Y] = \alpha \mathbb{E}[X] = \mathbb{E}[Y].$$

We now introduce an inequality to help us with the toy problem:

Theorem 5 (Markov's inequality)**.** *If X is a nonnegative $\mathbb{R}$-valued r.v., then for any $t > 0$ we have*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. If we set $\Pr[X \geq t] = \lambda$, then

$$\begin{aligned}\mathbb{E}[x] &= \sum_{x \leq t} \Pr[X = x] \cdot x + \sum_{x \geq t} E[x] \cdot x \\ &\geq 0 + \lambda \cdot t\end{aligned}$$

Thus,

$$\lambda \leq \frac{\mathbb{E}[X]}{t}.$$

$\square$

Returning to the toy problem, set $X = \sum_{i=1}^{n} X_i$. Then, by the linearity of expectation,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] \\ &= \frac{n}{2}.\end{aligned}$$

Applying Markov's inequality,

$$\Pr\left[X \geq \frac{3n}{2}\right] \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

We can get a better bound using Chebyshev's inequality, which we will state below. First, we introduce the variance of a random variable:

Definition 6 (Variance)**.** *Let X be an $\mathbb{R}$-valued random variable. The **variance** of X is*

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

Theorem 7 (Chebyshev's inequality)**.** *Let X be a $\mathbb{R}$-valued random variable. For any $t > 0$, we have*

$$\Pr\left[|X - \mathbb{E}[X]| > t\right] \leq \frac{Var(X)}{t^2}.$$

Proof. Define $Y = (X - \mathbb{E}[X])^2$. Applying Markov's inequality gives us

$$\Pr[Y \geq t^2] = \Pr[\sqrt{Y} \geq t] \leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\mathrm{Var}(X)}{t^2},$$

and as $\sqrt{Y} = |X - \mathbb{E}[X]|$, we get the desired inequality.

$\square$

Returning to the toy problem,

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mathbb{E}[X_i]\right)^2\right] \\ &= \mathbb{E}\left[\sum_{1\le i,j\le n}(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])\right] \\ &= \sum_{1\le i,j\le n}\mathbb{E}[(X_i-\mathbb{E}[X_i])(X_j-\mathbb{E}[X_j])] \\ &= \sum_{1\le i=j\le n}\mathbb{E}[(X_i-\mathbb{E}[X_i])^2] + \sum_{1\le i\ne j\le n}\mathbb{E}[(X_i-\mathbb{E}[X_i])]\mathbb{E}[(X_j-\mathbb{E}[X_j])] \\ &\qquad\qquad\text{using independence} \\ &= \sum_{i=1}^{n}\mathbb{E}\left[\left(X_i-\frac{1}{2}\right)^2\right]\end{aligned}$$

For any $1 \le i \le n$, we can compute

$$\left(X_i - \frac{1}{2}\right)^2 = \begin{cases}(0-\frac{1}{2})^2 = \frac{1}{4} & \text{With probability } \frac{1}{2} \\ (1-\frac{1}{2})^2 = \frac{1}{4} & \text{With probability } \frac{1}{2}\end{cases}$$

So taking the sum gives us $\mathrm{Var}(X) = \mathbb{E}[(X_i - 1/2)^2] = \frac{n}{4}$. Applying Chebyshev's inequality then gives

$$\Pr[|X - n/2| > t] \le \frac{\mathrm{Var}(X)}{t^2} = \frac{n/4}{t^2}$$

And setting $t = \frac{n}{4}$ gives a bound of $\Pr[|X-n/2| > t] \le \frac{4}{n}$. [1]

We can actually get an even stronger bound if we raised $(X - \mathbb{E}[X])$ to the power of 4 rather than 2; we would instead have a sum across $1 \le i, j, k, l \le n$ whose terms vanish unless an even number of i, j, k, l are equal. Plugging this in would give a bound on the order of $1/n^2$. We can use Chebyshev-style arguments to get even better bounds.

2 Chernoff Bounds

The aim of this section is to get a better inequality by applying Markov's inequality to fancier random variables. Using the random variables defined in the previous section's toy problem, define $Z_i = X_i - \frac{1}{2}$ and $Z = \sum_{i=1}^{n} Zi$. Fix $a > 0$, and define $Y = e^{aZ}$. Then, by Markov's inequality,

$$\Pr[Y > t] = \Pr\left[e^{aZ} > e^{at}\right] \le \frac{\mathbb{E}[e^{aZ}]}{e^{at}}$$

Now to compute $\mathbb{E}[e^{aZ}]$, we can use the fact that the Z_i's are independent to get

$$\mathbb{E}\left[e^{aZ}\right] = \mathbb{E}\left[e^{a(Z_1+\cdots+Z_n)}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{aZ_i}\right] = \prod_{i=1}^{n}\mathbb{E}\left[e^{aZ_i}\right]$$

[1] X is positive, so $|X - n/2| > n/4$ if and only if $X > 3n/4$.

For every $1 \le i \le n$, we can calculate the expectation of e^{aX_i}:

$$\mathbb{E}\left[e^{aX_i}\right] = \frac{1}{2} \cdot e^{-a/2} + \frac{1}{2} \cdot e^{a/2}.$$

Plugging in this expectation gives the bound

$$\Pr[Y > t] \le \frac{\left(\frac{e^{-a/2}}{2} + \frac{e^{a/2}}{2}\right)^n}{e^{at}}$$

We can then find the best bound possible by minimizing

$$r = \left(\frac{e^{-a/2}}{2} + \frac{e^{a/2}}{2}\right)^n$$

with respect to a, which we can estimate by the following lemma:

Lemma 8. *For any $x \in \mathbb{R}$, we have*

$$\cosh(x) \le e^{x^2/2}.$$

Proof. Comparing Taylor series,

$$\cosh(x) = \frac{e^x + e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{1}{(2k)!} x^{2k} \le \sum_{k=0}^{\infty} \frac{1}{2^k k!} x^{2k} = e^{x^2/2}.$$

$\square$

This above bound then gives us

$$\begin{aligned}\Pr[Y > t] &\le \frac{\left(\frac{e^{-a/2}}{2} + \frac{e^{a/2}}{2}\right)^n}{e^{at}} \\ &\le \frac{\left(e^{a^2/8}\right)^2}{e^{at}} \\ &= e^{a^2 n/8 - an/4}.\end{aligned}$$

Setting $a = 1$, we get that $Pr[Y > n/4] \le e^{n/8 - n/4} = e^{-n/8}$, which is significantly smaller than any of the polynomial bounds we got earlier.

Theorem 9 (Full Chernoff bound on n independent coins)**.** *If $X_1, \dots, X_n$ are independent random variables with $Pr[X_i = 1] = p$, then*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - np\right| > \varepsilon n\right] \le e^{\frac{\varepsilon^2 n}{3}}$$

This gets vanishingly small when $\varepsilon >> 1/\sqrt{n}$, i.e, for arbitrarily small $\delta > 0$, we can find large enough n such that

$$\Pr\left[\sum_{i=1}^{n} X_i \in [np - k\sqrt{n}, np + k\sqrt{n}]\right] = 1 - \delta.$$

3 Random graphs

Let $G(n,p)$ denote a graph with n vertices such that each edge shows up independently with probability p. We want to determine the average number of triangles that G has.

For each potential edge $\{i,j\}$, let $X_{i,j}$ be the indicator random variable for that edge. For each potential triangle $\{i,j\},\{j,k\},\{k,i\}$, let $Z_{i,j,k}$ be the indicator for that triangle appearing. Then define

$$Z = \sum_{\{i,j,k\}\in\binom{[n]}{3}} Z_{i,j,k}$$

The goal of this section is to understand what Z usually is. We first compute its expectation:

$$\mathbb{E}[Z] = \sum_{\{i,j,k\}\in\binom{[n]}{3}} \mathbb{E}[Z_{i,j,k}] = \sum_{\{i,j,k\}\in\binom{[n]}{3}} \mathbb{E}[X_{i,j}]\mathbb{E}[X_{j,k}]\mathbb{E}[X_{i,k}] = \binom{n}{3}p^3$$

We can also compute the variance of Z:

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathbb{E}\left[\left(Z - \binom{n}{3}p^3\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j,k} Z_{i,j,k} - p^3\right)^2\right] \\
&= \sum_{\substack{i,j,k \\ i',j',k'}} \mathbb{E}[(Z_{i,j,k}-p^3)(Z_{i',j',k'}-p^3)]
\end{aligned}$$

Notice that $Z_{i,j,k}$ and $Z_{i',j',k'}$ are independent whenever $|\{i,j,k\}\cap\{i',j',k'\}| < 2$, so

$$\begin{aligned}
Var(Z) &= \sum_{\substack{i,j,k \\ i',j',k' \\ |\{i,j,k\}\cap|\{i',j',k'\}|\geq 2}} \mathbb{E}[(Z_{i,j,k}-p^3)(Z_{i',j',k'}-p^3)] \\
&= \sum_{i,j,k} \mathbb{E}[(Z_{i,j,k}-p^3)^2] + \sum_{\substack{i,j,k \\ k'\notin\{i,j,k\}}} \mathbb{E}[(Z_{i,j,k}-p^3)(Z_{i',j',k'}-p^3)],
\end{aligned}$$

which is around $\binom{n}{k}p^3(1-p^3) + \binom{n}{4}\binom{4}{2}(p^5-p^6) \in \Theta(p^3n^3+n^4p^5)$. Using Chebyshev's inequality,

$$\Pr\left(\left|Z - \binom{n}{3}p^3\right| > t\right) \leq \frac{\mathcal{O}(n^3p^3+n^4p^5)}{t^3}$$

Setting $t = \varepsilon p^3\binom{n}{3}$ gives us that

$$\Pr\left[Z \notin (1\pm\varepsilon)\binom{n}{3}p^3\right] \leq \frac{n^3p^3+n^4p^5}{\epsilon^2p^6n^6} = \frac{1}{\varepsilon^2}\left(\frac{1}{n^3p^3}+\frac{1}{n^2p}\right),$$

and for $p >> \frac{1}{\sqrt{n}}$, we see that $Z \in (1\pm\varepsilon)\binom{n}{3}p^3$ with probability $1-\mathcal{O}(1)$.

4 Existence of $K_{2,2}$ in bipartite graphs

Definition 10 ($K_{2,2}$, $K_{2,1}$)**.** *A bipartite graph $G = (L \sqcup R, E)$ has a $K_{2,2}$ subgraph if there exist distinct $a, b \in L$ and $c, d \in R$ such that $\{(a, c), (a, d), (b, c), (b, d)\} \subseteq E$. G has a $K_{2,1}$ subgraph if there exists $a \in L$ and distinct $b, c \in R$ such that $\{(a, b), (a, c)\} \subseteq E$. $K_{2,2}$ and $K_{2,1}$ subgraphs are pictured in Figures 1 and 2.*

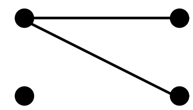

Figure 1: $K_{2,1}$ subgraph

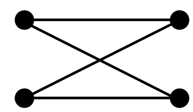

Figure 2: $K_{2,2}$ subgraph

Let $G = (L \sqcup R, E)$ be a bipartite graph with $|L| = |R| = n$. We want to determine how many edges guarantee the existence of $K_{2,2}$ and $K_{2,1}$ subgraphs in G.

First note that we can use the pigeonhole principle to show that any such bipartite graph with $n + 1$ edges is guaranteed to have a $K_{2,1}$ subgraph.

For $K_{2,2}$ subgraphs, the graph as in Figure 3 shows that the maximum number of edges a graph without having a $K_{2,2}$ is $\geq 2n$.

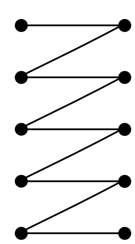

Figure 3: Bipartite graphs can have $2n$ edges and no $K_{2,2}$

We will construct our graph as follows: let $L = R = \mathbb{Z}/n\mathbb{Z}$, and let $S \subseteq \mathbb{Z}/n\mathbb{Z}$ be a subset that is to be determined. We join each $i \in L$ to $i + s \in R$ (with addition being done mod n) for each $s \in S$, so $E = \{(i, i + s) : i \in L, s \in S\}$. We want to find a suitable set S such that our graph does not have any $K_{2,2}$ subgraphs.

If there is a $K_{2,2}$, then there exist $i, j \in L$ and $s_1, s_2, s_3, s_4 \in S$ such that $\{(i, i+s_1), (i, i+s_2), (j, j+s_3), (j, j + s_4)\} \subseteq E$, and $i + s_1 = j + s_2$ and $j + s_3 = j + s_4$. Subtracting these equations gives us $s_1 - s_2 = s_3 - s_4$.

Question 1. *How big of a subset $S \subseteq \mathbb{Z}/n\mathbb{Z}$ exists such that for all $s_1, s_2, s_3, s_4 \in S$, we have $s_2 = s_1$ and $s_4 = s_3$ whenever $s_2 - s_1 = s_4 - s_3$?*

The true answer to Question 1 is about $|S| = \Theta(\sqrt{n})$, which we will not show. It's easy to find a set of size $|S| = \Theta(\log n)$ that satisfies this: consider the set consisting of powers of 2:

$$S = \{1, 2, 4, \dots, 2^k\}$$

Then S satisfies the condition mentioned above.

Theorem 11. *If a bipartite graph $G = (L \sqcup R, E)$ with $|L| = |R| = n$ has $m = \omega(n^{\frac{3}{2}})$ edges, then G has a $K_{2,2}$ subgraph.*

Proof. We want to find two vertices in L that have ≥ 2 common neighbours. We will count the number of tuples (i, j, k) where $i, j \in L$ and $k \in R$ such that $(i, k) \in E$ and $(j, k) \in E$, as in Figure 4.

If $\mathbb{E}[\text{degree}(k)]$ is the average degree of any vertex $k \in R$ and we assume that degree(k) is uniformly distributed then we denote $\overline{d} = \frac{m}{n} = \mathbb{E}[\text{degree}(k)]$ for any $k \in R$. Then,

$$\begin{aligned}
\mathbb{E}\left[\#\{(i,j,k) : (i,k) \in E \text{ and } (j,k) \in E\}\right] &= \sum_{k \in R} \mathbb{E}\left[\binom{\text{degree}(k)}{2}\right] \\
&\geq \sum_{k \in R} \binom{\overline{d}}{2} \\
&\qquad \text{by Jensen's inequality, as } f(x) = \binom{x}{2} = \tfrac{x(x-1)}{2} \text{ is convex} \\
&= n \cdot \binom{m/n}{2} \\
&\geq \frac{n}{2}\left(\frac{m}{n}\right)^2 \\
&= \frac{m^2}{2n}.
\end{aligned}$$

We can then use the pigeonhole principle: If the number of tuples $(i, j, k) \in L \times L \times R$ such that $\{(i, k), (j, k)\} \subseteq E$ exceeds the number of the 2-element subsets $\{i, j\} \subseteq L$, then G has a $K_{2,2}$ subgraph. Equivalently, if $\frac{m^2}{2n} > \binom{n}{2}$, then G has a $K_{2,2}$ subgraph, so $m = \omega(n^{\frac{3}{2}})$ guarantees the existence of a $K_{2,2}$ subgraph in G.

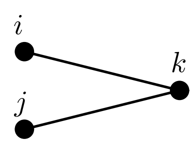

Figure 4: We count the number of vertices $i, j \in L$ and $k \in R$ such that $(i, k), (j, k) \in E$.

$\square$

5 Line-point incidence graph over $\mathbb{F}_q$

Now we show that the bound in Theorem 11 is tight by constructing a graph with $n^{3/2}$ edges that does not have a $K_{2,2}$. Let $\mathbb{F}_q$ be an arbitrary field. Let L denote the set of lines in $\mathbb{F}_q^2$, and let R denote the set of points in $\mathbb{F}_q^2$. Namely,

$$L = \{\text{lines in } \mathbb{F}_q^2\} = \{(m, b) \in \mathbb{F}_q^2\}$$
$$R = \{\text{points in } \mathbb{F}_q^2\} = \{(x, y) \in \mathbb{F}_q^2\}$$

Let

$$E = \{((m, b), (x, y)) \in L \times R : (x, y) \text{ is a solution to } y = mx + b\}.$$

Note that the graph $G = (L \sqcup R, E)$ has no $K_{2,2}$ subgraph, as two distinct lines can intersect in at most one point.

6 Sidon Sets

We now introduce Sidon sets, which are the formalization of the sets S we discussed in Section 4.

Definition 12 (Sidon Set)**.** *A set $S \subseteq \mathbb{Z}/n\mathbb{Z}$ is a **Sidon set** if for every $a, b, c, d \in S$, we have $\{a, b\} = \{c, d\}$ whenever $a + b = c + d$.*

Claim 13. *If $S \subseteq \mathbb{Z}/n\mathbb{Z}$ is a Sidon set, then $|S| \leq \mathcal{O}(\sqrt{n})$.*

Proof. Suppose that $S \subseteq \mathbb{Z}/n\mathbb{Z}$ is a Sidon set. For any distinct $a, b \in S$, the total number of all possible sums $a + b$ is at most n (as $a + b \in \mathbb{Z}/n\mathbb{Z}$), and as there are at most $\binom{|S|}{2}$ such distinct pairs (a, b), this implies that $\binom{|S|}{2} \leq n$, so $|S| = \mathcal{O}(\sqrt{n})$, as desired. □

Claim 14. *There exists a Sidon set $S \subseteq \mathbb{Z}/n\mathbb{Z}$ with $|S| = \Omega(n^{1/4})$.*

Proof. Let $p \in [0, 1]$ be a constant to be determined later. For each $x \in \mathbb{Z}/n\mathbb{Z}$, include x in S with probability p. We will get a bound for $\Pr[S \text{ is not Sidon}]$.

Fix $a, b, c, d \in \mathbb{Z}/n\mathbb{Z}$ such that $a+b = c+d$ and $\{a, b\} \neq \{c, d\}$. The probability that $\{a, b, c, d\} \subseteq S$ is p^4, as each a, b, c, d has a probability p of being in S. Consider the event $E_{a,b,c,d} =$ "$\{a, b, c, d\} \subseteq S$". By the above, we have $\Pr[E_{a,b,c,d}] = p^4$. Then,

$$\Pr[S \text{ is not Sidon}] = \Pr\left[\bigvee_{a,b,c,d \in S} E_{a,b,c,d}\right] = p^4 \cdot (\# \text{ of such } a, b, c, d) \leq p^4 \cdot n^3.$$

Setting $p = \frac{n^{-3/4}}{100}$ gives us that $\Pr[S \text{ is not Sidon}] \leq \frac{1}{10^8}$, and by Chebyshev's inequality we have $\Pr[|S| < \frac{pn}{2}] \leq \frac{1}{np} = \mathcal{O}\left(\frac{1}{n^{1/4}}\right)$. As $\frac{np}{2} = \Omega(n^{1/4})$ and $\Pr[S \text{ is not Sidon}] < 1$, this gives us that there exist Sidon sets S satisfying $|S| = \Omega(n^{1/4})$. Therefore, $\Pr[|S| \geq \frac{pn}{2} \text{ and } S \text{ is Sidon}]$ can be made arbitrarily large by adjusting p, so most subsets of $\mathbb{Z}/n\mathbb{Z}$ are Sidon.

□