Advanced Algorithms III (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Yujie Lu & Yangtian Zhang

Last modified on Sept 27, 2020

In this lecture, we first introduce the balls-into-bins model, a simple probabilistic model that we will meet many times in this course. We will see that the analysis of many randomized algorithms will eventually reduce to some basic questions about balls and bins. So we will develop tools to study on the model.

We will then introduce "concentration inequalities", namely a set of inequalities that provide bounds on how a random variable deviates from its expectation. We will develop the "second-order method" and see how it applies to analyze random graphs.

1 Balls into bins

Ball-into-bins is the following random process:

> Throw m balls into n bins uniformly at random.

Many interesting questions can be asked about the process and today we mainly investigate two of them.

1.1 Birthday paradox

Consider the probability that some bin has more than one ball. The problem can also be described as the probability that two persons in the class have the same birthday, hence is called **birthday paradox** . Since each ball is thrown independently, the probability that no collision occurs after k-th ball is thrown is $\frac{n-k+1}{n}$. Hence

$$\begin{aligned}
\mathbf{Pr}[\text{ no same birthday }] &= \prod_{k=1}^{m} \frac{n-k+1}{n} \\
&= \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right) \\
&\leq \exp\left(-\frac{\sum_{k=1}^{m-1} k}{n}\right) \quad (\text{ by } 1+x \leq e^x) \\
&= \exp\left(-\frac{m(m-1)}{2n}\right).
\end{aligned}$$

For $m = O\left(\sqrt{n}\right)$, the probability can be arbitrarily close to 0.

Remark. *This inequality is quite tight, since when n is sufficiently large, we have $k < m$ which implies that $\frac{k}{n} < \frac{\sqrt{n}}{n} \to 0$.*

1.2 Max Load

Max load is the number of balls in the fullest bin. X_i be the number of balls in the i-th bin. We need to compute $X = \max_{i \in [n]} X_i$. Today we assume that $m = n$. We try to find a number k such that $\mathbf{Pr}[X > k] = O(1)$. By the union bound, we have

$$\mathbf{Pr}\left[\max_i X_i > k\right] = \mathbf{Pr}[\exists i : X_i > k] \leq n \cdot \mathbf{Pr}[X_1 > k].$$

So it suffices to determine the k such that $\mathbf{Pr}[X_1 > k] = O\left(\frac{1}{n}\right)$.

Again by the union bound we have

$$\mathbf{Pr}[X_1 > k] \leq \binom{n}{k} \cdot n^{-k} \leq \frac{1}{k!} \leq \left(\frac{k}{e}\right)^k,$$

where the last inequality is due to the Stirling's formula [3] $k! \approx \sqrt{2\pi k}\left(\frac{k}{e}\right)^k$. So we need to choose k such that $\left(\frac{k}{e}\right)^k = O\left(\frac{1}{n}\right)$. Indeed, it is easy to verify that $k = O\left(\frac{\log n}{\log \log n}\right)$ is enough.

By the analysis above, we know the maximum load is $O\left(\frac{\log n}{\log \log n}\right)$ w.h.p. (with high probability). I will leave as an exercise to show that $\mathbf{E}[X] = \Theta\left(\frac{\log n}{\log \log n}\right)$.

Remark: Analysis for the case that $m \neq n$ is given in [5]. And a tight bound is given in [4], which is $\Gamma^{-1}(n) - \frac{3}{2} + o(1)$.

2 Concentration Inequality

We are often interested in how a random variable deviates from certain fixed value (typically 0 or its expected value). Concentration inequalities are inequalities of this form.

2.1 Markov's Inequality

Theorem 1 (Markov's Inequality). *For any non-negative random variable X and $a > 0$,*

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Proof. Just notice that

$$\mathbf{E}[X] \geq a \cdot \mathbf{Pr}[X \geq a] + 0 \cdot \mathbf{Pr}[X < a].$$

It follows that

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

□

2.2 Applications of Markov's Inequality

There are lots of applications of Markov's inequality. For example, we can apply Markov's inequality directly to the Max Load problem mentioned above. Since $\mathbf{E}[X_1] = 1$, we derive that

$$\mathbf{Pr}\left[X_1 > \frac{\log n}{\log \log n}\right] \leq \frac{\log \log n}{\log n}$$

The upper bound is weaker than the previous result $\frac{1}{n}$, mainly due to the fact that we only utilize $\mathbf{E}[X_1]$ to estimate the upper bound while lacking a lot of other information.

Here we give another simple application of the Markov's inequality. There are two types of randomized algorithms, Las Vegas and Monte Carlo. Assume the task is to find some correct answer k.

- Las Vegas Algorithm[2]: a random variable k is generated and checked, repeat this process until the right k is found.
- Monte Carlo Algorithm: the process is repeated for only N times and output certain k. The output might not be correct.

In other words, a Las Vegas randomized algorithm always outputs a correct answer when it terminates. However, the running time of the algorithm is random. In complexity theory, the family of problems solvable by a Las Vegas algorithm terminating in polynomial-time *in expectation* are called **ZPP** (zero-error probabilistic polynomial time).

On the other hand, the running time of a Monte Carlo algorithm is bounded by some fixed value, but the output might be wrong. In case a problem admits a Monte Carlo algorithm whose running time is a polynomial in the size of the input and the probability of correctness is at least $2/3$, we say the problem belongs to **BPP** (bounded-error probabilistic polynomial time).

We can use Markov inequality to show that **ZPP** $\subseteq$ **BPP**. Suppose we have a Las Vegas algorithm $\mathcal{A}$ who terminates in X steps on some input with $\mathbf{E}[X] = T$. We can turn it into a Monte Carlo algorithm by running A for $3T$ steps. If the $\mathcal{A}$ terminates before $3T$, we just output. Otherwise, we output an arbitrary answer. Then the probability that $\mathcal{A}$ makes a mistake is bounded by

$$\mathbf{Pr}[X > 3T] \leq \frac{1}{3}.$$

2.3 Chebyshev's Inequality

A common trick to improve concentration is to consider $\mathbf{E}[f(X)]$ instead $\mathbf{E}[X]$ for some nondecreasing $f : \mathbb{R} \to \mathbb{R}$, since

$$\mathbf{Pr}[X \geq a] = \mathbf{Pr}[f(X) \geq f(a)].$$

Take $f(x) = x^2$ we obtain another concentration inequality: Chebyshev's Inequality

Theorem 2 (Chebyshev's Inequality). *For any non-negative random variable X and $a > 0$,*

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}\left[X^2\right]}{a^2}.$$

Replace X with $X - \mathbf{E}[X]$, we get another form of this inequality

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq a] \leq \frac{\mathbf{Var}[X]}{a^2}.$$

2.4 Applications of Chebyshev's Inequality: Coupon-Collector's Problem, Revisited

Recall the coupon collector's problem: given n coupons, what's the expected number of coupons to draw with replacement before having drawn each coupon at least once?

Let X_i be the number of draws to get the i-th distinct coupon while exactly $i-1$ distinct coupons are already in hand. In the previous class we have computed the expectation using the linearity property of expectations:

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathbf{E}[X_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \cdot H(n) \overset{n\to\infty}{\longrightarrow} n(\ln n + \gamma).$$

For the above coupon collector problem, the Markov inequality only provides a very weak concentration. But we can yield a more tight concentration results by applying Chebyshev's Inequality:

$$\mathbf{Pr}[X \geq nH_n + cn] \leq \frac{\pi^2}{6c^2}.$$

In order to apply Chebyshev's inequality, we should first compute the variance of X. Since $X = \sum_{i=0}^{n-1} X_i$ and $X_0, \ldots, X_{n-1}$ are independent, we have

$$\mathbf{Var}[X] = \mathbf{Var}\left[\sum_{i=0}^{n-1} X_i\right] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i].$$

Recall that X_i follows geometric distribution $X_i \sim \text{Geom}(\frac{n-i}{n})$. We have the following lemma for the geometric distribution.

Lemma 3. *Assuming Y follows geometric distribution with parameter p, then* $\mathbf{Var}[Y] = \frac{1-p}{p^2}$

Proof. We first show that $\mathbf{E}\left[Y^2\right] = \frac{2-p}{p^2}$.

Since $\mathbf{Pr}[Y = i] = (1-p)^{i-1}p$, let

$$S \triangleq \mathbf{E}\left[Y^2\right] = \sum_{i=0}^{\infty} i^2(1-p)^{i-1}p. \tag{1}$$

Then we have

$$(1-p) \cdot S = \sum_{i=1}^{\infty} i^2 \cdot (1-p)^i \cdot p \tag{2}$$

Equation (1) - Equation (2) yields

$$S = 1 + \sum_{i=2}^{\infty}(1-p)^{i-1} \cdot (2i-1) = 1 + 2 \cdot \left(\frac{1}{p^2} - 1\right) - \frac{1-p}{p} = \frac{2-p}{p^2}$$

Since $\mathbf{Var}[Y] = \mathbf{E}\left[Y^2\right] - (\mathbf{E}[Y])^2$ and $\mathbf{E}[Y] = \frac{1}{p}$, we finally get

$$\mathbf{Var}[Y] = \mathbf{E}\left[Y^2\right] - (\mathbf{E}[Y])^2 = \frac{1-p}{p^2}.$$

□

Applying the lemma, we obtain $\mathbf{Var}[X_i] = \frac{i}{(n-i)^2}$. Recall that $\mathbf{Var}[X] = \sum_{i=0}^{n-1} \mathbf{Var}[X_i]$, then

$$\begin{aligned}\mathbf{Var}[X] &= \sum_{i=0}^{n-1} \mathbf{Var}[X_i] = \sum_{i=0}^{n-1} \frac{n \cdot i}{(n-i)^2} \\ &\leq n^2 \sum_{i=0}^{n-1} \frac{1}{(n-i)^2} = n^2 \left(\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \ldots + \frac{1}{n^2} \right) \\ &\leq \frac{\pi^2 n^2}{6}.\end{aligned}$$

Then by Chebyshev's inequality, we have

$$\mathbf{Pr}[|X - nH_n| \geq cn] \leq \frac{\mathbf{Var}[X]}{a^2} \leq \frac{\pi^2}{6c^2}.$$

Then

$$\mathbf{Pr}[X - nH_n \geq cn] \leq \mathbf{Pr}[|X - nH_n| \geq cn] \leq \frac{\pi^2}{6c^2}.$$

The use of Chebyshev's inequality is often referred as the "second-moment method" as it uses the variance of the random variable.

3 Threshold Behavior of Random Graph

Another application of Chebyshev's inequality is to establish the threshold behavior of Erdős-Rényi random graphs $G_{n,p}$.

The notation $G_{n,p}$ specify a distribution over all simple undirected graphs with n vertices. In the model, each of the $\binom{n}{2}$ possible edges exists with probability p independently. Therefore, the expected number of edges in the graph is $\binom{n}{2}p$ and each vertex has expected degree $(n-1)p$

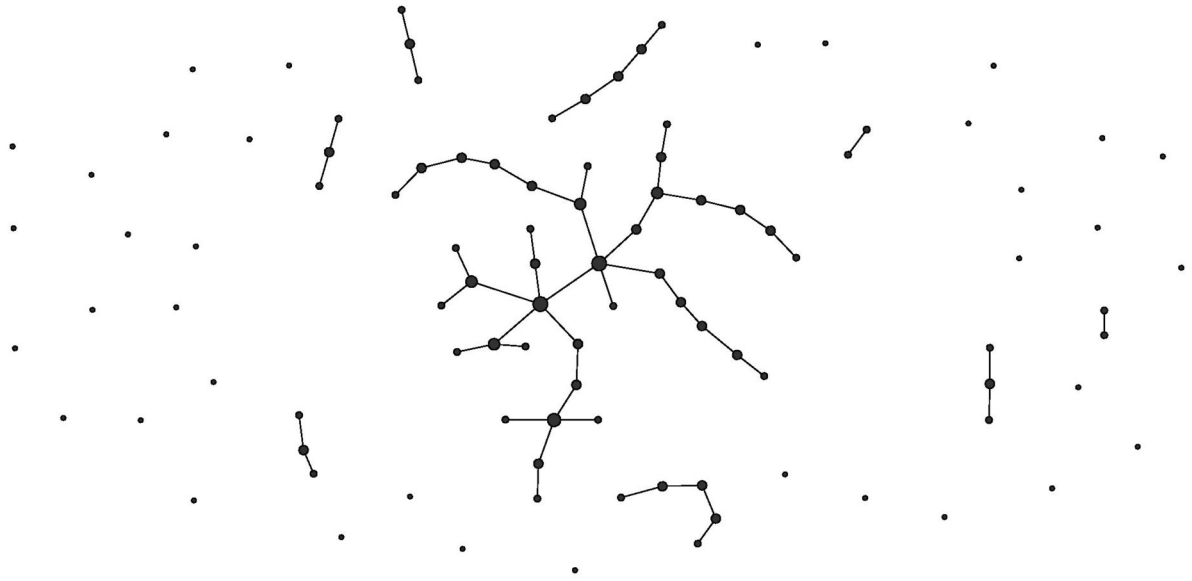

Figure 1: A graph generated with $p = 0.01$ [1]

Among many other interesting properties, random graphs establish the so-called "threshold behavior" for certain graph properties. That is, in the model $G_{n,p}$ it is often the case that there is a threshold function r such that: (a) when p is just less than the $r(n)$, almost no graph satisfies the desired property; (b) when p is just larger than $r(n)$, almost every graph has the desired property. Formally, we have

Definition 4 (Threshold function). *Given a graph property P, define its threshold function $r(n)$ as:*

- *if $p \ll r(n), G \sim G_{n,p}$ does not satisfy P w.h.p.;*
- *if $p \gg r(n), G \sim G_{n,p}$ satisfies P w.h.p.*

Now we will show the property P = "G contains a 4-clique" has the threshold function $n^{-\frac{2}{3}}$. Our proof will apply the Markov's inequality and Chebyshev's inequality we have just learned.

Theorem 5. *The property "G contains a 4-clique" has the threshold function $n^{-\frac{2}{3}}$.*

Proof. We need to verify the desired property for $p \ll n^{-\frac{2}{3}}$ and $p \gg n^{-\frac{2}{3}}$. The former case is easier.

For every $S \in \binom{[n]}{4}$, let X_S be the corresponding indicator variable, i.e.

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

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

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

If we consider the case $p \ll n^{\frac{-2}{3}}$, then $\mathbf{E}[X] = o(1)$. Since X is a nonnegative random variable, it follows by Markov inequality

$$\mathbf{Pr}[X \geq 1] \leq \frac{\mathbf{E}[X]}{1} = o(1).$$

Hence given any $\varepsilon > 0$, the probability that G has a clique 4 is less than ε for sufficiently large n.

However, we could not use the same argument to prove the $p \gg n^{-\frac{2}{3}}$ case. This is because the information of expectation does not imply w.h.p. results in general. 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 in all is large. Therefore, we have to look at the variance. First notice that

$$\mathbf{Pr}[X = 0] \leq \mathbf{Pr}[|X - \mathbf{E}[X]| \geq \mathbf{E}[X]]$$

By Chebyshev's Inequality we have

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

We only need to give bound on $\mathbf{Var}[X]$,

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

When $|S \cap T| = 2$ (the corresponding cliques share one edge), the probability that S, T are both 4-cique is p^{11} (since there are only 11 edges), hence we have

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

For $|S \cap T| = 3$ (the corresponding cliques share two edges), similarly the probability that S, T are both 4-clique is p^9, and it holds that

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

Now consider the last summation, just use the result above, we have

$$\sum_S \left(\mathbf{E}\left[X_S^2\right] - \mathbf{E}[X_S]^2\right) \le \sum_S \left(\mathbf{E}\left[X_S^2\right]\right) \le n^4 p^6.$$

To sum up, since $p \ll n^{-\frac{2}{3}}$, we have

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

Finally, we get

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

Thus P is not satisfied w.h.p. □

References

[1] *Erdős-Rényi model.* https://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model. Accessed Sep 25, 2020. 5

[2] *Las Vegas algorithm.* https://en.wikipedia.org/wiki/Las_Vegas_algorithm. Accessed Sep 25, 2020. 3

[3] *Stirling's approximation.* https://en.wikipedia.org/wiki/Stirling%27s_approximation. Accessed Sep 25, 2020. 2

[4] G. H. Gonnet, *Expected length of the longest probe sequence in hash code searching*, Journal of the ACM (JACM), 28 (1981), pp. 289–304. 2

[5] M. Raab and A. Steger, *Balls into bins - A simple and tight analysis*, in International Workshop on Randomization and Approximation Techniques in Computer Science, Springer, 1998, pp. 159–170. 2