Mathematical Toolkit **Autumn 2015**

Lecture 12: November 11, 2015

Lecturer: Madhur Tulsiani

1 Coupon Collection

Consider the following problem: There are n kinds of items/coupons and at each time step we get one coupon chosen to be from one of the n types at random. All types are equally likely at each step and the choices at different time steps are independent. We define a random variable, T which is the time when we first have all the n types of coupons. Find $\mathbb{E}[T]$.

We can make the following claim:

$$T = \sum_{i=1}^{n} X_i$$

Where, X_i is the time to get from the $i-1$ to the i types of coupons. Thus we have,

$$\mathbb{E}[T] = \sum_{i} \mathbb{E}[X_i]$$

Note that X_i is a geometric random variable with parameter $\frac{n-i+1}{n}$, since if we have $i-1$ type of coupons, X_i represents the time till we receive a coupon belonging to any one of the remaining $n-i+1$ types. Thus,

$$\mathbb{E}[X_i] \;=\; \frac{n}{n-i+1}\,.$$

Therefore,

$$\mathbb{E}[T] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n \cdot H(n)$$

where $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ is the n^{th} *harmonic number*. It is known (see Wikipedia for example) that $H_n = \ln n + \Theta(1)$. Thus, we have that $\mathbb{E}[T] = n \ln n + \Theta(n)$.

2 A randomized algorithm for Max 3-SAT

Recall that a 3-SAT formula φ is of the form

$$\varphi \;\equiv\; C_1 \wedge \cdots \wedge C_m\,,$$

where each C_i is a clause of the form $C_i = (l_{i_1} \vee l_{i_2} \vee l_{i_3})$ and each l_{i_j} is in turn x_{i_j} or its negation $\overline{x}_{i_j}$. We assume that each clause contains three *distinct* variables.

In the problem Max 3-SAT, the goal is not necessarily to satisfy all the clauses, but rather find an assignment to the variables which satisfies as many clauses as possible. We show that for any formula φ with m clauses, one can find an assignment satisfying $7m/8$ clauses.

Consider assigning each of the variables $x_1, \ldots, x_n$ a value in $\{0, 1\}$ independently at random. Let Z be a random variable equal to the number of clauses satisfied by the random assignment. We can write

$$Z \;=\; Y_1 + \cdots + Y_m \,,$$

where Y_i if the clause C_i is satisfied and 0 otherwise. By linearity of expectation $\mathbb{E}[Z] = \sum_{i=1}^{m} \mathbb{E}[Y_i]$. Note $C_i = (l_{i_1} \vee l_{i_2} \vee l_{i_3})$ is not satisfied if and only if $l_{i_1} = l_{i_2} = l_{i_3} = 0$ which happens with probability $1/8$ since the three literals correspond to three distict variables, which are assigned values 0 and 1 independently with probability $1/2$ each. Thus, $\mathbb{P}[Y_i = 0] = 1/8$, which gives

$$\mathbb{E}[Z] \;=\; \sum_{i=1}^{m} \mathbb{E}[Y_i] \;=\; \sum_{i=1}^{m} \left(1 - \frac{1}{8}\right) \;=\; \frac{7m}{8} \,.$$

Thus, there *exists* an assignment which satisfies at least $7m/8$ clauses. We now argue that it can be found efficiently. Note that

$$\mathbb{E}[Z] \;=\; \frac{1}{2} \cdot \mathbb{E}[Z \mid x_1 = 0] + \frac{1}{2} \cdot \mathbb{E}[Z \mid x_1 = 1] \,.$$

Thus, at least one of the expectations on the right hand side must be at least $7m/8$. Since each of these expectations can be computed efficienly (prove it!), we can find a value $b_1 \in \{0, 1\}$ such that

$$\mathbb{E}[Z \mid x_1 = b_1] \;\geq\; \frac{7m}{8} \,.$$

Continuing similarly by induction, we can find $b_1, \ldots, b_n$ such that

$$\mathbb{E}[Z \mid x_1 = b_1, \ldots, x_n = b_n] \;\geq\; \frac{7m}{8} \,.$$

Since Z is fixed given the values of all the variables, we get that the assignment $(b_1, \ldots, b_n)$ satisfies at least $7m/8$ clauses.

3 The Probabilistic Method: Independent Sets

Now we do one more application of expectations which is often called the *Probabilistic Method*. It is often used to show the existence of objects with certain properties without necessarily constructing them. In the previous section we used probabilistic reasoning to show that a cut exists, but then later also showed how to find such a cut.

Consider a graph $G = (V, E)$. Now, we want to define an independent set $S \subseteq V$, such that 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 $\deg(i) = |N(i)|$.

Theorem 3.1 *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 we will come back to it in a bit. We first prove the theorem using a different idea.

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. ■

We can now go back to our earlier idea. The problem here is: No matter what p is, we might end up picking up an edge, but we can always delete these edges (i.e., remove both vertices in the edge from our set) and obtain an independent set. This method is called the Method of Alterations (since we *alter* and object to make it satisfy the desired properties). Let T be a set obtained by picking each vertes individually with probability p and let S be the set obtained by deleting both endpoints of each edge contained in T. We have

$$\mathbb{E}[|S|] \ = \ \mathbb{E}[|T| - 2 \cdot (\text{number of edges contained in T})] \ .$$

We can use linearity to compute the above expectation. We have that $ex|S| = p \cdot n$. Also, each edge is deleted if and only if both its vertices are in S. This happens with probability p^2. Thus,

$$\mathbb{E}[|S|] \ = \ p \cdot n - 2 \cdot p^2 \cdot |E| \, .$$

Suppose the maximum degree is d, then the number of edges $|E| \leq \frac{dn}{2}$, thus we have:

$$\mathbb{E}[|S|] \ \geq \ p \cdot n - p^2 \cdot nd$$

By chosing $p = \frac{1}{2d}$, we get

$$\mathbb{E}[|S|] \ \geq \ \frac{n}{2d} - \frac{n}{4d} = \frac{n}{4d}$$

And thus, we still get that there exists an independent set with size at least $\frac{n}{4d}$.

4 Inequalities

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

4.1 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 4.1 (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 writing

$$\mathbb{E}[Z] \;=\; \mathbb{E}\left[Z \cdot \left(\mathbb{1}_{\{Z \geq t\}} + \mathbb{1}_{\{Z<t\}}\right)\right],$$

where $\mathbb{1}_{\{Z \geq t\}}$ denotes a function which is 1 when $Z \geq t$ and 0 otherwise (similarly for $\mathbb{1}_{\{Z<t\}}$). Using non-negativity of Z, we get

$$\mathbb{E}[Z] \;=\; \mathbb{E}\left[Z \cdot \left(\mathbb{1}_{\{Z \geq t\}} + \mathbb{1}_{\{Z<t\}}\right)\right] \;\geq\; \mathbb{E}\left[Z \cdot \mathbb{1}_{\{Z \geq t\}}\right] \;\geq\; \mathbb{E}\left[t \cdot \mathbb{1}_{\{Z \geq t\}}\right] \;=\; t \cdot \mathbb{P}[Z \geq t],$$

which completes the proof. ■

4.2 Chebyshev's Inequality

Recall that the variance of random variable X is defined as

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

Lemma 4.2 (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

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

■

5 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.

5.1 Application of Markov's inequality

Using Markov's inequality we have,

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

5.2 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,

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

Similarly,

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

So we have,

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

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

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

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

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

Setting $t = n/2$ 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 $\sqrt{\mathsf{Var}[Z]}$, a quantity which is referred to as the *standard deviation* of the random variable Z.