PRINCETON UNIVERSITY F'02 COS 597D: A THEORIST'S TOOLKIT

Lecture 1: Probabilistic Arguments

Lecturer: *Sanjeev Arora* Scribe: *Tony Wirth*

1 Introduction

Think of the topics in this course as a toolbox to rely upon during your research career. Today's lecture shows the application of simple probabilistic arguments to prove seemingly difficult results. Alon and Spencer's text goes into considerably more detail on this topic, as will Sudakov's course in the Math dept.

A random variable is a mapping from a probability space to $\mathbf{R}$. To give an example, the probability space could be that of all possible outcomes of n tosses of a fair coin, and X_i is the random variable that is 1 if the ith toss is a head, and is 0 otherwise.

Let $X_1, X_2, X_3, \ldots, X_n$ be a sequence of random variables. The first observation we make is that of the **Linearity of Expectation**, viz.

$$\mathbf{E}[\sum_i X_i] = \sum_i \mathbf{E}[X_i]$$

It is important to realize that linearity holds *regardless* of the whether or not the random variables are independent.

Can we say something about $\mathbf{E}[X_1 X_2]$? In general, nothing much but if X_1, X_2 are independent events (formally, this means that for all a, b $\mathbf{Pr}[X_1 = a, X_2 = b] = \mathbf{Pr}[X_1 = a]\mathbf{Pr}[X_2 = b]$) then $\mathbf{E}[X_1 X_2] = \mathbf{E}[X_1]\mathbf{E}[X_2]$.

The first of a number of inequalities presented today, **Markov's inequality** says that any *non-negative* random variable X satisfies

$$\mathbf{Pr}\left(X \geq k\mathbf{E}[X]\right) \leq \frac{1}{k}.$$

Note that this is just another way to write the trivial observation that $\mathbf{E}[X] \geq k \cdot \mathbf{Pr}[X \geq k]$.

Sometimes we refer to the application of Markov's inequality as an *averaging argument*.

Can we give any meaningful upperbound on $\mathbf{Pr}[X < c \cdot \mathbf{E}[X]]$ where $c < 1$, in other words the probability that X is a lot less than its expectation? In general we cannot. However, if we know an upperbound on X then we can. For example, if $X \in [0, 1]$ and $\mathbf{E}[X] = \mu$ then for any $c < 1$ we have (simple exercise)

$$\mathbf{Pr}[X \leq c\mu] \leq \frac{1 - \mu}{1 - c\mu}.$$

Sometimes this is also called an averaging argument.

EXAMPLE 1 Suppose you took a lot of exams, each scored from 1 to 100. If your average score was 90 then in at least half the exams you scored at least 80.

2 Independent sets in random graphs

We illustrate the power of the averaging argument by studying the size of the largest independent set in random graphs. Recall that an independent set is a collection of nodes between every pair of which there is no edge—an anti-clique, if you like. It is **NP**-hard to determine the size of the largest independent set in a graph in general.

Let $G(n, \frac{1}{2})$ stand for the distribution on graphs with n vertices in which the probability that each edge is included in the graph is $1/2$. In fact, $G(n, \frac{1}{2})$ is the uniform distribution on graphs with n nodes (verify this!). Thus one of the reasons for studying random graphs is to study the performance of some algorithms "on the average graph."

What is the size of the largest independent set in a random graph?

THEOREM 1
The probability that a graph drawn from $G(n, \frac{1}{2})$ has an independent set of size greater than $2\lceil \log n \rceil$ is tiny.

PROOF: For all subsets S of $\{1, 2, \ldots, n\}$, let X_S be an indicator random variable for S being an independent set. Now, let r.v. Y_k be the number of independent sets of size k; that is,

$$Y_k = \sum_{S:|S|=k} X_S.$$

Linearity of expectation tells us that

$$\mathbf{E}[Y_k] = \sum_{S:|S|=k} \mathbf{E}[X_S].$$

The X_S r.v.s are indicators and there are $\binom{|S|}{2}$ potential edges between $|S|$ vertices, hence

$$\mathbf{E}[X_S] = \mathbf{Pr}[X_S = 1] = \frac{1}{2^{\binom{|S|}{2}}},$$

Applying both the approximations

$$\binom{n}{k} \approx \left(\frac{ne}{k}\right)^k \quad \text{and} \quad \binom{n}{2} \approx \frac{n^2}{2}, \tag{1}$$

we can show that

$$\begin{aligned} \mathbf{E}[Y_k] &= \binom{n}{k} \frac{1}{2^{\binom{|S|}{2}}} \\ &= \left[\frac{ne}{k}\right]^k \frac{1}{2^{k^2/2}} \\ &= \left[\frac{ne}{k2^{k/2}}\right]^k \\ &= \left[\frac{e}{2\log n}\right]^{2\log n}, \end{aligned} \tag{2}$$

substituting $2\log n$ for k. Hence the mean value of Y_k tends to zero very rapidly.

In particular, Markov's inequality tells us that $\mathbf{Pr}[Y_k \geq 1] \leq \mathbf{E}[Y_k]$ and hence that the probability of an independent set of size $2\lceil \log n \rceil$ is *tiny.* □

Since the distribution $G(n, 1/2)$ picks each edge with probability $1/2$, it also leaves out each edge with probability $1/2$. Thus we can simultaneously bound the maximum size of a clique by $2\lceil \log n \rceil$.

Now we also show a lowerbound on the size of the largest independent set/clique. Note that substituting $k = 2\lceil \log n \rceil - \sqrt{\log n}$ in the above calculation for $\mathbf{E}[Y_k]$ shows that the expected number of independent sets of size $2\lceil \log n \rceil - \sqrt{\log n}$ is very large. However, we have to rule out the possibility that this number is large with probability ≈ 0.

For this we need a more powerful inequality, **Chebyshev's inequality**, which says

$$\mathbf{Pr}[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2},$$

where μ and σ^2 are the mean and variance of X. Recall that $\sigma^2 = \mathbf{E}[(X-\mu)^2] = \mathbf{E}[X^2] - \mu^2$. Actually, Chebyshev's inequality is just a special case of Markov's inequality: by definition,

$$\mathbf{E}\left[|X - \mu|^2\right] = \sigma^2,$$

and so,

$$\mathbf{Pr}\left[|X - \mu|^2 \geq k^2\sigma^2\right] \leq \frac{1}{k^2}.$$

THEOREM 2
If G is drawn from $G(n, \frac{1}{2})$ there is almost surely *an independent set of size $2\lceil \log n \rceil - \sqrt{\log n}$.*

PROOF: If $k = 2\log n - \sqrt{\log n}$, then by substituting into formula (2) we find that $\mathbf{E}[Y_k]$ is approximately

$$\left(\frac{ne}{2\log n \frac{n}{2^{\sqrt{\log n}/2}}}\right)^k \geq \left(\frac{2^{\sqrt{\log n}/2}}{2\log n}\right)^{\log n}.$$

This quantity tends to infinity, so the mean is *much* greater than one. However, the variance could be so large that the value of the mean tells us nothing.

The calculation of the variance is however quite messy. Let N stand for $\binom{n}{k}$ and p stand for $2^{-\binom{k}{2}}$. The expected value of Y_k is Np. We show that $\mathbf{E}[Y_k^2]$ is $N^2p^2 + \varepsilon$, where $\varepsilon \ll N^2p^2$, so that $\mathbf{Var}[Y_k] = \varepsilon$ is smaller than N^2p^2. Then concentration about the mean follows from Chebyshev's inequality. Observe that

$$\mathbf{E}[Y_k^2] = \mathbf{E}[(\sum_{S:|S|=k} X_s)^2] = \sum_{S,T:|S|=|T|=k} \mathbf{E}[X_S X_T],$$

by linearity of expectation.

Note that if $|S \cap T| \leq 1$ then X_S and X_T are independent events and $\mathbf{E}[X_S X_T] = \mathbf{E}[X_S]\mathbf{E}[X_T] = p^2$. Now we calculate the fraction of S, T that satisfy $|S \cap T| = i > 1$. There are $\binom{n}{k}$ ways to choose S and then $\binom{k}{i}\binom{n-k}{k-i}$ ways to choose T. Then the probability

that both are independent sets is $2^{-2\binom{k}{2}}2^{\binom{i}{2}} = p^2 2^{\binom{i}{2}}$. Since $X_S X_T = 1$ iff both S and T are independent sets and 0 otherwise, we have

$$\mathbf{E}[Y_k^2] = \binom{n}{k} p^2 \left(\binom{n-k}{k} + \binom{k}{1}\binom{n-k-1}{k-1} + \sum_{i \geq 2} \binom{k}{i}\binom{n-k}{k-i} 2^{\binom{i}{2}} \right).$$

This can be shown to be $N^2p^2 + \varepsilon$ for some $\varepsilon \ll N^2p^2$. □

In fact, in 1976 Bollobás and Erdős showed that for every n there is a number $k(n)$ (presumably around $2\log n$) such that the distribution of the size of the maximum independent set is very tightly concentrated on $k(n)$ and $k(n)+1$.

Algorithms: Somebody asked if we can find a clique or independent set of size $2\lceil \log n \rceil$ in polynomial time on such graphs. We describe a simple greedy polynomial time algorithm that finds a clique of size $\lceil \log n \rceil$ in a random graph with high probability. The greedy process is simply this: select a vertex v_1 at random, and then look at its set of neighbours, V_1. Next, Select a vertex v_2 in V_1 and restrict attention to neighbours of v_1 and v_2; call this set V_2. We continue this process until we reach a V_k that is empty. Thus $\{v_1, v_2, \ldots, v_k\}$ is a maximal clique.

Using the fact that in a random graph, all degrees are concentrated in $[\frac{n}{2} - O(\sqrt{n}\log n), \frac{n}{2} + O(\sqrt{n}\log n)]$ (see Chernoff bounds below), we can show that $k \geq \log n - O(1)$ with high probability.

Open Problem 1 Is it possible to find a clique of size $(1+\varepsilon)\log n$ in polynomial time? (The best algorithm we know of does exhaustive listing of sets of size $(1+\varepsilon)\log n$, which takes about $n^{(1+\varepsilon)\log n}$ time.

3 Graph colouring: local versus global

Imagine that one day you pick up a political map of some world and decide to colour in the countries. To keep the boundaries clear you decide to fill in adjacent countries with different colours. It turns out that no matter what map anyone gives you (a map is just a so-called *planar graph*), four colours will always suffice.

We can extend this notion of colouring to graphs: we assign each vertex a colour so that no two vertices with an edge between them share a colour. More formally, a ***k*-colouring** of a graph $G = (V, E)$ is a family of k independent sets whose union is V. Note that we can ask that these sets form a *partition* of V, but we don't have to.

Following convention, let $\alpha(G)$ stand for the size of the largest independent set in G and let $\chi(G)$ stand for the minimum k for which G admits a k-colouring. The formal definition of colouring tells us that

$$\chi(G) \geq \frac{n}{\alpha(G)}.$$

Can we provide an upper bound for the **chromatic number**, $\chi(G)$, using $\alpha(G)$, for instance $2n/\alpha(G)$? Such a bound is impossible: there exist graphs in which there is a large independent set, but the remainder of the graph is a clique. However, if the vertex degrees are bounded, then so is the chromatic number.

THEOREM 3
If the maximum degree of a graph is d then $\chi(G) \leq d+1$.

PROOF: Proof by induction: The base case is trivially true. For the inductive step, assume all graphs with n nodes satisfy the theorem. Given a graph G with $n+1$ nodes, identify one vertex v. The induced subgraph $G - \{v\}$ can be coloured with $d+1$ colours, by the induction assumption. The node v has links to at most d other nodes, due to the maximum degree constraint, and hence is adjacent to at most d colours. Hence the $d+1$th colour is available for v and so G can be $d+1$-coloured. □

Perhaps being able to colour small subgraphs with few colours might help us to colour larger graphs? Such conjectures can waste a lot of brain-cycles. But a simple probabilistic argument —doable on the back of an envelope— can help us dispose of such conjectures, as in the following 1962 result by Erdős. It shows that the chromatic number is truly a global property and cannot be deduced from looking locally.

THEOREM 4
For all k there exists a positive ε such that for all sufficiently large n there is a graph G on n vertices with $\chi(G) > k$, but every subgraph of G with at most εn vertices is 3-colourable.

PROOF: Let G be a graph selected from $G(n,p)$ where $p = c/n$. We show not only that there exists such a G that satisfies the theorem statement, but that selected in this way G satisfies it almost surely.

First, we show that with high probability $\alpha(G) < n/k$, which implies $\chi(G) > k$. We can approximate $\binom{n}{an}$ with $2^{H(a)n}$, where $H(a)$ is the *entropy* function

$$a \log\left(\frac{1}{a}\right) + (1-a)\log\left(\frac{1}{1-a}\right).$$

Using the fact that when p is small, $1 - p \approx e^{-p}$, we find that the expected number of independent sets of size n/k is

$$\binom{n}{n/k}(1-p)^{\binom{n/k}{2}} \approx 2^{nH(1/k)} e^{-pn^2/2k^2} = \exp\left[nH\left(\frac{1}{k}\right)(\ln 2) - \frac{cn}{k^2}\right].$$

If c is at least $2k^2 H(1/k)\ln 2$, this expectation drops to zero rapidly.

Second, we show that in every induced subgraph on at most εn nodes, the average degree is less than 3. Note that if there exists a subgraph on εn vertices that is not 3-colourable, then there is a minimal such subgraph. In the minimal such subgraph, every vertex must have degree at least 3 (Proof: Suppose a vertex has degree 2 yet omitting it gives a 3-colorable subgraph. Then putting the vertex back in, we can extend the 3-coloring to that vertex by giving it a color that has not been assigned to its 2 neighbors.) So we have a subgraph on $s \leq \varepsilon n$ vertices with at least $3s/2$ edges. The probability of such a subgraph existing is at most

$$\sum_{s \leq \varepsilon n} \binom{n}{s}\binom{\binom{s}{2}}{3s/2}\left(\frac{c}{n}\right)^{3s/2},$$

recalling that $p = c/n$. If s is $O(1)$, the terms tend to zero rapidly and the sum is negligible. Otherwise, we can use the approximations presented in line (1) to arrive at the quantity

$$\sum_{s \le \varepsilon n} \left[\frac{ne}{s} \left(\frac{se}{3} \right)^{3/2} \left(\frac{c}{n} \right)^{3/2} \right]^s \le \left[e^{5/2} 3^{-3/2} c^{3/2} \sqrt{s/n} \right]^s .$$

If ε is less than $e^{-5} 3^3 c^{-3}$, the summation terms are strictly less than 1, since $s \le \varepsilon n$. On the other hand we have eliminated the cases where s is $O(1)$ so the remainder of the terms form a geometric series and so their sum is bounded by a constant times the first term, which is tiny.

□

4 Bounding distribution tails

When we toss a coin many times, the expected number of heads is half the number of tosses. How tightly is this distribution concentrated? Should we be very surprised if after 1000 tosses we have 625 heads? We can provide bounds on how likely a sum of Poisson trials is to deviate from its mean. (A sequence of $\{0, 1\}$ random variables is Poisson, as against Bernoulli, if the expected values can vary between trials.) The **Chernoff bound** presented here was probably known before Chernoff published it in 1952.

THEOREM 5
Let $X_1, X_2, \ldots, X_n$ be independent Poisson trials and let $p_i = \mathbf{E}[X_i]$, where $0 < p_i < 1$. Then the sum $X = \sum_{i=1}^n X_i$, which has mean $\mu = \sum_{i=1}^n p_i$, satisfies

$$\mathbf{Pr}[X \ge (1+\delta)\mu] \le \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}} \right]^\mu .$$

PROOF: Surprisingly, this inequality also is proved using the Markov inequality.

We introduce a positive dummy variable t and observe that

$$\mathbf{E}[\exp(tX)] = \mathbf{E}[\exp(t \sum_i X_i)] = \mathbf{E}[\prod_i \exp(tX_i)] = \prod_i \mathbf{E}[\exp(tX_i)], \tag{3}$$

where the last equality holds because the X_i r.v.s are independent. Now,

$$\mathbf{E}[\exp(tX_i)] = (1 - p_i) + p_i e^t,$$

therefore,

$$\begin{aligned} \prod_i \mathbf{E}[\exp(tX_i)] = \prod_i [1 + p_i(e^t - 1)] &\le \prod_i \exp(p_i(e^t - 1)) \\ &= \exp(\sum_i p_i(e^t - 1)) = \exp(\mu(e^t - 1)), \end{aligned} \tag{4}$$

as $1 + x \le e^x$. Finally, apply Markov's inequality to the random variable $\exp(tX)$, viz.

$$\mathbf{Pr}[X \ge (1+\delta)\mu] = \mathbf{Pr}[\exp(tX) \ge \exp(t(1+\delta)\mu)] \le \frac{\mathbf{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} = \frac{\exp((e^t - 1)\mu)}{\exp(t(1+\delta)\mu)},$$

using lines (3) and (4) and the fact that t is positive. Since t is a dummy variable, we can choose any positive value we like for it. The right hand size is minimized if $t = \ln(1+\delta)$—just differentiate—and this leads to the theorem statement. □

A similar technique yields this result for the lower tail of the distribution

$$\mathbf{Pr}[X \leq (1-\delta)\mu] \leq \left[\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right]^{\mu}$$

By the way, if all n coin tosses are fair (Heads has probability 1/2) then the the probability of seeing N heads where $|N - n/2| > a\sqrt{n}$ is at most $e^{-a^2/2}$. The chance of seeing at least 625 heads in 1000 tosses of an unbiased coin is less than 5.3×10^{-7}.

§1 Prove that the greedy algorithm finds a clique of size $\log n - O(1)$ with high probability in a random graph.