Lecture 1

Various notations are used when comparing the rates of growth of different functions, and it is a good idea for us to get these out of the way before we start.

NOTATION : Let $f, g : \mathbf{N} \rightarrow \mathbf{R}$ be any functions. We write

(i) $f = O(g)$ if the quotient $|f(n)/g(n)|$ is bounded as $n \rightarrow \infty$.
(ii) $f = \Omega(g)$ if $g = O(f)$.
(iii) $f = o(g)$ if $f(n)/g(n) \rightarrow 0$ as $n \rightarrow \infty$.
(iv) $f \sim g$ if $f(n)/g(n) \rightarrow 1$ as $n \rightarrow \infty$.
(v) $f \preceq g$ if $\limsup |f(n)/g(n)| \leq 1$.
(vi) $f \succeq g$ if $g \preceq f$.

First, some general words of wisdom (or waffle) :

The basic application of probabilisitc techniques to combinatorics is to prove existence of a structure from amongst a certain class $\mathcal{X}$ of structures, which possesses some desired property $\mathcal{P}$.

One does so by introducing some appropriate probability measure μ on the collection $\mathcal{X}$ and showing (somehow) that, if one chooses at random, according to the distribution μ, an element of $\mathcal{X}$, then with positive probability one's choice possesses the property $\mathcal{P}$.

An important remark :

Usually, though not always, μ is just a simple uniform distribution. Also, since we're interested in combinatorial applications, $\mathcal{X}$ is usually (though not always) a finite collection[1] This means that

(i) there is usually no great mystery about how the probability theory is introduced to the problem under consideration. It is intuitively clear what is meant by 'choosing at random' and one doesn't need to be an expert in probability theory to understand what's going on.

[1] We will see some applications, for example in number theory, where $\mathcal{X}$ is infinite. But even here, the underlying set of interest, namely the natural numbers, is discrete.

(ii) also, since the sets under consideration are usually finite, one can in principle present most of the same arguments without ever mentioning probability theory at all, i.e.: by 'purely combinatorial' reasoning. Though this is the case, for more sophisticated applications, the advantages of using notions of probability in terms of the clarity of exposition outweigh the disadvantages of having to learn these notions.

Note that the probabilistic method is usually employed to show that some desired structure exists. It doesn't usually tell you how to actually find such a structure. This is an *algorithmic* problem, but obviously for real-world applications, one can conceive that it might be essential to actually be able to find what one is looking for. Sometimes the probabilistic method gives a good *randomized algorithm*, basically an algotirhm that is fast but has a certain probability of failure[2].

Intuitively, it is clear how this would work. One shows that a structure with property $\mathcal{P}$ exists by showing that if one chooses at random, then one finds something with property $\mathcal{P}$ with probability $\epsilon > 0$. Often it turns out that the proof yields a value of ϵ which is close to 1. This means that a random choice is very likely to be a good one.

The course is roughly divided into three parts :

I. Introduction to the basics of the probabilistic method by means of a variety of examples.
II. Some more sophisticated proabilistic techniques, in particular so-called *concentration inequalities.*
III. To be decided (depends on time considerations etc.).

We will discuss applications of the method to a variety of mathematical problems, for example in graph theory, Ramsey theory, number theory, discrepancy theory ...

Example 1 : Ramsey Numbers

Definition 1 : The *complete graph* on n vertices, denoted K_n, is the graph in which each pair of vertices is joined by a single edge. Thus K_n contains

[2] There is also a whole theory of *derandomization*, which deals with how to turn fast randomizsed algorithms into decent deterministic ones. We will not discuss this topic in our course. There is, however, a chapter devoted to it in the book of Alon and Spencer

$\begin{pmatrix} n \\ 2 \end{pmatrix}$ edges.

I will now state and prove an abridged form of what has become known as *Ramsey's theorem.* It is abridged in the sense that, in its' full generality, the number of colors in the statement below can be any finite number, not just two.

Theorem 1 *Let $k, l \geq 2$ be fixed positive integers. Then for all sufficiently large positive integers n (how large depends on k, l), the following holds :*

If each edge of K_n is colored either red or blue, then there must exist either a red K_k or a blue K_l.

Before proving this, we introduce some notation :

NOTATION : We denote by $R(k,l)$ the smallest integer n for which the above statement holds. It is called the *(k,l)-th Ramsey number.* Theorem 1 states that these numbers exist, for every $k, l \geq 2$.

PROOF OF THEOREM : We present the standard argument, which is basically an induction on $k + l$.

Step 0 : Note that $R(k,l) = R(l,k)$ by symmetry.

Step 1 : Observe that $R(2,l) = l$ since a K_2 is just a graph with a single edge, so if we're to avoid a red K_2 then we must color every edge of our graph blue. And then we'll have a blue K_l as soon as we have l or more vertices.

Step 2 : The general induction step involves verifying the following inequality :

$$R(k,l) \leq R(k,l-1) + R(k-1,l). \tag{1}$$

So we assume the two Ramsey numbers on the right hand side of (1) exist and consider a 2-coloring of the graph K_n, where $n = R(k-1,l)+R(k,l-1)$. We must prove the existence of either a red K_k or a blue K_l. Pick any one of the n vertices and give it a name, say v. Now v is joined by an edge to

$n - 1$ other vertices. Since

$$n - 1 > [R(k-1, l) - 1] + [R(k, l-1) - 1],$$

one of the following must occur :

(i) v is joined to at least $R(k-1, l)$ vertices by a red edge, or
(ii) v is joined to at least $R(k, l-1)$ vertices by a blue edge.

Suppose (i) occurs. By definition of the Ramsey numbers, amongst the vertices joined to v by a red edge, there must exist either a red K_{k-1} or a blue K_l. In the latter case we're done already. In the former case, adding on the vertex v gives a red K_k, and again we're done.

If instead (ii) holds, then the argument is similar. It is left to the reader to write out the details.

Corollary 2 *For every $k, l \geq 2$ we have that*

$$R(k, l) \leq \binom{k + l - 2}{k - 1}. \tag{2}$$

Proof : This follows from (1) and the well-known Pascal identity for binomial coefficients

$$\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}.$$

The details are left as an exercise.

It is natural to consider the special case $k = l$. Then (2) becomes

$$R(k, k) \leq \binom{2k - 2}{k - 1}. \tag{3}$$

Using simply the fact that

$$\sum_{r=0}^{n} \binom{n}{r} = 2^n$$

(both sides count the number of subsets of an n-element set), it follows that

$$R(k, k) \leq 4^{k-1}. \tag{4}$$

Using Stirling's formula[3] (details left as an exercise), we can obtain a slightly better estimate, namely

$$R(k,k) \preceq \frac{4^{k-1}}{\sqrt{\pi(k-1)}}. \tag{5}$$

But the important point is that (4) and (5) both say that the Ramsey numbers $R(k,k)$ grow at worst exponentially.

Now, finally, we intorduce probabilstic ideas to the discussion, in order to show that the numbers $R(k,k)$ do, in fact, exhibit exponential growth. We do this by proving

Theorem 3 *Let $k \geq 3$. If the integer n satisfies*

$$\left(\begin{array}{c} n \\ k \end{array}\right) 2^{1-\left(\begin{array}{c} k \\ 2 \end{array}\right)} < 1, \tag{6}$$

then $R(k,k) > n$.

For the moment, let us assume the theorem and prove what we're really after, namely

Corollary 4

$$R(k,k) > 2^{k/2}. \tag{7}$$

Proof of Corollary : We must show that if $k \geq 3$ and $n = 2^{k/2}$, then (6) is satisfied. Since $\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n(n-1)\cdots(n-k+1)}{k!}$ and, in particular, $\left(\begin{array}{c} k \\ 2 \end{array}\right) = \frac{k(k-1)}{2}$, the left-hand side of (6) is thus at most

$$n^k \cdot \frac{2^{1+\frac{k}{2}}}{k! 2^{\frac{k^2}{2}}}.$$

[3]

$$n! \sim n^n e^{-n}\sqrt{2\pi n}.$$

Taking $n = 2^{k/2}$ this becomes simply $2^{1+k/2}/k!$. It is then a simple exericse to verify that $2^{1+k/2}/k! < 1$ for all $k \geq 3$.

We remark that a more careful analysis, again based on Stirling's formula and left as an exercise for the reader, shows that

$$R(k,k) \succeq \frac{1}{e\sqrt{2}} k 2^{k/2}. \tag{8}$$

But again the main point is that both (7) and (8) say that the numbers $R(k,k)$ exhibit exponential growth. Combining all our results, the essence of what we have found is expressed in the following :

$$\sqrt{2} \leq \liminf_{k\to\infty} R(k,k)^{1/k} \leq \limsup_{k\to\infty} R(k,k)^{1/k} \leq 4. \tag{9}$$

BIG Open Problem *Does*

$$\lim_{k\to\infty} R(k,k)$$

exist and, if so, what is it ?

This problem has been open for 70 years without any progress whatsoever having been made beyond (9). An even more daunting task, however, is to compute Ramsey numbers $R(k,l)$ exactly. In fact, for $k, l > 2$ only a small (finite) collection of Ramsey numbers have been computed exactly.

We conclude this lecture by proving Theorem 3 :

PROOF OF THEOREM 3 : The proof will use the following simple facts about probabilities :

(I) For any two events A and B,

$$P(A \cup B) \leq P(A) + P(B), \tag{10}$$

with equality if the events are *mutually exclusive*, i.e.: if $P(A \cap B) = 0$.
(II) If A and B are *independent* events, then

$$P(A \cap B) = P(A) \cdot P(B). \tag{11}$$

Consider now a fixed n and k, and a random 2-coloring of the graph K_n. This means that each of the $\binom{n}{2}$ edges is colored independently red or blue, each with probability 1/2. We want to estimate the probability of obtaining a monochromatic K_k. We divide this procedure up into three steps :

(i) the probability of a given K_k being entirely red is $\left[\frac{1}{2}\right]^{\binom{k}{2}}$. This follows from (11) and the fact that the probability of any particular edge being red is 1/2. Obviously, we have the same expression for the probability of a given K_k being entirely blue.

(ii) hence, the probability of a given K_k being monochromatic is $2^{1-\binom{k}{2}}$. This follows from (i) and (10), since there are two mutually exclusive ways to have monochromaticity, namely redness or blueness, and these have equal probability.
(iii) hence, the probability of there being some monochromatic K_k is at most $\binom{n}{k}\cdot 2^{1-\binom{k}{2}}$. This follows from the previous steps and (10), since there are $\binom{n}{k}$ complete k-subgraphs in K_n.

From these estimates it follows that, if (6) holds, then there is a positive probability that a randomly chosen coloring of K_n will include no monocrhomatic K_k. In other words, at least one such good coloring exists, and thus $R(k,k) > n$. This completes the proof of the theorem.

Lecture 2

Today we will discuss two more examples, this time taken from number theory. One will notice an obvious similarity between the two examples, and between the first one especially and that discussed in the first lecture. In fact, the similarities between all three examples run quite deep, though it will not be our purpose to explore this issue in any great detail. One is referred to the book [1] for a more comprehensive treatment.

Example 1 : Van der Waerden numbers

DEFINITION 1 : Let $k \geq 1$. An increasing sequence $a_1 < a_2 < \cdots < a_k$ of k integers is said to be an *arithmetic progression of length k and common difference d* if $a_{i+1} - a_i = d$ for $i = 1, ..., k-1$. We will use the abbreviation 'k-AP' to denote an arithmetic progression of length k.

The following theorem was proven by the Dutch mathematician Bartel van der Waerden in the 1920s and has been given his name :

Theorem 5 (van der Waerden's Theorem) *Let $k, m \geq 1$ be given integers. Then for all sufficiently large positive integers n (depending on k and m), the following holds :*

If the integers $1, 2, ..., n$ are colored with at most m colors, then there must exist a monochromatic k-AP.

NOTATION : The *van der Waerden number $W(k, m)$* is the least integer n for which any m-coloring of $\{1, ..., n\}$ must yield a monochromatic k-AP. The theorem states that these numbers exist.

It is beyond the scope of this course to give a fully rigorous proof of Theorem 5. A full proof, taken from [1], was handed out in class. Basically, the proof involves two nice ideas and a lot of horrible notation. The two ideas are

(i) observe that $W(2, m) = m+1$ (why ?). This allows us to get an induction started. The induction proceeds by proving that the numbers $W(k+1, m)$ exist for all m and a fixed k, assuming that the numbers $W(k, m)$ all exist.
(ii) for a fixed m and k, the proof of the existence of $W(k, m)$ in this inductive manner involves an idea which has become called *color focusing.* It is basically the same idea for all m but because the numbers involved grow so

drastically with m, it becomes something of a technical nightmare to write down the details. The idea itself is quite beautiful, though.

We will be content to illustrate the method by proving that

$$W(3,2) \leq 325. \tag{12}$$

Note that, according to the program outlined above, our proof of this should at some point use the knowledge that $W(2,m) = m+1$. I'll leave it as an amusing exercise for you to spot where this is used, since it would be easy to miss it !

So let us suppose the numbers from 1 through 325 have been colored red or blue in some manner. We must prove the existence of a monochromatic 3-AP. The first step is to divide the 325 numbers into 65 blocks $B_1, ..., B_{65}$ of 5 consecutive numbers. So $B_1 = \{1,2,3,4,5\}$, $B_2 = \{6,7,8,9,10\}$ etc.

There are $2^5 = 32$ possible ways to color any block with 2 colors. Thus, amongst the first 33 blocks, there must be two which are colored in exactly the same pattern. Pick any two such blocks, say B_i and B_{i+j}. Since $i+j \leq 33$, it follows that $i+2j \leq 65$. Hence the block B_{i+2j} exists. We now focus our attention on the three blocks B_i, B_{i+j} and B_{i+2j}.

The rest of the proof is most easily understood with the help of pictures. I am not going to draw any pictures here, so I recommend that you look at the handout from [1] at the same time as you read this proof.

Note that

$$\begin{aligned} B_i &= \{5i-4, 5i-3, 5i-2, 5i-1, 5i\}, \\ B_{i+j} &= \{5(i+j)-4, 5(i+j)-3, 5(i+j)-2, 5(i+j)-1, 5(i+j)\}, \\ B_{i+2j} &= \{5(i+2j)-4, 5(i+2j)-3, 5(i+2j)-2, 5(i+2j)-1, 5(i+2j)\}. \end{aligned}$$

Amongst the first three elements of the block B_i, at least two must get the same color. Let's suppose that $5i-4$ and $5i-2$ are both colored red and complete the proof in this case. The argument is similar in the other five cases and I leave it to yourselves to become convinced of that.

If now $5i$ was also colored red, then we'd have a red 3-AP, namely $\{5i-4, 5i-2, 5i\}$. So we may assume $5i$ is colored blue. Next, we turn to the block B_{i+j}. Since it has exactly the same color pattern as B_i, we conclude that $5(i+j)-4$ and $5(i+j)-2$ are both colored red, whereas $5(i+j)$ is colored blue.

Finally, now, we focus on B_{i+2j} and, in particular, zone in on the number $5(i+2j)$. I claim that, no matter what color we give it, we can't avoid having a monochromatic 3-AP. For if this number is colored red, then $\{5i-4, 5(i+j)-2, 5(i+2j)\}$ is a red 3-AP. But if is colored blue, then $\{5i, 5(i+j), 5(i+2j)\}$ is a blue 3-AP. This completes the proof of (12).

The bounds on Van der Waerden numbers obtained by this kind of color focusing method are eeeeeeeenoooorrrrrmooouuussss[4]. We can see that the method is not optimal even for the example of $W(3,2)$. Our method gives that $W(3,2) \leq 325$. But, in fact, $W(3,2) = 9$. To see this, first check by hand that for every partition of $\{1, 2, ..., 9\}$ into two subsets, at least one contains a 3-AP. On the other hand, we can 2-color the integers 1,...,8 so that there are no monochromatic 3-APs. For example, let 1,3,6,8 be red and 2,4,5,7 be blue.

Even the best-known upper bounds on van der Waerden numbers (obtained by quite different and, I think it is safe to say, more sophisticated methods) are really, really big. We know that[5]

$$W(k,m) \leq e^{m^{2^{2^{k+9}}}}. \tag{13}$$

We finish our discussion by instead obtaining lowwr bounds for the numbers $W(k,m)$ via a probabilistic argument.

Theorem 6

$$W(k,m) > \sqrt{2(k-1)}\, m^{(k-1)/2}. \tag{14}$$

PROOF : Left as an exercise. The special case $k = 3$ was discussed in class.

[4] more precisely, they are not *primitive recursive*, for those of you who know what that means

[5] this bound was obtained by Timothy Gowers only a few years ago as a consequence of his proof of what is known as *Szemeredi's theorem*, which is in itself a strengthening of van der Waerden's result. See [1] and the notes in German I told you about. Gowers obtained the Fields Medal for this and other work.

The gap between (13) and (14) is an important open problem in combinatorial number theory/Ramsey theory. The gap is obviously enormous. I think it is fair to say that most people believe that the lower bound (14), which gives exponential growth in k for a fixed m, is closer to the truth. But noone knows ...

Example 2 : Sum-free sets

DEFINITION 2 : A subset A of an abelian group G is said to be *sum-free* if there are no solutions in A to the equation $x = y + z$.

The abelian groups which are of most interest to number theorists are $\mathbf{Z}$ and $\mathbf{Z}_p$, where p is a prime.

Theorem 7 (Alon, Kleitman ?) *If A is any finite subset of $\mathbf{Z}$, then there exists a sum-free subset B of A such that $|B| > |A|/3$.*

PROOF : Deferred to the next lecture.

REFERENCE

[1] R. Graham, B. Rothschild and J. Spencer, *Ramsey Theory*, Wiley 1980.

Lecture 3

We now prove Theorem 7. First I will present the argument without mentioning probability. Then I will introduce some very basic general notions from probability, which we will need for later on in the course, and then present the same argument again, but 'dressed up' in this more fancy terminology.

(COMBINATORIAL) PROOF OF THEOREM 7 : Let A be given. First choose a prime number p with the following two properties[6] :

(I) $p > 2 \cdot \max_{a \in A} |a|$,
(II) $p \equiv 2 \pmod 3$.

Write $p = 3k + 2$ and let $C := \{k+1, k+2, ..., 2k+1\}$. Observe that the set C is not only sum-free, but is in fact sum-free modulo p, i.e.: there are no solutions in C to the equation $x \equiv y+z \pmod p$. The other important thing is that

$$|C| > \frac{p-1}{3}. \tag{15}$$

The crucial idea in the proof is now to count, in two different ways, the number of ordered pairs[7](x, a) of integers such that
(i) $x \in \{1, ..., p-1\}$,
(ii) $a \in A$,
(iii) $xa \pmod p \in C$.

Call the collection of such pairs $\mathcal{S}$ for convenience. On the one hand, for each fixed a, the number of x such that $(x, a) \in \mathcal{S}$ is just $|C|$. This is because, as x ranges over the integers from 1 to $p-1$, $xa \pmod p$ also does so, though perhaps in a different order. To see this, just observe that the choice of p means that p does not divide a (by (I)) and hence, for any two

[6] The existence of such a prime is guaranteed by the fact that there are infinitely many primes $p \equiv 2 \pmod 3$. This can be proven by an elementary argument exactly analogous to Euclid's basic proof that there are infinitely many primes period. Note, though, that there is a very general result, due to Dirichlet (1829), which states that if a, b are any two relatively prime integers, i.e.: $\text{GCD}(a, b) = 1$, then there exist infinitely many primes $p \equiv b \pmod a$.

[7] the trick of counting ordered pairs in two different ways is basically the idea of interchanging the order of summation in a double sum (discrete setting), or changing the order of integration (continuous setting), i.e.: Fubini's theorem.

distinct $x_1, x_2 \in \{1, ..., p-1\}$, $x_1 a$ and $x_2 a$ are also distinct and non-zero modulo p.

We thus conclude, using (15), that

$$|\mathcal{S}| = |C| \cdot |A| > (p-1) \cdot \frac{|A|}{3}. \tag{16}$$

On the other hand

$$|\mathcal{S}| = \sum_{x=1}^{p-1} B_x, \tag{17}$$

where $B_x := \{a \in A : xa \pmod{p} \in C\}$. From (16) and (17) it follows that there is at least one x for which $|B_x| > |A|/3$. But the set $\{xa : a \in B_x\}$ is sum-free, since C is. But then B_x is itself a sum-free subset of A and we're done.

Remark Alon and Kleitman showed in their paper that the constant 1/3 in the statement of Theorem 7 cannot be replaced by anything bigger than 12/29. It remains an open problem, however, as to what the largest possible constant is for which the theorem holds. Here I wish to note that the argument given above cannot be directly modified to go beyond 1/3. What one would need in order to be able to do this is to construct a subset of $\{1, ..., p-1\}$ which is larger than the set C but is still sum-free modulo p. But this is not possible. Given any prime p, there is no sum-free subset of the abelian group $\mathbf{Z}_p$ of size larger than $\frac{p+1}{3}$. This follows (exercise !) from the so-called *Cauchy-Davenport Theorem*, which states the following :

Let p be a prime and A, B be any two subsets of $\mathbf{Z}_p$. Then

$$|A + B| \geq \min\{p, |A| + |B| - 1\}.$$

Here $A + B := \{a + b : a \in A, b \in B\}$.

We now introduce some basic terminology from probability theory and then rewrite the proof of Theorem 7 in this fancy language. So for the moment it will be a bunch of definitions !

Definition 3 : Let Ω be a finite[8] set. A *probability measure* μ on Ω is a function $\mu : 2^{\Omega} \to [0, 1]$ such that

[8] as this is not meant to be a course in general probability theory, we choose to avoid the technicalities encountered when extending our concepts to sets of arbitrary cardinality.

(i) $\mu(\Omega) = 1$,

(ii) for any two disjoint subsets A, B of Ω we have that $\mu(A \cup B) = \mu(A) + \mu(B)$.

The property (ii) is called *additivity.* Note that it implies that the measure μ is completely determined by its value on singelton subsets of Ω. If the set Ω consists of n elements, then a standard notation is $\Omega := \{1, ..., n\}$ and $\mu(\{i\}) := p_i$. We think of Ω as the set of possible outcomes of some random process, and p_i is the probability of the outcome being i.

The simplest probability measure on a space is the one that assigns equal probability to each outcome, i.e.: $p_i = 1/n$ for all i in the above notation. This is called *uniform measure* and corresponds most intuitively to the notion of the outcome being 'random'.

Definition 4 : A set Ω endowed with a probability measure μ is called a *probability space* and denoted (Ω, μ). A subset of Ω is then called an *event.*

Definition 5 : Let (Ω, μ) be a probability space. A function $X : \Omega \to \mathbf{R}$ is called a *(real-valued) random variable* on Ω[9].

If $\Omega = \{1, ..., n\}$, then the standard notation is $X(i) := x_i$.

Definition 6 : With the standard notations above, if X is a random variable on the space (Ω, μ) then the *expected value/expectation* of X, denoted $E[X]$, is the quantity

$$E[X] := \sum_{i=1}^{n} x_i p_i.$$

The following trivial fact is often used :

Proposition 8 *Let X be a random variable on a space (Ω, μ). Then*

$$P(X \geq E[X]) > 0. \tag{18}$$

Proof : The proof is trivial once one understands the notation. First, it is common to write $P(\cdots)$ instead of $\mu(\cdots)$ when there can be no confusion as to what probability measure is being used. Second, the expression

[9] the term *random variable* is an unfortunate historical accident. A more accurate term would be *random function.* But the former term has become so conventional that no-one dares change it. It also explains why the letter X, rather than say f, is used to denote a random variable.

'$X \geq E[X]$' is shorthand for the event $\{w \in \Omega : X(\omega) \geq E[X]\}$. This kind of sloppy notation has become standard, so we will use it from now on without further comment.

A second property, informally referred to as *linearity of expectation*, is also simple but very useful :

Proposition 9 (Linearity of expectation) *Let $X_1, ..., X_k$ be random variables on the same probability space (Ω, μ). Then*

$$E[X_1 + \cdots + X_k] = \sum_{i=1}^{k} E[X_i]. \tag{19}$$

PROOF : Exercise. Note that the sum of RV:s on the left of (19) means just what one would expect, namely the pointwise sum of functions.

One particular class of RV:s which is especially useful in applications is the class of so-called *indicator* variables.

DEFINITION 7 : Let (Ω, μ) be a probability space and $A \subseteq \Omega$. The *indicator random variable of the event A*, denoted $\mathcal{X}_A$, is the random variable given by

$$\mathcal{X}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \notin A. \end{cases}$$

Note that it is an immediate consequence of the definition that

$$E[\mathcal{X}_A] = \mu(A). \tag{20}$$

More generally, let $f : \Omega \to \Omega$ be any function. The *indicator random variable of the event '$f \in A$'*, denoted $\mathcal{X}_{f,A}$, is the random variable given by

$$\mathcal{X}_{f,A}(\omega) = \begin{cases} 1, & \text{if } f(\omega) \in A, \\ 0, & \text{if } f(\omega) \notin A. \end{cases}$$

The analogue of (20) is then

$$E[\mathcal{X}_{f,A}] = \mu[f^{-1}(A)], \tag{21}$$

where $f^{-1}(A) = \{\omega : f(\omega) \in A\}$. Note that (20) is the special case where f is the *identity map* on Ω, i.e.: $f(\omega) = \omega \ \forall \ \omega$. Also note that if f is a 1-1 mapping and μ is uniform measure, then

$$E[\mathcal{X}_{f,A}] = \mu(A) = \frac{|A|}{|\Omega|}. \tag{22}$$

This fact wil be used in the redoing of the proof of Theorem 7, which we are now ready to present :

(PROBABILISTIC) PROOF OF THEOREM 7 : Let A be given and choose a prime p and the set C as in the first proof. We shall work in the probability space (Ω, μ) where $\Omega = \{1, 2, ..., p-1\}$ and μ is uniform measure. For each $a \in A$ let $f_a : \Omega \to \Omega$ be the map given by

$$f_a : \omega \mapsto \omega a \ (\text{mod } p).$$

As explained in the previous proof, the maps f_a are each one-to-one. Let $X_a := \mathcal{X}_{f_a,C}$. Then, by (22), for every a we have

$$E[X_a] = \frac{|C|}{p-1} > \frac{1}{3}.$$

Let $X = \sum_{a \in A} X_a$. By linearity of expectation,

$$E[X] > \frac{|A|}{3}.$$

Thus, by Proposition 8, there exists some $\omega \in \Omega$ such that $X(\omega) > |A|/3$. But, unwinding the definitions, we see that

$$X(\omega) = \#\{a \in A : \omega a \ (\text{mod } p) \in C\}.$$

This is a sum-free subset of A, by the same argument as before, so we're done !

To complete the first part of the course I have chosen three further examples from graph theory to illustrate the use of the basic probabilistic method. The arguments in our three examples will become successively more intricate, though in all cases, the amount of probability used is, in principle, no more than what is contained in Propositions 8 and 9 above.

Example 1 : Tournaments

DEFINITION 8 : A *Hamilton*[10] *path* in a graph is a path that visits every vertex exactly once.

DEFINITION 9 : A *tournament* on n players is the complete graph K_n in which every edge is given a direction.

Given a tournament T, a simple-minded way to try to rank the players is to look for a Hamilton path that respects the directions of the edges. We call such a path a Hamilton path IN the tournament.

As we all know, this doesn't usually lead to an unambiguous ranking. The simplest example which illustrates this is when $n = 3$ and the edges are directed $1 \to 2 \to 3 \to 1$. There are three different Hamilton paths in this tournament, hence no unambiguous ranking. The proof of the following result (which is also of historical interest as it is generally recognised as the first published use of a probabilistic argument in combinatorics) shows that this is in fact the usual situation :

Theorem 10 (Szele 1943) *For every $n \geq 1$ there exists a tournament on n players in which there are at least $n!/2^{n-1}$ different Hamilton paths.*

PROOF : We consider a random tournament on n players, i.e.: the probability space (Ω, μ) under consideration is

$$\Omega = \{\text{all possible } n\text{-player tournaments}\}, \quad \text{hence } |\Omega| = 2^{\left(\begin{array}{c} n \\ 2 \end{array}\right)},$$
$$\mu = \text{uniform measure.}$$

More intuitively, what this means is that the direction of each edge in K_n is chosen independently at random by tossing a fair coin. For each Hamilton path H in K_n, we let X_H be the indicator random variable of the event that H is a Hamilton path in our randomly chosen tournament. Since this event depends on the outcome of $n-1$ independent coin tosses (the path contains $n-1$ edges), we have that

$$E[X_H] = \frac{1}{2^{n-1}}.$$

[10] This is the most important name you will encounter in this course. Why ? Because he was Irish !!!!

Let $X = \sum X_H$, the sum being taken over all possible Hamilton paths in K_n. Since there are $n!$ such paths (one for each ordering of the n vertices), linearity of expectation implies that

$$E[X] = \frac{n!}{2^{n-1}}. \tag{23}$$

The theorem now follows from (18) as, unwinding the definitions, we see that the r.v. X just counts the total number of Hamilton paths in our randomly chosen tournament.

Remark What is more interesting than the theorem itself is the equation (23), since this says that 'on average' a tournament has an awful lot of Hamilton paths. This is what I meant by it being the usual situation that one cannot get an unambiguous ranking in this manner.

Exercise Show that in every tournament there is at least one Hamilton path.

Lecture 4

We continue with our examples from graph theory :

Example 2 : Turán's Theorem

Let's first go back to van der Waerden's theorem. It is natural to conjecture, but apparently much harder to prove, the following stronger result :

Theorem 11 (Szemerédi's Theorem) *Let $k \geq 3$ and $\epsilon > 0$. Then for all sufficiently large n, depending on k and ϵ, if A is any subset of $\{1, ..., n\}$ such that $|A| > \epsilon n$, then A contains a k-AP.*

This theorem was first proven by the Hungarian mathematician Endre Szemerédi in 1975, in a 50-page paper which is generally considered 'a masterpiece of combinatorial reasoning'. The theorem had been conjectured by Erdős and Turán in the 1930s already when they worked (more or less unsuccessfully) on strengthening van der Waerden's result. The special case $k = 3$ was proven by Roth in 1952 using Fourier analysis, and this work was cited when Roth received the Fields Medal in 1956.

For our present purposes, what is of interest to us is the comparison with the situation for graphs. Ramsey's theorem (for an arbitrary number of colors - we just stated it for 2 colors earlier) may be considered the analogue of van der Waerden's theorem for graphs. The analogue of Szemerédi's theorem would then be the following :

Let $k \geq 3$ and $\epsilon > 0$. For all sufficiently large n, depending on k and ϵ, if G is a graph on n vertices and with more than $\epsilon \cdot \left(\begin{array}{c} n \\ 2 \end{array} \right)$ edges, then G must contain a K_k.

It is pretty easy to see, however, that this statement is false. In fact it is already false for $k = 3$ and $\epsilon = 1/2$. For let n be any even integer. Consider a graph on n vertices in which the vertices are partitioned into two subsets of size $n/2$ and in which two vertices are joined by an edge if and only if they lie in opposite halves of the partition. Such a graph is called *bipartite*. Now G has $\frac{n}{2} \cdot \frac{n}{2} = \frac{n^2}{4}$ edges, which is more than half of $\left(\begin{array}{c} n \\ 2 \end{array} \right) = \frac{n(n-1)}{2}$. But G contains no K_3, indeed no cycle of any odd length, since any path of odd length takes one from one side of the partition to the other.

We can generalise this example to any $k \geq 3$. For simplicity suppose that n is a multiple of $k-1$. A $(k-1)$-*partite* graph on n vertices is obtained by partitioning the vertices into $k-1$ subsets of equal size, and joining two vertices by an edge if and only if they do not lie in the same part. The total number of edges in this graph is

$$\binom{k}{2} \cdot \left(\frac{n}{k-1}\right)^2 = \frac{k-2}{k-1} \cdot \frac{n^2}{2},$$

which, as k gets bigger, heads towards 100 procent of all possible edges ! But the graph has no K_k since, at the very least, a K_k contains k vertices, hence (by the pigeonhole principle) at least two would have to come from the same part of the graph. But then they are not joined to one another - contradiction !

Turán's theorem, proven in 1941, is the statement that the above examples can't be improved upon.

Theorem 12 (Turán's Theorem) *Let $k \geq 3$ and n be a multiple*[11] *of $k-1$. Then any graph with n vertices and strictly more than $\frac{k-2}{k-1} \cdot \frac{n^2}{2}$ edges contains a K_k.*

We will find it more convenient to prove an equivalent formulation of the theorem, where one replaces a graph by its *complement*, i.e.: the graph consisting of the same vertices and those edges missing from the original. We require a definition :

DEFINITION 10 : A collection of vertices in a graph are said to be *independent*, if no two amongst them are joined by an edge. The *independence number* of a graph G, denoted $\alpha(G)$, is the maximum size of an independent set of vertices in G.

The following is then equivalent to Theorem 12 :

[11] If n is not a mutiple of k then one can prove a correspoding result anyway, but I wish to avoid the associated technicalities in this presentation. If $n = (k-1)q + r$ say, where $0 < r < k-1$, then the optimal way to avoid a K_k is to take a $(k-1)$-partite graph, where r of the parts have $q+1$ vertices each and the remaining $k-1-r$ parts have q vertices each.

Theorem 12' *Let $k \geq 3$ and n be a multiple of $k-1$. Then any graph G with n vertices and strictly fewer than* $\left(\begin{array}{c} n \\ 2 \end{array} \right) - \frac{k-2}{k-1} \cdot \frac{n^2}{2} = \frac{n^2}{2(k-1)} - \frac{n}{2}$ *edges satisfies $\alpha(G) \geq k$.*

Our proof of this will require three lemmas. The probabilistic component[12] is the first (and most interesting) one, for which we need some more terminology :

DEFINITION 11 : For a vertex v in a graph G, the vertices to which it is joined by an edge are called its *neighbours*. The number of its neighbours is called the *degree* of the vertex v, and is denoted d_v. Two neighbours in a graph are also said to be *adjacent*.

Lemma 13 *For any graph G we have that*

$$\alpha(G) \geq \sum_{v \in V(G)} \frac{1}{d_v + 1}. \tag{24}$$

PROOF : Suppose G has n vertices. We consider the probability space (Ω, μ), where Ω is the collection of all possible orderings of the n vertices, hence $|\Omega| = n!$, and μ is uniform measure. For each vertex v, we let X_v be the indicator random variable of the event that v appears before all its neighbours in a randomly chosen ordering. Now since v and its neighbours form a collection of $d_v + 1$ vertices in all, and each of them is equally likely to appear first, it is clear that $E[X_v] = \frac{1}{d_v+1}$. Let $X = \sum_v X_v$. By linearity of expectation, $E[X] = \sum_v \frac{1}{d_v+1}$. By (18), there is thus at least one ordering of the vertices, call it $\mathcal{O}$, such that $X(\mathcal{O}) \geq \sum_v \frac{1}{d_v+1}$. But now one just needs to observe that, in any ordering whatsoever, those vertices which appear before all their neighbours must form an independent set. This proves (24).

We will need one more simple general fact about graphs.

Lemma 14 *For any graph G we have that*

$$\#edges\ in\ G = \frac{1}{2} \sum_{v \in V(G)} d_v. \tag{25}$$

[12] there are other ways to prove this theorem, the standard proof being a kind of double induction on k and n.

PROOF : When we sum up the degress of the vertices, we are summing up the edges emanating from each vertex, and then each edge will be counted twice.

Finally, we need a third fact which is pure algebra/calculus :

Lemma 15 *Let $x_1, x_2, ..., x_n, t$ be positive real numbers. If*

$$x_1 + \cdots + x_n \geq t,$$

then

$$\frac{1}{x_1} + \cdots + \frac{1}{x_n} \geq \frac{n^2}{t},$$

with equality in the latter if and only if $x_1 = \cdots = x_n = t/n$.

PROOF : Exercise.

PROOF OF THEOREM 12' : By Lemma 14, the assumption in the statement of the theorem about the number of edges in G can be written as

$$\frac{1}{2} \sum_v d_v < \frac{n^2}{2(k-1)} - \frac{n}{2},$$

which can be rewritten as

$$\sum_v (d_v + 1) < \frac{n^2}{k-1}.$$

Hence, by Lemma 15,

$$\sum_v \frac{1}{d_v + 1} > \frac{n^2}{\frac{n^2}{k-1}} = k - 1.$$

So, by Lemma 13, $\alpha(G) > k - 1$. But $\alpha(G)$ is an integer, thus $\alpha(G) \geq k$, V.S.V.

Example 3 : Girth and chromatic number

DEFINITION 12 : The *girth* of a graph G is the minimum length of a cycle in it (where the *length* of a cycle means the number of vertices/edges in it), and is denoted $\text{girth}(G)$. If G contains no cycles at all (a so-called *forest*), then we set $\text{girth}(G) := \infty$.

DEFINITION 13 : The *chromatic number* of a graph G, denoted $\chi(G)$, is the minimum number of colors needed to color the vertices of G, if no two neighbours can get the same color.

One of the best known problems in graph theory is that of finding an efficient algorithm for computing the chromatic numbers of graphs. In its full generality, this is known to be an *NP-complete problem.*

On the other hand, consider the following examples : A graph has chromatic number 1 if and only if it consists of a bunch of isolated vertices. The chromatic number is 2 if and only if the graph is bipartite. Every forest has chromatic number 2 (and hence is bipartite !). A cycle has chromatic number 2 or 3 depending on whether its length is even or odd respectively.

These examples suggest that large graphs with low chromatic number should in general be 'fat', i.e.: have large girth. In fact, for many years around the middle of the last century, the following was an open problem :

Do there exist graphs which simoultaneously have large girth and large chromatic number ?

If you try to construct such graphs by hand, you probably won't get very far. However, such graphs do exist in abundance, though they are very large (i.e.: have a lot of vertices). This was first proven by Paul Erdős in 1959, and was a wake-up call to the world about the usefulness of the probabilistic method in combinatorics !

Theorem 16 (Erdős) *Given any positive integers k, l, there exists a graph G with $\chi(G) > k$ and girth$(G) > l$.*

In proving this theorem, we might as well introduce, for the first time in the course, the *standard random graph model.* First, an informal definition :

DEFINITION 14 : Let n be a positive integer and $p \in [0, 1]$. The *random graph* $G(n, p)$ has n vertices and is obtained by choosing each of the $\left(\begin{array}{c} n \\ 2 \end{array}\right)$ possible edges randomly and independently with probability p.

The informal terminology 'random graph' is a bit misleading, because strictly speaking a random graph is not a graph at all, it is a probability space. To be able to say what space, we first need to introduce the notion of a *product measure* :

DEFINITON 15 : Let (Ω, μ) be a probability space and suppose the underlying set Ω is a Cartesian product $\Omega = \Omega_1 \times \cdots \times \Omega_k$ of k sets. Then the measure μ is called a *product measure* if there exist probability measures μ_i on Ω_i, for $i = 1, ..., k$, such that, for any point $(\omega_1, ..., \omega_k) \in \Omega$,

$$\mu[\{(\omega_1, ..., \omega_k)\}] = \prod_{i=1}^{k} \mu_i[\{\omega_i\}].$$

In this case, we write that

$$\mu = \prod_{i=1}^{k} \mu_i.$$

DEFINITION 16 : The *random graph* $G(n, p)$ is the probability space (Ω, μ), where $\Omega = \{0, 1\}^{\left(\begin{array}{c} n \\ 2 \end{array}\right)}$, i.e.: Ω is the Cartesian product of $\left(\begin{array}{c} n \\ 2 \end{array}\right)$ copies of the two-element set $\{0, 1\}$, and $\mu = \prod \mu_p$, i.e.: the product of the same number of copies of the measure μ_p on $\{0, 1\}$ given by

$$\mu_p(\{0\}) = 1 - p, \qquad \mu_p(\{1\}) = p.$$

One simple lemma and some useful notation now before we begin the proof. The lemma connects independence number with chromatic number :

Lemma 17 *For any graph G we have*

$$\chi(G) \cdot \alpha(G) \geq |V(G)|. \tag{26}$$

PROOF OF LEMMA : Consider an optimal coloring of G, i.e.: one using $\chi(G)$ colors. Each color class must be an independent set of vertices, hence has size at most $\alpha(G)$. Since there are $\chi(G)$ color classes, (26) follows.

NOTATION : Let n, i be positive integers with $n \geq i$. We denote $(n)_i := n(n-1)\cdots(n-i+1)$.

PROOF OF THEOREM 16 : Since this is definitely the most intircate argument we have encountered to date, we divide it up into steps.

Step 1 : Let k, l be fixed from now on. We are going to work with a random graph $G(n,p)$. The thing we have to get right is the choice of the parameter p. It turns out that, for things to work, p will have to depend on n. To be precise, let θ be any real number such that $0 < \theta < 1/l$. Then we can take $p = n^{\theta - 1}$. Note that $\theta - 1 < 0$ so $p \in [0,1]$ at the very least, and in fact $p \to 0$ as $n \to \infty$. So far, so good.

Step 2 : The purpose of this step is to show that, for the above choice of p, the expected number of cycles of length at most l in $G(n,p)$ is $o(n)$. Now let's be precise :

For any cycle C in the complete graph K_n, let X_C denote the indicator random variable of C in $G(n,p)$. Clearly,

$$E[X_C] = p^{\text{length}(C)},$$

since the presence or absence of C depends on as many independent (biased) coin tosses as there are edges in it. We now set

$$X := \sum_{3 \leq \text{length}(C) \leq l} X_C, \tag{27}$$

so that the random variable X counts the total number of cycles in $G(n,p)$ of length at most l. We can compute $E[X]$ using linearity of expectation, but we first need a formula for the number of cycles in K_n of any given length. I claim that

$$\#\text{cycles in } K_n \text{ of length } i = \frac{(n)_i}{2i} \tag{28}$$

To see this, note that there are $(n)_i$ choices for the i vertices in the cycle in order, but we must divide by $2i$ because, for any given cycle, there are i possible starting points as well as 2 possible orientations.

OK, so we've explained (28). Then (27), (28) and linearity of expectation yield that

$$E[X] = \sum_{i=3}^{l} \frac{(n)_i}{2i} p^i. \tag{29}$$

It remains to show that the right-hand side of (29) is $o(n)$. But this is easy. Recalling the choice of p, we have the following sequence of estimates :

$$\sum_{i=3}^{l} \frac{(n)_i}{2i} p^i \leq \sum_{i=3}^{l} n^i p^i = \sum_{i=3}^{l} n^{i\theta} \leq (l-2) n^{l\theta}.$$

Since l is fixed and $l\theta < 1$ by definition, it is clear that the last term above is $o(n)$.

Step 3 : In this step we show that

$$P\left[\alpha(G(n,p)) \geq \lceil \frac{3 \ln n}{p} \rceil \right] = o(1). \tag{30}$$

First let t be any fixed positive integer. Then, for any graph G, $\alpha(G) \geq t$ if and only if there exists at least one independent set of size t in G. Now consider $G(n,p)$. The probabiity of a given collection of t vertices being independent is $(1-p)^{\left(\begin{array}{c} t \\ 2 \end{array}\right)} = (1-p)^{t(t-1)/2}$. Since there are $\left(\begin{array}{c} n \\ t \end{array}\right)$ possible choices of the t vertices, it follows that

$$P\left[\alpha(G(n,p)) \geq t\right] \leq \left(\begin{array}{c} n \\ t \end{array}\right) \cdot (1-p)^{t(t-1)/2}.$$

We bound the right-hand side conveniently using the simple estimates

$$\left(\begin{array}{c} n \\ t \end{array}\right) \leq n^t, \qquad 1 - p \leq e^{-p}.$$

These yield that

$$P\left[\alpha(G(n,p)) \geq t\right] \leq \left(n e^{-\frac{p(t-1)}{2}}\right)^t.$$

If we now insert the value $t = \lceil \frac{3 \ln n}{p} \rceil$, then we'll obtain

$$P\left[\alpha(G(n,p)) \geq t\right] \leq (1 + o(1)) \left(n \cdot n^{-3/2}\right)^t,$$

which evidently goes to zero as $n \to \infty$. This proves (30).

Step 4 : Now we just need to gather our thoughts - all the hard work is done. According to the first step, the expected number of cycles of length at most l in $G(n,p)$ is $o(n)$. I'll leave it as an exercise for you to prove the following :

'*If X is any positive-valued random variable, depending on some parameter n, and $E[X] = o(n)$, then for any $\epsilon > 0$, $P(X > \epsilon n) \to 0$ as $n \to \infty$.*'

Let's just take $\epsilon = 1/2$ for simplicity. Then with very high probability (going to 1 as $n \to \infty$), $G(n,p)$ contains at most $n/2$ cycles of length at most l. Let G^* be the (random) graph obtained by removing one vertex, and all its adjoining edges, from each cycle of length at most l in $G(n,p)$. Then $\text{girth}(G^*) > l$ and

$$P\left[|V(G^*)| \geq n/2\right] = 1 - o(1). \tag{31}$$

Furthermore, removing vertices and their adjoining edges from a graph cannot increase the independence number, so $\alpha(G^*) \leq \alpha(G(n,p))$. Hence, by (30),

$$P\left[\alpha(G^*) \geq \lceil \frac{3 \ln n}{p} \rceil\right] = o(1). \tag{32}$$

Now (32), (31), Lemma 17 and a bit of computation give that

$$P\left[\chi(G^*) \geq \frac{n^\theta}{6 \ln n}\right] = 1 - o(1). \tag{33}$$

But no matter how small θ is, $\frac{n^\theta}{\ln n} \to \infty$ as n does, so for sufficiently large n it will be greater than k.

The proof is thus complete. What we have actually shown is that with the choice $p = n^\theta$, then with probability tending to 1 as $n \to \infty$ the following holds : the random graph $G(n,p)$ contains a subgraph G^* which has at least $n/2$ vertices, has girth greater than l and has chromatic number greater than k.

Lecture 5

Today will mostly involve developing some general, though quite elementary, tools from probability theory. At the end of the lecture, we will define the notion of a threshold function for (a property of) random graphs. Specific applications will follow in the next lecture. In subsequent lectures we will develop and apply further, and in some cases more sophisticated, general tools.

The general setting is the following : from the point of view of general probability theory, we have so far in this course been engaged in the computation of expectation values of random variables. And not just any old random variables. The finite, combinatorial nature of the applications meant that our random variables X were 'counting something'. More precisely, they usually had the following properties :

(i) they were non-negative integer valued

(ii) they could be expressed as sums of identically distributed indicator variables.

From such computations we have been able to deduce interesting existence results, using essentially nothing more complicated than things like

(I) $\mathrm{P}(X \geq E[X]) > 0$,

(II) If X is non-negative integer valued and $E[X] < 1$, then $\mathrm{P}(X = 0) > 0$.

From now on, the nature of our applications will be characterised by the following types of requirements :

(A) we will be interested in proving that certain events occur with high probability, not just non-zero probability

(B) it will not be enough to be able to compute $E[X]$, we will also require information on how much X is 'spread out' around its mean.

The following example illustrates features of both (A) and (B) :

As indicated above in (II), if you have a non-negative integer valued RV, then if you want to prove that $X = 0$ with high probability, it suffices to show that $E[X]$ is 'much smaller than 1'. Actually, what one is using here is a very simple, but very useful general result, whose trivial proof we leave as an exercise :

Proposition 18 (Markov's Inequality) *Let X be a non-negative real valued RV, and $\alpha \geq 1$. Then*

$$P(X \geq \alpha E[X]) \leq \frac{1}{\alpha}. \tag{34}$$

This is the simplest example of a so-called *concentration inequality.* For applications to come it is by itself far too weak, though it is used all over the place in the proofs of much stronger results. The other problem is that it is 'one-sided', i.e.: it only bounds the probability of X being too large. For the application to showing that $X = 0$ with high probability, when $E[X] << 1$, that's fine. But suppose now instead you're interested in showing that $X > 0$ with high probability. The natural thing to do is to first show that $E[X]$ is large. But this is, by itself, not enough.

EXAMPLE : Suppose $X = 0$ with probability $0,99$ and $X = 10,000,000$ with probability $0,01$. Then $E[X] = 100,000$ is still very large, but the event $X > 0$ is highly unlikely. In the book of Alon and Spencer (which was written when the Cold War still hadn't quite ended), X is the number of deaths from nuclear war in the next 12 months.

The problem with the X in the above example is obviously that it is too spread out. Our first task in the coming lectures will be to develop tools which allow us to determine that certain random variables of interest are not too spread out, and therefore attain values in certain ranges with high probability. Sometimes we'll have an application where it's enough to know that $X > 0$ with high probability given that $E[X]$ is large. Other times, we'll want X to be located in a narrow range around its mean value with high probability.

In our analyses we will make very full use of the simplifying properties (i) and (ii) of the kinds of RV:s we encounter in combinatorial applications. The main obstacle to obtaining stronger results will be that, in most cases, the indicator variables in question are not *independent* of one another. This is a crucial point. On the one hand, independence simplifies lots of probabilistic analysis immensely. On the other hand, even with the current state of knowledge (we're talking 2006 !), effective tools for dealing with dependent events are few and far between. The techniques that have been developed all basically rely on knowing that either the amount of interdependence is 'small' in some precisely quantifiable manner, or that the dependencies are 'correlated' (we avoid defining this term precisely for the moment). Otherwise, you're probably screwed in terms of getting proofs : you might as well get out your computer and run simulations.

The first method we discuss is the simplest but most important one :

Second Moment Method

Basically this involves studying the *variance* of a RV as a measure of how far it is spread out. To simplify matters, unless otherwise stated, all RV:s are henceforth assumed to be non-negative integer valued, even if some of the things we prove hold more generally, and even with the same proofs (left to the reader to investigate these matters). At a later point we will specialise to the case of sums of indicator variables.

DEFINITION 17 : Let X be a RV. The *variance* of X, written as $\mathrm{Var}[X]$, is defined as

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

The square root of the variance is called the *standard deviation.*

Using linearity of expectation, it's easy to show that (exercise, if you have never done it before !)

$$\mathrm{Var}[X] = E[X^2] - E[X]^2. \tag{35}$$

NOTATION : $E[X] := \mu_X$, $\sqrt{\mathrm{Var}[X]} := \sigma_X$. We drop the subscripts when there can be no confusion about what RV is being considered.

Remark At this point it is worth clarifying the terminology *second moment method.* Let X be a RV. The *exponential generating function* of X is the RV e^X. Thus

$$e^X = \sum_{k=0}^{\infty} \frac{X^k}{k!}.$$

Under suitable convergence conditions, linearity of expectation yields that

$$E[e^X] = \sum_{k=0}^{\infty} \frac{E[X^k]}{k!}.$$

The quantity $E[X^k]/k!$ in this expression is called the *k:th moment* of the r.v. X. From (35) we see that the variance of X involves its second moment, hence the name.

A rough analogy to studying the 2nd moment of a r.v. is to study the second derivative of a smooth function in calculus. And just as it is pretty hard to find a real-life situation where one is interested in the third derivative of a smooth function, so in probability theory it is pretty rare to study the third moment of a r.v. Basically, if you can't get a handle on the second moment, then you're probably in a whole lot of trouble !

Finally, it should now not come as a great shock that the term *first moment method* is applied when one just studies the expectation of a r.v. itself. So this is the method we've been using in the whole of the first week. See Alon and Spencer.

The basic concentration estimate involving variance is

Proposition 19 (Chebyshev's Inequality) *Let X be a r.v. with mean μ and standard deviation σ. Let $\lambda \geq 1$. Then*

$$P(|X-\mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}. \tag{36}$$

PROOF : Define a new r.v. Y by $Y := |X-\mu|^2$. Then the left-hand side of (36) is just, by definition of variance, $P(Y \geq \lambda^2 E[Y])$. Markov's inequality (34) now gives the result immediately.

Corollary 20 *Let X be a r.v., $\epsilon > 0$. Then*

$$P(|X-\mu| \geq \epsilon\mu) \leq \frac{\sigma^2}{\epsilon^2\mu^2}. \tag{37}$$

In particular,

$$P(X=0) \leq \frac{\sigma^2}{\mu^2}. \tag{38}$$

PROOF : For the first part, take $\lambda = \epsilon\mu/\sigma$ in (36). For the second part, set $\epsilon = 1$.

According to this corollary, we get good concentration of X around its mean provided that $\mathrm{Var}[X]$ is small compared to $E[X]^2$. We now specialise to the case where

$$X = X_1 + \cdots + X_n$$

is a sum of indicator RV:s. We do not assume the X_i to be identically distributed though. Indeed let us denote by A_i the event indicated by X_i and $p_i := P(A_i)$. Thus

$$X_i = \begin{cases} 1, & \text{with probability } p_i, \\ 0, & \text{with probability } 1 - p_i. \end{cases}$$

Also denote $\mu_i := E[X_i]$, $\sigma_i^2 := \mathrm{Var}[X_i]$. Clearly, $\mu_i = p_i$. Also, by (35) and the fact that $X_i^2 = X_i$ since X_i only takes on the values 0 and 1, we have $\sigma_i^2 = p_i - p_i^2 = p_i(1-p_i)$. We thus have the inequality

$$\sigma_i^2 \leq \mu_i. \tag{39}$$

Since in applications the individual probabilities p_i are usually very small (even if the number of events A_i is usually very large), we are not losing much information in using (39).

We want an expression for the variance of X. Using (35) and several applications of linearity of expectation (LOE from now on), we obtain that

$$\sigma^2 = \sum_{i=1}^{n} \sigma_i^2 + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j), \tag{40}$$

where the *covariance* of X_i and X_j is defined by

$$\mathrm{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j].$$

By (39) and LOE, the first sum on the right of (40) is at most μ. This is good, since we are interested in having σ^2 much smaller than μ^2 in situations where μ is large. So we can focus in on the sum of covariances. Since the X_i are indicator variables, we have

$$E[X_i X_j] - E[X_i]E[X_j] = P(A_i \cap A_j) - P(A_i)P(A_j).$$

Hence $\mathrm{Cov}(X_i, X_j) = 0$ if and only if the events A_i and A_j are independent[13] So independent pairs don't contribute anything to the sum. Let $i \sim j$ denote

[13] More generally, for any two random variables X and Y, if X and Y are *independent* then $E[XY] = E[X]E[Y]$, though the converse need not hold (find an example !). What does it mean for two random variables to be independent in general ? It means simply what one would expect, namely that knowledge of the value of one variable does not give any information on the value of the other. There are several equivalent ways to express this formally. In the finite setting the following definition suffices : we say that real-valued RV:s X and Y are *independent* if, for all real numbers r, s, $P(X = r|Y = s) = P(X = r)$

that events A_i and A_j are not independent. We have at the very least the bound

$$\sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \leq \sum_{i \sim j} P(A_i \cap A_j).$$

Since $P(A_i \cap A_j) = P(A_i) \cdot P(A_j | A_i)$, we can rewrite the last sum as a double-sum, namely

$$\sum_{i \sim j} P(A_i \cap A_j) = \sum_i P(A_i) \sum_{j \sim i} P(A_j | A_i).$$

Let us now make one further simplifying assumption, namely that the inner sum above is independent of i. This is a kind of 'symmetry' requirement which holds for most applications. Following standard practice, we now denote the inner sum Δ^*. Thus we have

$$\sum_{i \neq j} \mathrm{Cov}(X_i, X_j) \leq \Delta^* \cdot \sum_i P(A_i) = \Delta^* \cdot \sum_i \mu_i = \Delta^* \cdot \mu.$$

So let's summarise where we stand : assuming that our r.v. X is a sum of indicator variables, and that a certain symmetry condition is fulfilled, we have that

$$\mathrm{Var}[X] \leq (1 + \Delta^*) E[X].$$

Hence, to show that $\mathrm{Var}[X]$ is much smaller than $E[X]^2$, it suffices to show that Δ^* is much smaller than $E[X]$. This is the crux of the second moment method.

We will start to see some applications in the next lecture. We close this lecture, which has been about setting things up properly, in the same spirit, with a definition of the important concept of a threshold function for random graphs.

and $P(Y = r | X = s) = P(Y = r)$. In words, the probability that X (resp. Y) attains the value r given that Y (resp. X) is known to have the value s, is the same as it was before the value of Y (resp. X) was known.

Note that we've defined what it means for two variables to be independent OF ONE ANOTHER. It could happen that X is independent of Y but not vice versa. This is intuitively clear as the following toy example illustrates : let X be the mood of a teacher on the day (s)he is preparing an exam, and Y be the mood of one of his/her students. I leave it as an exercise to give a more formal example.

NOTATION : Let A be a graph property and G a graph. We write $G \models A$ to denote the fact that G has the property A. For example, if A is the property 'is connected', then $G \models A$ means that G is connected.

DEFINITION 18 : Let A be a graph property and $t : \mathbf{N} \to [0,1]$ a function. Then t is said to be a *threshold function* for the property A if two conditions hold :

(I) If $p(n) = o[t(n)]$ then $P[G(n,p(n)) \models A] \to 0$ as $n \to \infty$,
(II) If $t(n) = o[p(n)]$ then $P[G(n,p(n)) \models A] \to 1$ as $n \to \infty$.

Remark If t is a threshold for some property A, then so is $c \cdot t$ for any constant c such that $||c \cdot t||_\infty \leq 1$.

We postpone further waffle to the next lecture's notes.

Lecture 6

Today we present some applications of the second moment method.

First application : Subgraph threshold

Given a graph property A, there are basically three stages in the analysis of the threshold phenomenon for that property :

(I) Prove that a threshold exists.
(II) Compute the threshold[14]
(III) Investigate more closely what happens as the threshold is crossed.

We will concentrate in this course on stage (II). There are some very general theorems about existence of thresholds, but to do justice to these would require too long a detour toward mathematical logic. Stage (III) obviously is likely to be more technical, and it cannot be undertaken before stage (II) anyway. Note however that, in speaking of 'crossing the threshold' we are adopting a *dynamic* model of random graphs $G(n,p)$, where we think of the parameter p as growing, and the edges of the graph 'growing' accordingly.

The graph property we have chosen to exhibit how the second moment method can be used to compute thresholds is that of *subgraph containment.* So let H be any fixed graph. The graph property $\mathcal{A} = \mathcal{A}_H$ under consideration is 'contains a copy of H', so that $G \models \mathcal{A}_H$ means that the graph G contains a copy of the graph H. For example, $K_n \models \mathcal{A}_{K_m}$ if and only if $n \geq m$, in which case K_n in fact contains $\left(\begin{array}{c} n \\ m \end{array} \right)$ or $(n)_m$ different copies of K_m, depending on how one counts.

We need some definitions before stating our main result :

Definition 19 : Let H be a graph, with e edges and v vertices. The *density* of H, denoted $\rho(H)$, is the quantity

$$\rho(H) := \frac{e}{v}.$$

[14] Here we are once again deliberately sloppy with our language. Since a threshold function can never be unique (one can always multiply by a constant, for example), one shouldn't speak of 'the' threshold. But this is common practice.

The graph H is said to be *balanced* if $\rho(H) \geq \rho(H')$ for every subgraph H' of H.

Theorem 21 *Let H be a balanced graph. Then the function*

$$t(n) := n^{-1/\rho(H)}$$

is a threshold function for the property $\mathcal{A}_H$.

PROOF : We need to prove two things, namely :

(I) If $p(n) = o[t(n)]$ then $P[G(n, p(n)) \models \mathcal{A}_H] = o(1)$.
(II) If $t(n) = o[p(n)]$ then $P[G(n, p(n)) \models \mathcal{A}_H] = 1 - o(1)$.

PROOF OF (I) : This part does not require the knowledge that H is balanced. Let e, v denote the number of edges and vertices of H respectively. These quantities are thus constants and do not affect any estimates of orders of magnitude of quantities as $n \to \infty$. Fix an n and $p \in [0, 1]$. For every subset S of the vertices of K_n of size v, let X_S be the indicator variable of the event that, in $G(n, p)$, at least one copy of H appears on the vertices in S. Note that, a priori, many different copies of a single graph may appear on the same set of vertices. But since H is fixed, the number of copies of it which may appear on any set of v vertices is bounded by a function depending only on v. All of this implies that

$$E[X_S] = \Theta(p^e).$$

Let $X := \sum X_S$, the sum being over all v-element sets of vertices in K_n. Then

$$E[X] = \sum E[X_S] = \left(\begin{array}{c} n \\ v \end{array} \right) \cdot \Theta(p^e) = \Theta(n^v p^e). \tag{41}$$

But X just counts the total number of copies of H in $G(n, p)$. So from (41) it is already clear that if $p = p(n) = o[t(n)]$, then $E[X] = o(1)$, implying that $P(X = 0) = 1 - o(1)$. This proves part (I).

PROOF OF (II) : Similarly, (41) implies that if $t(n) = o[p(n)]$ then $E[X] \to \infty$. All we need to show is that $P(X > 0) = 1 - o(1)$. We apply the second moment method. To simplify notation, let A_S denote the event indicated by X_S. Clearly, the various conditions introduced in our discussion of the second moment method apply to the present situation, so that it suffices for us to show that $\Delta^* = o(E[X])$, where

$$\Delta^* = \sum_{T \sim S} P(A_T | A_S).$$

Here S is a fixed set of vertices of size v, and the sum runs over all sets T of vertices of size v so that the event A_T is not independent of the event A_S. Now in the random graph setting, two events are independent if they are defined on disjoint sets of edges. So A_T is dependent on A_S if and only if the edge-sets defined by T and S are not disjoint, which is the case if and only if T and S share at least two vertices. Hence we can write

$$\Delta^* = \sum_{T:2 \leq |T \cap S| \leq k-1} P(A_T | A_S) = \sum_{i=2}^{k-1} \sum_{T:|T \cap S| = i} P(A_T | A_S). \tag{42}$$

In the inner sum, the quantity $P(A_T|A_S)$ must be the same for every choice of T. The number of such choices is $\left(\begin{array}{c} v \\ i \end{array} \right) \left(\begin{array}{c} n-v \\ v-i \end{array} \right)$, since T must have i vertices in common with S and $v-i$ other vertices. This number is $\Theta(n^{v-i})$.

Now fix an i and a T. We need an estimate for $P(A_T|A_S)$. Here it is assumed that at least one copy of H appears on S and want to estimate the probability of at least one copy of H also appearing on T. Up to a constant factor, as before, we may consider a fixed copy of H on S. Let H' be the part of it on $T \cap S$. Once again, up to a constant factor, we may consider a fixed extension of H' to a copy of H on the vertices of T.

At this point we use the fact that H is balanced. It implies that $\rho(H') \leq \rho(H)$, thus H' contains at most ie/v edges. This means that the appearance or otherwise of a fixed extension of H' on T depends on the presence or otherwise in $G(n,p)$ of at least $e - ie/v$ edges.

Putting all this together, what we have shown is that, for a fixed i and T,

$$P(A_T|A_S) = \Theta(p^{e-ie/v}).$$

Substituting this and the estimate for the number of different T:s into (42) we find that

$$\Delta^* = \sum_{i=2}^{k-1} \Theta(n^{v-i}) \cdot \Theta(p^{e-ie/v}) = \sum_{i=2}^{k-1} \Theta\left[(n^v p^e)^{1-i/v}\right].$$

Since $1 - i/v \leq 1 - 2/v < 1$ for every value of i, it is thus clear that the sum is $o(n^v p^e) = o(E[X])$, which completes the proof of the theorem.

The proof of the following completely general result is more technical.

Theorem 22 *Let H be any graph, not necessarily balanced. Let H' be a subgraph of H of maxmimal density. Then the function $t(n) = n^{-1/\rho(H')}$ is a threshold for the property $\mathcal{A}_H$.*

A full proof is not contained in Alon and Spencer (henceforth referred to as [AS]), but there are some technical extensions of Theorem 21 above from which Theorem 22 can be deduced without too much pain and is left as an exercise. Copies of these were handed out in class.

Second application : Concentration of random graph invariants

A graph *invariant* just means any numerical quantity which may be associated to an arbitrary graph. Examples of graph invariants are chromatic number, girth, number of connected components, number of Hamilton cycles etc. Invariants of random graphs $G(n,p)$ are thus (non-negative integer valued) functions of two variables, n and p.

The computation of random graph invariants is a natural counterpart to the problem of computing thresholds. In the former type of problem, one considers a fixed p (the most natural and interesting choice often being $p = 1/2$, as it corresponds to the edges of the graph being chosen by independent tosses of a fair coin) and wants to estimate the value of the invariant as $n \to \infty$. This basically involves estimating the expectation of some random variable X. Of more interest, though, is the degree of predictability of the invariant's value, in other words, how well concentrated the variable X is around its mean. The second moment method sometimes gets us quite strong results, a particularly nice example being the following '*2-values theorem*' :

Theorem 23 *There is an integer-valued function $k(n)$ such that*

$$P[\omega(G(n,1/2)) = k(n) \text{ or } k(n)+1] \to 1 \text{ as } n \to \infty.$$

. *In addition, $k(n) \sim 2\log_2 n$.*

Here $\omega(G)$ denotes the *clique number* of a graph G, which is the maxmimal number of vertices in a complete subgraph of G. Note that the theorem does not tell us exactly what two values $\omega(G(n,1/2))$ is concentrated on, for any given large n, just that there are two such values, and they are of the order of magnitude of $2\log_2 n$. However, it will be clear from the proof of the theorem that the function $k(n)$, and the amount of concentration, is easily[15] computable for any particular n.

SKETCH PROOF OF THEOREM 23 : I did not go through all the details of the proof, but gave the main ideas and handed out pages from Chapter 4 of [AS] for the full computations. What is interesting is that, in [AS], they defer a final proof of this theorem to Chapter 10, and there use some more advanced probabilistic machinery, namely the so-called *Janson inequalities.* I think this is unnecessary, however, and that the second moment method suffices to get a full proof. I leave it for yourselves to check this !

Anyway, here is the sketch :

Fix an n and a k and let X be a r.v. which counts the number of cliques of size k in $G(n,1/2)$. We can write (should by now be getting used to this !) $X = \sum X_S$, the sum being taken over all vertex sets S of size k, where X_S indicates that all $\left(\begin{array}{c} k \\ 2 \end{array}\right)$ edges between the vertices of S are, so to speak, 'turned on'. Thus

$$E[X] = \left(\begin{array}{c} n \\ k \end{array}\right) \cdot 2^{-\left(\begin{array}{c} k \\ 2 \end{array}\right)}. \tag{43}$$

Denote the quantity on the right hand side of (43) as $f(n,k)$. We have already seen in the very first lecture that $f(n,k)$ becomes less than 1 when k is in the vicinity of $2\log_2(n)$. Since the event $X = 0$ is the same as the event $\omega[G(n,1/2)] < k$, this is the crucial transition as long as we can show that $\Delta^* = o(f(n,k))$ when the latter is large. The exceptionally high concentration of the clique number comes from the fact that the function $f(n,k)$, which is a decreasing function of k for fixed n, is decreasing very rapidly when k is close to $2\log_2 n$. In fact, direct insertion into the formula

[15] i.e.: in polynomial time, at least.

for $f(n,k)$ gives that

$$\frac{f(n,k+1)}{f(n,k)} = 2^{-k}\frac{n-k}{k+1},$$

so that when $k \sim 2\log_2 n$, $\frac{f(n,k+1)}{f(n,k)} = n^{-1+o(1)}$.

Some remarks on the estimation of Δ^* : It can be broken up exactly as in (42) above. The analysis is even simpler than in Theorem 21, however, as one doesn't need to worry about those annoying 'up to a constant factor' estimates here. One finds (see the handout from [AS]) that

$$\frac{\Delta^*}{E[X]} = \sum_{i=2}^{k-1} g(i),$$

where

$$g(i) = \frac{\begin{pmatrix} k \\ i \end{pmatrix}\begin{pmatrix} n-k \\ k-i \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \cdot 2^{\begin{pmatrix} i \\ 2 \end{pmatrix}}.$$

One needs to show that each term in the sum is $o(1)$, when $E[X]$ is large and $k \sim 2\log_2 n$. In [AS] they show by direct computation that

$$g(2) \sim \frac{k^4}{n^2} = n^{-2+o(1)},$$
$$g(k-1) \sim \frac{2kn2^{-k}}{E[X]} = \frac{n^{-1+o(1)}}{E[X]},$$

and leave the remaining cases to the reader. In fact, one can see (again just by direct insertion into the formula for $g(i)$) that, up to a $1+o(1)$ factor, the function $g(i)$ starts off by decreasing as i increases, and then starts increasing again as i approaches the order of magnitude of $\log_2 n$. What this implies is that, for every i,

$$g(i) \leq (1+o(1))\max\{g(2), g(k-1)\}.$$

It is this estimate which I think allows one to finish off the proof of Theorem 23 without needing to resort to any more advanced techniques. But don't take my word for it : check it yourselves !

Third application : Distinct subset sums

Here we make a detour back to number theory.

DEFINITION 20 : Let $A = \{a_1, ..., a_k\}$ be a finite set of positive integers. A is said to have *distinct subset sums* if, for every two distinct subsets X, Y of $\{1, ..., k\}$, the sums $\sum_{i \in X} a_i$ and $\sum_{i \in Y} a_i$ have different values[16].

Let $f(n)$ be the maximum possible size of a subset of $\{1, ..., n\}$ which has distinct subset sums.

LOWER BOUNDS :

Take $n = 2^k$ and $A = \{2^i : 0 \leq i \leq k\}$. This example shows that $f(n) \geq 1 + \lfloor \log_2 n \rfloor$. Erdős offered 300 dollars for a proof that there exists a universal constant C such that $f(n) \leq \log_2 n + C$. Note that he's not asking here for a computation of the optimal C or even a decent estimate of it, just a proof that some such constant exists, in other words that $f(n) = \log_2 n + O(1)$. The base-2 example shows that $C \geq 1$. If we confine ourselves to integer C, then an example constructed by John Conway and Richard Guy in 1969 shows that $C \geq 2$. There have since been a few papers presenting modest improvements on that construction, which lead to the conclusion that $C \geq 3$ (I think !). Note that, in order to get a better lower bound on C, it suffices to do so for a single n, because of the following trick : if $A = \{a_1, ..., a_k\}$ is a subset of $\{1, ..., n\}$ with distinct subset sums, and u is any odd number s.t. $1 \leq u \leq 2n$, then $A' = \{2a_1, ..., 2a_k, u\}$ is a subset of $\{1, ..., 2n\}$ with distinct subset sums and one additional element. This means that if $f(n) > \log_2 n + C$ then $f(N) > \log_2 N + C$ for every N of the form $N = 2^t n$.

One can then use a computer to help find individual examples ...

UPPER BOUNDS :

If A has size k and is contained in $\{1, ..., n\}$ then there are 2^k distinct subset sums and each is among $\{0, ..., nk - \frac{k(k-1)}{2}\}$. Thus

$$2^{f(n)} \leq 1 + nk - \frac{k(k-1)}{2},$$

[16] If X is the empty set, the sum is assigned the value zero.

which leads to a bound of the form

$$f(n) \leq \log_2 n + \log_2 \log_2 n + O(1).$$

Erdős improved this to

$$f(n) \leq \log_2 n + \frac{1}{2} \log_2 \log_2 n + O(1) \tag{44}$$

using a probabilistic argument involving Chebyshev's inequality. Since the proof of (44) is more intersting than the result itself, and since we have too many other things to cover, I decided not to go through the proof, but instead handed out the relevant text from [AS].

Lecture 7

A central class of results in probability theory are so-called *Central Limit Theorems.* A weaker set of results, called *Laws of Large Numbers*, capture the layman's notion that things tend to average out over time. The Central Limit Theorems are more precise : they tell you that random variables which are long-term averages tend to have normal distributions. Recall that the normal distribution with mean μ and standard deviation σ is the real-valued random variable $X = N(\mu, \sigma)$ such that, for every $z \in \mathbf{R}_+$,

$$P(|X - \mu| \geq z) = 2 \cdot \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2\sigma} \, dt. \tag{45}$$

The normal distribution is thus well concentrated about its mean. For example, (45) implies that, for any $\lambda > 0$,

$$P(|X - \mu| \geq \lambda\sigma) \leq e^{-\lambda^2/2}.$$

This should be compared with the totally general Chebyshev inequality.

Classically, 'the' Central Limit Theorem is about sums of independent, identically distributed (i.i.d.) random variables. It is an old result which says, basically, that if $X_1, X_2, ...$ is a sequence of i.i.d. random variables, each with mean μ and variance σ, and $Y_n = (X_1 + \cdots + X_n)/n$ is the average of the first n of them, then Y_n approaches $N(\mu, \sigma)$. What do we mean here by 'approaches' ? Well, there are different possibilities, but the simplest notion, which is also the weakest and thus the easiest to get results about, is that, for every positive real number z,

$$\lim_{n \to \infty} P(|Y_n - \mu| \geq z) = 2 \cdot \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-t^2/2\sigma} \, dt. \tag{46}$$

While the CLT is a fundamental theoretical result, there are several problems associated with its application :

(I) it assumes identical distributions
(II) it assumes independence
(III) it is qualitiative, not a quantitative result. In other words, it doesn't say anything about the rate of convergence to the limit in (46).

Problem (I) is not serious : the CLT can be extended to sums of variables with different distributions. (II) and (III) are much more serious, though. There are CL Theorems that concern dependent variables, but results are limited. In a seminal paper, Chernoff (1952) dealt significantly with problem (III). His results concern sums of independent indicator variables. Chernoff was interested in statistics, and his results are of great importance in that field. We've already seen in this course that sums of indicator variables are also ubiquitous in combinatorial applications, so Chernoff's results deserve attention. On the other hand, the fact that they don't address the issue of independence limits their applicability. Nevertheless, the methods employed by Chernoff lay the foundation for much subsequent work on addressing the independence issue and knowledge of his method (and of the various qualitative CL Theorems) is a prerequisite for appreciating these later developments. We thus present a detailed proof of the following result

Theorem 24 (Chernoff's bound) *Let X be a random variable which is a sum of independent indicator variables. Let $E[X] := \mu$. Then for any $\epsilon > 0$ there exists a positive constant c_ϵ, depending only on ϵ, such that*

$$P(|X - \mu| > \epsilon\mu) < 2e^{-c_\epsilon \mu}. \tag{47}$$

In fact one can take

$$c_\epsilon = \min\left\{\frac{\epsilon^2}{2}, (1+\epsilon)\ln(1+\epsilon) - \epsilon\right\}. \tag{48}$$

The crucial point here is that c_ϵ does not depend on X, i.e.: it doesn't depend on how many indicator variables X is the sum of, nor on the distributions of these.

We will deduce Theorem 24 from a *normalised* version of it. Let X_i be an indicator variable, say

$$X_i = \begin{cases} 1, & \text{with probability } p_i, \\ 0, & \text{with probability } 1 - p_i. \end{cases}$$

The *normalisation* of X_i, which we denote $\hat{X}_i$, is the variable $X_i - p_i$, i.e.:

$$\hat{X}_i = \begin{cases} 1 - p_i, & \text{with probability } p_i, \\ -p_i, & \text{with probability } 1 - p_i. \end{cases} \tag{49}$$

Thus $\hat{X}_i$ has mean zero. It has the same variance as X_i, namely $p_i(1-p_i)$.

Now let $\hat{X}$ be a r.v. which is a sum of n normalised indicator variables, for some fixed n. Write $\hat{X} = \hat{X}_1 + \cdots + \hat{X}_n$, with the $\hat{X}_i$ as above, and define the number p by $np = p_1 + \cdots + p_n$. Finally, let a be any positive real number. We will prove the following two inequalities :

$$P(\hat{X} > a) < \exp\left[a - pn\ln\left(1+\frac{a}{pn}\right) - a\ln\left(1+\frac{a}{pn}\right)\right], \tag{50}$$

$$P(\hat{X} < -a) < \exp\left[-\frac{a^2}{2pn}\right]. \tag{51}$$

Note that the theorem follows from (50) and (51) upon setting $a = \epsilon pn$. We will prove (50) in detail. The proof of (51) is very similar and thus omitted, but the proof from [AS] will be handed out in class. First, though, a couple of remarks are in order :

(i) there is an obvious asymmetry in the estimates (50) and (51), depending on whether $\hat{X}$ is positive or negative. Unfortunately, this is a feature of Chernoff's method.
(ii) the connection to the normal distribution is clear in (51), as the variance of $\hat{X}$ is about np if the individual p_i are small, as is usually the case in applications. With (50), the connection is not so obvious. However, if a is small compared to pn and we use the fact that $\ln(1+u) \geq u - u^2/2$ when $0 < u < 1$, then we can deduce from (50) that

$$P(\hat{X} > a) < \exp\left[-\frac{a^2}{2pn} + \frac{a^3}{2(pn)^2}\right]. \tag{52}$$

Note that (52) gives no information when a is large compared to np as then the cubic term dominates. Again, this is a feature of Chernoff's method, but is not important, since we're only interested in having concentration close to the mean anyway.

Proof of (50) : The proof uses the *exponential generating function* of $\hat{X}$, namely : Let $\lambda > 0$. Then we will consider the r.v.

$$e^{\lambda\hat{X}} := \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\hat{X}^k.$$

Now $\hat{X} > a$ if and only if $e^{\lambda \hat{X}} > e^{\lambda a}$. The simple Markov inequality gives a bound

$$P(e^{\lambda \hat{X}} > e^{\lambda a}) < \frac{E[e^{\lambda \hat{X}}]}{e^{\lambda a}}. \tag{53}$$

We will estimate the expectation and then the clever part of the proof is that λ, which at this point is still some arbitrary positive real number, will be chosen so as to minimise the right-hand side of (53). The estimate of the expectation will use the concavity of the logarithm. Let us begin by formally defining what this means :

Definition 21 : A function f on the positive reals is said to be *concave* if, for any n, any positive reals $x_1 \leq x_2 \leq \cdots \leq x_n$ and any positive reals $a_1, ..., a_n$ satisfying $\sum a_i = 1$, it holds that

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \geq \sum_{i=1}^{n} a_i f(x_i).$$

Concavity has a simple geometric interpretation, namely that the graph of f lies on or above the straight line drawn between any two points on it.

Lemma 25 *Let $C > 0$. Then the function $f(x) = \ln(Cx+1)$ is concave.*

Proof of Lemma : Exercise.

Now let us return to the proof of (49). Since $\hat{X} = \sum \hat{X}_i$, one easily sees that

$$e^{\lambda \hat{X}} = \prod_{i=1}^{n} e^{\lambda \hat{X}_i}.$$

Now we use the independence of the $\hat{X}_i$. Recall that if A, B are independent random variables, then $E[AB] = E[A]E[B]$. Thus, by induction,

$$E[e^{\lambda \hat{X}}] = \prod_{i=1}^{n} E[e^{\lambda \hat{X}_i}]. \tag{54}$$

But from (49), the definition of e.g.f. and linearity of expectation (convergence is not a problem), one easily computes that

$$E[e^{\lambda \hat{X}_i}] = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} = e^{-\lambda p_i}\left[p_i(e^{\lambda}-1)+1\right].$$

Substituting into (54) and recalling the definition of p, we thus have

$$E[e^{\lambda\hat{X}}] = e^{-\lambda pn}\prod_{i=1}^{n}[p_i(e^{\lambda}-1)+1]. \tag{55}$$

But

$$\prod_{i=1}^{n}[p_i(e^{\lambda}-1)+1] \leq [p(e^{\lambda}-1)+1]^n. \tag{56}$$

Indeed this follows from taking logarithms and using Lemma 25. So substituting (56) back into (55) and in turn back into (53), we have the estimate

$$P(\hat{X} > a) < e^{-\lambda pn}[pe^{\lambda}+(1-p)]^n e^{-\lambda a}. \tag{57}$$

It is now a horrid calculus exercise to compute the precise value of λ which minimises the right hand side of (57)[17]. However, a good approximation when $a << np$ is to take $\lambda = \ln(1+a/pn)$. Substituting this into (57) we get the desired relation (50) upon noticing that, with this choice of λ,

$$[pe^{\lambda}+(1-p)]^n = (1+a/n)^n \leq e^a.$$

This completes the proof of Theorem 24. Applications will follow in the next lecture(s).

[17] The right value turns out to be

$$\lambda = \ln\left[\left(\frac{1-p}{p}\right)\left(\frac{a+np}{n-(a+np)}\right)\right].$$

Lecture 8 : Thin bases

This whole lecture was concerned with presenting a single, particularly beautiful application of the Chernoff bounds to number theory. We start with the requisite definitions :

Definition 22 : Let h be a positive integer and A a subset of $\mathbf{N}$. The *representation function of A of order h*, denoted $r_{h,A}$, is the non-negative integer valued function on $\mathbf{N}$ such that $r_{h,A}(n)$ is the number of solutions in A to

$$a_1 + \cdots + a_h = n.$$

Here we are considering unordered solutions and repititions are allowed. So, for example, if $A = \{1, 2, 3, 4, 6, 9\}$ then $r_{2,A}(6) = 2$ since we have the two solutions $2 + 4 = 3 + 3 = 6$.

Definition 23 : Let h be a positive integer and $A \subseteq \mathbf{N}$. A is said to be a *basis of order h* if $r_{h,A}(n) > 0$ for all sufficiently large n.

The case $h = 1$ is totally uninteresting, since then a subset of $\mathbf{N}$ is a basis if and only if its complement is finite. But as soon as $h > 1$ things get interesting.

In that part of classical *analytic* number theory which deals with bases, the type of question posed is whether some particularly interesting subset A of $\mathbf{N}$ is a basis of a certain order. There are 2 examples which everyone likes to quote :

(i) $A = \{\text{set of primes}\}$,

(ii) $A = \{n^k : n \in \mathbf{N}\}$, for any fixed $k > 1$.

Regarding (i), the state of the art is

Theorem 26 (Vinogradov 1937) *Every sufficiently large odd number is a sum of at most three primes. Hence, the primes are a basis of order 4.*

If you want to become rich and famous then you solve

Goldbach Conjecture *Every even number greater than two is the sum of two primes. Hence, the primes are a basis of order 3.*

Regarding (ii),

Theorem 27 (Hilbert 1909, Hardy-Littlewood 192x) *For every $k > 1$ the set of k:th powers is a basis of some order.*

The problem to which Theorem 27 is the solution is commonly known as *Waring's Problem.* One denotes by $G(k)$ the smallest integer such that the k:th powers are a basis of order $G(k)$. The case $k = 2$ dates back to Lagrange, who showed that every positive integer (not just every sufficiently large one) is a sum of at most four squares. On the other hand, it's easy to see (exercise !) that there are infinitely many integers which are not sums of three or fewer squares, so $G(2) = 4$. It is known that $G(3) \leq 7$ and that $G(4) = 16$. The exact value of $G(k)$ is not known for any $k > 4$, and finding improved upper bounds continues to be an active research topic.

Problems like (i) and (ii) are tackled using Fourier analysis, or what number theorists refer to as the *Hardy-Littlewood circle method.* A standard reference if you're interested is [1].

An ovverriding feature of *combinatorial* number theory is that one is interested in properties of general sets of integers rather than of individual ones with a special arithmetical structure. This is pretty wafflish, and there is no real dividing line between the ranges of applicability of analytic and combinatorial methods. However, regarding bases, the following curious result from the 1940s was the starting point of another line of investigation :

Proposition 28 *There is no infinite subset A of* $\mathbf{N}$ *for which the representation function $r_{2,A}(n)$ is constant for all sufficiently large n.*

Proof : Suppose the contrary and let A be a basis of order 2 such that $r_{2,n}(A) = k$ for all sufficiently large n and some constant $k > 0$. We consider the *generating function* of the set A, which is the power series[18]

$$F(z) := \sum_{a \in A} z^a \qquad (z \in \mathbf{C}).$$

The power series certainly converges when $|z| < 1$, so we will work in this region so that all our algebraic manipulations will be valid. The connection between the generating function and the representation function is that

$$[F(z)]^2 + F(z^2) = 2 \cdot \sum_{n=1}^{\infty} r_{2,A}(n) z^n. \tag{58}$$

Suppose now that $r_{2,A}(n) = k$ for all $n \geq n_0$. Then (58) can be written as

$$[F(z)]^2 + F(z^2) = \sum_{n=1}^{n_0 - 1} r_{2,A}(n) z^n + 2k \cdot \sum_{n=n_0}^{\infty} z^n. \tag{59}$$

The first sum on the right of (59) is some polynomial in z. We denote it as $P(z)$. The second sum is a geometric series, so has a simple formula. We thus obtain that

$$[F(z)]^2 + F(z^2) = P(z) + 2k \cdot \frac{z^{n_0}}{1-z}. \tag{60}$$

The desired contradiction is obtained by seeing what happens as $z \rightarrow -1$ from the right along the real axis. Because of all the squares present, the left hand side heads inexorably toward positive infinity. But the right hand side heads toward some finite value, namely $P(1) + k$. This contradiction completes the proof.

The following problem, originally posed by Erdős in [2], remains after 50-plus years the biggest unresolved issue in the combinatorial theory of bases :

Open Problem *Does there exist a constant $C > 0$ and a basis A of order*

[18] We've encountered generating functions once already in this course, namely we used the exponential generating function of a random variable in the proof of the Chernoff bounds. Still, if you're not familiar with the use of generating functions, proofs like the present one may strike you as coming out of the blue. However, it is standard practice to invoke generating functions of sequences when one wants to apply analytical methods to combinatorial or arithmetical problems. There are many, many illustrations of this. See [1] for applications in number theory.

2 such that $r_{2,A}(n) < C$ for all n ?

Erdős actually conjectured that the answer is 'No', and this is still generally believed to be the case, even though progress on the problem has been exactly zero. Notice that, if this is the case, it says something kind of weird, namely : if the set A 'covers' $\mathbf{N}$ at least once, then it has to do so an unbounded number of times.

Informally, a basis of a certain order is called *thin* if its representation function is a slowly growing function of n. The thinnest bases known to exist have been identified by probabilistic arguments. It would be a major achievement to give an explicit construction which comes anywhere close to matching the following :

Theorem 29 (Erdős 1956) *There exist bases A of order 2 for which*

$$r_{2,A}(n) = \Theta(\ln n). \tag{61}$$

This theorem, and its subsequent extension to higher orders which we will remark on later, are very much state of the art. The gap between it and the Open Problem above is a gaping black hole in our current understanding of bases. The proof of Theorem 29 is a beautiful application of the Chernoff bounds.

PROOF : Let K be a fixed positive constant whose value will be determined later. We consider a random subset A of $\mathbf{N}$ such that each positive integer x is chosen independently of all others with probability p_x given by

$$p_x := \min \left\{ K\sqrt{\frac{\ln x}{x}}, 1 \right\}.$$

We will show that, for an appropriate choice of K, the representation function of A satisfies (61) with probability one[19]. For each $n > 0$, let X_n denote the random variable $r_{2,A}(n)$. Note that

$$X_n = \sum_{x=1}^{\lfloor n/2 \rfloor} X_{n,x},$$

[19] which is not the same thing as saying 'with certainty', since we are no longer in a finite setting. Indeed, A is a subset of $\mathbf{N}$, hence there are uncountably many possibilities for it.

where $X_{n,x}$ is the indicator variable of the event that both x and $n-x$ lie in A. Let $\mu_n := E[X_n]$. Thus,

$$\mu_n = \sum_{x=1}^{\lfloor n/2 \rfloor} \min\left\{ K\sqrt{\frac{\ln x}{x}}, 1 \right\} \cdot \min\left\{ K\sqrt{\frac{\ln(n-x)}{n-x}}, 1 \right\}.$$

The main technical challenge in the proof is to prove an estimate for μ_n. But, conceptually, the crucial point is that, for each fixed n, the variables $X_{n,x}$ are mutually independent, hence we will eventually be able to apply the Chernoff bounds to get good concentration of the X_n. For higher order bases, this is where the present line of reasoning breaks down and more sophisticated concentration results are needed to get around the problem. We defer further discussion of this issue until we're done with the current proof.

OK, so we need to estimate the μ_n. The claim is that

$$\mu_n \sim \frac{K^2\pi}{2} \ln n. \tag{62}$$

The verification of (62) is a challenging Calculus 101 exercise. So as not to obscure the probabilistic ideas being employed here, we relegate the proof to Appendix 1 and continue with the main thrust of the argument. So we assume (62). Fix any choice of real number $\epsilon \in (0,1)$, and let A_n denote the event that $r_{2,A}(n)$ does not lie between $(1-\epsilon)\frac{K^2\pi}{2}\ln n$ and $(1+\epsilon)\frac{K^2\pi}{2}\ln n$. Chernoff's Inequality now tells us that

$$P(A_n) < 2\exp\left(-c_\epsilon \frac{K^2\pi}{2} \ln n\right) = 2 \cdot n^{-c_\epsilon \frac{K^2\pi}{2}}.$$

If K is now chosen so that

$$c_\epsilon \frac{K^2\pi}{2} > 1,$$

then

$$\sum_{n=1}^{\infty} P(A_n) < \infty. \tag{63}$$

The theorem will then follow directly from

Lemma 30 (Borel-Cantelli Lemma) *Let $(A_n)_{n=1}^{\infty}$ be a sequence of events in a probability space, and suppose that (63) holds. Then with probability one, only finitely many of the A_n occur.*

Proof of Lemma : Let $\epsilon > 0$. We will show that the probability of infinitely many A_n occurring is less than ϵ. Eq. (63) implies that there exists an n_0 such that

$$\sum_{n=n_0}^{\infty} P(A_n) < \epsilon. \tag{64}$$

But the left hand side of (64) is an upper bound for the probability of at least one A_n occurring for $n \geq n_0$, hence in turn an upper bound for the probability of infinitely many A_n occurring. So we're done !

Remark 1 To prove the theorem, it would have sufficed to show that our random choice of A satisfied (61) with non-zero probability. We actually succeeded in showing that this was achieved with probability one, so that in some sense bases of this thinness are abundant. However, as previously noted, no-one has a clue how to construct one explicitly.

Remark 2 We remarked above where the argument breaks down for higher order bases. It took 34 years to overcome this obstacle and prove

Theorem 31 (Erdős, Tetali 1990 [3]) *Let $h \geq 2$. Then there exists a basis A of order h for which $r_{h,A}(n) = \Theta(\ln n)$.*

Though this was not the original approach of Erdős and Tetali, the quickest known way to get around the obstacles presented by non-independence is to use what are called the *Janson inequalities*, proven by Svante Janson in the late 1980s. These are discussed in Chapter 8 of [AS], and in [3] itself, but we probably won't have time to get that far in this course. Hopefully, though, I have provided sufficient motivation for you to study them on your own !

References

[1] R.C. Vaughan, The Hardy.Littlewood Method (2nd edition). Cambridge University Press (1997).

[2] P. Erdős and P. Turan, On a problem of Sidon in additive number theory and some related problems, *J. London Math. Soc.* **16** (1941), 212-215.

[3] P. Erdős and P. Tetali, Representations of integers as a sum of k terms, *Random Structures and Algorithms* **1** (1990), 245-261.

Lecture 9

We begin with two further applications of the Chernoff bounds.

EXAMPLE 1 : DISCREPANCY THEORY

DEFINITION 24 : Let S be a finite set. A map $\chi : S \to \{\pm 1\}$ is called a 2-*coloring* of S. The elements $s \in S$ s.t. $\chi(s) = -1$ will be said to be colored *blue*, and the other points colored *red*.

DEFINITION 25 : Let S be a finite set, χ a 2-coloring of S and A a subset of S. The *discrepancy* of A with respect to χ, denoted $\mathrm{disc}(A, \chi)$, is defined as

$$\mathrm{disc}(A, \chi) := \left| \sum_{s \in A} \chi(s) \right| .$$

In words, it's the difference between the number of red and blue points in A.

DEFINITION 26 : Let $\mathcal{F}$ be a family of subsets of the finite set S. The *discrepancy* of $\mathcal{F}$ w.r.t. a 2-coloring χ of S, denoted $\mathrm{disc}(\mathcal{F}, \chi)$, is defined as

$$\mathrm{disc}(\mathcal{F}, \chi) := \max_{A \in \mathcal{F}} \mathrm{disc}(A, \chi).$$

The 2-*color discrepancy* of $\mathcal{F}$, denoted simply $\mathrm{disc}_2(\mathcal{F})$, is defined as

$$\mathrm{disc}_2(\mathcal{F}) := \min_{\chi} \mathrm{disc}(\mathcal{F}, \chi),$$

the minimum being taken over all possible 2-colorings of the set S.

The Chernoff estimates give upper bounds on 2-color discrepancies.

Theorem 32 *Let S be a finite set of m elements and $\mathcal{F}$ a collection of n subsets of S. Then*

$$disc_2(\mathcal{F}) = O\left(\sqrt{m \ln n}\right).$$

PROOF : A random 2-coloring of an m-set can obviously be thought of as a sequence of m independent coin tosses. Thus we have a very simple instance where the Chernoff bounds apply. Now let's be more precise :

Denote $S = \{1, ..., m\}$ for simplicity. For each $i = 1, ..., m$, let X_i be the random variable for which

$$P(X_i = +1) = P(X_i = -1) = \frac{1}{2}.$$

Thus the X_i are i.i.d. and by a random 2-coloring of S we mean any such i.i.d. sequence of m random variables. For any subset A of S, we set $X_A := \sum_{i \in A} X_i$. Thus the absolute value of X_A records the discrepancy of A w.r.t. a random 2-coloring of S. We shall show that, for an appropriately chosen constant $C > 0$, and for any fixed n and A,

$$P(|X_A| > C\sqrt{m \ln n}) < \frac{1}{n}. \tag{65}$$

This implies that, given a family $\mathcal{F}$ of n subsets, the total probabilility that $|X_A| > C\sqrt{m \ln n}$ for at least one $A \in \mathcal{F}$ is strictly less than one. In other words, there is a positive probability that a random 2-coloring χ of S satisfies $\mathrm{disc}(\mathcal{F}, \chi) \leq C\sqrt{m \ln n}$, as desired. So it suffices to verify (65).

We use (51) in the special case where, in the notation of (49), each $p_i = 1/2$. Notice that each X_i above is twice such a normalised indicator variable, so (51) implies that

$$P(X_A < -a) < \exp\left[-\frac{(a/2)^2}{2 \cdot \frac{1}{2} \cdot |A|}\right] = \exp\left(-\frac{a^2}{4|A|}\right) \leq \exp\left(-\frac{a^2}{4m}\right).$$

But here everything is symmetric about zero, so the same inequality must hold for $P(X_A > +a)$, even if this is not generally the case in the Chernoff estimates. We conclude that, for any positive real number a,

$$P(|X_A| > a) < 2 \cdot \exp\left(-\frac{a^2}{4m}\right).$$

Setting $a := C\sqrt{m \ln n}$, this becomes

$$P(|X_A| > C\sqrt{m \ln n}) < \exp\left(-\frac{C^2 m \ln n}{4m}\right) = n^{\frac{-C^2}{4}}.$$

Thus (65) will be satisfied if $C > 2$ and the theorem is proved.

A particular case of interest in Theorem 32 is when $m = n$, in which case it bounds the discrepancy by $O(\sqrt{n \ln n})$. In Chapter 12 of [AS], Spencer reproduces his argument which improves this to $O(\sqrt{n})$. It is a highly non-trivial argument running over several pages, so we don't go through it here. Note, however, that there are examples known, involving so-called *Hadamard matrices*, which show that this order of magnitude cannot in general be beaten. The best-possible constant is, I think, still unknown. Active research areas within discrepancy theory include, for example :

(I) studying the discrepancy of specific families of sets, not just general ones. This is somewhat analogous to studying specific subsets of the natural numbers in the theory of bases. There is an old, famous result of this type due to Roth, which states that if $\mathcal{F}$ is the family of all arithmetic progressions (of all lengths) in $\{1, ..., n\}$, then $\text{disc}(\mathcal{F}) = \Omega(n^{1/4})$. More recently, Spencer and Matousek proved the reverse estimate, namely $\text{disc}(\mathcal{F}) = O(n^{1/4})$. See their paper [1] for details and references.

(II) extending the notion of discrepancy to when there are more than 2 colors involved, so-called *multi-colored discrepancies.* Of course here it's not even obvious what the right definitions should be. Search for 'multi-colored discrepancies' on Google if you're interested.

Example 2 : Degrees in random graphs

For any n and p, the degree of any vertex in $G(n,p)$ is the sum of $n-1$ i.i.d. indicator variables X_i such that $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Indeed each such indicator corresponds to an edge from the given vertex to one of the other $n-1$ vertices. Thus the expected value of the degree of any vertex is $(n-1)p$ and we expect that the Chernoff bounds would supply some kind of concentration estimate for the degrees about this average. Given n and p, and $\epsilon > 0$, let A_ϵ denote the event that the degree of every vertex in $G(n,p)$ lies between $(1-\epsilon)(n-1)p$ and $(1+\epsilon)(n-1)p$. Then we can prove the following :

Theorem 33 : *For any* $\epsilon > 0$, *if* $\frac{\ln n}{n} = o[p = p(n)]$ *then*
$P[G(n, p(n)) \models A_\epsilon] = 1 - o(1)$.

Remark This is kind of a 'threshold result'. It says that if $p(n)$ is above the threshold $\frac{\ln n}{n}$ then we get good concentration of the degrees. It says nothing, however, about what's going on below the threshold.

PROOF OF THEOREM 33 : Let ϵ be given and for any vertex v of K_n let X_v be the random variable which records the degree of v in $G(n,p)$. As explained above, X_v is a sum of $n-1$ indicator variables and has mean $\mu = (n-1)p$. Thus, by Theorem 24,

$$P(|X_v - \mu| > \epsilon\mu) < 2 \cdot e^{-c_\epsilon \mu}, \tag{66}$$

where c_ϵ s a fixed positive constant. Now A_ϵ is the event that $|X_v - \mu| \leq \epsilon\mu$ for every vertex v. Thus in order for the probability of this event to be $1 - o(1)$, it suffices for the right hand side of (66) to be $o(1/n)$. But this is the case if $\frac{\ln n}{n} = o(p)$, as one verifies by direct insertion.

How to deal with non-independence ?

We now come to the last part of the course, where we introduce some techniques for dealing with non-independent events. We will only have time for two topics, namely :

(I) The Lovasz Local Lemma.
(II) Martingales.

The two topics are quite diefferent in spirit, though, so give a good flavour of the range of techniques at one's disposal.

The Local Lemma is basically an attempt to generalise the following simple observation :

Proposition 34 *Let $A_1, ..., A_n$ be a finite sequence of events in an arbitrary probability space. If each event has non-zero probability and they are independent of one another, then with positive probability, all occur simoultaneously.*

PROOF : By independence,

$$P\left(\bigwedge_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) > 0.$$

In the probabilistic approach to many combinatorial problems, as we have seen, we are interested in having none of a sequence of events occurring. If we call the events A_i and $P(A_i) = x_i$, then if the events were independent, the above proof would yield that

$$P\left(\bigwedge_i \overline{A}_i\right) = \prod_i (1 - x_i). \tag{67}$$

So as long as each $x_i < 1$, the probability of none of these 'bad' events occurring is non-zero. Moreover, in applications, the x_i are usually very small, so sometimes this would even give a fairly large probability of none of the bad events occurring. But lack of independence is a major problem which has to be dealt with.

The idea behind the local lemma is that we can recover something close to (67) if the dependencies between events are *localised*, which usually means that each event is independent of 'most' of the others. For the precise statement, we require a definition :

DEFINITION 27 : Let $A_1, ..., A_n$ be events in a probability space. The *dependency digraph* of these events is the directed graph G on n vertices, such that the directed edge (i, j) is present if and only if event A_i depends on event A_j, that is if and only if $P(A_i|A_j) \neq P(A_i)$.

Theorem 35 (Lovasz Local Lemma 1975) *Let $A_1, ..., A_n$ be a finite sequence of events in an arbitrary probability space with dependency digraph G. Suppose we can find real numbers $x_1, ..., x_n$ such that*
(i)

$$0 < x_i < 1, \quad i = 1, ..., n,$$

(ii)

$$P(A_i) \leq x_i \cdot \prod_{j:(i,j)\in E(G)} (1 - x_j), \quad i = 1, ..., n.$$

Then

$$P\left(\bigwedge_{i=1}^n \overline{A}_i\right) \geq \prod_{i=1}^n (1 - x_i) > 0. \tag{68}$$

The idea of localised dependencies becomes very clear in the following speical case of the theorem :

Corollary 36 (Symmetric Local Lemma) *Let $A_1, ..., A_n$ be events in an arbitrary probability space. If $P(A_i) \leq p$ for every i, each event depends on at most d others and*

$$e \cdot p \cdot (d+1) < 1, \tag{69}$$

then $P(\wedge_i \overline{A}_i) > 0$.

Proof of Corollary : Take each $x_i = \frac{1}{d+1}$ in the statement of Theorem 35 and verify that (69) implies that conditions (i) and (ii) of the theorem are satisfied.

We will prove Theorem 35 and discuss applications next day. One final remark : perhaps the best way to think of it is that the product over edges of G in condition (ii) is a 'correction term' which one inserts in order to recover (67) when one has localised dependencies.

Reference

[1] J. Matousek and J. Spencer, Discrepancy in arithmetic progressions, *J. Amer. Math. Soc.* **9** (1996), 195-204.

Lecture 10

PROOF OF THEOREM 35 : The proof is quite elementary, but a bit of a notational nightmare. We need a lemma which is basically a reformulation of the inclusion-exclusion principle.

Lemma 37 *Let $A_1, ..., A_n$ be events in a probability space. Then*

$$P\left(\bigwedge_{i=1}^{n} \overline{A}_i\right) = \prod_{i=1}^{n} \left[1 - P\left(A_i | \bigwedge_{j=1}^{i-1} \overline{A}_j\right)\right]. \tag{70}$$

PROOF OF LEMMA : Induction on n, starting with $n = 2$. For two events, (70) becomes

$$P(\overline{A}_1 \wedge \overline{A}_2) = [1 - P(A_1)][1 - P(A_2|\overline{A}_1)]. \tag{71}$$

Now the induction step basically just consists of applying (71) over and over, so we just prove that relation and leave the rest as an exercise.

irst note that the left hand side is $1 - P(A_1 \vee A_2)$ which in turn, by the inclusion-exclusion principle, is $1 - P(A_1) - P(A_2) + P(A_1 \wedge A_2)$. Expanding the right hand side of (71) and cancelling like terms, we see that what's left to be shown is that

$$P(A_2) - P(A_1 \wedge A_2) = [1 - P(A_1)] \cdot P(A_2|\overline{A}_1).$$

But this is true since

$$P(A_2) - P(A_1 \wedge A_2) = P(\overline{A}_1 \wedge A_2) = P(\overline{A}_1) \cdot P(A_2|\overline{A}_1) = [1 - P(A_1)] \cdot P(A_2|\overline{A}_1), \quad \text{v.s.v.}$$

Now back to the theorem. We suppose that we have located real numbers $x_1, ..., x_n$ s.t. conditions (i) and (ii) are satisfied, and must prove (68). The strategy will be to prove the following statement, which I'll call (*) :

'For every subset S of $\{1, ..., n\}$ other than the whole set, and each $i \notin S$, it holds that

$$P\left(A_i | \bigwedge_{j \in S} \overline{A}_j\right) \leq x_i. \tag{72}$$

First suppose we have (*). Then (69) follows directly from this and (70). The left hand side of (70) is what we're interested in and, by (*), for each term on the right hand side of (70) we have, upon taking $S = \{1, ..., i-1\}$, the inequality

$$1 - P\left(A_i | \bigwedge_{j=1}^{i-1} \overline{A}_j\right) \geq 1 - x_i.$$

So we're left having to verify (*). This will be done by induction on $|S|$. If S is the empty set, then it is follows from condition (ii). Suppose now $|S| > 0$ and that (*) holds for all smaller S. Fix also an $i \notin S$. We form a partition $S = S_1 \sqcup S_2$ where $S_1 := \{j \in S : (i,j) \in E(G)\}$. The left hand side of (72) is thus

$$P\left(A_i | \bigwedge_{j \in S_1} \overline{A}_j \wedge \bigwedge_{j \in S_2} \overline{A}_j\right).$$

Now for any three events X, Y and Z we have the relation[20]

$$P(X|Y \wedge Z) = \frac{P(X \wedge Y|Z)}{P(Y|Z)} \leq \frac{P(X|Z)}{P(Y|Z)}.$$

Taking

$$X = A_i, \quad Y = \bigwedge_{j \in S_1} \overline{A}_j, \quad Z = \bigwedge_{j \in S_2} \overline{A}_j,$$

we thus obtain the inequaliity

$$P\left(A_i | \bigwedge_{j \in S} \overline{A}_j\right) \leq \frac{P\left(A_i | \bigwedge_{j \in S_2} \overline{A}_j\right)}{P\left(\bigwedge_{j \in S_1} \overline{A}_j | \bigwedge_{j \in S_2} \overline{A}_j\right)}. \tag{73}$$

We consider separately the numerator and denominator on the right of (73). Regarding the numerator, since A_i is, by definition, independent of all events A_j for $j \in S_2$, the probability of A_i occurring is unaffected by the non-occurrence of these other events. Thus the numerator is simply $P(A_i)$ and hence, by assumption (ii),

$$\text{Numerator} \leq x_i \cdot \prod_{j:(i,j) \in E(G)} (1 - x_j). \tag{74}$$

[20] it follows from the relation $P(X \wedge Y) = P(Y) \cdot P(X|Y)$, after conditioning everything on Z.

Now for the denominator. If the set S_1 is empty then the denominator is just one (strictly speaking, the event Y can be ignored completely) and we're done. So we may suppose S_1 is non-empty. This will allow us to employ the induction hypothesis on $|S|$. Let $S_1 := \{j_1, ..., j_r\}$ say. Then Lemma 37 implies that the denominator can be written as

$$\prod_{k=1}^{r} \left[1 - P\left(A_{j_k} | \bigwedge_{l=1}^{k-1} \overline{A}_{j_l} \wedge \bigwedge_{j \in S_2} \overline{A}_j \right) \right].$$

But this is a product of the form

$$\prod_{k=1}^{r} \left[1 - P\left(A_{j_k} | \bigwedge_{j \in S^k} \overline{A}_j \right) \right],$$

where each S^k is a set of smaller cardinality than S and $j_k \notin S^k$. Thus the induction hypothesis implies that the product is greater than or equal to $\prod_{k=1}^{r}(1 - x_{j_k})$. But this is a subset of the product $\prod_{j:(i,j) \in E(G)}(1 - x_j)$, hence bigger than or equal to that in turn. We conclude that

$$\text{Denominator} \geq \prod_{j:(i,j) \in E(G)} (1 - x_j). \tag{75}$$

From (74) and (75) we see that the induction step is complete, and hence the proof of Theorem 35.

We now give two applications of Theorem 35. The first applies the symmetric version (Corollary 36), which is indeed the version often used, but the second illustrates the use of the full theorem. Not surprisingly, the latter example is far more intricate.

First Application : Van der Waerden numbers

Earlier (Theorem 6) we applied a basic probabilistic argument to get a lower bound on Van der Waerden numbers. Thus bound can be substantially improved using the Local Lemma. In fact, the following result is essentially the best general lower bound for Van der Waerden numbers that is currently known.

Theorem 38 *Let $k, m \geq 1$. Then there exists a function $f(k)$, depending on k only, such that*

$$W(k,m) \geq \frac{k-1}{ek^2}m^{k-1} - f(k). \qquad (76)$$

PROOF : Fix an $n > 0$ and consider a random m-coloring of $\{1, ..., n\}$. Let $S_1, ..., S_t$ denote all the k-term AP:s in $\{1, ..., n\}$. Note that $t = \Theta(n^2)$. For $i = 1, ..., t$ let A_i be the event that S_i is monochromatic. Then $P(A_i) = m^{-(k-1)}$ for every i. So set $p := m^{-(k-1)}$. Now consider any event A_i. We need an upper bound on the number of other events A_j on which it depends. Clearly, A_i depends on A_j if and only if, as sets, S_i and S_j are not disjoint. Thus we need an upper bound on the number of k-term AP:s in $\{1, ..., n\}$ which intersect a given one. I claim that there is such an upper bound of the form $\frac{k^2}{k-1}n + f(k)$, where $f(k)$ is a function of k only, and not of n. To see this, fix a k-term AP - call it S - and consider separately

(i) the number of k-term AP:s which intersect S in exactly one point
(ii) the number of k-term AP:s which intersect S in at least two points.

First, what about (i) ? Well, there are k choices for the point of intersection. Given that point, there are k choices for its position in the intersecting AP, which we call T. There are also at most $\lfloor \frac{n}{k-1} \rfloor$ choices for the common difference between the terms of T. These three choices determine T uniquely, hence the number of possible T is no more than $k \cdot k \cdot \lfloor \frac{n}{k-1} \rfloor \leq \frac{k^2}{k-1}n$.

Next, what about (ii) ? Again denote an intersecting AP by T. There are certainly no more than 2^k choices for the points of intersection of S with T. But an AP is completely determined by specifying two or more of its points and their positions within it. For each choice of points in $S \cap T$ there are again certainly no more than 2^k possible choices for the positions of these points within T. Thus the total number of possibilities for T is certainly no more than $2^k \cdot 2^k = 4^k$. The important point is that this bound is a function of k only, so our claim is proved.

The theorem now follows immediately from Corollary 36. We've already chosen p and the above discussion means we can take $d := \frac{k^2}{k-1}n + f(k)$. Then (69) will be satisfied if $n \leq \frac{k-1}{ek^2}m^{k-1} + f_1(k)$, where f_1 is some (other) function of k only. Thus for any such n, there is a non-zero probability that a random m-coloring satisfies $P(\wedge \overline{A}_i) > 0$, i.e.: that there are no monochromatic AP:s. Hence this must be a lower bound for $W(k,m)$, v.s.v.

Second Application : Off-diagonal Ramsey numbers

We've also applied a basic probabilistic argument to obtain a lower bound (8) for the diagonal Ramsey numbers $R(k,k)$. It turns out that the Local Lemma can only improve on this by a constant factor (specifically, a factor of two), because each bad event (a red or blue K_k) depends on too many others. Where the Local Lemma does give a big improvement on the basic approach is for computing lower bounds for so-called *off-diagonal* Ramsey numbers. Here we are interested in the numbers $R(k,l)$, where k is considered fixed and l is allowed to grow. We thus consider $R(k,l)$ as a function of l. Now (2) gives an upper bound $R(k,l) = O_k(l^{k-1})$. It is conjectured that this is not far from the truth, namely :

Conjecture 39 *Fix $k \geq 3$. Then for each $\epsilon > 0$, as $l \to \infty$, $R(k,l) = \Omega(l^{k-1-\epsilon})$.*

Note that this is weaker than the assertion that $R(k,l) = \Theta_k(l^{k-1})$. In fact, the latter assertion is false since, for example, Erdős showed that

$$R(3,l) = O\left(\frac{l^2}{\ln l}\right). \tag{77}$$

Regarding lower bounds, Spencer [1] proved the following result using the Local Lemma :

Theorem 40 (Spencer 1977) *For each fixed $k \geq 3$, as $l \to \infty$ we have that*

$$R(k,l) = \Omega\left(\frac{l}{\ln l}\right)^{\alpha(k)}, \tag{78}$$

where

$$\alpha(k) = \frac{\binom{k}{2} - 1}{k-2} = \frac{k^2-k-2}{2(k-2)}. \tag{79}$$

Essentially $\alpha(k) \approx k/2$, which is about half of what Conjecture 39 says should be the right power. Note that for $k = 3$ we have $\alpha(3) = 2$ so Conjecture 39 has been proven in that case. But even here we still have the tantalising gap

$$c_1 \frac{l^2}{\ln^2 l} < R(3,l) < c_2 \frac{l^2}{\ln l}.$$

Indeed it is pretty weird that the Local Lemma leaves such a small gap when one considers that a basic probabilistic argument bascially gets you nowhere in terms of lower bounds. Let me explain what I mean.

OK, fix an n and consider a random 2-coloring of K_n. We want to avoid both blue triangles and red K_l. To get a lower bound for $R(3,l)$ we have to determine for which n such configurations are avoided with non-zero probability. Because of the asymmetry here between blue and red, it seems intuitively reasonable that our random coloring should in this case be biased in favour of red, i.e.: we color each edge of K_n blue with some probability p which is small and perhaps dependant on n. For any choice of p, the usual basic probabilistic argument tells us that the expected number of bad configurations (blue triangles or red K_l) is

$$\binom{n}{3} \cdot p^3 + \binom{n}{l} \cdot (1-p)^{l(l-1)/2}.$$

So what one now wants to do is to choose p so that the above quantity is less than one for as large n as possible, as a function of l, ideally for n of the order of l^2. Hoever, I leave it as an exercise for you to prove the following assertion (I went through it roughly in class) :

Proposition 41 *Fix $\epsilon > 0$. Then for any choice of $p = p(n)$ it will be the case that, if l is sufficiently large, then*

$$\binom{n}{3} \cdot p^3 + \binom{n}{l} \cdot (1-p)^{l(l-1)/2} > 1$$

when $n = l^{1+\epsilon}$.

Since the trivial lower bound is $R(3,l) \geq l$ we thus see the severe limitations of applying only simple-minded probabilistic techniques to this problem. Theorem 40 is therefore a very powerful illustration of the Local Lemma. The proof follows in the next lecture.

Reference

[1] J. Spencer, *Discrete Math.* **20** (1977), 69-76.

Lecture 11

PROOF OF THEOREM 40 : We give the proof for $k = 3$ only. The proof for general k is similar. Fix l and n, though think of these as parameters which are going to infinity. We consider a random 2-coloring of K_n where each edge is colored blue with probability $p = p(n)$, to be determined later, and red otherwise (if you prefer, we are working with the random graph $G(n, p(n))$). For each 3-element subset S of the vertices, let A_S denote the event that S is a blue triangle. For each l-element subset T of the vertices, let B_T denote the event that T is a red K_l. There are $\left(\begin{array}{c} n \\ 3 \end{array} \right)$ A-events and $\left(\begin{array}{c} n \\ l \end{array} \right)$ B-events. Each of the former has probability p^3, and each of the latter has probability $(1-p)^{l(l-1)/2}$. Denote by

N_{AA} = number of A-events on which a given A-event depends,
N_{AB} = number of B-events on which a given A-event depends,
N_{BA} = number of A-events on which a given B-event depends,
N_{BB} = number of B-events on which a given B-event depends.

We use the estimates

$$\begin{aligned} N_{AA} &= 3(n-3) < 3n, \\ N_{AB} &\leq \left(\begin{array}{c} n \\ l \end{array} \right) \leq \left(\frac{ne}{l}\right)^l, \\ N_{BA} = \left(\begin{array}{c} l \\ 2 \end{array} \right)(n-l) + \left(\begin{array}{c} l \\ 3 \end{array} \right) &< \frac{1}{2}l^2 n, \\ N_{BB} &\leq \left(\begin{array}{c} n \\ l \end{array} \right) \leq \left(\frac{ne}{l}\right)^l, \end{aligned}$$

where in the second and fourth estimates we have used Stirling's formula. We seek positive real numbers $x_i \in (0, 1)$, one for each A- and B-event, so that the hypotheses of Theorem 35 are satisfied when $n = \Theta\left(\frac{l^2}{\ln^2 l}\right)$. Since the A-events are identically distributed, it makes sense to confine the search to when all corresponding x_i are equal, to x say. Similarly, we assume that the x_i corresponding to all B-events are equal, to y say. Then the conditions that need to be satisfied can be summarised as follows :

$$0 < p < 1, \quad 0 < x < 1, \quad 0 < y < 1, \tag{80}$$

$$p^3 \leq x \cdot (1-x)^{N_{AA}} \cdot (1-y)^{N_{AB}}, \tag{81}$$

$$(1-p)^{l(l-1)/2} \leq y \cdot (1-x)^{N_{BA}} \cdot (1-y)^{N_{BB}}. \tag{82}$$

From our estimates for the various N's, it suffices to replace (81) and (82) respectively by

$$p^3 \leq x \cdot (1-x)^{3n} \cdot (1-y)^{\left(\frac{ne}{l}\right)^l}, \tag{83}$$

$$(1-p)^{l(l-1)/2} \leq y \cdot (1-x)^{\frac{l^2 n}{2}} \cdot (1-y)^{\left(\frac{ne}{l}\right)^l}. \tag{84}$$

Now the claim is that, for an appropriate choice of absolute constants c_1, c_2, c_2, c_3, c_4, then (80), (83) and (84) will be satisfied for all sufficiently large n and the following choice of parameters :

$$p = c_1 n^{-1/2}, \tag{85}$$

$$l = c_2 n^{1/2} \ln n, \tag{86}$$

$$x = c_3 n^{-3/2}, \tag{87}$$

$$y = \exp(-c_4 n^{1/2} \ln^2 n). \tag{88}$$

It is left as a (tedious, but worthwhile) exercise to verify this. The important point is that (86) implies the theorem.

Martingales

A martingale is, from a certain point of view, a generalisation of a sequence of i.i.d. variables, but the concept is far more general. In 1968 Azuma observed that for martingales that satisfy a certain so-called *Lipschitz condition*, the final term in the martingale satisfies the same type of Chernoff concentration estimate as a sum of i.i.d. variables. The result has applications to random graphs, as there is a natural way to associate a martingale to any random graph invariant, and for some invariants, the best-known being the chromatic number, the Lipschitz condition is satisfied.

In order to be able to present this material, we need to introduce the notion of *conditional expectation* for random variables. In keeping with our general philosophy in this course, we keep the abstract probability theory to a minimum sufficient for our requirements.

DEFINITION 28 : Let (Ω, μ) be a finite probability space, X a real-valued random variable on Ω. For each $r \in \mathbf{R}$, the *level set of X at level r*, denoted $\mathcal{B}_r$, is defined as

$$\mathcal{B}_r := \{\omega \in \Omega : X(\omega) = r\}.$$

Definition 29 : Now let Y be another random variable on the same space. We can define a third r.v. Z, called the *conditional expectation of* Y *w.r.t.* X, and usually denoted $E(Y|X)$, as follows : for each $\omega \in \Omega$, we have

$$Z(\omega) := \frac{1}{\mu(\mathcal{B}_{X(\omega)})} \sum_{\tau \in \mathcal{B}_{X(\omega)}} \mu(\tau) Y(\tau).$$

In words, the value of the r.v. $E(Y|X)$ at any point ω in the probability space is the μ-weighted average of the values of Y at the points of the level set of $X(\omega)$. Thus $E(Y|X)$ is constant on each level set of X. If each level set of X is a single point, then $E(Y|X) = Y$. Otherwise, $E(Y|X)$ is a 'partial revelation' of the r.v. Y.

Exercises (i) Describe $E(Y|X)$ more explicitly when X and Y are indicator variables of events, say A and B respectively.
(ii) Show that $E[E(Y|X)] = E(Y)$.
(iii) Show that (we'll use this later on)

$$E[X \cdot E(Y|X)] = E[X \cdot Y]. \tag{89}$$

Definition 30 : A sequence $X_0, ..., X_n$ of random variables, all defined on the same probability space, is called a *martingale* if

$$E(X_{i+1}|X_i) = X_i, \quad \text{for } i = 0, ..., n-1.$$

Example 1 : Let $Y_0, ..., Y_n$ be i.i.d. variables on a space (Ω, μ). For each $i = 0, ..., n$ set $X_i := \sum_{j=0}^{i} Y_j$. The X_i may all be considered as defined on the same space, namely Ω^{n+1} with the product measure. Then the X_i form a martingale (exercise !).

Example 2 : The mathematical use of the term 'martingale' historically comes from the following example : consider a game which consists of an unlimited (i.e.: continue until you get fed up) sequence of coin tosses, where the amount bet on the outcome of each toss is decided independently just before it takes place. Consider the following strategy for winning : 'double the bet until I win'. So, for example, you could start by betting 1 euro. If you win, stop. Otherwise, bet 2 euro on the next toss. If you win then, stop. Otherwise, bet 4 euro on the next toss etc. One might reason that since one must surely win a bet at some point, this is a guaranteed money-making strategy.

Exercise : Show how to model this game with a martingale. What's the flaw in the reasoning above ?

One type of martingale which arises in many contexts is where the last term X_n is a r.v. whose distribution is being 'gradually revealed' by the terms in the martingale. An example, which is the central example of interest for applications to random graphs, will hopefully make this idea clear :

DEFINITION 31 : We work in the probability space $G(n,p)$ for any fixed n and p. Let f be any graph invariant. Let $e_1, ..., e_{n(n-1)/2}$ be any ordering of the edges of K_n. We define a corresponding martingale $X_0, ..., X_{n(n-1)/2}$, called the *edge exposure martingale* of f in $G(n,p)$, as follows :

For each $i = 0, ..., n(n-1)/2$, X_i is the random variable on $G(n,p)$ whose value at any graph H on n vertices is the average value of the function f taken over all graphs G on n vertices which coincide with H amongst the edges $e_1, ..., e_i$.

The *vertex exposure martingale* is defined similarly. Here we order the vertices of K_n in any order, say $v_1, ..., v_n$. Then our martingale is $X_0, ..., X_{n-1}$, where $X_i(H)$ is the average of $f(G)$ taken over all G which coincide with H on the subgraph induced by $v_1, ..., v_{i+1}$.

Note that the vertex esposure martingale may be considered as a subsequence of the edge exposure martingale.

N.B.: In either the edge- or vertex exposure martingale, the first term X_0 is a constant, namely $E[f(G(n,p))]$, whereas the last term (either $X_{n(n-1)/2}$ or X_{n-1} as appropriate) is $f(G(n,p))$ itself, i.e.: the random graph invariant in its full glory !!

EXAMPLE : $f(G) = \chi(G)$, the chromatic number. We computed the corresponding martingales for $G(3, 1/2)$. Left as an exercise for you to do so again.

In the next lecture we will formulate the key property of a random graph invariant which means that we can get a Chernoff-type concentration estimate for it from the corresponding edge- or vertex exposure martingale.

Lecture 12

Our purpose in this last lecture is to show how the symmetric case of Chernoff's inequality (see the proof of Theorem 32) can be extended to a certain class of martingales, with basically the same proof. The important concept is the following :

DEFINITION 32 : A martingale $X_0, ..., X_n$ is said to satisfy a *Lipschitz condition* if there exists a constant $c > 0$ such that $|X_i - X_{i-1}| \leq c$ for all $i = 1, ..., n$.

DEFINITION 33 : Let f be a graph invariant. Then f is said to satisfy an *edge (resp. vertex) Lipschitz condition* if there exists a constant $c > 0$ such that, whenever G_1 and G_2 are two graphs that differ only at one edge (resp. vertex), then $|f(G_1) - f(G_2)| \leq c$.

Note that, in this definition, when we say that two graphs differ only at one edge, then we mean that the two graphs have the same number of vertices, and that they share exactly the same edges but one, which is present in one graph but not the other. Thus one of the graphs is a subgraph of the other in this case. When we say that two graphs differ at one vertex, we mean that, when that vertex and all its adjacent edges are removed, then the remaining graphs are identical. Thus, if two graphs only differ at one edge then they also only differ at one vertex, though not always vice versa.

Exercise Show that if f is a graph invariant satisfying an edge (resp. vertex) Lipschitz condition, then the corresponding edge (resp. vertex) exposure martingale satisfies a Lipschitz condition with the same constant.

The point of these definitions is the following :

Theorem 42 (Azuma's inequality 1968) *Let $\mu = X_0, ..., X_n$ be a martingale satisfying a Lipschitz condition with constant $c > 0$. Then, for any $a > 0$,*

$$P(|X_n - \mu| > a) \leq 2 \cdot \exp\left(-\frac{a^2}{2nc^2}\right). \tag{90}$$

PROOF : We prove the result in the case where $c = 1$ and $\mu = 0$. The general result just follows by simple change of variables. We will need the following lemma :

Lemma 43 *Let Y be a r.v. satisfying*

(i) $E[Y] = 0$,

(ii) $|Y| \leq 1$.

Then for any $\lambda > 0$,

$$E[e^{\lambda Y}] \leq e^{\frac{\lambda^2}{2}}.$$

PROOF OF LEMMA : The function $f(x) = e^{\lambda x}$ is convex, hence its graph in the interval $[-1, 1]$ lies on or below the line joining the points $(-1, f(-1)) = (-1, e^{-\lambda})$ and $(1, f(1)) = (1, e^{\lambda})$. In other words, for $x \in [-1, 1]$,

$$e^{\lambda x} \leq \cosh \lambda + \sinh \lambda \cdot x.$$

Hence, by assumptions (i) and (ii) and linearity of expectation,

$$E[e^{\lambda Y}] \leq E[\cosh \lambda + \sinh \lambda \cdot Y] = \cosh \lambda.$$

But it's simple to check that $\cosh \lambda \leq e^{\frac{\lambda^2}{2}}$ for all $\lambda > 0$, which completes the proof of the lemma.

So back to the theorem. For each $i = 1, ..., n$ let $Y_i := X_i - X_{i-1}$. Then the martingale condition implies that $E(Y_i | X_{i-1}) = 0$ and the Lipschitz condition that $|Y_i| \leq 1$. Thus, by Lemma 42, if $\lambda > 0$ then

$$E[e^{\lambda Y_i} | X_{i-1}] \leq e^{\frac{\lambda^2}{2}}, \quad \text{for } i = 1, ..., n. \tag{91}$$

We are, of course, interested in X_n so, in the spirit of Chernoff's method, we consider $E[e^{\lambda X_n}]$ for some λ to be chosen appropriately later. Observe that

$$X_i = Y_1 + \cdots + Y_i, \quad \text{for } i = 1, ..., n. \tag{92}$$

Thus

$$E[e^{\lambda X_n}] = E\left[\prod_{j=1}^{n} e^{\lambda Y_j}\right].$$

Now applying (89), (92) and Lemma 42 we have that

$$E\left[\prod_{j=1}^{n} e^{\lambda Y_j}\right] = E\left[\prod_{j=1}^{n-1} e^{\lambda Y_j} \cdot e^{\lambda Y_n}\right] = E\left[\prod_{j=1}^{n-1} e^{\lambda Y_j} \cdot E[e^{\lambda Y_n} | X_{n-1}]\right] \leq E\left[\prod_{j=1}^{n-1} e^{\lambda Y_j}\right] \cdot e^{\frac{\lambda^2}{2}}.$$

Now just apply the same argument a further $n-1$ times to get

$$E[e^{\lambda X_n}] \leq e^{\frac{n\lambda^2}{2}}.$$

Then, by Markov's inequality, if $a > 0$ we have

$$P(|X_n| > a) = 2 \cdot P(X_n > a) = 2 \cdot P(e^{\lambda X_n} > e^{\lambda a}) \leq 2 \cdot e^{\frac{n\lambda^2}{2} - \lambda a}.$$

The exponent is minimised when $\lambda = a/n$, which yields (90) when $c = 1, \mu = 0$. This completes the proof.

The chromatic number is a classic example of a graph invariant which satisfies a Lipschitz condition : clearly, it satisfies a vertex Lipschitz condition with $c = 1$. Various applications of Azuma's inequality to the computation of the chromatic numbers of random graphs are given in Chapter 7 of [AS]. Unfortunately, time ran out before we could do justice to these in the lecture, so you'll have to read the stuff yourselves !