Lecture 11 (Dec. 6, 2011)

We now present a selection of arithmetic and combinatorial applications of Chernoff bounds.

Example 1 : Thin bases

We begin with the requisite definitions :

DEFINITION 22 : Let h be a positive integer and A a subset of $\mathbb{N}_0$. The *representation function of A of order h*, denoted $r_{h,A}$, is the non-negative integer valued function on $\mathbb{N}_0$ 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 $A \subseteq \mathbb{N}_0$. A is said to be a *basis of order h* if $r_{h,A}(n) > 0$ for all $n \in \mathbb{N}_0$. More usefully, A is said to be an *asymptotic basis* of order h if $r_{A,h}(n) > 0$ for all sufficiently large n.

The case $h = 1$ is totally uninteresting, since then a subset of $\mathbb{N}_0$ is a (asymptotic) basis if and only if its complement is empty (resp. 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 $\mathbb{N}_0$ is a (asymptotic) basis of a certain order. There are two examples which everyone likes to quote :

(i) $A_1 := \{\text{set of primes}\} \cup \{0, 1\}$,

(ii) $A_{2,k} := \{n^k : n \in \mathbb{N}_0\}$, for any fixed $k > 1$.

Regarding (i), the state of the art is

Theorem 33 (Vinogradov 1937) *Every sufficiently large odd number is a sum of at most three primes. Hence, the set A_1 is an asymptotic 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, A_1 is a basis of order 3.*

Regarding (ii),

Theorem 34 (Hilbert 1909, Hardy-Littlewood 192x) *For every $k > 1$ the set $A_{2,k}$ is a basis of some order, depending on k.*

The problem to which Theorem 34 is the solution is commonly known as *Waring's Problem*. One denotes by $g(k)$ (resp. $G(k)$) the smallest integer such that the k:th powers are a basis (resp. asymptotic basis) of order $g(k)$ (resp. $G(k)$). It turns out that $G(k)$ is much smaller than $g(k)$ for large k and it is the more interesting function of the two. 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 $G(2) = g(2) = 4$. It is known that $4 \leq G(3) \leq 7$ 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 35 *There is no infinite subset A of $\mathbb{N}$ for which the representation function $r_{2,A}(n)$ is constant for all sufficiently large n.*

Proof : Suppose the contrary and let A be an asymptotic 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[1]

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

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

[1] 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.

between the generating function and the representation function is that

$$[F(z)]^2 + F(z^2) = 2 \cdot \sum_{n=1}^{\infty} r_{2,A}(n) z^n. \tag{73}$$

Suppose now that $r_{2,A}(n) = k$ for all $n \geq n_0$. Then (73) can be written as

$$[F(z)]^2 + F(z^2) = 2 \cdot \sum_{n=1}^{n_0-1} r_{2,A}(n) z^n + 2k \cdot \sum_{n=n_0}^{\infty} z^n. \tag{74}$$

The first sum on the right of (74) 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

$$[F(z)]^2 + F(z^2) = P(z) + 2k \cdot \frac{z^{n_0}}{1-z}. \tag{75}$$

The desired contradiction is obtained by seeing what happens as $z \to -1$ 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 $P(-1) + (-1)^{n_0} \cdot k$. This contradiction completes the proof.

The following problem, originally posed by Erdős in [2], remains after 70 years the most important (in my opinion) unsolved problem in additive number theory :

Open Problem *Does there exist a constant $C > 0$ and a basis A of order 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 less than ϵ. Informally, the conjecture asserts that if the set A "covers" $\mathbb{N}$ at least once under addition, then it has to do so with a considerable amount of room to spare.

Informally, a basis of a certain order is called *thin* if its representation function is a slowly growing function[2] of n. Classical bases like the sets A_1 and $A_{2,k}$ from earlier are very thick (exercise !), there is an awful lot of redundancy. 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 :

[2]There is another way to define thinness, namely in terms of the density of the set A itself. The two points of view are not entirely equivalent, indeed there are subtle differences which can make the same type of question much easier or harder depending on the viewpoint taken. However, we won't waste time here getting deeper into these matters.

Theorem 36 (Erdős 1956) *There exist bases A of order 2 for which*

$$r_{2,A}(n) = \Theta(\log n). \tag{76}$$

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 36 is a beautiful application of the Chernoff bounds.

PROOF : First of all, note that it suffices to prove the existence of an asymptotic basis A with property (76), as then we can just add a finite number of elements to A to make it into a basis, without affecting the order of magnitude of the representation function.

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

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

We will show that, for an appropriate choice of K, the representation function of A satisfies (76) with probability one[3]. For each $n > 0$, let X_n denote the random variable $r_{2,A}(n)$. Note that

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

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

$$\mu_n = \sum_{x=1}^{\lfloor n/2 \rfloor} \min\left\{K\sqrt{\frac{\log x}{x}}, 1\right\} \cdot \min\left\{K\sqrt{\frac{\log(n-x)}{n-x}}, 1\right\}. \tag{77}$$

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.

[3]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 $\mathbb{N}$, hence there are uncountably many possibilities for it.

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

$$\mu_n \sim \frac{K^2\pi}{2}\log n. \tag{78}$$

The verification of (78) 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 (78). 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 $(1-\epsilon)\frac{K^2\pi}{2}\log n$ and $(1+\epsilon)\frac{K^2\pi}{2}\log n$. Theorem 31 now tells us that

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

If K is now chosen so that

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

then

$$\sum_{n=1}^{\infty}\mathbb{P}(A_n) < \infty. \tag{79}$$

The theorem will then follow directly from

Lemma 37 (Borel-Cantelli Lemma) *Let $(A_n)_{n=1}^{\infty}$ be a sequence of events in a probability space, and suppose that (79) 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. (79) implies that there exists an n_0 such that

$$\sum_{n=n_0}^{\infty}\mathbb{P}(A_n) < \epsilon. \tag{80}$$

But the left hand side of (80) 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 (76) 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 38 (Erdős, Tetali 1990 [3]) *Let $h \geq 2$. Then there exists a basis A of order h for which $r_{h,A}(n) = \Theta(\log n)$.*

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 *Jansson inequalities*, proven by Svante Jansson in the late 1980s. These are discussed in Chapter 8 of [AS], and in [3] itself, but we won't have time to get that far in this course. The Janson inequalities will be a topic for the follow-up course in the spring.

References

[1] R.C. Vaughan, *The Hardy-Littlewood Method (2nd edition)*, Cambridge University Press (1997).
[2] P. Erdős and P. Turán, *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 Algorithms **1** (1990), 245-261.

Appendix 1

We need to prove (78). We start from (77). Clearly, there will be a bounded number of terms in this sum which are equal to 1. In all other terms, the minimum in (77) will be a function of x. Hence,

$$\mu_n \sim K^2 \sum_{x=1}^{\lfloor n/2 \rfloor} \sqrt{\frac{\log x \log(n-x)}{x(n-x)}}. \tag{81}$$

Note that the summand above is symmetric about $n/2$. Hence, in order to establish (78), it remains to prove that

$$\sum_{x=1}^{n-1} \sqrt{\frac{\log x \log(n-x)}{x(n-x)}} \sim \pi \log n. \tag{82}$$

Applying the standard trick of replacing the sum by an integral, we will show instead that[4]

$$\int_1^{n-1} \sqrt{\frac{\log x \log(n-x)}{x(n-x)}}\, dx \sim \pi \log n. \tag{83}$$

[4]It needs to be justified that replacing the sum by the integral does not lead to a significant error in our estimates. It is easy to see that the error will be $o(\log n)$. A rigorous proof is technical, and hence I omit it, though if you read through the rest of the calculations presented here, you should be able to see how to do it.

We change variables $x := \xi n$, and are left with having to show that

$$\int_{1/n}^{1-1/n} \sqrt{\frac{\log(\xi n)\log[(1-\xi)n]}{\xi(1-\xi)}}\, d\xi \sim \pi \log n. \tag{84}$$

At this point, we need a bit of calculus :

Lemma 39

$$\int_0^1 \frac{d\xi}{\sqrt{\xi(1-\xi)}} = \pi. \tag{85}$$

In particular, the integral converges, hence

$$\lim_{\delta\to 0}\int_0^\delta \frac{d\xi}{\sqrt{\xi(1-\xi)}} = \lim_{\delta\to 0}\int_{1-\delta}^1 \frac{d\xi}{\sqrt{\xi(1-\xi)}} = 0. \tag{86}$$

The second assertion of the lemma follows from the first (note that the integrand is symmetric about $\xi = 1/2$), and the first is proven by making the trigonometric substitution $\xi := \sin^2\theta$.

So back to proving (84). Let δ be a small positive number. At the end we will let $\delta \to 0$. Divide up the integral in (84) into three parts, (i) from 0 to δ, (ii) from δ to $1-\delta$, (iii) from $1-\delta$ to 1. Call these three sub-integrals I_1, I_2 and I_3 respectively. Now, for any fixed $\xi \in (0,1)$, we have

$$\begin{aligned}\log(\xi n)\log[(1-\xi)n] &= (\log n + \log\xi)(\log n + \log(1-\xi)) \\ &= (\log n + O(1))(\log n + O(1)) \sim (\log n)^2,\end{aligned} \tag{87}$$

so that the numerator of the integrand in (84) is $\sim \log n$. From this and Lemma 39, it follows easily that, as $\delta \to 0$,

$$I_1 \preceq (\log n)\cdot \int_0^\delta \frac{d\xi}{\sqrt{\xi(1-\xi)}} = o(\log n), \tag{88}$$

$$I_2 \sim (\log n)\cdot \int_\delta^{1-\delta} \frac{d\xi}{\sqrt{\xi(1-\xi)}} \sim \pi \log n, \tag{89}$$

$$I_3 \preceq (\log n)\cdot \int_{1-\delta}^1 \frac{d\xi}{\sqrt{\xi(1-\xi)}} = o(\log n). \tag{90}$$

Eq. (84) follows immediately from these estimates.

Example 2 : 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 40 *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

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

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,

$$\mathbb{P}(|X_A| > C\sqrt{m \ln n}) < \frac{1}{n}. \tag{91}$$

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 (91).

We use (66) in the special case where, in the notation of (64), each $p_i = 1/2$. Notice that each X_i above is twice such a normalised indicator variable, so (66) implies that

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

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

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

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

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

Thus (91) will be satisfied if $C > 2$ and the theorem is proved.

A particular case of interest in Theorem 40 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 $\mathrm{disc}(\mathcal{F}) = \Omega(n^{1/4})$. More recently, Spencer and Matousek proved the reverse estimate, namely $\mathrm{disc}(\mathcal{F}) = O(n^{1/4})$. See their paper [4] for details and references.

(II) extending the notion of discrepancy to when there are more than two 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.

(III) *Geometric discrepancy theory.* There are more geometrical analogues of the notion of discrepancy, where one is interested in random distributions of points in space. We will not say any more about this topic here.

Reference

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

Example 3 : 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 $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1-p$. 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 41 *For any $\epsilon > 0$, if $\frac{\ln n}{n} = o[p = p(n)]$ then*
$\mathbb{P}[G(n,p(n)) \models A_\epsilon] = 1 - o(1)$.

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 41 : 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 31,

$$\mathbb{P}(|X_v - \mu| > \epsilon\mu) < 2 \cdot e^{-c_\epsilon \mu}, \tag{92}$$

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 (92) to be $o(1/n)$. But this is the case if $\frac{\ln n}{n} = o(p)$, as one verifies by direct insertion.

Lecture 12 (Dec. 8, 2011)

This lecture is concerned with martingales, which are formally defined below (Definition 29). 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 27 : Let (Ω, μ) be a finite probability space, X a real-valued random variable on Ω. For each $r \in \mathbb{R}$, 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 28 : 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[5] $\mathcal{E}(Y|X)$, 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. $\mathcal{E}(Y|X)$ 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 $\mathcal{E}(Y|X)$ is constant on each level set of X. If each level set of X is a single point, then $\mathcal{E}(Y|X) = Y$. At the other extreme, if X is a constant function, then $\mathcal{E}(Y|X)$ is also constant, equal to $\mathbb{E}(Y)$. In between these two extremes, $\mathcal{E}(Y|X)$ is a "partial revelation" of the r.v. Y.

Proposition 42 *(i) Suppose X and Y are indicator variables of the events A and B respectively. Let $Z := \mathcal{E}(Y|X)$. Then for $\omega \in \Omega$,*

$$Z(\omega) = \begin{cases} \mathbb{P}(B|A) = \frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(A)}, & \text{if } \omega \in A, \\ \mathbb{P}(B|A^c) = \frac{\mathbb{P}(A^c \wedge B)}{1 - \mathbb{P}(A)}, & \text{if } \omega \notin A. \end{cases} \tag{93}$$

(ii) For any X and Y one has

$$\mathbb{E}[\mathcal{E}(Y|X)] = \mathbb{E}(Y). \tag{94}$$

[5] To avoid confusion, we use different notation to distinguish between an expected value $\mathbb{E}$, which is a number, and a conditional expectation $\mathcal{E}$, which is a random variable. The two things are related, obviously - see Proposition 42 and the discussion preceeding it.

(iii) For any X and Y, one has

$$\mathbb{E}[X \cdot \mathcal{E}(Y|X)] = \mathbb{E}[X \cdot Y]. \tag{95}$$

PROOF : Left as an exercise. Note that the $\cdot$ in part (iii) just means an ordinary product of functions. This part of the proposition will be used in the proof of Theorem 43 below.

DEFINITION 29 : A sequence $X_0, ..., X_n$ of random variables, all defined on the same probability space, is called a *martingale* if

$$\mathcal{E}(X_{i+1}|X_i, X_{i-1}, ..., X_0) = X_i, \quad \text{for } i = 0, ..., n-1.$$

EXAMPLE 1 : Let $Y_0, ..., Y_n$ be i.i.d. variables on a space (Ω, μ), such that $\mathbb{E}(Y_0) = 0$. 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 30 : 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 exposure 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 $\mathbb{E}[f(G(n,p))]$, 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. As an exercise, compute the corresponding edge- and vertex-exposure martingales for $G(3, 1/2)$.

We shall now show how the symmetric case of Chernoff's inequality, concerning a sum of ± 1 i.i.d. indicator variables (see the proof of Theorem 40), can be extended to a certain class of martingales, with basically the same proof. The important concept is the following :

DEFINITION 31 : 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 32 : 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 43 (Azuma's inequality 1968) *Let $\mu = X_0, ..., X_n$ be a martingale satisfying a Lipschitz condition with constant $c > 0$. Then, for any*

$a > 0$,

$$\mathbb{P}(|X_n - \mu| > a) \leq 2 \cdot \exp\left(-\frac{a^2}{2nc^2}\right). \tag{96}$$

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 lemmas :

Lemma 44 *If* $\lambda > 0$ *then* $\cosh\lambda \leq e^{\lambda^2/2}$, *where* $\cosh\lambda = \frac{1}{2}(e^{\lambda} + e^{-\lambda})$.

PROOF OF LEMMA 44 : Taylor expansions.

Lemma 45 *Let* Y *be a r.v. satisfying*
(i) $\mathbb{E}[Y] = 0$,
(ii) $|Y| \leq 1$.
Then for any $\lambda > 0$,

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

PROOF OF LEMMA 45 : 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,

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

But now use Lemma 44 to complete the proof.

So back to the theorem. For each $i = 1, ..., n$ let $Y_i := X_i - X_{i-1}$. Then the martingale condition implies that $\mathcal{E}(Y_i|X_{i-1}) = 0$ and the Lipschitz condition that $|Y_i| \leq 1$. Thus, by Lemma 45, if $\lambda > 0$ then

$$\mathcal{E}[e^{\lambda Y_i}|X_{i-1}] \leq e^{\frac{\lambda^2}{2}}, \quad \text{for } i = 1, ..., n. \tag{97}$$

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

$$X_i = Y_1 + \cdots + Y_i, \quad \text{for } i = 1, ..., n. \tag{98}$$

Thus

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

Now applying (95), (98) and Lemma 45 we have that

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

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

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

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

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

The exponent is minimised when $\lambda = a/n$, hence

$$\mathbb{P}(X_n > a) \leq e^{-a^2/2n}.$$

Now $(-X_k)$ is a martingale whenever (X_k) is, so by symmetry we have the same upper bound for $\mathbb{P}(X_n < -a)$. This yields (96) when $c = 1, \mu = 0$ and completes the proof of the theorem.

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]. Next week, Jeff will present a proof of the following result (Theorem 7.3.3 in [AS]):

Theorem 46 *Let $p = n^{-\alpha}$ where α is fixed, $\alpha > 5/6$. Let $G = G(n,p)$. Then there exists $u = u(n,p)$ such that*

$$\mathbb{P}(u \leq \chi(G) \leq u + 3) \to 1, \quad \text{as } n \to \infty. \tag{99}$$