MATH 159 NOTES: DISCRETE PROBABILISTIC METHODS

ARUN DEBRAY
FEBRUARY 21, 2014

CONTENTS

1. RAMSEY THEORY AND TOURNAMENT GRAPHS: 1/6/14

"There is no homework, but there is also homework. How can this be?"

This course is assumed to be a second course in probability, i.e. that you have already taken a probability class at the level of Stats 116 or Math 151. The course website is `http://www-stat.stanford.edu/~adembo/math-159/`, and the textbook is Alon & Spencer, with a few supplements.

This material is at once elementary and very deep, so giving tests or homework on it is difficult; you won't finish it.[1] Instead, the grading is based on student presentations in class in the last two weeks of the quarter. Each student is to give a twenty-minute presentation, e.g. an example or a subsection of the book, understood and then explained to the rest of the class. Students are also required to attend almost all of the lectures.

This course is about an idea pioneered by Erdős: to use probability to construct, find, or show the existence of certain discrete objects with a desired property. These objects could be graphs, hypergraphs, directed graphs, vectors, subsets with a given property, and so on. This is done by assigning a probability distribution to the finite set (so there's no need to invoke analysis or measure theory), and then showing that the probability of something happening is positive. The methods and tools may be crude, but the will be used in clever ways.

The basic principle: if S is a finite set and $\mathbb{P}$ is a probability measure on S such that $\mathbb{P}(A) > 0$ for some $A \subseteq S$, then $A \neq \emptyset$. (Equivalently, $\mathbb{P}(A^c) < 1$.)

One could also take a random variable X taking non-negative integer values, and let $A = \{X \geq 1\}$.

Markov's inequality states that $\mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$, which is true because $\mathbb{P}(X \geq 1) = \mathbb{E}[f(X)]$, where $f(X) = 1$ if $X \geq 1$ and is 0 otherwise. Since $X \geq f(X)$ and expectation is an integral, then this inequality follows. Then, if $\mathbb{E}[X] < 1$, then there exists an $\omega \in S$ such that $X(\omega) = 0$. In some sense, we count the average number of violations of some property; if it is less than 1 and is positive-integer-valued, it must be zero somewhere.

[1] No kidding: the professor doesn't know how to solve some of the harder exercises in the book, and there's a good chance a few are open problems.

Expectation is particularly useful because it is linear: if $X = \sum_{i=1}^{n} c_i X_i$, then

$$\mathbb{E}[X] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i],$$

and, most importantly, this holds true no matter the relationships between the X_i. Thus, as long as you can get the $\mathbb{E}[X_i]$, then you're in very good shape.

These are the ideas that underlie the first two chapters of the text; but of course, the trick lies in putting them to use.

Example 1.1. The first example comes from Ramsey theory, on the subject of graphs with non-monochromatic cliques. Specifically, if one wishes to two-color the edges of the complete graph on n vertices, K_n.[2] Thus, there are $2^{\binom{n}{2}}$ such 2-colorings. Then, $R(k,\ell)$ is the smallest n such that any 2-coloring of the edges of K_n contains either a red K_k or a bleu K_ℓ.

For example, K_3 is just a triangle. Thus, for very large graphs, it is impossible to find a coloring that avoids triangles, and the smallest such number would be $R(3,3)$. A lower bound for $R(k,k)$ can be found with the probabilistic method:

Proposition 1.1. *If* $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$, *then* $R(k,k) > n$, *and therefore* $R(k,k) \geq \lfloor 2^{k/2} \rfloor$.

Proof. The thing we want to prove is that if $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$, then there exists a two-coloring with no monochromatic K_k. Call this property A, so that A^c is the property that a two-coloring does have a monochromatic K_k. Thus, the goal is to prove that $\mathbb{P}(A^c) < 1$ under some probability distribution.

Let $\mathbb{P}$ be the uniform distribution on colorings of K_n: randomly color each edge with probability 1/2. There are exactly $\binom{n}{k}$ ways to embed K_k in K_n, since every possible subset of k vertices out of the n forms a K_k. More interestingly, the probability that a specific K_k is monochromatic is $2/(2^{\binom{k}{2}}) = 2^{1-\binom{k}{2}}$, because there are two choices (all red and all blue) out of all possible options for the colorings. Then, let E_i be the event that the i^{th} coloring of K_k is monochromatic. Then, $A^c = \bigcup_{i=1}^{\binom{n}{k}} E_i$. Using the union bound from probability,

$$\mathbb{P}(A^c) = \mathbb{P}\left(\bigcup_{i=1}^{\binom{n}{k}} E_i\right) \leq \sum_{i=1}^{\binom{n}{2}} \mathbb{P}(E_i) = \binom{n}{k} 2^{1-\binom{k}{2}} < 1. \qquad \boxtimes$$

One could try to do something more clever with the distribution, but the result is useful enough already.

Definition. A tournament of size n is a directed graph $G = (V,E)$ that is an orientation of K_n (i.e. for every two vertices v and w, there is either an edge $v \to w$ or an edge $w \to v$; the intuition that there is a tournament for some game, and every two players played each other, and someone won).

There are $2^{\binom{n}{2}}$ tournaments of size n, because there are n edges, and each has two possibilities. These can be thought of as colorings, where one color corresponds to an edge going one way, and blue to the edge going the other way.

Definition. A Hamiltonian path on some directed graph G is a directed path visiting each vertex in G exactly once.

Theorem 1.2 (1943, Szele)**.** *For any n, there exists a tournament on n players with at least $n!2^{-(n-1)}$ Hamiltonian paths.*

Proof. The clever part is to restate this in a way amenable to probabilistic methods: a Hamiltonian path in a tournament is associated with a permutation σ of vertices $\{1,\dots,n\}$ with $X_\sigma = 1$ if σ is a Hamiltonian path in the tournament T and $X_\sigma = 0$ otherwise. In other words, $X_\sigma = 1$ iff $(\sigma(i),\sigma(i+1)) \in E$ for all $i = 1,\dots,n-1$: σ is a suggestion for a tournament and X indicates whether it is one.

The number of Hamiltonian paths on T is $X = \sum_\sigma X_\sigma$, so $\mathbb{E}[X] = \sum_\sigma \mathbb{E}[X_\sigma]$, thanks to linearity of expectation. Take this expectation under the random uniform probability distribution again. Then, $\mathbb{E}[X_\sigma] = P(X_\sigma) = 1 = 2/2^n = 2^{-(n-1)}$, so $\mathbb{E}[X] = n!2^{-(n-1)}$.

If X is a uniformly distributed random variable, then it is a fact that there exist ω_1, ω_2 such that $X(\omega_1) \geq \mathbb{E}[X]$ and $X(\omega_2) \leq \mathbb{E}[X]$. Thus, there is a tournament with at least that many Hamiltonian paths. $\boxtimes$

These proofs tend to be existence proofs, though without construction. Sometimes this is not as helpful as one might like, but it still says useful things. However, sometimes you get an algorithm: notice that for Proposition 1.1 the odds that a given coloring doesn't contain a monochromatic K_k is very small as n gets large; thus, one good algorithm is just to randomly choose a graph and check!

The next example has an elegant and short non-probabilistic proof, but there is also a nice proof by probabilistic methods.

[2] The complete graph on n vertices is the graph where there is exactly one edge between v_i and v_j for all $1 \leq i < j \leq n$. Then, one assigns each edge to either the color red or blue.

Theorem 1.3 (Balancing Vectors)**.** *Let $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ such that $|\mathbf{v}_i| = 1$ for all $I = 1, \dots, n$. Then, there exist $\varepsilon_i \in \{\pm 1\}$ such that $\left|\sum_{i=1}^n \varepsilon_i \mathbf{v}_i\right| \leq \sqrt{n}$, and there exist $\varepsilon_i \in \{\pm 1\}$ such that $\left|\sum_{i=1}^n \varepsilon_i \mathbf{v}_i\right| \geq \sqrt{n}$.*

Proof. Assign ε_i to 1 or -1 with equal probability (so once again the uniform distribution is used). Then, let $X = \left|\sum_{i=1}^n \varepsilon_i \mathbf{v}_i\right|^2$. Thus,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\left|\sum_{i=1}^n \varepsilon_i \mathbf{v}_i\right|^2\right] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n \varepsilon_i \varepsilon_j (\mathbf{v}_i, \mathbf{v}_j)\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n (\mathbf{v}_i, \mathbf{v}_j)\mathbb{E}(\varepsilon_i \varepsilon_j). \\ &= \sum_{i=1}^n (\mathbf{v}_i, \mathbf{v}_i)\mathbb{E}[\varepsilon_i \varepsilon_i] = \sum_{i=1}^n |\mathbf{v}_i| = n.\end{aligned}$$

Then, take the square root and apply the fact that, since $\mathbb{E}[X] = n$, then there exists $\underline{\varepsilon}$ such that $X(\underline{\varepsilon}) \leq n$ and $\overline{\varepsilon}$ such that $X(\overline{\varepsilon}) \geq n$. ⊠

There are lots of applications of these methods, but in a lot of cases they require a more intimate understanding of the application, e.g. group theory, where one needs a better understanding of the structure of groups, or elsewhere with Fourier analysis, and so on. But these involve much less obvious things than taking the norm as above.

2. Expectation and Variance: 1/8/14

"Blah, blah, blah… analysis… blah, blah, blah… "

Theorem 2.1 (Alon & Spencer, Theorem 2.2.1)**.** *Suppose $G = (V, E)$ is a graph with n vertices and e edges. Then, there exists a partition of V into T and $B = V \setminus T$ such that there are at least $e/2$ (i.e. at least half) of the edges between T and V.*

Though this isn't optimal for all graphs, it's a sharp bound: if $G = K_n$, then $e = n(n-1)/2$; if $|T| = \ell$, then there are $(n - ell)\ell$ edges in the bipartite subgraph, which is maximal when $\ell = n/2$, yielding $(n-1)(n+1)/4$ when n is odd and $(n/2)^2$ when n is even. These are very close to $e/2$, so it's not possible to do better. Of course, in more complicated graphs, finding the optimal partition, which might be better than $e/2$, can be much harder.

Proof of Theorem 2.1. Let $\mathscr{S} = 2^{|V|}$ with T a random subset given by $\mathbb{P}(x \in T) = 1/2$, independently and identically distributed. Then, let

$$X_{\{x,y\}} = \begin{cases} 1, & \text{if } x \in T \text{ and } y \in B \text{ or } x \in B \text{ and } y \in T, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Then, let $X = \sum_{\{x,y\} \in E} X_{\{x,y\}}$. Then,

$$\mathbb{E}[X] = \sum_{\{x,y\} \in E} \mathbb{E}[X_{\{x,y\}}] = \sum_E \frac{1}{2} = \frac{e}{2},$$

so there exists a T such that $X(T) \geq e/2$. ⊠

This suggests an algorithm for obtaining such a partition by just picking a random partition and checking, but how effective it is depends on the variance of such a distribution.

Theorem 2.2 (Alon & Spencer, Theorem 2.3.1)**.** *There exists a two-coloring of the edges of K_n with at most $\binom{n}{k} 2^{1-\binom{k}{2}}$ monochromatic subgraphs K_k.*

Proof. Take a random coloring, where each edge is colored red or blue with i.i.d. probability $1/2$, and let X be the number of monochromatic K_k in this graph. In Proposition 1.1, we saw that $\mathbb{E}[X] = \binom{n}{k} 2^{1-\binom{k}{2}}$, so there must exist an ω such that $X(\omega) \geq \mathbb{E}[X]$. ⊠

The above bound might be non-optimal.

Remark. In the example of Szele's theorem of tournaments with $p(n)$ Hamiltonian paths, it was seen that $p(n) \geq n! 2^{1-n}$. Another argument shows that $p(n) \leq c n^{3/2} n! 2^{1-n}$, so the bound is actually pretty good. However, this bound was found non-probabilistically. Often, the probabilistic method only works well for one of the upper or lower bounds.

Now, it's worth understanding how far values on the distribution can be from the expectation. This involves using the second moment (variance), which is unsurprisingly known as the second moment method. The principle involves the following observations:

(1) Chebyshev's or Markov's inequality: $\mathbb{P}(|X-\mu| \geq \lambda\sigma) \leq 1/\lambda^2$, where $\sigma^2 = \operatorname{Var}(X)$ and $\mu = \mathbb{E}[X]$. The proof of this involves constructing $f(x) = (x-\mu)^2/(\lambda\sigma)^2$, which is necessarily greater than the indicator function $g(x) = \mathbb{1}_{|x-\mu|\geq\lambda\sigma}$, so $\mathbb{E}[f] \geq \mathbb{E}[g]$.

(2) If $X = \sum_{i=1}^{m} \mathbb{1}_{A_i}$, where the A_i are events, then

$$\operatorname{Var}(X) \leq \mathbb{E}[X] + \sum_{i\sim j} \mathbb{P}(A_i \cap A_j),$$

where $i \sim j$ means that $i \neq j$ and A_i and A_j aren't independent. This is because

$$\operatorname{Var}(X) = \sum_{i=1}^{m}\sum_{j=1}^{m} \operatorname{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_j}),$$

and if A_i and A_j are independent, then their covariance is zero. Furthermore, $\mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] - \mathbb{E}[\mathbb{1}_{A_i}]\mathbb{E}[\mathbb{1}_{A_j}] \leq P(A_i \cap A_j)$.

The consequence is that if $X_n \geq 0$ is integer-valued and $\mathbb{E}[X_n] \to \infty$ as $n \to \infty$ and $\operatorname{Var}(X_n)/(\mathbb{E}[X_n])^2 \to 0$, then $\mathbb{P}(X_n = 0) \to 0$ and $X_n/\mathbb{E}[X^n] \to 1$. These follow from (1) with $\lambda = \mu/\sigma$ and $\lambda = \varepsilon\mu/\sigma$, respectively.

Anothr consequence is that by rearranging the sum in the second item: let

$$\Delta = \sum_i \mathbb{P}(A_i) \sum_{\{j:j\sim i\}} \mathbb{P}(A_j \mid A_i).$$

This involves summing over the dependency graph of these variables, in which two variables are connected if they aren't independent. Then, $X_1, \ldots, X_n$ are called symmetric if these probabilities are invariant under permutations. Then, the quantity $\Delta^* = \sum_{\{j:j\sim i\}} \mathbb{P}(A_j \mid A_i)$ is independent of i.

Corollary 2.3. *$\Delta_n = (\mathbb{E}[X_n])\Delta_n^*$, so if $\Delta_n^*/\mathbb{E}(X_n) \to 0$, then $\operatorname{Var}(X_n)/(\mathbb{E}[X_n]) \to 0$ as well.*

From number theory, $\nu(n)$ is the number of primes that divide n for an $n \in \mathbb{N}$. The following theorem wasn't originally proven with the probabilistic method, but admits a nice, short proof using it. that is not the original proof of Hardy and Ramanjuan.

Theorem 2.4 (Hardy & Ramanjaun, 1920)**.** *For all functions $\omega(n) \to \infty$, $(1/n)$(the number of x such that $|\nu(x) - \ln\ln n| \geq \omega(n)\sqrt{\ln\ln n}$) $\to 0$ as $n \to \infty$.*

Proof. The proof, due to Tuvan in 1934, depends on the following facts from elementary number theory: first, that

$$\sum_{\substack{p\leq n\\ p \text{ prime}}} \frac{1}{p} = \ln\ln n + o(1),$$

and second, that $\pi(n) \approx n/\ln n$, where $\pi(n)$ is the number of primes less than n. The first follows from Abel's summation formula and the Stirling approximation, but this isn't a number theory class; the point is probability.

For all prime $p \leq n$, define X_p to be the indicator for $p \mid x$, where x is uniformly chosen from $\{1, \ldots, n\}$. Then, $\nu(x) = X = \sum_p X_p$, so one wants to show that as $n \to \infty$, $\mathbb{P}(|X - \ln\ln n| > \omega(n)\sqrt{\ln\ln n}) \to 0$, which involves computing the mean and variance of X.

Using the two observations established above, the mean is

$$\mathbb{E}[X] = \sum_p \mathbb{E}[X_p] = \sum_{\substack{p\leq n\\ p \text{ prime}}} \frac{1}{n}\left\lfloor \frac{n}{p} \right\rfloor = \sum_{\substack{p\leq n\\ p \text{ prime}}} \left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right) = \ln\ln n + o(1),$$

and the variance satisfies

$$\begin{aligned}
\operatorname{Var} X &\leq \mathbb{E}[X] + \sum_{p\neq q}\left\{\mathbb{E}[X_pX_q] - \frac{\lfloor n/p\rfloor}{n}\cdot\frac{\lfloor n/q\rfloor}{n}\right\}\\
&= \mathbb{E}[X] + \sum_{p\neq q}\left(\frac{\lfloor n/(pq)\rfloor}{n} - \frac{\lfloor n/p\rfloor\lfloor n/q\rfloor}{n^2}\right)\\
&\leq \mathbb{E}[X] + \frac{2}{n}\sum_{p\leq n}\frac{1}{p}\pi(n) = \mathbb{E}[X](1 + 2\pi(n)/n) = \lg\lg n + o(1).
\end{aligned}$$

This is because $\frac{\lfloor n/(pq)\rfloor}{n} - \frac{\lfloor n/p\rfloor\lfloor n/q\rfloor}{n^2}a \leq \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right)$ and some various other number-theoretic magic. ⊠

As it happens, a better result can be obtained with judicious use of the Central Limit Theorem and a bunch of unpleasant calculation. To wit:

Theorem 2.5 (Erdős-Kac, 1940)**.** $(X - \mathbb{E}[X]/\sqrt{\operatorname{Var} X}) \to \mathrm{N}(0,1)$*, where* X *is as in the previous proof.*

The proof idea is to replace X_p by Y_p, independent Bernoulli random variables with probability $1/p$, and then use the Central imit Theorem, if $Y = \sum Y_p$ and $\hat{Y} = (Y - \mathbb{E}[Y])/(\sqrt{\operatorname{Var} Y})$, then one can check that all of the moments of $\hat{X}$ and $\hat{Y}$ converse to the same limit as $n \to \infty$.

The way in which probability theory got its hand into the metaphorical number-theoretic cookie jar is in the averaging in the result; not all results in number theory admit such an interpretation.

Definition. The Erdős-Renye random graph, denoted $G(n,p)$, is the random graph on n vertices, where each edge is present independently with probability p.

Denote $\omega(G)$ to be the size of the maximal clique in G (i.e. the maximal complete graph that is a subgraph of G). The expected number of k-cliques is $f(k) = \mathbb{E}[k\text{-cliques}] = \binom{n}{k} p^{\binom{k}{2}}$.

Theorem 2.6. *If* $p = 1/2$ *and* $k(n) \sim 2\log_2 n$ *is such that* $f(k) \to \infty$*, then* $\mathbb{P}[\omega(G) < k(n)] \to 0$.

The proof will be given next time, and involves computing the variance.

3. More Uses of Variance: 1/10/14

Proof of Theorem 2.6. The proof uses the second moment method. For every k-set S (i.e. subset of V of size k), assign $X_S = \mathbb{1}_{S \text{ is a clique}}$, and let $X = \sum_{|S|=k} X_S$, which counts the number of k-cliques in G. Then, $\omega(0) \geq k$ iff $X > 0$, since X is integer-valued.

We already computed $\mathbb{E}[X] = f(k) \to \infty$ last lecture, so it suffices to show that $\Delta^* = o(\mathbb{E}[X])$, where

$$\Delta^* = \sum_{\substack{T \neq S \\ X_T \not\perp X_S}} \mathbb{P}(X_T \mid X_S) = \sum_{T \sim S} \mathbb{P}(X_T \mid X_S).$$

FIx S; then $T \sim S$ iff $|T \cap S| = i$, for $2 \leq i \leq k-1$. Then,

$$\Delta^* = \sum_{i=1}^{k-1} \underbrace{\binom{k}{i}\binom{n-k}{k-i}}_{\text{number of choices of such } T} 2^{-\binom{k}{2}+\binom{i}{2}}.$$

Thus, $\Delta^*/\mathbb{E}[X] = \sum_{i=2}^{k-1} g(i)$, where $g(i) = (\binom{k}{i}\binom{n-k}{k-i}/\binom{n}{k}) 2^{\binom{i}{2}}$.

Supposing k is constant, $g(2) = 2\binom{k}{2}\binom{n-k}{k-2}/\binom{n}{k} \sim k$, and if k grows slowly this is still $k < n^{\varepsilon}$ for some not terribly large ε, and

$$g(k-1) = \frac{k\binom{n-k}{1} 2^{\binom{k-1}{2}}}{\binom{n}{k}} \approx \frac{2^{(k-1)^2/2}}{n^{k-1}} = \frac{(2^{k-1})^{(k-1)/2}}{n^{k-1}} = \frac{n^{(k-1)/2}}{n^{k-1}} \to 0,$$

and so on. The point is that k doesn't grow too quickly, $\log_2 n$ is about as fast as it can grow for this to work. ⊠

Remark. This is a good bound, but not optimal; thanks to Boll-Erdős and Matura in 1976, there exists a formula $k(n)$ such that $\mathbb{P}(\omega(G) = k \text{ or } \omega(G) = k+1) \to 1$ as $n \to \infty$. Though we don't know what the numbers are, the clique numbers stay extremely concentrated as n becomes large.

The second moment method also has applications to the k-SAT problem in theoretical computer science, which involves determining whether some Boolean functions can be satisfied. After some threshold, this becomes very difficult, and a recent paper established a bound for such a threshold using the second moment method. The method also pops up in random walks, and trees, and so on.

Theorem 3.1 (Bollobas, 1965)**.** *Let* $\mathscr{F}$ *be a collection of subsets* $\mathscr{F} = \{(A_i, B_i)\}_{i=1}^h$*, where* $A_i, B_i \subset X$ *for some finite set* X*.* $\mathscr{F}$ *is called a* (k,ℓ)*-system if:*

(1) $|A_i| = k$ *and* $|B_i| = \ell$*, and*
(2) $A_i \cap B_i = \emptyset$*, but* $A_i \cap B \neq \emptyset$ *if* $i \neq j$*.*

Then, in a (k,ℓ)*-system,* $h \leq \binom{k+\ell}{k}$.

Notice that the bound has nothing to do with $|X|$, which is pretty impressive.

For an example of such a (k,ℓ)-system, consider $\mathscr{F} = \{(A, X \setminus A) : A \subset X, |A| = k\}$, where $X = \{1, \ldots, k\}$. Here, $h = \binom{k+\ell}{k}$. Thus, the bound is tight.

Proof of Theorem 3.1. Without loss of generality, let $X = \bigcup_{i=1}^{h}(A_i \cup B_i)$, since we don't need to worry about anything else. Then, put on X a random order π, uniformly chosen (really a permutation over the entries of X). For all i, let

$$\mathscr{E}_i = \{\text{all elements of } A_i \text{ precede all elements of } B_i \text{ in the order } \pi\}.$$

Thus, $\mathbb{P}(\mathscr{E}_i) = 1/\binom{k+\ell}{k}$, since the goal is to make sure out of the $k+\ell$ elements of $A_i \cup B_i$, the k in A_i come first.

More interestingly, $P(\mathscr{E}_i \cap \mathscr{E}_j) = 0$ whenever $i \neq j$: suppose without loss of generality that the last element of A_i doesn't appear after the last element of A_j (if not, then switch them); if this is true, then A_i precedes B_i and therefore B_j as well. But recall that $A_i \cap B_j$ is nonempty, so this event cannot happen.

Now,

$$1 \geq \mathbb{P}\left(\bigcup_i \mathscr{E}_i\right) = \sum_{i=1}^{h} \mathbb{P}(\mathscr{E}_i) = \frac{h}{\binom{k+\ell}{k}}. \qquad \boxtimes$$

This is one of the cleverest elementary proofs anyone in the class has seen in a long time. Where did the last line come from? Proof often have a level of novelty or depth (how surprising it is), and how comptuational or cumbersome it is. These two proofs illustrate some of the grave differences between the two!

Theorem 3.2 (Erdős-Ko-Rado)**.** *A family $\mathscr{F}$ of sets is called intersecting if all of its elements are pairwise intersecting, i.e. for all $A < B \in \mathscr{F}$, $A \cap B \neq \emptyset$. Then, if $\mathscr{F}$ is an intersecting family of k-subsets of $\{0, \dots, n-1\}$ and $n \geq 2k$, then $|\mathscr{F}| \leq \binom{n-1}{k-1}$.*[3]

Proof. This proof is due to Katona, in 1972. It depends on a small result in number theory: for $n \geq 2k$ and $0 \leq s \leq n-1$, let $A_S = \{s, \dots, (s+k-1) \bmod n\}$, then such an $\mathscr{F}$ as discussed in the theorem statments contains at most k of the sets A_s. The proof will be given in just a bit.

For the theorem statement, choose a permutation σ of $\{0, \dots, n-1\}$ uniformly, and choose an $i \in \{0, \dots, n-1\}$ uniformly and independently from σ. Consider the set $A = \{\sigma(i), \sigma(i+1), \dots, \sigma(i+k-1 \bmod n)\}$. This is like A_s, but after the permutation has been applied. By the result claimed above, $\mathbb{P}(A \in \mathscr{F} \mid \sigma) \leq k/n$, and therefore that $P(A \in \mathscr{F}) \leq k/n$, by averaging over the σ. Thus, A has been obtained as a uniformly chosen k-subset of $\mathscr{F}$. But this simplifies to $k/n \geq |\mathscr{F}|/\binom{n}{k}$, and then you're done.

The first claim admits a simple geometric proof: suppose $A_\ell \in \mathscr{F}$; then, of the sets A_s, only those that intersect A_ℓ can be in $\mathscr{F}$. But these can be arranged into disjoint pairs, e.g. if $A_1 \cap A_\ell$, then pair A_1 the set A_2 given by $A_2 \cap A_\ell = A_\ell \setminus A_1$; thus, A_1 and A_2 are disjoint. Thus, one can only pick at most one from each pair, giving $k-1$, plus the original, giving k. This part doesn't depend on k. $\boxtimes$

Once again, the proof is really short, but really clever. It's not exactly something one would cook up on a problem set.

One way to refine this method is to use alterations; this technique combines probabilistic and non-probabilistic ideas to construct a probability measure P and a random variable X on a discrete set S, as in the classical probabilistic method, but it's difficult to coerce X into being large enough to make the property work. if it "almost works," in some sense that only a few constraints are violated, then one might be able to alter X a little bit to obtain the needed structure. The loss that results might be nonzero, but not large enough to be important.

As an example, an alternate formulation of the idea of Ramsey number is that for all graphs G of $R(k,k)$ vertices, either $\omega(G) \geq k$, or $\alpha(G) \geq k$, where $\omega(G)$ is the size of the largest clique and $\alpha(G)$ is the size of its largest independent set (a set of vertices which have no edges between any of them). This can be shown by taking a given graph G and coloring K_n such that an edge is blue if it's in G and red if it isn't. Then, cliques and independent sets correspond to monochromatic subgraphs K_k.

We have seen that in a random two-coloring, the expected number of monochromatic K_k is $s = \binom{n}{k}2^{1-\binom{k}{2}}$. Thus, we saw that if $s < 1$, then there exists a two-coloring with no monochromatic K_k, so $R(k,k) > n$.

Theorem 3.3 (Alon & Spencer, Theorem 3.1.1)**.** *For all n, $R(k,k) > n - s$, even if $s > 1$.*

Proof. Fix n and by the basic methods, there exists a two-coloring of K_n with at most s monochromatic subgraphs K_k. Then, remove from K_n one vertex from each such monochromatic K_k. Now, this graph has at least $n-s$ vertices and no monochromatic K_k. $\boxtimes$

While this illustrates the general principle, the actual improvement was not large. Yet it was pretty easy to follow.

[3] This is also a tight bound: take all of the sets containing 0, which are clearly intersecting, and there are $\binom{n-1}{k-1}$ of them. Once again, the simplest construction is optimal.

Not everything in this class is discrete or random; sometimes the examples are more interesting.

Let $C \subseteq \mathbb{R}^d$ be a bounded, convex, centrally symmetric set centered at the origin (i.e. if $\mathbf{c} \in C$, then $-\mathbf{c} \in C$ as well), and let $\mu(C)$ denote the volume (or Lebesgue measure) of C. Let $B(x) = [0,x]^d$. Then, the packing problem is to fit as many copies of C into $B(x)$ by translation. In other words, let $f(x) = \max\{N : \mathbf{x}_1, \dots, \mathbf{x}_N \text{ with } \mathbf{x}_i + C \subseteq B(x) \text{ and } (\mathbf{x}_i + C) \cap (\mathbf{x}_j + C) = \emptyset, i \neq j\}$, and define the packing constant to be $\delta(C) = \mu(C) \liminf_{x\to\infty} f(x)x^{-d}$, which is the fraction of volume occupied by the pakcing, and represents the efficiency of the packing, and is always at most 1.

Impressively, the following result admits a probabilistic proof.

Theorem 4.1. $\delta(C) \geq 2^{-(d+1)}$.

Proof. The standard probabilistic method will cause there to be some overlapping copies of C, so they will be removed; this is an example of alterations.

Choose P and Q independently uniformly on $B(x)$. Then, $(C+P) \cap (C+Q) \neq \emptyset$ iff there exist $c_1, c_2 \in C$ such that $P - Q = c_1 - c_2$, and therefore $P - Q = 2((c_1 - c_2/2)) \in 2C$ (since $c_1, c_2 \in C$ and C is convex, so their midpoint is in C). Thus, $\mathbb{P}((C+P) \cap (C+Q) \neq \emptyset) \leq \mathbb{P}(P \in Q + 2C) \leq \mu(2C)/x^d$.

Choose $P_1, \dots, P_n$ i.i.d. and uniformly at random in $B(x)$. Then, compue the expected number of intersections: let $X = \#\{i < j : (P_i + C) \cap (P_j + C) \neq \emptyset\}$. By linearity of expectation,

$$\mathbb{E}[X] \leq \binom{n}{2} \mu(2C)x^{-d} \leq \frac{n^2}{2} 2^d \mu(C) x^{-d}.$$

Thus, there exists a choice of n points $x_1, \dots, x_n$ having no more than $n^2\mu(C)2^{d-1}x^{-d}$ intersections among $\{(X_i + C)\}_{i=1}^n$. Thus, by removing exactly this number of points, one obtains a packing of $n - n^2\mu(C)2^{d-1}x^{-d}$. However, the centers are in the packing, but the edges of C might not be; it's a packing of $B(x+2w)$, where $w = \max_{1\leq i\leq d} |i^{\text{th}} \text{ coordinate of a point in } C|$. Thus, $f(x+2w) \geq n = (n^2/2)\mu(C)2^d x^{-d}$.

This can be explicitly solved for the optimal n: $n^* = x^d 2^{-d}/\mu(C)$, so $f(x+2w) \geq x^d 2^{-(d+1)}/\mu(C)$. And since $((x + 2w)/x)^d \to 1$ as $x \to \infty$, then in the limit this is correct. ⊠

Unfortunately, this proof doesn't provide an algorithm other than the standard one (i.e. choose a random set of points and see what happens). Moreover, it's not at all optimal; the textbook gives an easy improvement to 2^{-d} and a much harder improvement to $2^{-(d-1)}$ (eleven chapters later!); it is not clear what can be said beyond that. Interestingly, though, this proof doesn't require too much cleverness.

The following theorem is in a similar spirit.

Theorem 4.2 (Erdős)**.** *If n is prime, then there exist n points on $[0,1]^2$ such that any triangle drawn between three points chosen from these n is of area at least $1/(2(n-1)^2)$.*

The probabilistic proof is easy enough to be given in the textbook, though it requires some number theory, but a non-probabilistic proof is much harder, was only seen in 1982, and isn't even that much better: $(c \lg n)/n^2$. There's also an even simpler probabilistic proof which gives a bound of $1/Cn^2$, even when n is not prime. This is once again of the form that in the average case, something happens, and then there must be something better than average.

Definition. $\chi(G)$ denotes the chromatic number of a graph G, the minimum number of colors of vertices of the graph such that no two vertices of the same color are connected. Then, girth(G), known as the girth of the graph, is the length of the minimal circuit.

Intuitively, though the chromatic number is complicated, it ought to be related to girth somehow. Unfortunately, the truth isn't as nice.

Theorem 4.3 (Erdős, 1959)**.** *For all k, ℓ there exists a graph G with* girth$(G) \geq \ell$ *and $\chi(G) \geq k$.*

Proof. This proof depends on the following graph-theoretic inequality, which is not hard to show: that $\chi(G)\alpha(G) \geq |V|$, where $\alpha(G)$ is again the size of the maximal independent set.

This is because one can make a proper coloring of the graph using at least $\chi(G)$ colors, such that no edge connects vertices of the same color. Then, every set of vertices of the same color is independent, so if α_{c_i} is the number of vertices of color i, then $\alpha_{c_i} \leq \alpha(G)$ and

$$|V| = \sum_{i=1}^{\chi(G)} \alpha_{c_i} \leq \sum_{i=1}^{\chi(G)} \alpha(G) = \chi(G)\alpha(G).$$

The idea of the general proof is to construct G with n vertices and less than $n/2$ cycles of length at most ℓ, with $\alpha(G) = o(n)$. Then, one can remove up to $n/2$ vertices, each from such cycles, to obtain a slightly smaller graph G^* with at least $n/2$ vertices and such that girth$(G^*) > \ell$ and $\alpha(G^*) \le \alpha(G) = o(n)$. Now, use the given inequality to show that $\chi(G^*) \ge |G^*|/\alpha(G^*) \to \infty$ as $n \to \infty$. Thus, all one has to do is pick n large enough. In some sense, dealing with independent sets is easier than dealing with the chromatic number, so the inequality is used to translate between them.

The construction is as follows: take $G = G(n,p)$, the Erdős-Renye random graph, where $p = n^{-(1-\theta)}$, where $\theta < 1/\ell$. It will become clear in the proof why p and θ are as given. Then, it happens (though it has to be shown) that if X is the number of cycles with length at most ℓ, then $\mathbb{E}[X] = o(n)$, and $\mathbb{P}(\alpha(G) \ge (3/p)\ln n + 1) < 1/2$.

To show the first, sum over the lengths of all possible cycles, the choice of the vertices in the cycle, and the choices of the cycle ordering. Thus,

$$\mathbb{E}[X] = \sum_{i=3}^{\ell} \binom{n}{i} \frac{i!}{2i} p^i \le \sum_{i=3}^{\ell} \frac{(np)^i}{2i} = o(n),$$

because $(np)^\ell = n^{\theta\ell}$ for $\theta < 1$. Thus, by the Markov inequality, if $X \ge 0$, then $\mathbb{P}(X \ge n/2) \le \mathbb{E}[X]/(n/2)$, which goes to zero as $n \to \infty$. Using the union bound over all possible choices of y vertices out of n,

$$\mathbb{P}(\alpha(G) \ge X) \le \binom{n}{y}(1-p)^{\binom{y}{2}} \le \left[n(1-p)^{(y-1)/2}\right]^y \le \left[ne^{-p(y-1)/2}\right]^y = o(1),$$

because $1 - p \le e^{-p}$. The asymptotic bound uses the fact that $y = (3/p)\ln n + 1$, so in the end the factor of n is dwarfed by $n^{-3/2}$. ⊠

This proof was difficult, but the trickiness wasn't in the probabilistic part! The trickery is in being as clever as Erdős in reasoning about graphs, though the probabilistic part could even be assigned as an exercise.

Another graph-theoretic example of alterations is Turán's theorem.

Theorem 4.4 (Turán)**.** *If G is a graph on n vertices with $nd/2$ edges and $d \ge 1$, then $\alpha(G) \ge n/2d$.*

To motivate this theorem, take n/d disjoint copies of K_d; then $\alpha(G) = n/d$, which is a factor of two off. Then, the proof is by alterations.

5. The Lovász Local Lemma I: 1/15/14

Lemma 5.1 (Lovász Local Lemma)**.** *Suppose that $X_1, \dots, X_n$ are indicator random variables and $W = \sum_{i=1}^n X_i$. Suppose that for every i, there exists a subset $B_i \subset \{1, \dots, n\}$ such that* $\sup\{\mathbb{P}(X_i = 1 \mid \{X_k\}_{k \notin B_i})\} = p_i$.[4] *We are interested in when these A_i don't happen; specifically, if there exist $x_i \in [0,1)$ such that for all i,*

$$p_i \le x_i \prod_{j \in B_i} (1 - x_j),$$

then

$$\mathbb{P}(W = 0) \ge \prod_{j=1}^{n} (1 - x_j).$$

Proof. The proof consists of two parts, the first of which is non-obvious (and deferred), and the second of which is more obvious. Thus, assume for now that for all $S \subseteq \{1, \dots, n\}$ with $i \notin S$,

$$\mathbb{P}\left(X_i = 1 \mid \sum_{j \in S} X_j = 0\right) \le x_i. \tag{1}$$

Thus, assuming this, one can write

$$\begin{aligned} P(W = 0) &= P(X_1 = 0)P(X_1 = 0 \mid X_1 = 0) \cdots \left(\mathbb{P}\left(X_i = 1 \mid \sum_{j \in S} X_j = 0\right)\right) \\ &\ge (1 - x_1)(1 - x_2) \cdots (1 - x_j). \end{aligned}$$

To prove (1), proceed by induction on the size of S. If $|S| = 0$, then it's trivial, because the probabilities are unconditional: since p_i is the supremum of the conditional probabilities, then it's certainly true that $x_j \ne 1$.

[4] The textbook deals with a special case in which $X_i = \mathbb{1}_{A_i}$, $p_i = \mathbb{P}(X_i = 1)$, and $i \sim j$ if $j \in B_i$ and A_i is independent of $\{A_j\}_{j \notin B_i}$.

Otherwise, use Bayes' rule. Let $T = \sum_{j \in S \cap B_i^c} X_j = 0$, which intuitively represents a part of B_i over which we have some control. Then,

$$\mathbb{P}\left(X_i = 1 \mid \sum_{j \in S} X_j = 0\right) = \frac{\mathbb{P}\left(X_i = 1, \sum_{j \in S \cap B_i} X_j = 0 \mid T\right)}{\mathbb{P}\left(\sum_{j \in S \cap B_i} X_j = 0 \mid T\right)}$$
$$\leq \frac{\mathbb{P}(X_i = 1 \mid T)}{(1 - \mathbb{P}(X_{j_1} = 1 \mid T))(1 - \mathbb{P}(X_{j_2} = 1 \mid X_{j_1}, T)) \cdots (1 - \mathbb{P}(X_{j_r} = 1 \mid X_{j_1}, \ldots, X_{j_r}, T))},$$

where $S \cap B_i = \{j_1, \ldots, j_r\}$. The case of $r = 0$ is trivial, since $p_i \leq x_i$. Thus, more generally, by the inductive hypothesis,

$$\leq \frac{x_i \prod_{j \in B_i} (1 = x_j)}{\prod_{j \in S \cap B_i} (1 - x_j)} \leq x_i.$$

The last step follows because there are more terms in the product in the denominator than in the numerator. ⊠

One special case is given when the probabilities are symmetric.

Corollary 5.2. *Let $p = \max_{i=1}^n \{p_i\}$ and $d = \max_{i=1}^n |B_i|$; then, $\mathbb{P}(W = 0) \geq (1 - 1/(d+1))^n$ whenever $p \leq (1/(d+1))(1 - 1/(d+1))^d \leq 1/e(d+1)$ for all $d \geq 1$.*

Proof. Take $x_i = x = 1/(d+1)$, and the conditions to check from the lemma are satisfied. ⊠

This is pretty nice in that it doesn't even mention n.

Intuitively, this lemma says that when computing the probabilities of intersections of events, even if they're not independent, as long as they aren't "too" dependent, then they are approximately independent.

Interestingly, there was a more recent constructive proof of Lemma 5.1, which actually provides an algorithm for constructing the specified objects. But it's much longer than this proof.

For an example, consider two-coloring hypergraphs.

Definition. A hypergraph is a set $H = (V, E)$, where V is some finite set and E is a set of subsets of V. A hypergraph is 2-colorable is there exists a two-coloring of V such that no (hyper)-edge is monochromatic.

The edges are now allowed to connect more than two vertices; if this requirement is imposed, one obtains a regular graph. Notice that the coloring condition is not that any two vertices connected by a hyperedge must be different colors, but that in some hyperedge there must be some two vertices with different colors.

Theorem 5.3 (Alon & Spencer, Theorem 5.2.1)**.** *If $H = (V, E)$ is a hypergraph with e hyperedges, then each edge has at least k elements and intersects (as sets) at most d other edges, such that $e(d+1) \leq 2^{k-1}$, then there exists a 2-coloring of V with non-monochromatic edges.*

Proof. Color V by assigning each vertex to each color randomly with probability $1/2$ (Bernoulli distribution). Let $X_f = 1$ if the edge f is monochromatic, so that $\mathbb{P}(X_f) = 2/2^{|f|} \leq 2^{-(k-1)}$. Let $p = 2^{-(k-1)}$, and note that X_f and X_g are independent when $f \cap g = \emptyset$, so we have the condition of the symmetric version of the local lemma. ⊠

Theorem 5.4 (Erdős & Lovász)**.** *A function $c : \mathbb{R} \to \{1, \ldots, k\}$ is called a k-coloring, and $T \subseteq \mathbb{R}$ is called multicolored if $c(T) = \{1, 2, \ldots, k\}$. Then, let $m, k \in \mathbb{N}$ such that $e(m(m-1)+1)k(1-1/k)^m \leq 1$, and let $d = m(m-1)$ and $p = k(1-1/k)^m$. Then, for all $S \subseteq \mathbb{R}$ of cardinality m, there exists a k-coloring such that for all $x \in \mathbb{R}$, $x + S$ is multicolored.*

This is impressive in that it works even despite uncountably many shifts. It doesn't seem like it should at all be possible to tackle with a finiteness condition.

Remark. The condition $e(m(m-1)+1)k(1-1/k)^m \leq 1$ seems arbitrary, but the point is that it's sufficient to have $m > (3 + o(1))k \lg k$.

Proof of Theorem 5.4. First, take $\mathcal{X} \subseteq \mathbb{R}$ to be a finite set of arbitrary size. Then, the theorem will be shown for $\{x + S\}_{x \in \mathcal{X}}$ by using the symmetric version of Lemma 5.1. Then, let $\mathcal{Y} = \bigcup_{x \in \mathcal{X}} (x + S)$, and choose $c : \mathcal{Y} \to \{1, \ldots, k\}$ at random choosing $c(y)$ i.i.d. and uniformly in $\{1, \ldots, k\}$.

Define $X_x = 1$ iff $|c(x + S)| < k$; then, $p = \mathbb{P}(X_x = 1) \leq k(1-1/k)^m$. Then, X_x is indepednent of $\{X_y\}$ for all Y such that $(x + S) \cap (y + S) = \emptyset$, so there are at most $d = m(m-1)$ numbers y such that $x - y \in S$. Then, apply the symmetric version of Lemma 5.1, which shows the theorem for $\mathcal{X}$.

To extend this to $\mathbb{R}$, some analysis is needed, but no probability. $\{1, \ldots, k\}$ is a compact space, and by Tychonov's theorem, the space of functions $\{c : \mathbb{R} \to \{1, \ldots, k\}\}$ is also compact, with respect to the topology of pointwise convergence. Thirdly,

for any fixed $x \in \mathbb{R}$, the set $C_x = \{c : \mathbb{R} \to \{1, \dots, k\}$ such that $x + S$ is multicolored$\}$ is closed with respect to the topology of pointwise convergence (since if it converges to something, then the colors ought to stabilize too).

Thus, the C_x are compact, since they're closed subsets of a compact space. But $\bigcap_{x \in \mathscr{X}} C_x \neq \emptyset$ for any finite collection $\mathscr{X}$, which was the point of the probabilistic argument. Thus, by compactness, $\bigcap_{x \in \mathbb{R}} C_x \neq \emptyset$.[5] ⊠

6. The Lovász Local Lemma II: 1/17/14

Recall the symmetric version of the Lovász local lemma: that if $W = \sum_i X_i$ where the X_i are independent variables, then if $\mathbb{P}(X_i = 2) \leq p$ for all I, then $\mathbb{P}(W = 0) > 0$, provided $X_i \coprod \{X_j\}_{j \notin B_i}$ for all i, and for all j $|B_j| \leq d$ such that $e(d+1)p \leq 1$.

Theorem 6.1 (Alon & Linial, 1989)**.** *If D is a simple (i.e. no multiple edges) directed graph with a minimum outdegree δ and maximum indegree Δ such that $e(\Delta\delta + 1)(1 - 1/k)^{\delta} \leq 1$. Then, there exists a directed, simple cycle in D (i.e. there are no subcycles) of length $0 \pmod k$.*

Proof. Without loss of generality, assume that every outdegree is δ, because this only makes it harder to find cycles. Now, we put a k-coloring of the vertices of the graph D, uniformly at random (and i.i.d., and so on); call this coloring $f(v)$ for $v \in V$, i.e. $f : V \to \{1, \dots, k\}$, where V is the set of vertices of D.

For $v \in V$, let X_v be an indicator for the event that for all $u \in V$ such that $v \to u$ is an edge in D, we have $f(u) \neq (f(v)+1) \pmod k$. By the local lemma, there exists a coloring f with $W = 0$, i.e. $X_v = 0$ for all v. Thus, for all $v \in V$, there exists a $u \in V$ such that $v \to u \in D$ and $f(u) = f(v) + 1 \pmod k$. Explore D along these edges;given a $v_0 \in V$, there exists a v_1 with $f(v_1) = f(v_0) + 1 \bmod k$, and then a v_2 with that property for v_1, and so on. Since D is finite, eventually this will create a cycle $v_i, v_{i+1}, \dots, v_{i+n} = v_i$, which is simple. But since $f(v_i)$ increments by 1 mod k each time, but ends up at the same place mod k, so $k \mid n$, or the length of the cycle is $0 \pmod k$.

Now, we need to actually invoke the probabilistic argument, which requires somewhat less cleverness. Firstly, $\mathbb{P}(X_v = 2) = (1 - 1/k)^{\delta}$, because there are δ edges out of v leading to different vertices, so this is the probability that a given u has $f(u) \neq f(v) + 1 \pmod k$ given $f(v)$.

Then, X_v is independent from $\{X_u\}_{u \notin B_v}$, because B_v consists of all of all u such that $u \to w$ and $v \to w$ as well as all u such that $v \to u$. Its size is at most $\delta\Delta$, where δ comes from the outdegree of v and Δ the indegrees of the possible u. Then, the conditions placed in the theorem allow one to invoke the Lovász local lemma. ⊠

Definition. A family $\mathscr{F}$ of open unit balls in $\mathbb{R}^3$ is called a k-fold covering (of $\mathbb{R}^3$) if for all $\mathbf{x} \in \mathbb{R}^3$, there exist at least k balls in $\mathscr{F}$ which $\mathbf{x}$ belongs to. $\mathscr{F}$ is decmomposable if ther exist disjoint subsets $\mathscr{F}_1, \mathscr{F}_2$ of $\mathscr{F}$ such that each $\mathscr{F}_i$ is a covering of $\mathbb{R}^3$.

Theorem 6.2 (Mani-Levitska & Pach, 1988)**.** *For all $k \geq 1$, there exists a non-decomposable k-fold covering of $\mathbb{R}^3$.*

This is hard to prove, but they also proved the following theorem, which can be attacked with the local lemma in a completely non-obvious way. The result itself is also somewhat counterintuitive.

Theorem 6.3. *If no point in $\mathbb{R}^3$ is contained in more than t balls of the k-fold covering of $\mathscr{F}$ and $et2^{18} \leq 2^{k-1}$, then $\mathscr{F}$ is decomposable.*

Proof. Let $\{C_j\}_{j \in J}$ be the set of connected components obtained when removing the boundaries of the balls from $\mathbb{R}^3$, or, in other words, subsets not separated by the boundary of any of the balls. Then, construct an infinite hypergraph $H = (V, E)$, where $V = \{B_i\}_{i \in I}$ is the original $\mathscr{F}$, and an edge $E_j(H) = \{B_i : i \in I, C_j \subseteq B_i\}$. Several edges are connected if they contain some common C_j. Then, because $\mathscr{F}$ is a k-fold covering, each hyperedge E_j contains at least k vertices.

Claim. The statement that no point in $\mathbb{R}^3$ is contained in more than t balls implies that each E_j intersects at most $t^3 2^{18}$ other E_ℓ.

The proof will be deferred and isn't probabilistic anyways.

Consider any finite sub-hypergraph L of H (i.e. there are only finitely many edges). Then, each edge of L has at least k vertices, and intersects at most $d < t^3 2^{18}$ other edges. Thus, $e(d_1) \leq 2^{k-1}$, and we saw in Theorem 5.3 that this means there exists a 2-coloring of L in which no edge is monochromatic. Then, the same compactness argument used in the packing problem for Theorem 5.4 allows us to claim that H is 2-colorable. Let $\mathscr{F}_1$ be the set of blue balls and $\mathscr{F}_2$ the set of red balls; then, each $\mathscr{F}_i$ covers every connected component C_i (if not, the corresponding edge in the hypergraph would be monochromatic), and since the balls in question are open, then they must also cover the boundaries, so $\mathscr{F}$ is decomposable.

Now, for the claim check. Fix an edge E_ℓ corresponding to the connected component C_ℓ. Let E_j be any other edge of H (for the component C_j) that intersects E_ℓ. Thus, there exists a B_i such that $B_i \supseteq C_\ell$ and $B_i \supseteq C_j$. Thus, any ball containing C_j must intersect B_i, so all *closed* unit balls that contain or touch C_j must intersect B_i. All of them have to be included within a

[5] This uses a theorem that if $\{A_i\}_{i \in I}$ is a family of compact subsets of X such that for any finite collection $J \subset I$, $\bigcap_{j \in J} A_j \neq \emptyset$, then $\bigcap_{i \in I} A_i \neq \emptyset$ as well.

fixed ball of radius 4, and no point of this ball is covered more than t times, to by a volume argument there are at most $t4^3$ such balls. Then, there are at most $m = t2^6$ such balls, so one can cut $\mathbb{R}^3$ into at most m^3 connected components, and each C_j must be one of them. Thus, $\#C_j \le (t2^6)^3 = t^3 2^{18}$. ⊠

The more diverse tricks you need to prove such statements like this one, the fewer people are able to discover it…

7. Poisson Approximation: Jensen's Inequality: 1/22/14

Suppose $X_1, \dots, X_n$ are indicator random variables (so that they take values 0 or 1) and come with a neighborhood of dependences $B_1, \dots, B_n$, which is to say that X_i is independent of $\{X_j\}_{j \notin B_i}$ (so if the X_i are all independent, each $B_i = \emptyset$). Then, for any $J \subseteq I$, write $X_j = \sum_{i \in J} X_i$.

The idea behind the following theorem is to have a result not too unlike independence in the case of variables that aren't too strongly dependent. Δ represents how dependent they are, and is ideally not too large. $\tilde{M}$ represents the independent case.

Theorem 7.1. *If*

$$\tilde{M} = -\sum_{i \in I} \log(1 - \mathbb{P}(X_i = 1)),$$

then

$$\tilde{M} \ge -\log \mathbb{P}(X_I = 0) \ge \sup_{\theta \in [0,1]} (\theta \tilde{M} - \theta^2 \Delta / 2(1 - \varepsilon)), \tag{2}$$

where

$$\Delta = \sum_{i \in I} \mathbb{P}(X_i = 1) \sum_{\substack{j \in B_i \\ j \ne i}} \mathbb{P}(X_j = 1 \mid X_i = 1)$$

and $\varepsilon = \max_{i \in I}\{\mathbb{P}(X_i = 1)\}$, *under the conditions of positive dependence, i.e. that for all* $J \subseteq I$ *and* $i \notin J$, $\mathbb{P}(X_J = 0 \mid X_i = 1) \le \mathbb{P}(X_J \le 0)$, *and that (another kind of positive dependence) for every* $j \in B_i^c$ *and any* $k \ne i$ *such that* $k \in B_i$, *but* $k \notin J$, *then* $\mathbb{P}(X_j = 0 \mid X_i, X_k = 1) \le \mathbb{P}(X_j = 0 \mid X_k = 1)$.

(2) is known as Jesen's inequality. Notice that the local lemma just showed that the middle term is positive; this states that with a little more information, one can obtain a nicer bound. It's also useful to observe that this is approximately a Poisson distribution, e.g. if X_I is Poission as $|P(X_i = 0) = e^{-\lambda} = \tilde{M}$, then $\lambda = -\log \tilde{M}$. Then, using linearity of expectation,

$$\mu = \mathbb{E}[X_I] = \sum_{i \in I} \mathbb{P}(X_i = 1),$$

and we expect $\lambda = \mu$. The point of Poisson approximation is that since $-\lg(1 - x) \simeq x$ when x is small, one can approximate $\mathbb{P}(X_I = k) \approx e^{-\lambda} \lambda^k / k!$ for $k = 1, 2, \dots$

Proof of Theorem 7.1. For the left-side inequality, it's equivalent to showing that

$$\mathbb{P}(X_I = 0) \ge \prod_{i \in I} \mathbb{P}(X_i = 0) = e^{-\tilde{M}}.$$

Well, $P(X_I = 0) = P(X_1 = 0) P(X_2 = 0 \mid X_1 = 0) P(X_3 = 0 \mid X_1 + X_2 = 0) \cdots$, and so on. Let $J_1 = \{1\}$, $J_2 = \{1, 2\}$, up to $J_{n-1} = \{1, \dots, n-1\}$. Then, rewrite this as

$$\mathbb{P}(X_I = 0) = \prod_{i \in I} \mathbb{P}(X_i = 0) \left(\prod_{i=1}^{n-1} \frac{\mathbb{P}(X_{i+1} = 0 \mid X_{J_i} = 0)}{\mathbb{P}(X_{i+1} = 0)} \right),$$

so it's sufficient to show that $\mathbb{P}(X_{i+1} = 0 \mid X_{J_i} = 0) \ge \mathbb{P}(X_{i+1} \ge 0)$ for $i = 1, 2, \dots, n-1$. Rewriting the left-hand side as $\mathbb{P}(X_{J_i} = 0 \mid X_{i+1} = 0) \mathbb{P}(X_{i+1} = 0) / \mathbb{P}(X_{J_i} = 0)$, this is equivalent to $\mathbb{P}(X_{J_i} = 0 \mid X_{i+1} = 0) \ge \mathbb{P}(X_{J_i} = 0)$, and therefore that $\mathbb{P}(X_{J_i} = 0 \mid X_{i+1} = 1) \le \mathbb{P}(X_{J_i} = 0)$, which is true by the first positive-correlation assumption.

Now, write

$$\tilde{M} + \log \mathbb{P}(X_I = 0) = \sum_{i=1}^{n-1} \log \left(\frac{\mathbb{P}(X_{i+1} = 0 \mid X_J = 0)}{\mathbb{P}(X_{i+1} = 0)} \right) \ge 0, \tag{3}$$

so separate J_i into $J_i^+ = J_i \cap B_{i+1}$ and $J_i^- = J_i \setminus J_i^+$. Then, by the obvious inequality $\mathbb{P}(A \mid B, C) \ge \mathbb{P}(A, B \mid C) = \mathbb{P}(B \mid C) \mathbb{P}(A \mid B, C)$, then

$$\begin{aligned} \mathbb{P}(X_i = 1 \mid X_{J_i} = 0) &\ge \mathbb{P}(X_{i+1} = 1, X_{J_i^+} = 0 \mid X_{J_i^-} = 0) \\ &= \mathbb{P}(X_{i+1} = 1 \mid X_{J_i^-} = 0) \mathbb{P}(X_{J_i^+} = 0 \mid X_{i+1} = 1, X_{J_i^-} = 0). \end{aligned}$$

Since $J_i^- \cap B_{i+1} = \emptyset$, then $\mathbb{P}(X_{i+1} = 0 \mid X_{J_i^-} = 0) = \mathbb{P}(X_{i+1} = 1)$, and so

$$= \mathbb{P}(X_{i+1} = 1)\left(1 - \sum_{j \in J_i^+} \mathbb{P}(X_j = 1 \mid X_{i+1} = 1, X_{J_i -} = 0)\right).$$

By the second positive-correlation condition $\mathbb{P}(X_j = 1 \mid X_{i+1} = 1, X_{J_i -} = 0) \leq \mathbb{P}(X_j = 1 \mid X_{i+1} = 1$, because the conditions work with $J = J_i^+$, $k = i+1$, and $\ell = j \in J_i^+$. Thus, $\mathbb{P}(X_{j_i^-} = 0 \mid X_{i+1} X_j = 1) \leq \mathbb{P}(X_{J_i^-} = 0 \mid X_{i+1} = 1)$, and the result follows because $\mathbb{P}(A \mid B, C) \leq \mathbb{P}(A \mid B)$ iff $\mathbb{P}(C \mid A, B) \leq \mathbb{P}(C \mid B)<$ and then, dividing, $\mathbb{P}(A, B, C)/\mathbb{P}(B, C) \leq \mathbb{P}(A, B)/\mathbb{P}(B)$ iff $\mathbb{P}(C, A, B)/\mathbb{P}(A, B) \leq \mathbb{P}(C, B)/\mathbb{P}(B)$.

Then, we see

$$\mathbb{P}(X_{i+1} = 1 \mid X_{J_i} = 0) \geq \mathbb{P}(X_{i+1} = 1) - \sum_{j \in J_i^+} \mathbb{P}(X_{i+1} X_j = 1),$$

so taking 1 minus the above,

$$\mathbb{P}(X_{i+!} = 0 \mid X_{J_i} = 0) \leq \mathbb{P}(X_{i+1} = 0) + \sum_{j \in J_i^+} \mathbb{P}(X_{i+1} X_j = 1),$$

and therefore by (3),

$$\tilde{M} + \lg \mathbb{P}(X_I = 0) \leq \sum_{i=1}^{n-1} \log\left(1 + \sum_{j \in J_i^+} \frac{\mathbb{P}(X_{i+1} X_j = 1)}{1 - \mathbb{P}(X_{i+1} = 1)}\right).$$

Then, since $\log(1 + x) \leq x$,

$$\leq \sum_{i=1}^{n-1} \mathbb{P}(X_{i+1} = 1) \sum_{\substack{j \in J_i^+ \subseteq B_{i+1} \\ j \leq i+1}} \frac{\mathbb{P}(X_j = 1 \mid X_{i+1} = 1)}{1 - \mathbb{P}(X_{i+1} = 1)} \leq \frac{\Delta}{2(1-\varepsilon)}.$$

This was frighteningly computational, but doesn't require too much thinking or insight. The next part is a little more clever. Clearly, $-\lg \mathbb{P}(X_I = 0) \geq \sup_{S \subseteq I}\{-\lg \mathbb{P}(X_S = 0)\}$ (since the probability that all of them are zero cannot be greater than the probability that some subset of them are zero), and this is greater than or equal to $\mathbb{E}_S[-\lg \mathbb{P}(X_S = 0)]$ over any distribution over subsets S of I which is independent of the X_i. Then, for every such S, by the same proof, $-\lg \mathbb{P}(X_S = 0) \geq \tilde{M}_S - \Delta_S/(1-\varepsilon)$, where

$$M_S = \sum_{i \in S} \lg(1 - \mathbb{P}(X_i = 1))$$

and

$$\Delta_s = \sum_{i \in S} \sum_{\substack{j \in B_i, j \in S \\ j \neq 1}} \mathbb{P}(X_i X_j = 1).$$

Let $\{i \in S\}$ be i.i.d Bernoulli random variables parameterized by θ. Then,

$$\mathbb{E}_S[-\log \mathbb{P}(X_S = 0)] \geq \mathbb{E}_S[\tilde{M}_S] - \frac{1}{2(1-\varepsilon)} \mathbb{E}[\Delta_S].$$

Since this works for any $\theta \in [0, 1]$, one can take their supremum. ⊠

Note that one could instead improve the bound slightly by taking $i \in S$ to be Bernoulli with probability θ_i to get the lower bound

$$\sup_{1 \geq \theta_1, \ldots, \theta_n \geq 0} \left\{ \sum_i \theta_i d_i - \sum_{i,j} \theta_i \theta_j d_{ij} \right\}.$$

This is a complicated expression, but it's quadratic, so you can at least solve it. The given lower bound is less complicated, however.

8. Poisson Approximation to the Binomial: 1/24/14

Let X_i be i.i.d. random variables and $N \gg 1$ and $p \ll 1$, so that $np = \lambda$. Then, let $X_I = \sum_{i \in I} X_i \sim \text{Binomial}(n, p)$, and thus $\mathbb{P}(X_I = s) = \binom{n}{s} p^s (1-p)^{n-s}$ for $s = 0, 1, 2, \ldots, n$.

Then, as $n \to \infty$ and λ, s are fixed, then this is approximately $e^{-\lambda}\lambda^s/s!$, and, as a specia case where $s = 0$, $\mathbb{P}(X_I = 0) \approx e^{-\lambda}$. This also applies with independent variables Bernoulli(p_i), where $i = 1, \ldots, n$, as long as $\varepsilon = \max_{i=1}^n \{p_i\} \to 0$ and $\mu = \sum p_i$ is fixed.

To see why Poisson approximation works (where μ is used to denote λ, to be consistent with the previous lecture),

$$\begin{aligned}
\mathbb{P}(X_I = S) &= \binom{n}{s} p^s (1-p)^{n-s} \\
&= \frac{n!}{(n-s)!s!} \left(\frac{\mu}{n}\right)^s \left(1 - \frac{\mu}{n}\right)^{n-s} \\
&= 1 \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{s-1}{n}\right) \frac{\mu^s}{s!} \left(1 - \frac{\mu}{n}\right)^n \frac{1}{(1-\mu/n)^s},
\end{aligned}$$

but everything goes to 1 except $\frac{\mu^s}{s!}\left(1 - \frac{\mu}{n}\right)^n$, which goes to the Poisson formula.

The point of Jensen's inequality was to use this approximation:

$$\mathbb{P}(X_I = 0) \approx \tilde{M} = \prod_{i=1}^n (1 - p_i) = e^{\sum_{i=1}^n \lg(1-p_i)}.$$

It also used the first-order Taylor approximation $x - x^2/2 \le \log(1-x) \le x$ when $x \to 0$.

Then, the bound on $\mathbb{P}(X_I = 0)$ is $e^{-\mu+\Delta/2} \le \mathbb{P}(X_I = 0) \le e^{-\mu}$, where Δ is roughly the sum of the squares of the p_i, and the overall bounds are $e^{-\tilde{M}}$ and for a lower bound, $e^{\sup_{\theta \in [0,1]}(\theta\tilde{M} - \theta^2\Delta/2(1-\varepsilon)}$. Note that in Alon and Spencer, the positive dependence conditions we gave aren't mentioned, and they use a special case which we haven't discussed yet.

A lot of classical approximations (e.g. the Central Limit theorem) depend on the random variables being independent, but many problems are time-ordered, so the dependencies are nontrivial, but simple. This lends itself to approximation by independent variables. In other applications, such as graph theory, the dependence network might be much more complicated, so these sorts of approximations (such as the ones developed in these few lectures) are better if they aren't time-dependent.

Take $X_i = \mathbb{1}_{A_i \subseteq R}$ for $i \in I$, where $A_i \subseteq \Omega$ are non-random subsets and the random set R is obtained by putting each $r \in \Omega$ inside R independently with probability $p_r \in [0, 1]$ (hopefully farther away from 1, because the maximal value appears in the bound). Then, require the following conditions:

(A) $\mathbb{P}(X_J = 0 \mid X_i =) \le \mathbb{P}(X_J = 0)$, or equivalently, $\mathbb{P}(X_I = 1 \mid X_J = 0) \varepsilon \mathbb{P}(X_i = 1)$ for all $J \subseteq I$ and $i \notin J$.

(B) $\mathbb{P}(X_J = 0 \mid X_i X_k = 1) \le \mathbb{P}(X_J = 0 \mid X_k = 1)$ for any $i \ne j$ not in J. Equivalently, one has $\mathbb{P}(X_i = 1 \mid X_J = 0, X_k = 1) \le \mathbb{P}(X_i = 1 \mid X_k = 1)$.

These equivalences are shown by moving conditional probabilities around and doing algebra.

These follow from something called the FKG property: that if one has two increasing functions, the expectation of their product is greater than the product of their expectations. This is applied here where the function just increases from 0 to 1, since the X_i are indicators. In Greek letters, the inequality states that

$$\mathbb{E}[f(\xi_1, \ldots, \xi_n) g(\xi_1, \ldots, \xi_n)] \ge \mathbb{E}[f(\xi_1, \ldots, \xi_n)] \mathbb{E}[g(\xi_1, \ldots, \xi_n)],$$

where f and g are monotone and the ξ_i are independent. Alon & Spencer prove this in Chapter 6. Another way of understanding this inequality is that monotone functions over independent variables are positively correlated.

Then, one can show (A) from this inequality: using the definition of conditional probability, $\mathbb{P}(X_i = 1, X_J = 0) \le \mathbb{P}(X_i = 1)\mathbb{P}(X_J = 0)$. Then, $\mathbb{E}[\mathbb{1}_{A_i \subseteq R} \mathbb{1}_{A_j \not\subseteq R \text{ for all } j \in J}] \le \mathbb{E}[\mathbb{1}_{A_i \subseteq R}] \mathbb{E}[\mathbb{1}_{A_j \not\subseteq R \text{ for all } j \in J}]$, where $f(\xi) = \mathbb{1}_{A_i \subseteq R}$, $g(\xi) = \mathbb{1}_{A_j \not\subseteq R \text{ for all } j \in J}$, and $\xi = 1\{r \in R\}$. This makes the proof easier, but showing the FKG property is nontrivial, and it's quite possibly easier to just assume the result of the property as a condition.

Definition. A set $J \subseteq I$ is a disjoint family if for all $i, j \in J$, $i \notin B_j$, so that X_i and $\{X_j\}_{j \in J}$ are indepedent for all $i \in I$.

Lemma 8.1. $\mathbb{P}$(*there exists a disjoint family J with $X_J \ge s$*) $\le \mu^s/s!$

Lemma 8.2. *The probabiity that there exists a disjoint family J with $X_J = s$, and if $X_i = 1$ for some $i \notin J$, then $J \cup \{i\}$ is not disjoint is at most* $(\mu^s e^{-\mu}/s!) e^{sr} e^{\Delta/(2(1-\varepsilon))}$.

The whole reason for taking a disjoint family is that one can take points outside of a given point's neighborhood of dependence. This is easier than carrying around a lot of words about whether it's possible to have counterexamples or multiples.

Proof of Lemma 8.1. Notice that if J is joint and $\tilde{J} \subseteq J$, then $\tilde{J}$ doesn't have any new dependence relations, so it's also disjoint. Thus,

$$\mathbb{P}(\text{there exists a disjoint } J \text{ such that } X_J \geq s) \leq \sum_{\substack{|J|=s \\ J \text{ disjoint}}} \mathbb{P}(X_J = s) = \sum_{\substack{|J|=s \\ J \text{ disjoint}}} \prod_{i \in J} \mathbb{P}(X_i = 1),$$

because the X_i are independent on $\{X_j \mid j \neq i\}$. But since there are $s!$ possible permutations of $\{j_1, \dots, j_s\}$, then this is

$$\begin{aligned} &\leq \frac{1}{s!} \sum_{j_1 \neq j_2 \neq \cdots \neq j_s} \prod_{\ell=1}^{s} \mathbb{P}(X_{j_\ell} = 1) \\ &\leq \frac{1}{s!} \left(\sum_r \mathbb{P}(X_r = 1) \right)^s = \frac{\mu^s}{s!}. \end{aligned}$$

⊠

Notice that the $e^{-\mu}$ term vanishes into the union bound, and in some cases this is fine, but in others it causes the bound to be greater than 1, which is less useful. This is why Lemma 8.2 exists, though its derivation is slightly more involved.

Proof of Lemma 8.2. Continuing with the reasoning from the proof of Lemma 8.1,

$$\mathbb{P}(\text{there exists a disjoint } J \text{ such that } X_J = s) \leq \sum_{\substack{|J|=s \\ J \text{ disjoint}}} \prod_{i=1}^{s} \mathbb{P}(X_{j_i} = 1) \mathbb{P}(X_{I \setminus \bigcup_{i \in J} B_i} = 0).$$

Then, the rightmost term, not present in the previous proof, is what we want to end up in a term like $e^{-\mu}$. Then, applying Jensen's inequality,

$$\leq e^{-\sum_{\theta \in [0,1]} \left(\theta \tilde{M}_J - \theta^2 \Delta_J / (2(1-\varepsilon)) \right)}.$$

Here, $\tilde{M}_J$ and Δ_J are given by replacing I with $I \setminus \bigcup_{i \in J} B_i$. It then happens that $\Delta_J \leq \Delta$ and $\tilde{M}_j \geq \tilde{M} - sr$, so once this is shown take $\theta = 1$. For Δ it is at least clear: Δ is formed of sums of pairs of positive things, so removing some terms makes it smaller, or at least not larger. ⊠

9. The Chen-Stein Method: 1/27/14

Let $Z_\lambda = \text{Po}(\lambda)$, i.e. Z_λ is a Poisson random variable, so that $\mathbb{P}(Z_\lambda = k) = e^{-\lambda} \lambda^{-k} / k!$ Then, let W be a random variable on $\mathbb{Z}^+$ coming from some other application (e.g. graph theory) which is believed to be close to Z_λ (or some sequence W_n that approaches Z_λ). Then, one can use the Chen-Stein method (also the Stein-Chen method) to approximate W with a Poisson distribution, which is a special case of the more general Stein method for approximating one distribution with another.

Claim. For any $f : \mathbb{Z}^+ \to \mathbb{R}$, let $(Tf)(w) = wf(w) - \lambda f(w+1)$. Then, if

$$\sum_{w=0}^{\infty} (wf(w) - \lambda f(w+1)) p(w) = 0$$

for all $f : \mathbb{Z}^+ \to \mathbb{R}$, then $p(w) = e^{-\lambda} \lambda^w / w!$

This is an interesting way to show that a function has a particular distribution, and if one can show this is close to zero, then the distribution can be approximated by a Poisson distribution. Interestingly, this was discovered only 30 years ago, but is no longer a subject of active research; it's been dealt with.

It's fairly easy to check that the Poisson distribution satisfies the claim, but showing that anything else doesn't is a little harder.

Let S be any linear operator such that $TS = I$ (where I is the identity), and consider the family of functons $\mathscr{H} = \{h : \mathbb{Z}^+ \to \{0,1\}\}$ and $\mathscr{F} = \{f : f = S(h - \overline{h})\}$, where $\overline{h} = \mathbb{E}[h(Z_\lambda)]$. Then, $\mathbb{E}[Tf(W)] = \mathbb{E}[h(W)] - \mathbb{E}[h(Z_\lambda)]$ for all $f \in \mathscr{F}$. Then, there is a quantity called the total variation distance

$$d(W, Z_\lambda) = ||W - Z_\lambda||_{\text{t,v,}} = \sup_{f \in \mathscr{F}} \mathbb{E}[Tf(W)] = \sup_{h \in \mathscr{H}} \mathbb{E}[h(W)] - \mathbb{E}[h(Z_\lambda)] = \sup_{A \subseteq \mathbb{Z}^+} \{\mathbb{P}(W \in A) - \mathbb{P}(Z_\lambda \in A)\}.$$

Now, one can define two constants that only depend on λ:

$$c_1 = \frac{1}{\lambda}(1-e^{-\lambda}) \geq \sup_{f\in\mathscr{F}} \|f(\cdot+1)-f(\cdot)\|_\infty$$
$$c_2 = \min\left(1, \sqrt{\frac{2}{\lambda}}\right) \geq \sup_{f\in\mathscr{F}} \|f\|_\infty .$$

Now, the goal is to bound the total variation distance in terms of c_1, c_2, and some conditions on W. This will allow one to make a Poisson approximation. Specifically, assume $W = \sum_{\alpha\in I} X_\alpha$ and $\lambda = \sum_{\alpha\in I} p_\alpha = \mathbb{E}[W]$, where $X_\alpha \in \{0,1\}$ an $p_\alpha = \mathbb{P}(X_\alpha = 1)$.[6] Then, we want the following to be small:

$$\begin{aligned}\mathbb{E}[Tf(W)] &= \mathbb{E}\left[\sum_{\alpha\in I}(X_\alpha f(W) - p_\alpha f(W+1))\right] \\ &= \sum_{\alpha\in I}\mathbb{E}[X_\alpha f(W) - p_\alpha f(W+1)]\end{aligned}$$

Let $a_\alpha = \mathbb{E}[X_\alpha f(W) - p_\alpha f(W+1)]$ for any $\alpha \in I$, so the above sum becomes $\sum_\alpha a_\alpha$. Let also $W = X_\alpha + W_\alpha$, so that $W_\alpha = \sum_{\beta\neq\alpha} X_\beta$. Then,

$$\begin{aligned}a_\alpha &= \mathbb{E}[X_\alpha f(W) - p_\alpha f(W+1)] \\ &= p_\alpha \mathbb{E}[f(W_\alpha+1)\,|\,X_\alpha = 1] - p_\alpha^2\mathbb{E}[f(W_\alpha+2)\,|\,X_\alpha=1] - (1-p_\alpha)p_\alpha\mathbb{E}[f(W_\alpha+1)\,|\,X_\alpha=0] \\ &= p_\alpha(1-p_\alpha)(\mathbb{E}[f(W_\alpha+1)\,|\,X_\alpha=1] - \mathbb{E}[f(W_\alpha+1)\,|\,X_\alpha=0)]) + p_\alpha^2\mathbb{E}[(f(W_\alpha+1)-f(W_\alpha+2))\,|\,X_\alpha=1].\end{aligned}$$

Notice that if X_α and W_α are indepedent, then the first term vanishes and the second is bounded above by c_1. Thus, one might want a random variable V_α in the same probability space which is "more independent" of X_α and "not too far from" W_α. In this case,

$$\begin{aligned}a_\alpha = {} & p_\alpha(1-p_\alpha)(\mathbb{E}[f(W_\alpha+1)-f(V_\alpha+1)\,|\,X_\alpha=1] - \mathbb{E}[f(W_\alpha+1)-f(V_\alpha+1)]) \\ & + p_\alpha(1-p_\alpha)(\mathbb{E}[f(V_\alpha+1)\,|\,X_\alpha=1] - \mathbb{E}[f(V_\alpha+1)\,|\,X_\alpha=0]) \\ & + p_\alpha^2\mathbb{E}[f(W_\alpha+2)-f(W_\alpha+1)\,|\,X_\alpha=0]\end{aligned}$$

Let I denote the first term, II the second term, and III denote the third term. Then, since $|f(x)-f(y)| \leq c_1|x-y|$ (the Lipschitz norm of f is at most c_1),

$$\begin{aligned}I &\leq c_1 p_\alpha(1-p_\alpha)(\mathbb{E}[|W_\alpha - V_\alpha|\,|\,X_\alpha=1] + \mathbb{E}[|W_\alpha - V_\alpha|\,|\,X_\alpha=0]), \\ III &\leq c_1 p_\alpha^2, \text{ and} \\ II &= p_\alpha(1-p_\alpha)\left|\sum_{k=0}^\infty f(k+1)(\mathbb{P}(V_\alpha=k\mid X_\alpha=1) - \mathbb{P}(V_\alpha=k\mid X_\alpha=0))\right| \\ &\leq c_2 p_\alpha(1-p_\alpha) d_{\mathrm{tv}}(V_\alpha|_{X_\alpha=1}, V_\alpha|_{X_\alpha=0}).\end{aligned}$$

This last inequality follows because for any distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$\sum_{k=0}^\infty |\mathbb{P}(k)-\mathbb{Q}(k)| = 2\sup_{A\subseteq\mathbb{Z}^+}(\mathbb{P}(A)-\mathbb{Q}(A)) = 2(\mathbb{P}(A^*)-\mathbb{Q}(A^*)),$$

where $A^* = \{k : \mathbb{P}(k) \geq \mathbb{Q}(k)\}$. Thus, one has the following theorem.

Theorem 9.1. *For all random variables V_α,*

$$\begin{aligned}d_{\mathrm{tv}}(W, Z_\lambda) \leq {} & c_1\sum_{\alpha\in I} p_\alpha^2 + c_1\sum_{\alpha\in I} p_\alpha(1-p_\alpha)(\mathbb{E}[|W_\alpha - V_\alpha|\,|\,X_\alpha=1] + \mathbb{E}[|W_\alpha - V_\alpha|\,|\,X_\alpha=0]) \\ & + 2c_2\sum_{\alpha\in I} p_\alpha(1-p_\alpha) d_{\mathrm{tv}}(V_\alpha|_{X_\alpha=1}, V_\alpha|_{X_\alpha=0}).\end{aligned}$$

[6] A lot of this will be the same in the case of a normal distribution, i.e. for the Central Limit theorem. However, this means that the high-level ideas, including that of Taylor approximation, are the same, but the equations and constants are different.

Thus, the idea that V_α should be not too far from W_α is refined into a condition on the L_1-norm. For example, if X_α and V_α are independent for all α, the Poisson approximation if for all α, $\mathbb{E}[|W_\alpha - V_\alpha| \,|\, X_\alpha = 1] \to 0$ and $\mathbb{E}[|W_\alpha - V_\alpha|] \to 0$. In general, there may be neighborhoods of dependence, but in relatively small ways, which makes approximation nicer.

Another interesting case is the symmetric one, where for all α, $p_\alpha = \lambda/n$. Then,

$$d(W, Z_\lambda) \le (1-e^{-\lambda})\frac{\lambda}{n} + (1-e^{-\lambda})\left(1-\frac{1}{n}\right)(\mathbb{E}[|W_1 - V_1| \,|\, X_1]1] + \mathbb{E}[|W_1 - V_1| \,|\, X_1 = 0])$$
$$+ 2\min\left(1, \sqrt{\frac{2}{\lambda}}\right)\left(1-\frac{\lambda}{n}\right)\lambda d_{\text{tv}}(V_1|_{X_1=1}, V_1|_{X_1=0}).$$

In this sense, you only need one of the α to know all of them.

If $X \sim \text{Bernoulli}(p)$ and Γ is any random variable, then one has the identity

$$\mathbb{E}(p + (1-2p)X)\Gamma] = p(1-p)\mathbb{E}[\Gamma \mid X = 0] + p(1-p)\mathbb{E}[\Gamma \mid X = 1],$$

so one can use a slightly different bound in Theorem 9.1:

$$\sum_{\alpha \in I} p_\alpha(1-p_\alpha)(\mathbb{E}[|W_\alpha - V_\alpha| \,|\, X_\alpha = 1] + \mathbb{E}[|W_\alpha - V_\alpha| \,|\, X_\alpha = 0]) = \sum_\alpha p_\alpha \mathbb{E}[|W_\alpha - V_\alpha|] + (1-2p_\alpha)\mathbb{E}[X_\alpha | W_\alpha - V_\alpha|].$$

Then, if $V_\alpha = \sum_{\gamma \notin B_\alpha} X_\gamma$ and X_α is independent to the X_γ where $\gamma \notin B_\alpha$, then

$$I = \sum_\alpha p_\alpha \mathbb{E}[W_\alpha - V_\alpha] + (1-2p_\alpha)\mathbb{E}[(W_\alpha - V_\alpha)X_\alpha]$$
$$= \sum_\alpha p_\alpha \left(\sum_{\beta \in B_\alpha} p_\beta + (1-2p_\alpha)\sum_{\beta \in B_\alpha} p_{\alpha\beta}\right),$$

where $p_{\alpha\beta} = \mathbb{P}(X_\alpha = X_\beta = 1)$, their correlation. Hopefully this is on the order of $1/n^2$, so that the approximation is easier: $I \approx |B_\alpha|/n + n|B_\alpha| p_{\alpha\beta}$. If $p_{\alpha\beta} \approx c/n^{1+\gamma}$, the second term is just $|B_\alpha|/n^{1-\gamma}$.

If all one wants is $|\mathbb{P}(W=0) - e^{-\lambda}|$, one doesn't need all of this; it evaluates as $|\mathbb{P}(W=0) - e^{-\lambda}| \le c_1$.

10. The Probabilistic and Coupling Methods: 1/29/14

Recall the Chen-Stein method introduced last lecture, and in chapter 1 of [2]. Interestingly enough, this was done in Chen's PhD thesis while he was a graduate student at Stanford!

Let $W = \sum_\alpha X_\alpha$, where $X_\alpha \in \{0,1\}$, $p_\alpha = \mathbb{P}(X_\alpha) = 1$, and $\lambda = \mathbb{E}[W] = \sum_\alpha p_\alpha$. Then, $Z_\lambda \sim \text{Po}(\lambda)$. The total variation distance was shown to be

$$d_{\text{tv}}(W, Z_\lambda) \le \left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_\alpha (b_\alpha^{(1)} + b_\alpha^{(2)}) + \min\left(1, \sqrt{\frac{2}{\lambda}}\right)\sum_\alpha b_\alpha^{(3)}$$

and

$$|\mathbb{P}(W=0) - e^{-\lambda}| \le \left(\frac{1-e^{-\lambda}}{\lambda}\right)\sum_\alpha (b_\alpha^{(1)} + b_\alpha^{(2)} + b_\alpha^{(3)}),$$

where

$$b_\alpha^{(1)} = p_\alpha^2 + p_\alpha \mathbb{E}[|W_\alpha - V_\alpha|]$$
$$b_\alpha^{(2)} = (1-2p_\alpha)\mathbb{E}[X_\alpha | W_\alpha - V_\alpha|]$$
$$b_\alpha^{(3)} = 2p_\alpha(1-p_\alpha)d_{\text{tv}}(V_\alpha|_{X_\alpha=0}, V_\alpha|_{X_\alpha=1})$$
$$= \mathbb{E}[|\mathbb{E}[X_\alpha - p_\alpha \mid V_\alpha]|].$$

It's also convenient to write $b_1 = \sum b_\alpha^{(1)}$, and so on.

Example 10.1. Let $Y_1, \dots, Y_n$ be i.i.d., nonnegative random variables. Set $X_1 = 1$ iff $Y_1 + \cdots + Y_r \le \varepsilon_{n,r}$, $X_2 = 1$ iff $Y_2 + \cdots + Y_{r+1} \le e_{n,r}$, and so on up to $X_{n-r+1} = 1$ iff $Y_{n-r+1} + \cdots + Y_n \le \varepsilon_{n,r}$. This has a nice small neighborhood of dependence, and in effect counts how often these sums are very small.

Let $W = \sum_{i=1}^{n-r+1} X_i$ and $\lambda = (n-r+1)\mathbb{P}(Y_1, \dots, Y_r \le \varepsilon_{n,r})$. Then, the neighborhood of dependence is $B_i = \{i-(r-1), \dots, i+(r-1)\}$ and thus $b_3 = 0$, $|B_i \setminus i| = 2(r-1)$, $|B_i| = 2r-1$, and $p_i = \lambda/(r-r+1)$. Thus, one can compute $b_1 = p_i^2|B_i|(n-r+1) = \lambda^2(2r-1)/(n-r+1)$.

Then, calculating b_2 is a little harder:

$$b_2 \le \lambda(2r-1)ax_{s=1,\dots,r-1}\mathbb{P}(Y_{s+1}+\cdots+Y_{s+r}\le \varepsilon_{n,r} \mid Y_1,\dots,Y_r \le \varepsilon_{n,r}).$$

Since the Y_i are nonnegative, one has the upper bound

$$b_2 \le \mathbb{P}(Y_{r+1}+\cdots+Y_{r+s}\le \varepsilon_{n,r}) \le \mathbb{P}(Y_1 \le \varepsilon_{n,r}).$$

Thus, $b_1 \to 0$ as long as $r/n \to 0$, so as long as $\varepsilon_{n,r} \to 0$, then $b_2 \to 0$ and r is bounded, making this Poisson approximation possible. For example, if these are independent arrivals given by some Poisson process, then $Y_i \sim \text{Exp}(1)$ and $Y_1+\cdots+Y_r \sim \Gamma(r,1)$. Here,

$$p_i = 1-e^{-\varepsilon_{n,r}}\sum_{k=0}^{r-1}\frac{\varepsilon_{n,r}^k}{k!} = e^{-\varepsilon_{n,r}}\sum_{k=r}^{\infty}\frac{\varepsilon_{n,r}^k}{k!}.$$

Then, one can calculate that $\varepsilon = \varepsilon_{n,r} \approx \sqrt[r]{\lambda r!/n}$, and thus that $\mathbb{P}(Y_1 \le \varepsilon) \approx \varepsilon$, with error about $O(n^{-1/r})$.

A more elaborate version of that: let $X_1,\dots,X_n$ be i.i.d $U[0,1]$ (i.e. uniformly distributed on $[0,1]$) random variables. Then, the goal is to generate order statistics $0 \le X_1^* \le X_2^* \le \cdots \le X_n^* \le 1$.

A known result from classical probability states that

$$(X_1^*, X_2^*-X_1^*, X_3^*-X_2^*, \dots, X_n^*-X_{n-1}^*, 1-X_n^*) = \left(\frac{Y_1}{X_{n+1}}, \frac{Y_2}{S_{n+1}}, \dots, \frac{Y_n}{S_{n+1}}, \frac{Y_{n+1}}{S_{n+1}}\right),$$

where $S_{n+1} = \sum_{i=1}^{n+1} Y_i$, where the Y_i are i.i.d. with distribution Exp(1). This seems kind of magical — what doe the uniform and exponential distributions have to do with each other?

Let

$$W^* = \sum_{i=0}^{n+1-r} \mathbb{1}_{\{X_{i+r}^*-X_i^* \le \varepsilon_{n,r}/n\}},$$

so that W^* counts how many groups of r samples are within an interval of size $\varepsilon_{n,r}/n$. Thus, it can be rewritten as

$$W^* = \sum_{i=0}^{n+1-r} \mathbb{1}_{\{(Y_i+\cdots+Y_{i+r-1}) \le \varepsilon_{n,r}(S_{n+1}/n\}}).$$

So since $S_{n+1}/n \to 1$ as $n \to \infty$ quite rapidly, then in approximation $d_{\text{tv}}(W^*, Z_\lambda) \le d_{\text{tv}}(W, Z_\lambda) + \mathbb{P}(|S_{n+1}/n - 1|) \ge \delta_n$.

This local method lends itself to some of the examples from [1], e.g. the graph theoretic examples. But there's another way, called the coupling method, which chooses $V_\alpha = \sum_{\beta\neq\alpha} J_{\beta\alpha}$ and $W_\alpha = \sum_{\beta\neq\alpha} X_\beta$ for some $J_{\beta\alpha} \in \{0,1\}$. Here, $V_\alpha = [W_\alpha \mid X_\alpha = 1]$.

Then, define

$$\begin{aligned} a_\alpha &= \mathbb{E}[X - \alpha f(W) - p_\alpha f(W+1)] \\ &= p_\alpha \mathbb{E}[f(W) \mid X_\alpha = 1] - p_\alpha \mathbb{E}[f(W+1)] \\ &= p_\alpha \mathbb{E}[f(V_\alpha+1)] - p_\alpha \mathbb{E}[f(W_\alpha+X_\alpha+1)] \\ &= p_\alpha \mathbb{E}[f(V_\alpha+1)0f(W_\alpha+X_\alpha+1)] \\ |a_\alpha| &\le c_1 p_\alpha(p_\alpha + \mathbb{E}[|V_\alpha - W_\alpha|]). \end{aligned}$$

if it's possible to choose $J_{\beta\alpha} \le X_\beta$, which is called a monotone coupling. Then,

$$\begin{aligned} \sum_\alpha |a_\alpha| &\le c_1 \sum_\alpha p_\alpha \mathbb{E}[W_\alpha + X_\alpha - V_\alpha] \\ &= c_1\left(\sum_\alpha p_\alpha^2 + \sum_{\substack{a\in I \\ \beta\neq\alpha}} p_\alpha p_\beta - \sum_{\substack{\alpha\in I \\ \beta\neq\alpha}} p_{\alpha\beta}\right). \end{aligned}$$

In this case, the Poisson approximation is particularly nice, iff we have the nice bound on V and $W\dots$ but this doesn't work in the regular case.

Theorem 10.1 (Theorem 2C of [2]). *Partition $I \setminus \alpha$ as follows: if $J_{\beta\alpha} \ge X_\beta$, let $\beta \in I_\alpha^+$; if $J_{\beta\alpha} \le X_\beta$, then $\beta \in I_\alpha^-$; and if neither, then $\beta \in I_\alpha^0$.*

Then, $b_\alpha^{(1)} = p_\alpha^2$,

$$b_\alpha^{(2)} = \sum_{\beta\notin I_\alpha^0} \left|\text{Cov}(X_\alpha, X_\beta)\right|,$$

and

$$b_\alpha^{(3)} = \sum_{\beta \in I_\alpha^0} (p_\alpha p_\beta + p_{\alpha\beta}).$$

Proof.

$$\mathbb{E}\left[\left|\sum_{\beta\neq\alpha}(X_\beta - J_{\beta\alpha})\right|\right] \le \mathbb{E}\left[\sum_{\beta\in I_\alpha^-}(X_\beta - J_{\beta\alpha})\right] - \mathbb{E}\left[\sum_{\beta\in I_\alpha^+}(X_\beta - J_{\beta\alpha})\right] + \mathbb{E}\left[\sum_{\beta\in I_\alpha^0}(X_\beta - J_{\beta\alpha})\right].$$

Thus, $p_\alpha \mathbb{E}[X_\alpha + J_{\beta\alpha}] = p_\alpha p_\beta + p_{\alpha\beta}$ and

$$\mathrm{Cov}(X_\alpha, X_\beta) = \mathbb{E}[X_\alpha X_\beta - p_\alpha p_\beta] = p_\alpha(\mathbb{E}[X_\beta \mid X_\alpha = 1] - \mathbb{E}[X_\beta]) = p_\alpha \mathbb{E}(J_{\alpha\beta} - X_\beta),$$

which is nonnegative in I_α^+ and nonpositive in I_α^-. ⊠

let σ be a uniformly random permutation of $\{1,\dots,n\}$ and $X_{(i,j)} = 1$ if $\sigma(i) = j$ snd 0 otherwise. Let $c(i,j) \in \{0,1\}$ be nonrandom, and let

$$W = \sum_{i=1}^n c(i,j) X_{(i,j)}.$$

For example, one might take $c(i,j) = 1$ iff $j \in F(i)$ for some function F. Finally, define

$$\pi_i = \frac{1}{n}\sum_{j=1}^n c_{ij}$$
$$\rho_j = \frac{1}{n}\sum_{i=1}^n c_{ij},$$

so that π_i is the number of $c(i,j) = 1$ over n, and similarly with ρ_j.

Theorem 10.2 (Theorem 4A in [2])**.**

$$d_{tv}(W, Z_\lambda) \le \frac{3}{2} c_1 \left(\sum_{i=1}^n \pi_i^2 + \sum_{j=1}^n \rho_j^2 - \frac{2}{3}\frac{\lambda}{n}\right).$$

Proof. Let

$$\lambda = \frac{1}{n}\sum_{i,j=1}^n c(i,j) = \sum_i \pi_i = \sum_j \rho_j,$$

and then take the coupling $\alpha = (i,j)$ and $I = \{(i,j) : i,j = 1,\dots,n\}$. If $X_\alpha = 1$, set $J_{\beta\alpha} = X_\beta$, and if $X_\alpha = 0$, then modify σ to σ^* by transposing j and σ_i, i.e. $\sigma_i^* = j$ and letting $J_{\beta\alpha} = X_\beta(\sigma^*)$.

Then, the rest of the proof is computation, and will be computed next time. ⊠

11. More on the Coupling Method: 1/31/14

Recall that last time we derived the coupling approach to Poisson approximation: if $W = \sum_\alpha X_\alpha$, $\lambda = \sum_\alpha p_\alpha$, and $Z_\lambda \sim \mathrm{Po}(\lambda)$, where $p_\alpha = \mathbb{E}[X_\alpha]$ and $p_{\alpha\beta} = \mathbb{E}[X_\alpha X_\beta]$, with the $X_\alpha \in \{0,1\}$. Then, et $k_2 = (1 - e^{-\lambda})/\lambda$. The idea is to couple these in some space with a set of $\{X_\beta\}$, with $\beta \neq \alpha$ and lying in I_α^+ if $J_{\beta\alpha} \ge X_\beta$, and so on. und derived was

$$d_{\mathrm{tv}}(W, Z_\lambda) \le k_2 \left(\sum_\alpha p_\alpha^2 + \sum_{\alpha\neq\beta\notin I_\alpha^0} \left|\mathrm{Cov}(X_\alpha, X_\beta)\right| + \sum_{\alpha\neq\beta\in I_\alpha^0} (p_\alpha p_\beta + p_{\alpha\beta})\right). \tag{4}$$

Looking back at Theorem 10.2 from last time, there is still some computation left in order to finish the proof. The coupling was to choose σ uniformly at random, and if $X_\alpha = 1$ (i.e. $\sigma(i) = j$), then let $J_{\alpha\beta} = X_\alpha$, and if $X_\alpha = 0$, then let σ^* be obtained from σ by adding an extra transposition between $\sigma(i)$ and j, forcing $X_\alpha = 1$ (which bubbles back to the first case). This means that $I_\alpha^0 = \emptyset$, and $I_\alpha^- = \{(i,\ell),(k,j) : \ell \neq j, k \neq i\}$. Then, $I_\alpha^+ = \{(k,\ell) : k \neq i, \ell \neq j\}$. Thus, the last term in (4) drops out, and

one has that $p_\alpha = 1/n$, so if $\beta \in I_\alpha^-$, then $J_{\beta\alpha} = 0$, because $\mathbb{E}[X_\alpha X_\beta] = \mathbb{E}[X_\alpha J_{\beta\alpha}] = 0$. Then, $\mathrm{Cov}(X_\alpha, X_\beta) = -1/n^2$ if $\beta \in I_\alpha^-$, and if $\beta \in I_\alpha^+$, then $\mathbb{E}[X_\alpha X_\beta] = 1/(n(n-1))$, so $\mathrm{Cov}(X_\alpha, X_\beta) = 1/(n^2(n-1))$. Thus, (4) becomes

$$\begin{aligned} d_{\mathrm{tv}}(W, Z_\lambda) &\le k_2 \left(\frac{\lambda}{n} + \sum_{\alpha \neq \beta} c_\alpha c_\beta \left| \mathrm{Cov}(X_\alpha, X_\beta) \right| \right) \\ &\le k_2 \left(\frac{\lambda}{n} + \sum_{\alpha \neq \beta} c_\alpha c_\beta \left(\frac{1}{n-1} \mathbb{1}_{\beta \in I_\alpha^+} + \mathbb{1}_{\beta \notin I_\alpha^+} \right) \right). \end{aligned}$$

Notice this is different than the Chen-Stein method, in which everything eventually (after a bunch of algebra) goes to zero.

Let $K_{n,p} = G_{n,p}$ be the Erdős-Rényi random graph on n vertices, i.e. where each edge is added independently with probability p. Let Γ be some collection of subgraphs of K_n, e.g. all of the triangles. In this question, one wants to count how many elements of Γ are present. Let α be a specific subgraph, and $W = \sum_{\alpha \in \Gamma} X_\alpha$, where $X_\alpha = 1$ if $\alpha \subseteq G_{n,p}$ and 0 otherwise.

Thus, one can compute $p_\alpha = \mathbb{E}[X_\alpha] = p^{|\alpha|}$. Let $J_{\beta\alpha} = X_\beta(G_{n,p} \cup \{\alpha\})$, so that $J_{\beta\alpha} \ge X_\beta$ and $I_\alpha^- = I_\alpha^0 = \emptyset$. Thus, the total variation is

$$d_{\mathrm{tv}}(W, Z_\lambda) \le k_2 \left(\sum_{\alpha \in \Gamma} p^{2|\alpha|} + \sum_{\substack{\alpha \neq \beta \\ \alpha, \beta \in \Gamma \\ \alpha \cap \beta \neq \emptyset}} \left(p^{|\alpha \cup \beta|} - p^{|\alpha| + |\beta|} \right) \right).$$

Special choices of Gamma make these bounds simpler or more explicit; for example, if Γ is the set of triangles in the graph,

Moving into the ideas of large deviations and concentration of measure, let $X_1, \dots, X_n$ be i.i.d. random variables in $\mathbb{R}^d$, and let $S_n = \sum_{i=1}^n X_i$. Then, the goal is to provide an upper bound $\mathbb{P}(S_n/n \in K)$ for some closed, convex set K and large n. In some sense, this will go to zero, but what matters is how quickly it does this.

Claim.

$$\mathbb{P}\left(\frac{S_n}{n} \in K \right) \le e^{-n \inf_{x \in K} \Lambda^*(x)},$$

where

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} \left(\langle \lambda, x \rangle - \log \mathbb{E}[e^{\langle \lambda, X \rangle_1}] \right).$$

In some special cases, one can explicitly solve this optimization problem and determine $\Lambda^*(x)$.

Proof. The idea is to use Markov's inequality, albeit in an unusual way. First, we have the bound that if $S_n/n \in K$, then there is a point θ in K at least the size of the infimum (in $\langle \theta, S_n/n \rangle$):

$$\begin{aligned} \mathbb{P}\left(\frac{S_n}{n} \in K \right) &\le \mathbb{E}\left[e^{\langle \theta, S_n/n \rangle - \inf_{x \in K} \langle \theta, x \rangle} \right] \\ &= e^{-n \inf_{x \in K} \langle \lambda, x \rangle} \mathbb{E}\left[e^{\left\langle \lambda, \sum_{i=1}^n X_i \right\rangle} \right] \\ &= e^{-n \inf_{xK} \langle , x \rangle} \mathbb{E}\left[\prod_{i=1}^n e^{\lambda, X_i} \right]. \end{aligned}$$

Since these X_i are i.i.d., the expectation and product commute:

$$\begin{aligned} &= e^{-n \inf_{x \in K} \langle \lambda, x \rangle} \prod_{i=1}^n \underbrace{\mathbb{E}\left[e^{\lambda, X_i} \right]}_{e^{\Lambda(\lambda)}} \\ &= \exp(-n \inf_{x \in K} \langle \lambda, x \rangle + n\Lambda(\lambda)). \end{aligned}$$

This uses the fact that $f(y) = (\langle \theta, y \rangle - \inf_{x \in K} \langle \theta, x \rangle) \ge 0$ for any $y \in K$, so

$$\mathbb{P}\left(\frac{S_n}{n} \in K \right) = \mathbb{E}[\mathbb{1}_{\{S_n/n \in K\}}] \le \mathbb{E}[e^{f(S_n/n)}].$$

Now, one can take the best possible bound:

$$\mathbb{P}\left(\frac{S_n}{n} \in K \right) \le \exp\left(-n \sup_{\lambda \in \mathbb{R}^d} (\inf_{x \in K} \langle \lambda, x \rangle - \Lambda(\lambda)) \right).$$

In order to progress further, one needs the following result from analysis, invoking it where $g(\theta, y) = \langle \theta, y \rangle - \Lambda(\theta)$. Then, $\theta \mapsto \log \mathbb{E}[e^{\langle \theta, x \rangle}]$ can be shown to be convex (just differentiate it twice) and lower semicontinuous.

Definition. A continous function is one for which $g(\theta_n) \log(\theta)$ when $\theta_n \to \theta$. To generalize this, a *lower semicontinuous* function is one for which $\liminf_{\theta_n \to \theta} g(\theta_n) \geq \theta$, and upper semicontinuous is defined in the analogous way.

An intreresting example, related to the exponential distribution is

$$\Lambda(\theta) = \log \int_0^\infty e^{-x} e^{\theta x} \, \mathrm{d}x = \begin{cases} \log(1/(1-\theta)), & \theta < 1 \\ \infty, & \theta \geq 1. \end{cases}$$

Theorem 11.1 (Min-Max)**.** *If C is a convex, compact set, $g(y, \theta)$ is lower semicontinuous convex in y and upper semicontinuous in θ, then*

$$\inf_{y \in C} \sup_{\theta} g(\theta, y) = \sup_{\theta} \inf_{y \in C} g(\theta, y).$$

Now, using this theorem (with $C = K \cap H_p$, where H_p is a hypercube of size p in order to guarantee compactness), the intended bound has been shown. ⊠

This theorem statement has not very much regularity, but strong notions of convexity and compactness. Notice also that almost all of the proof works in any topolgical vector space.

As an example, if $X \sim \mathrm{Po}(\theta)$, then $\Lambda^*(x) = \theta - x + x \lg(x/\theta)$ if $x > 0$ (and is infinite when $x < 0$), and if $X \sim \text{Bernoulli}(p)$, then $\Lambda^*(x) = x \lg x/p + (1-x) \lg((1-x)/(1-p)) \triangleq H(x \mid p)$ when $0 \leq x \leq 1$. If X is normally distributed with mean 0 and variance σ^2, $\Lambda^*(x) = x^2/(2\sigma^2)$. For more on this, see [3], though this is just calculus to find the optimal solution. Yet in the more intricate cases, it can be much harer to find a solution.

It's useful to have the Hoeffding bound for this, where we have $a \leq X \leq b$ and $\overline{x} = \mathbb{E}[X]$, but no idea what the actual distribution is. Some general bounds can be given based on this information, relying on the general fact that

$$\log \mathbb{E}[e^{\lambda x}] \leq \log \left(\frac{b - \overline{x}}{b-a} e^{\lambda a} + \frac{\overline{x} - a}{b-a} e^{\lambda b} \right),$$

where $\mathbb{P}(X = b) = p$ and $\mathbb{P}(X = a) = q$, where $p = (\overline{x} - a)/(b-a)$ and $q = (b - \overline{x})/(b-a$. Then, the goal is to find $pb + qa = \overline{x}$ and $p + q = 1$.

12. Large Deviations: 2/3/14

First, some notes about the student presentations:

- Plan for 60 minutes per group of 3, split up as 3 twenty-minute presentations per person.
- Focus on the probabilistic part of the material; technical aspects of economics or discrete mathematics or such should be minimized, or stated with reference to a location for the proof.
- Coordinate the presentations, so that it comes across as one presentation, rather than three. Coordinate notation and content; practice on each other.
- Provide one page of essentials in writing. This is what you believe people might not remember.
- Aim the presentation at our peers, not too high-level nor too basic.
- Of course, make sure to understand the proof before you present it.

The grading is based on two things: clarity and command of the material. Know more than strictly what is presented, in case someone asks a question.

Now, on to concentration inequalities and large deviations, from Chapter 7 of [1] and Chapters 2.4 and 2.1 of [3]. Recall that if K is closed and convex and $\{X_i\}$ are i.i.d $\mathbb{R}^d$-valued random variables, then let $\Lambda(\lambda) = \lg \mathbb{E}[e^{\langle \lambda, X_1 \rangle}]$ and $\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}^d} (\langle \lambda, x \rangle - \Lambda(\lambda))$. Then, $\mathbb{P}(S_n/n \in J) \leq \exp(-n \inf_{x \in K} \Lambda^*(x))$, where S_n is the sum of the X_i.

This isn't yet a concentration inequality, but by choosing $K = [b, \infty)$, $b > \mathbb{E}[X_1]$, and $d = 1$, then $\mathbb{P}(S_n/n \geq b) \leq e^{-n\Lambda^*(b)}$, and if $K = (-\infty, a]$ and $a < \mathbb{E}[X_1]$ (with $d = 1$ again), then $\mathbb{P}(S_n/n \leq a) \leq e^{-b\Lambda^*(a)}$. These can be grouped together: if $\overline{x} = \mathbb{E}[X_1] \in \mathbb{R}$, then

$$\mathbb{P}\left(\left| \frac{1}{n} S_n - \overline{x} \right| \geq \delta \right) \leq e^{-nI(s)}.$$

There are interesting things one can do with this, for example determining a J_n such that $J_n \mathbb{P}(S_n/n \geq b) \to 1$ as $n \to \infty$; a delicate calculation shows that

$$J_n = \frac{1}{\sqrt{2\pi n}\, c(b)} e^{-n\Lambda^*(b)}.$$

If the X_i take values in a finite set, one can show that

$$\lim_{n\to\infty} \frac{1}{n} \lg \mathbb{P}\left(\frac{1}{n} S_n \in K\right) = -\inf_{x\in K} \Lambda^*(x).$$

The method of moment-generating differences, proven in Alon & Spencer, has that

$$\mathbb{P}\left(\left|\frac{1}{n} S_n - \overline{x}\right| \geq \delta\right) \leq e^{-n\delta^2}.$$

Turning to the first example above, suppose $a \leq X_1 \leq b$ and $\mathbb{E}[X_1] = \overline{x}$. Then, the goal is to obtain a bound $\Lambda^*(x) \geq \Lambda^*_{\text{MIN}}(x)$ and $\Lambda(\lambda) \leq \Lambda_{\text{MAX}}(\lambda)$. The upper bound in particular is

$$\Lambda_{\text{MAX}}(\lambda) = \lg\left(\frac{b-\overline{x}}{b-a} e^{\lambda a} + \frac{\overline{x}-a}{b-a} e^{\lambda b}\right),$$

so $\mathbb{P}(X_1 = b) = p$ and $\mathbb{P}(X_1 = a) = q$, because $p+q = 1$ and $bp + aq = \overline{x}$. Then, $\Lambda^*(\cdot)$ is up to affine transformation the relative entropy corresponding to $Y_i \sim \text{Bernoulli}(p)$. Thus, $\Gamma_n \sim \text{Binomial}(n, p)$, where $\Gamma_n = (S_n - na)/(b-a)$ is the sum of the $Y_i = (X_i - a)/(b-a)$. The relaive entropy is defined as

$$H(y \mid p) = y \lg \frac{y}{p} + (1-y) \lg\left(\frac{1-y}{1-p}\right);$$

then,

$$\Lambda^*_{\text{MIN}}(x) = H\left(\frac{x-a}{b-a} \mid \frac{\overline{x}-a}{b-a}\right).$$

One side of the bound makes sense, because we have an example, but how do we know it's the worst case? This can be proven by invoking a more general fact called the Tchebyceff system (from approximation theory, in some sense an optimization across probability measures). This says that if $U_0, U_1, \ldots, U_n, \Omega$ are functions from $T \subseteq \mathbb{R}^k \to \mathbb{R}$, then,

$$\sup_{\sigma\in V_c} \int_T \Omega \, \mathrm{d}\sigma = \inf_{x\in\mathscr{P}_+} \sum_{i=0}^n x_i c_i. \tag{5}$$

$$\inf_{\sigma\in V_c} \int_T \Omega \, \mathrm{d}\sigma = \sup_{x\in\mathscr{P}_-} \sum_{i=0}^n x_i c_i. \tag{6}$$

where

$$\mathscr{P}_+ = \left\{x_0, \ldots, x_n \mid \sum_{i=0}^n X_i U_i \geq \Omega \text{ for all } t\right\},$$

$$\mathscr{P}_- = \left\{x_0, \ldots, x_n \mid \sum_{i=0}^n X_i U_i \leq \Omega \text{ for all } t\right\},$$

amd V_c is the set of all finite non-negative measures σ on T such that $\int_T |U_i| \, \mathrm{d}\sigma$ is finite or all i and $\int_T U_i \, \mathrm{d}\sigma = c_i$ for each i. (See [4] for a more detailed discussion on this.) If $V_{c+\Delta c} \neq \emptyset$, for all sufficiently small Δc and $\mathscr{P}_+$ and $\mathscr{P}_-$ are both nonempty (as in the application we care about), the best or worst case will occur when σ is atomic, with $n+1$ atoms.

This can be reduced to the specific case we care about using Jensen's inequality and the fact that

$$e^{\lambda x} \leq \left(\frac{b-x}{b-a}\right) e^{\lambda a} + \left(\frac{x-a}{b-a}\right) e^{\lambda b}$$

whenever $a \leq x \leq b$, which follows because

$$x = \left(\frac{x-a}{b-a}\right) a + \left(\frac{x-a}{b-a}\right) b.$$

These are convex combiantions of a and b, or previously $e^{\lambda a}$ and $e^{\lambda b}$, so Jensen's inequality applies.

The next thing to invoke is Hoeffding's bound. If $X_1, \ldots, X_n$ are i.i.d. random variables on $[a, b]$ where $\mathbb{E}[X_1] = \overline{x}$, then

$$\mathbb{P}\left(\frac{1}{n} S_n \geq x\right) \leq e^{-nH((x-a)/(b-a) \mid (\overline{x}-a)/(b-a))}$$

$$\mathbb{P}\left(\frac{1}{n} S_n \leq x\right) \leq e^{-nH((x-a)/(b-a) \mid (\overline{x}-a)/(b-a))}.$$

In the specific case $a=-1$, $b=1$, and $\overline{x}=0$, one obtains a function

$$f(x)=H\left(\frac{x+1}{2}\mid\frac{1}{2}\right)=\frac{x+1}{2}\lg(x+1)+\frac{1-x}{2}\lg(1-x).$$

The bound is $\mathbb{P}(S_n/n\geq x)\leq e^{-nf(x)}$. Then, using calculus, one can show that $f(x)\geq x^2/2$, because $f(0)=f'(0)=0$ and $f;'(x)=1/(2(1+x))+1/(2(1-x))\geq 1$. Thus, the bound can be greatly simplified:

$$\mathbb{P}\left(\frac{1}{n}S_n\geq x\right)\leq e^{-nH((x+1)/2|1/2)}\leq e^{-nx^2/2}.$$

Thus, writing $nx=\sqrt{n}\,y$,

$$\mathbb{P}(|S_n|\geq nx)\leq 2e^{-nx^2/2}=2e^{-y^2/2}.$$

Thus, as long as the mean is 0 and the bounds are ± 1, then fluctuation in the mean decreases of order n^2. This is a concentration inequality, and is often called the Hoeffding bound.

Another observation, called the Azuma-Hoeffding inequality, shows that this also holds for martingales, rather than just independent random variables.

Definition. A martingale is a sequence $Y_m=\sum_{i=1}^m X_i+Y_0$ such that $\mathbb{E}[Y_m\mid Y_0,\dots,Y_{m-1}]=Y_{m-1}$, or, equivalently, $\mathbb{E}[X_m\mid X_1,\dots,X_{m-1}]=0$; these conditions are known as the martingale differences.

Then, the observation is that the Azuma-Hoeffding bound

$$\mathbb{P}(|Y_n|\geq nx)\leq 2^{-nH((x+1)/2|1/2)},$$

or, equivalently,

$$\mathbb{P}(|Y_n|\geq\sqrt{n}\,y)\leq 2e^{-y^2/2},$$

holds whenever $\{Y_m\}$ is a martingale of bounded differences and $|X_i|\leq 1$.

This is shown by bounding by the expectation: if $K=[x,\infty)$, then $\inf_{y\in K}\lambda y=\lambda x$ whenever $\lambda\geq 0$.

$$\begin{aligned}\mathbb{P}\left(\frac{1}{m}(Y_m-Y_0)\geq x\right)&\leq\mathbb{E}[e^{m\lambda((1/m)(Y_m-Y_0)-x)}]\\&=e^{-m\lambda x}\mathbb{E}[e^{\lambda\sum_{i=1}^n X_i}]\\&=\mathbb{E}[\mathbb{E}[e^{\lambda\sum_{i=1}^n X_i}\mid X_0,\dots,X_{m-1}]]\\&=\mathbb{E}[e^{\lambda\sum_{i=1}^{m-1}X_i}\mathbb{E}[e^{\lambda X_m}\mid X_0,\dots,X_{m-1}]]\end{aligned}$$

But $\mathbb{E}[e^{\lambda X_m}\mid X_0,\dots,X_{m-1}]$ is just the conditional law of X_m given $|X_m|=1$ and $\mathbb{E}[X_m\mid X_1,\dots,X_{m-1}]=0$, so

$$\begin{aligned}&\leq\frac{1}{2}(e^{\lambda}+e^{-\lambda})\mathbb{E}[e^{\lambda\sum_{i=1}^{m-1}X_i}]\\&\leq e^{-m\lambda x}\left(\frac{1}{2}(e^{\lambda}+e^{-\lambda})\right)^m.\end{aligned}$$

In many corners of discrete mathematics, constructing these martingales is very easy, and so these bounds aren't hard to show, but the actual variables themselves are hard to deal with.

13. The Azuma-Hoeffding Bound: 2/5/14

"My notes here are a little ambitious… I tried to prove something which is wrong!"

Recall that last time, we proved the Azuma-Hoeffng bound: if $\{Y_m\}$ is a martingale of bounded differences, i.e. $\left|Y_m-Y_{m-1}\right|\leq 1$, then for all $y\geq 0$ and all m, $\mathbb{P}(|Y_m-y_0|\geq\sqrt{m}\,y)\leq 2e^{-y^2/2}$. (A martingale is a sequence of random variables $\{Y_n\}$ such that $\mathbb{E}[Y_n\mid Y_0,\dots,Y_{n-1}]=Y_{n-1}$ for all n.)

Suppose $\Omega=A^B$ is the set of functions $g:B\to A$, with some random measure assigned to Ω such that $\{g|_{B_{i+1}\setminus B_i}\}$ are independent of each other. Consider a functional $L:\Omega\to\mathbb{R}$ and an increasing sequence of subsets $\emptyset=B_0\subseteq B-1\subseteq B_2\subseteq\cdots\subseteq B_m=B$.

Definition. L is Lipschitz when $\left|L(g)-L(g')\right|\leq 1$ for any $g,g'\in\Omega$ and $0\leq i\leq m-1$, such that $g\neq g'$ at most on $B_{i-1}\setminus B_i$.

Claim. If L is Lipschitz, then for any $y>0$,

$$\mathbb{P}(L(g)-\mathbb{E}[L(g)]\geq y\sqrt{m})\leq e^{-y^2/2},\text{ and}$$
$$\mathbb{P}(L(g)-\mathbb{E}[L(g)]\leq -y\sqrt{m})\leq e^{-y^2/2}.$$

Proof. Apply the Azuma-Hoeffding inequality for Doob's martingale of $L(g)$, i.e. $Y_i(h) = \mathbb{E}[L(g) \mid g = h \text{ on } B_i]$, so that $Y_0(h) = \mathbb{E}[L(g)]$ and $Y_m(h) = L(g)$ (since g is uniquely determined on all of B). This is a fairly general way to obtain martingales (conditioning on more and more information). Then, to check that it has bounded differences,

$$\begin{aligned} Y_{i+1}(h) - Y_i(h) &= \mathbb{E}[L(g) \mid g = h \text{ on } B_{i+1}] - \mathbb{E}[L(g) \mid g = h \text{ on } B_i] \\ &= \mathbb{E}[\mathbb{E}[L(g') \mid g' = h \text{ on } B_{i+1}] \mid g' = h \text{ on } B_i] \\ &= \mathbb{E}[L(g) - L(g') \mid g' = g = h \text{ on } B_i \text{ and } g = h \text{ on } B_{i+1} \setminus B_i]. \end{aligned}$$

This uses a principle called the law of iterated expectation, which says that if $\mathcal{G} \supseteq \mathcal{H}$, then $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}] \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]$. Then, since $g' = g$ on B_i, then $\left|Y_{i+1} - Y_i\right| \le \mathbb{E}[\left|L(g) - L(g)'\right|] \le 1$ by the Lipschitz condition, since the distribution is independent on $B_j \setminus B_{j-1}$, so without loss of generality, one may also take $g = g'$ on $B_{j+1} \setminus B_j$ whenever $j > i$.

That $\{Y_i\}$ is in fact a martingale follows because

$$\begin{aligned} Y_i(h) &= \mathbb{E}[L(g) \mid g = h \text{ on } B_i] \\ &= \mathbb{E}[\mathbb{E}[L(g') \mid g' = h \text{ on } B_{i+1}] \mid g' = h \text{ on } B_i] = \mathbb{E}[Y_i \mid g|_{B_1}, g|_{B_2}, \dots, g|_{B_{i-1}}]. \end{aligned}$$

⊠

This has an application to randomg graphs: let $G = G(n,p) = G_{n,p}$ be the Erdős-Rényi random graph as usual.

Definition. A functional L of $G_{n,p}$ is called edge-Lipschitz if $\left|L(g) - L(g')\right| \le 1$ whenever $g \ne g'$ on at most one edge.

The assumption is that $L : A \to \mathbb{R}$, where A is the set of functions out of the edges of the graph. Then, for any edge-Lipschitz functional L and any probability p we have

$$\mathbb{P}\left(|L(G_{n,p}) - \mathbb{E}[L(G_{n,p})]| \ge y\sqrt{\binom{n}{2}}\right) \varepsilon 2e^{-y^2/2}. \tag{7}$$

This just says that in these contexts, the value tends to be very close to its expectation.

This is given by the construction of the edge exposure martingale: let $\Omega = \{0,1\}^{\binom{n}{2}}$ and B_i be the set of the first i edges (in some chosen order). let $m = \binom{n}{2}$; then, the idea is that one is looking at pieces of the graph, adding one edge at a time.

There's a related concept called the vertex exposure matringale, where $m = n$ and B_i is the set of all edges out of the first i vertices. This time, the new information in each step is the set of edges connected to the next vertex. This is also a partition of the graph respecting the independence condition outlined about. In this case, one can call L vertex-Lipschitz if the analogous idea holds: $\left|L(g) - L(g')\right| \le 1$ whenever $g \ne g'$ on at most one neighborhood of a vertex. Thus, this is a stronger condition, and it implies a stronger bound in (7), where the $\binom{n}{2}$ can be replaced with n.

Theorem 13.1 (Shamir & Spencer, 1987)**.** *Let $\chi(G)$ denote the chromatic number of G, i.e. the minimal number of colors in a vertex coloring of G such that G has no monochromatic edge. Then, the chromatic number is a vertex-Lipschitz functional, i.e.*

$$\mathbb{P}(|\chi(G_{n,p}) - \mathbb{E}[\chi(G_{n,p})]| \ge y\sqrt{n}) \le e^{-y^2/2}.$$

Proof. Any single vertex can always be given a new color, so the vertex-exposing Lipschitz property holds. Thus, the chromatic number can differ by at most 1 in the aboselce of a given vertex. ⊠

Remark. In [1], it is shown that if $p = 1/2$, then $\mathbb{E}[\chi(G)] \sim n/(2\log_2 n)$, and the deviation is about $\sqrt{n}$.

Theorem 13.2 (Alon & Specner, Theorem 7.3.3)**.** *Suppose $p = n^{-\alpha}$ and $\alpha > 5/6$ is fixed. Then, there exists a number $u(n.p)$ such that $\mathbb{P}(u(n,p) \le \chi(G) \le u(n,p) + 3) \to 1$ as $n \to \infty$.*[7] *That is, the probability is at least $1 - 2\varepsilon$ for all $n \ge n_0(\varepsilon, \alpha)$.*

Proof. Fix an $\varepsilon > 0$ and let $u(n,p,\varepsilon)$ be the minimal number such that $\mathbb{P}(\chi(G) \le u) > \varepsilon$. Let $Y(G)$ be the minimal size of a set of vertices S such tat $G \setminus S$ is u-colorable. Then, by the same argument as before, the vertex-Lipschitz property holds, so take $\mu = \mathbb{E}[Y(G)]$. Choose λ such that $e^{-\lambda^2/2} = \varepsilon$, so that

$$\begin{aligned} \mathbb{P}(Y \le \mu - \sqrt{n}\,\lambda) &\le e^{-\lambda^2/2} \\ \mathbb{P}(Y \ge \mu + \sqrt{n}\,\lambda) &\le e^{-\lambda^2/2}. \end{aligned}$$

By the choice of u, it's known that $\mathbb{P}(Y = 0) > \varepsilon = e^{-\lambda^2/2} = \mathbb{P}(Y \le \mu - \lambda\sqrt{n})$, so $\mu \le \lambda\sqrt{n}$, or else there's a contradiction. By the same idea, $\mathbb{P}(Y \ge 2\mu) \le e^{-\lambda^2/2}$. Thus, with probability at least $1 - \varepsilon$, there is a u-coloring of all but at most $c'\sqrt{n}$ vertices (here, $c' = 2\lambda$). But by Lemma 7.4.3 of [1], then by the first moments method, under the conditions of the theorem statement (i.e. restrictions on p and α), every $c'\sqrt{n}$ set of vertices in $G(n,P)$ are 3-colorable. finishing the proof. ⊠

[7] There is a much harder proof that this is 1 rather than 3, and of course it's not constructive.

Theorem 13.3 (Alon & Spencer, Theorem 7.5.1)**.** *Consider functions $g : \{1,\dots,n\} \to \{1,\dots,n\}$, with all n^n choices equally likely. Then, as in a typical coupon-collecting problem, let $L(g) = |n - g(\{1,\dots,n\})|$ (i.e. the number of missing elements in the range of g). Then, $\mathbb{E}[L(g)] = n\mathbb{P}(g(x) \neq 1 \text{ for all } x) = n(1-1/n)^n \sim n/e$, and for all λ,*

$$\mathbb{P}\left(\left|L(g) - n\left(1-\frac{1}{n}\right)^n\right| \geq \lambda\sqrt{n}\right) \leq 2e^{-\lambda^2/2}$$

(which is the Lipschitz property with respect to the Doob martingale again).

In other words, the mean is of order n, and the fluctuation is of order $\sqrt{n}$.

14. Concentration Inequalities: 2/7/14

First, there will be one more example of concentration inequalities applied to martingales, where we once again use that $\mathbb{P}(X - \mathbb{E}[X] \geq \lambda\sqrt{m}) \leq e^{-\lambda^2/2}$ and similarly for the other side of $X - \mathbb{E}[X]$, as long as on can write Doob's martingale for X, composed of at most m differences bounded by 1.

Example 14.1 (Alon & Spencer, 7.5.2)**.** Fix $v_i \in \mathbb{R}^d$ with $|v_i| \leq 1$, for $i = 1,\dots,n$, and let $\varepsilon_i \in \{1,-1\}$ be i.i.d. random variables. If

$$X = X(\varepsilon) = \left|\sum_{i=1}^n \varepsilon_i v_i\right|,$$

then

$$\mathbb{P}(X - \mathbb{E}[X] \geq \quad \lambda\sqrt{n}) \leq e^{-\lambda^2/2}, \text{ and}$$
$$\mathbb{P}(X - \mathbb{E}[X] \leq -\lambda\sqrt{n}) \leq e^{-\lambda^2/2}.$$

Proof. Let $\Omega = \{-1,1\}^n$, with a martingale defined by exposing the ε_i one at a time: $Y_i(\varepsilon) = \mathbb{E}[X \mid \varepsilon_1,\dots,\varepsilon_i]$. Thus, $Y_n = X$, $Y_0 = \mathbb{E}[X]$, and for all ε, $Y(\varepsilon) = (1/2)(Y_{i+1}(\varepsilon) + Y_{i+1}(\varepsilon'))$, where $\varepsilon' \neq \varepsilon$ at only the i^{th} coordinate. Thus, $Y_{i+1}(\varepsilon) - Y_i(\varepsilon) = (1/2)(Y_{i+1}(\varepsilon) - Y_{i+1}(\varepsilon'))$, so

$$\begin{aligned}
\left|Y_{i+1}(\varepsilon) - Y_i(\varepsilon)\right| &= \frac{1}{2}\left|\mathbb{E}_\eta\left|\sum_{j=1}^{i-1}\varepsilon_j v_j + \varepsilon_i v_i + \sum_{j=1}^n \eta_j v_j\right| - \mathbb{E}\left|\sum j = 1^i \varepsilon_j v_j - \varepsilon_i v_i + \sum_{j=i+1}^n \eta_j v_j\right|\right| \\
&= \mathbb{E}_\eta\left[\frac{1}{2}\left|\sum_{j=1}^{i-1}\varepsilon_j v_j + \varepsilon_i v_i + \sum_{j=1}^n \eta_j v_j\right| - \frac{1}{2}\left|\sum j = 1^i \varepsilon_j v_j - \varepsilon_i v_i + \sum_{j=i+1}^n \eta_j v_j\right|\right] \\
&\leq |v_i| \leq 1,
\end{aligned}$$

using the triangle inequality and the fact that $|\varepsilon_i| = 1$. Thus, we have the required martingale. ⊠

Method of Types. This part, referenced in Chapters 2.1.1 and 2.1.2 of [3]. The goal is to prove Sanov's theorem and Croner's theorem[8] for a finite alphabet.

Suppose $Y_1,\dots,Y_n$ are i.i.d. random variables taking values in a finite set $\Sigma = \{a_1,\dots,a_N\}$. Let $\mu(a_i) = \mathbb{P}(Y_1 = a_i) \in M_1(\Sigma)$ (i.e. μ is a measure of some sort). Assume without loss of generality that $\mu(a_i) > 0$ for all i.

Definition. The type of a vector $\underline{y} = (y_1,\dots,y_n) \in \Sigma^n$ is the empirical law

$$L_n^y(a_i) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{y_j = a_i}.$$

Then, $L_n^y(\cdot) = (L_n^y(a_1),\dots,L_n^y(a_N))$ is a probability vector with N coordinates. The set $\mathscr{L}_n$ of all possible types is

$$\left\{\left(\frac{k_1}{n},\frac{k_2}{n},\dots,\frac{k_N}{n}\right), 0 \leq k_i \leq n, \text{ and } \sum_{i=1}^N k_i = n\right\}.$$

Thus, $|\mathscr{L}_n| \leq (n+1)^N$.

Lemma 14.1 ([3])**.** *For all $\nu \in M_1(\Sigma)$, there exists a $\nu' \in \mathscr{L}_n$ such that*

$$\frac{1}{2}\sum_{i=1}^N \left|\nu(a_i) - \nu'(a_i)\right| \leq \frac{N}{2n}.$$

Thus, $\mathscr{L}_n$ is dense in $M_1(\Sigma)$.

[8] Does anyone know how to spell this? I couldn't parse the name from the pronunciation and can't parse the professor's cursive sometimes.

Now, one can talk about type classes: $T_n(\nu) = \{y \in \Sigma^n : L_n^y = \nu\}$.

Definition. The Shannon entropy of a $\nu \in M_1(\Sigma)$ is

$$H(\nu) = -\sum_{i=1}^{N} \nu(a_i) \lg \nu(a_i).$$

This measures the randomness of ν, so to speak.

Definition. The relative entropy, or K-L distance, of one measure given another is

$$H(\nu \mid \mu) = \sum_{i=1}^{n} \nu(a_i) \log\left(\frac{\nu(a_i)}{\mu(a_i)}\right) \geq 0.$$

This measures the distance between μ and ν, in some sense.

Lemma 14.2. *For all $\underline{y} \in T_n(r)$ that have the same probability of $\underline{Y}$,*

$$\mathbb{P}_\mu(\underline{Y} = \underline{y}) = e^{-n(H(\nu)+H(\nu|\mu))}.$$

This can be proven directly by substituting in the definitions.

Lemma 14.3 (Stirling's approximation)**.**

$$(n+1)^{-|\Sigma|} e^{nH(r)} \leq \binom{n}{n\mu_1, \ldots, n\mu_n} = |T_n(\nu)| \leq e^{nH(\nu)}.$$

Together, these imply the following lemma:

Lemma 14.4. *For all $\nu \in \mathscr{L}_n$,*

$$(n+1)^{-N} e^{-nH(\nu|\mu)} \leq \mathbb{P}_\mu(L_n^{\underline{Y}} = \nu) \leq e^{-nH(\mu|\mu)}.$$

Then, using the probabilistic method (finally!), one obtain the following theorem, which views the probability distribution as chosen from another probability distribution. It's the large deviation principle to this measure-theoretic issue.

Theorem 14.5 (Sanov)**.** *For all $\Gamma \subset M_1(\Sigma)$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu(L_n^Y \in \Gamma) \leq \inf_{\nu\in\Gamma}\{H(\nu \mid \mu)\}$$
$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_\mu(L_n^Y \in \Gamma) \geq \inf_{\nu\in\Gamma^0}\{H(\nu \mid \mu)\},$$

where Γ^0 is the interior of Γ.[9]

Proof. For the upper bound, use Lemma 14.4:

$$\mathbb{P}_\mu(L_n^Y \in \Gamma) = \sum_{\nu\in\Gamma\cap\alpha_n} \mathbb{P}_\mu(L_n^Y = \nu) \leq |\mathscr{L}_n| \, e^{-n \inf_{\nu\in\Gamma} H(\nu|\mu)},$$

but when $n \to \infty$, $\lg|L_n|/n \to 0$.

For the lower bound, if $\nu \in \Gamma^0$, then there exists a $\nu_n \in \mathscr{L}_n \cap \Gamma$ such that $\nu_n \to \nu$ as $n \to \infty$. Thus, $\nu \mapsto H(\nu \mid \mu)$ is a continuous function on $M_1(\Sigma) \subseteq \mathbb{R}^n$. Thus, using Lemma 14.4,

$$\mathbb{P}_\mu(L_n^Y \in \Gamma) \geq \mathbb{P}_\mu(L_n^Y \in \nu_n) \geq (n+1)^{-N} e^{-nH(\nu_n|\mu)},$$

but as $n \to \infty$, $\nu_n \to \nu$. ⊠

This can be used to get Cramer's theorem out of Sanov's in the case of a finite alphabet. Let $X_j = f(Y_j)$ for some non-random f and the Y_j are i.i.d. over a finite Σ, such that $\mu(a_i) > 0$ for $i = 1, \ldots |\Sigma| = N$. Without loss of generality, order the a_i such that $f(a_1) < \cdots < f(a_N)$. If $\hat{S}_n = (1/n)\sum_{j=1}^n X_j$, then there is a large deviation principle for $\hat{S}$ Let $\Lambda^*(x) = \sup_{\lambda\in|R}(\lambda x - \Lambda(\lambda))$, where

$$\Lambda(\lambda) = \log \mathbb{E}_\mu\left[e^{\lambda X_1}\right] = \mathbb{E}\left[\sum_{i=1}^{N} e^{\lambda f(a_i)} \mu(a_i)\right].$$

[9]This only works in the combinatorial case. In the more general case, the upper bound is the infimum over the closure of Γ, and the definitions of the random entropy have to be updated for the infinite case, i.e. using expected values.

Let $I(x) = \inf_{\{\nu:\langle f,\nu\rangle = x\}}(H(\nu \mid \mu))$, where the inner product is given by

$$\langle f, \nu\rangle = \sum_{i=1}^{N} f(a_i)\nu(a_i).$$

Then, the bounds are

$$\overline{\lim_{n\to\infty}} \frac{1}{n}\mathbb{P}(\hat{S}_n \in A) \le -\inf_{x\in A} I(x)$$
$$\underline{\lim_{n\to\infty}} \frac{1}{n}\mathbb{P}(\hat{S}_n \in A) \ge -\inf_{x\in A^\circ} I(x)$$

Proof. Note that $\hat{S}_n \in A$ iff $L_n^Y \in \Gamma = \{\nu : \langle \nu, f\rangle \in A\}$. This is because $\hat{S}_n$ just averages f over the Y_i.

$$\hat{S}_n = \frac{1}{n}\sum_{j=1}^{n} f(Y_j) = \sum_{i=1}^{N} f(a_i)\overbrace{\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Y_j=a_i\}}}^{L_n^Y(a_i)} = \left\langle f, L_n^Y\right\rangle.$$

Now, take the limits; for concision, do both at once (which is what the notation "lim" will represent).

$$\lim \frac{1}{n}\log\mathbb{P}(\hat{S}_n \in A) = \lim \frac{1}{n}\log\mathbb{P}(L_n^Y \in \Gamma) \le -\inf_{\Gamma} H(\nu \mid \mu) = \inf_{x\in A}\inf_{\{\langle f,\nu\rangle = x\}} \cdots I(x)$$
$$\ge -\inf_{\Gamma^\circ} H(\nu \mid \mu) \ge -\inf_{x\in A^\circ}\inf \cdots \qquad \boxtimes$$

The identity $\Lambda^* = I$ can be proven, for example by solving the finite-dimensional Lagrange multiplier problem

$$\min_{\nu}\left\{\sum_{i=1}^{N}\nu(a_i)\log\frac{\nu(a_i)}{\mu(a_i)} - \lambda\sum_{i=1}^{N} f(a_i)\nu(a_i)\right\}.$$

Thus, one explicitly can obtain $\nu_\lambda^*(a_i) = \mu(a_i)e^{\lambda f(a_i)-\Lambda(\lambda)}$ for some λ. Then, for any $f(a_1) < x < f(a_N)$, there is a unique λ^* such that $\Lambda'(\lambda^*) = \langle f, \nu_{\lambda^*}\rangle = x$.

There's still stuff to be done for $x \notin [f(a_1), f(a_N)]$, but when $x = f(a_N)$, $\lambda^* \to \infty$, and when $x = f(a_1)$, then $\lambda^* \to -\infty$. Thus, the correct bound is obtained by solving the conceptually simple but scary-looking calculus problem.

15. Talagrand's Inequality: 2/10/14

We've already seen the Azuma-Hoeffding inequality for large deviations of bounded i.i.d. random variables, in which the concentration extends to a martingale of bonded differences.

Talagrand's concentration inequality deals with product spaces; the most important case is $\Omega = \{0,1\}$, so that the product space is $\{0,1\}^N$. Introduce the notation $I(x,y) = \{i \le N : x_i \ne y_i\}$.

The Azuma-Hoeffding inequality states that for all $A \subseteq \Omega^N$ such that $\mathbb{P}(A) \ge 1/2$, $\mathbb{P}(d(\cdot, A) \ge u) \le 2\exp(-u^2/N)$, where $d(x,A) = \inf_{y\in A} d_H(x,y)$ and $d_H(x,y) = \mathrm{Card}(I(x,y))$ (i.e. the number of points they have different; the Huffman distance).

Definition. A subset $I \subset \{1,\dots,N\}$ is a pattern for x, A if there exists a $y \in A$ such that $I(x,y) = I$.

One way to measure the distance from x to A is by looking at the set of patterns for x, A. Then, the Azuma-Hoeffding inequality tells us that for generic x, there exists a pattern of cardinality at most $O(\sqrt{N})$.

Example 15.1. The most fundamental example is

$$A = \left\{y : \sum_{i=1}^{N} y_i \le p_N\right\},$$

so that $\mathbb{P}(A) \simeq 1/2$ if $\mu(1) = p$. Then, define an indicator $X = \mathbb{1}_J$ where $|J| = pN + m$. Then, any $I \subseteq J$ with $|I| = m$ is a pattern: by the Central Limit Theorem, a generic x has $m \simeq 10\sqrt{pN}$, and there are many patterns of size $\sqrt{pN}$.

Then, one wants a new notion of distance to represent this property, so define

$$g(x,A) = \sup_{\{\beta\in\mathbb{R}^N \mid \|\beta\|\le 1\}}\inf_{y\in A}\left[\sum_{i=1}^{N}\beta_i\mathbb{1}_{x_i\ne y_i}\right].$$

Then, $g(\cdot, A) \ge (1/\sqrt{N})d_H(\cdot, A)$. Finally, let $f(x,A) = (1/2)g(x,A)^2$.

Theorem 15.1 (Talagrand's Concentration Inequality)**.** *If $A \subseteq \Omega^n$ and $u \geq 0$, then*

$$\mathbb{P}(f(X,A) \geq u)e^u \leq \mathbb{E}\left[e^{f(x,A)}\right] \leq \frac{1}{\mathbb{P}(A)}.$$

Thus, $\mathbb{P}(A)(1-\mathbb{P}(A_t)) \leq e^{-t^2/4}$*, where* $A_t = \{z : g(z,A) < t\}$.

This makes it a little clearer why we used the second distance notion: it allows for a more general result while still providing control over the inequality.

As an example, suppose one has a random variable $Z_N(\underline{x})$ of interest, and let $A = \{\underline{x} : Z_N(\underline{x}) \geq \text{Med}(Z_N)\}$ (where $\text{Med}(x)$ denotes its median). Then, $\mathbb{P}(A) \geq 1/2$ automatically, so if $Z_N(\underline{x}) \geq \text{Med}(Z_N) + v$ implies that $g(\underline{x},A) \geq t$ for some $t = t(v,N)$, then $\mathbb{P}(Z_N(\underline{x}) \geq \text{Med}(Z_n) + v) \leq 2e^{-t^2/4}$.

For example, one problem that can't be solved with just Hamming distance is the longest increasing subsequence problem on (i.i.d. random variables on) the uniform distribution on $[0,1]$. That is,

$$Z_n(\underline{x}) = \max\{m : x_{k_1} < x_{k_2} < \cdots < x_{k_m} \text{ for some } 1 \leq k_1 < k_2 < \cdots < k_m \leq n\}.$$

How much does a permutation fluctuate from the mean?

One can take $A(j) = \underline{y} : Z_N(\underline{y}) \leq j\}$ and suppose $Z_N(\underline{x}) \geq j + v$ for some $v \in \mathbb{N}$. Then,

$$\sum_{i=1}^{v+j} \mathbb{1}_{x_{k_i} \leq y_{k_i}} \geq v$$

for all $y \in A_j$, so $g(\underline{x},A_j)) \geq v/\sqrt{j+v} = t$. Letting $M_N = \text{Med}(Z_N(\underline{x}))$, one can show for all $v \in \mathbb{N}$ that

$$\mathbb{P}(Z_n(\underline{x}) \geq M_N + v) \leq 2e^{-v^2/4(M_N+v)}$$
$$\mathbb{P}(Z_N(\underline{x}) \leq M_N - v) \leq 2e^{-v^2/4M_N}.$$

Thus, $M_N = \text{Med}(Z_N(\underline{X})) = O(\sqrt{N})$. The concentration is in a window of size $N^{1/4}$, which is better than we would have guessed. The true answer is conjectured to be $N^{1/6}$, but this is open.

Now, one can replace means and expectations, up to this error term: $|\mathbb{E}[Z_N(\underline{x}) - \text{Med}(Z_N(\underline{x}))| = O(N^{1/4})$. But this doesn't require sums, so even if there isn't a martingale lying around to help, this inequality works.

Proof of Theorem 15.1. The proof will be by induction on N; when $N = 1$, it boils down to

$$f(x,A) = \frac{1}{4}\inf_{\{y \in A\}}\{\mathbb{1}_{x \neq y}\} = \frac{1}{4}\mathbb{1}_{x \notin A}.$$

Then,

$$\mathbb{P}(A)\mathbb{E}\left[e^{(1/4)\mathbb{1}_{x \neq A}}\right] = \mathbb{P}(A)(e^{1/4}\mathbb{P}(A^c) + \mathbb{P}(A)) = (p + e^{1/4}(1-p)) \leq 1$$

by some not too interesting analysis.

Now, in the general case, suppose it's true for N, and consider $N+1$. Then, let $\Omega = \Omega' \times \Omega_{N+1}$, where $\Omega' = \prod_{i=1}^N \Omega_i$ was the old set, and Ω_{N+1} is the new one. Write $z = (x,w)$ with $x \in \Omega'$ and $w \in \Omega_{N+1}$, and define $A_w = \{x \in \Omega' : z(x,w) \in A\}$ and $B = \{x \in \Omega' : z(x,w) \in A \text{ for some } w \in \Omega_{N+1}\} = \bigcup_w A_w$. Let

$$f(\underline{x},A) = \inf_{\mathbf{u} \in V(\underline{x},A)}\left\{\frac{1}{4}\sum_{i=1}^N u_i^2\right\},$$

where $V(\underline{x},A)$ is the conve hull of $U(\underline{x},A)$. There's some non-probabilistic analysis here, which is encapsulated as Theorem 7.6.2 of [1]. Now, we can induct: given $z = (x,w)$ and an A, consider $\mathbf{s} \in U(\underline{x},B)$ and $\mathbf{t} \in U(\underline{x},A_w)$, so that $(\mathbf{s},1),(\mathbf{t},0) \in U(z,A)$. By convexity, $\lambda\mathbf{s} + (1-\lambda)\mathbf{t} \in V(\underline{x},A)$, so using the Cauchy-Schwartz inequality,

$$f(\underline{x},A) \leq \frac{1}{4}(1-\lambda)^2 + \frac{1}{4}|(1-\lambda)\mathbf{s} + \lambda\mathbf{t}|^2 \leq \frac{1}{4}(1-\lambda)^2 + \frac{1}{4}(1-\lambda)|\mathbf{s}|^2 + \frac{1}{4}\lambda|\mathbf{t}|^2.$$

So now I guess we take the maximal values of $\mathbf{s}$ and $\mathbf{t}$, which will allow one to write

$$z \leq e^{(1/4)(1-\lambda)^2}\frac{1}{\mathbb{P}(B)^{1-\lambda}}\frac{1}{\mathbb{P}(A_w)^\lambda} = e^{(1/4)(1-\lambda^2)}r^{-\lambda}\frac{1}{\mathbb{P}(B)}.$$

Optimizing over $\lambda = 1 + 2\ln r$, and so on… (here, the professor had to hand over the room to another class). ⊠

16. Correlation Inequalities: 2/12/14

The beginning of the notion of correlation inequalities comes from a result in algebra.

Definition.

- A lattice L is a partially ordered set (with $x \leq x$), such that for all $x, y \in L$ there exists a unique minimal upper bound $x \vee y$ and a unique maximal lower bound $x \wedge y$ (sometimes called meet and join, respectively); that is, $x \vee y \geq x, y$, and if $z \geq x, y$, then $z \leq x \vee y$, and analogously for $x \wedge y$.
- A lattice L is finite distributive if for all $x, y, z \in L$, $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ and $(x \vee y \wedge z) = (x \vee y) \wedge (x \vee z)$, i.e. join and meet distribute.

As a matter of notation, if L is a lattice and $X, Y \subseteq L$, then $X \vee Y = \{x \vee y : x \in X, y \in Y\}$ and $X \wedge Y = \{x \wedge y : x \in X, y \in Y\}$.

Theorem 16.1 (The Four-Function Theorem; Alswede & Daykin, 1978)**.** *Suppose L is a finite distributive lattice and $\alpha, \beta, \gamma, \delta : L \to \mathbb{R}^+$ are such that $\alpha(x)\beta(y) \leq \gamma(x \vee y)\delta(x \wedge y)$ for all $x, y \in L$. Then, for all $X, Y \subseteq L$,*

$$\left(\sum_{x \in X} \alpha(x)\right)\left(\sum_{y \in Y} \beta(y)\right) \leq \left(\sum_{z \in X \vee Y} z\right)\left(\sum_{w \in X \wedge Y} w\right).$$

Corollary 16.2 (Alon & Spencer, Corollary 6.1.3)**.**

(1) Take $\alpha = \beta = \gamma = \delta = 1$, then, for all $X, Y \subseteq L$, $|X||Y| \, l \, e \, |X \wedge Y||X \vee Y|$.

(2) If N is some finite set, $L = 2^N$ is a lattice under inclusion, $\cup$, and $\cap$ (in some sense, just round off all of the operators). Then, if $X \subseteq 2^N$, let $Y = \{G^c : G \in X\}$, so that $|X \vee Y| = |X \wedge Y| = |\{F \cup G^c : F, G \in X\}|$, so $|X| \leq |\{F \cap G^c : F, G \in X\}|$.

Definition. On a finite distributive lattice L, a function $\mu : L \to \mathbb{R}^+$ is called log-supermodular if the condition in Theorem 16.1 holds when $\alpha = \beta = \gamma = \delta = \mu$.

Theorem 16.3 (FKG Inequality)**.** *Let $\mu : L \to \mathbb{R}^+$ be a log-supermodular function. Then, for all $f, g : L \to \mathbb{R}^+$ that are increasing with respect to the partial order,*

$$\left(\sum_{x \in L} \mu(x) f(x)\right)\left(\sum_{x \in L} \mu(x) g(x)\right) \leq \left(\sum_{x \in L} \mu(x) f(x) g(x)\right)\left(\sum_{x \in L} \mu(x)\right).$$

Already this is starting to look more like probability! For example, if μ is a probability measure, this can be used to say things about the expected values of f and g.

Proof of Theorem 16.3. The proof will apply the Four-Function theorem to $\alpha = \mu f$, $\beta = \mu g$, $\gamma = \mu f g$, and $\delta = \mu$. Ten, since $x, y \leq x \vee y$ and f and g are increasing, the conditions for it to be applicable. That is, $\mu(x)\mu(y) \leq \mu(x \wedge y)\mu(x \vee y)$ and $f(x) \leq f(x \vee y)$ and $g(x) \leq g(x \vee y)$, so

$$\mu(x) f(x) \mu(x) g(x) \leq \mu(x \vee y) f(x \vee y) g(x \vee y) \mu(x \wedge y).$$

Then, the general result follows from the Four-Function theorem. ⊠

The same inequality holds when both f and g are decreasing, and when one is increasing but the other is decreasing, the reverse inequality holds.

Then, one has that $\mathbb{E}[f g] \geq \mathbb{E}[f]\mathbb{E}[g]$ with respect to the probability distribution $\mu(x)/\sum_{z \in L} \mu(z)$.

The name of the FKG theorem comes from the names of three physicists, Fortrin, Kastelan, and Ginibre, who proved it in a more concrete context of statistican mechanics in 1971.

To discuss some more applications, the following lemma will be useful.

Lemma 16.4 (Kleitman, 1966)**.** *Let $\underline{X} = (X_1, \ldots, X_n)$ be n i.i.d. Bernoulli rabdom variables with probability p.*[10] *Then, put a partial order on coordinates, and define*

- *A is a monotone going-down event if for all $\underline{w} \in A$ and for all $\underline{v} \leq \underline{w}$, $\underline{v} \in A$, and*
- *A is a monotone going-up event if for all $\underline{w} \in A$ and for all $\underline{v} \geq \underline{w}$, $\underline{v} \in A$.*

Then, let A and B be monotone going-up events and C and D be monotone going-down events. Then, $\mathbb{P}(A \mid B) \geq P(A)$, $\mathbb{P}(C \mid D) \geq P(C)$, and $\mathbb{P}(A \mid C) \leq \mathbb{P}(A)$.

Proof. This is just the application of Theorem 16.3 with $f = \mathbb{1}_A$ and $g = \mathbb{1}_B$, so that $f g = \mathbb{1}_{A \cap B}$. Then, $\mathbb{E}[f g] = \mathbb{P}(A \cap B)$ and $\mathbb{E}[f]\mathbb{E}[g] = \mathbb{P}(A)\mathbb{P}(B)$. It remains to show that the Bernoulli product measure $\mu(\underline{x})$ (i.e. the number of 1s in the vector) is log-supermodular, but since everything only takes values on $\{0, 1\}^n$, this is not too bad. ⊠

[10] The result holds if $X_i \sim \text{Bernoulli}(p_i)$, which is more general, but this makes for harder and less enlightening notation.

Notice the dates of all of these proofs: most of them were developed independently, and then discovered to be special cases of the more general theorem.

Example 16.1.

- Let N be a finite set, $A_1,\dots,A_k \subset N$, and let $A \subseteq N$ be a random set, where $X_i = \mathbb{1}_{\{i\in A\}}$ are $|N|$ independent Bernoulli trials with probability p. Then,

$$\mathbb{P}(\text{ for all } j, A\cap A_j \neq \emptyset) \geq \prod_{j=1}^{k} \mathbb{P}(A\cap A_j \neq \emptyset).$$

  This allows us to fill the gap earlier in the course: when stating janson's inequality, this was exactly the condition required.

  This follows from Lemma 16.4, where

$$A = \left\{ \sum_{i\in A_j} X_i \neq 0 \right\} \quad \text{and} \quad B = \left\{ \sum_{i\in A_j'} X_i \neq 0 \right\}.$$

  Well, technically, this shows it in the case $k = 2$. For the full result, one can extend Theorem 16.3 to the case of k sets.
- In the subject of random graphs, consider $G_{n,p}$ again.
  - A property Q is monotone up if whenever G has Q and $G \subseteq H$, then H has Q.
  - A property Q is monotone down if whenever G has Q and $G \supseteq H$, then H has Q.

  Then, for $G_{n,p}$, suppoe Q_1 and Q_2 are monotone up and Q_3 and Q_4 are monotone down. There are many such properties, so having inequalities tend to be particularly useful. Then, by the above theorems, $\mathbb{P}(Q_3 \mid Q_4) \geq \mathbb{P}(Q_3)$, $\mathbb{P}(Q_|Q_2) \geq \mathbb{P}(Q_1)$, and $\mathbb{P}(Q_1 \mid Q_3) \leq \mathbb{P}(Q_1)$.

  For example, what's the probability of the event A, that G has a Hamiltonian circuit given C, that G can be drawn in the plane? Then, A is monotone up and C is monotone down. Then, this is at most $\mathbb{P}(C)$, which is much nicer than conditioning on C, a harder global property.

Theorem 16.5 (The XYZ Theorem; Shepp 1982)**.** *If $\{a_1,\dots,a_n\}$ is a partially ordered set, a bijecion $\sigma : \{a_1,\dots,a_n\} \to \{1,\dots,n\}$ is a linear extension if for all i, j, $a_i \leq a_j$ implies that $\sigma(a_i) \leq \sigma(a_j)$.*

Consider the space of all linear extensions of $\{a_1,\dots,a_n\}$, with each chosen equally likely. Then,

$$\mathbb{P}(\sigma(a_1) \leq \sigma(a_2) \mid \sigma(a_1) \leq \sigma(a_3)) \geq \mathbb{P}(\sigma(a_1) \leq \sigma(a_2)).$$

We still actually haven't proven the Four-Function theorem, which will be fixed next lecture.

17. Proof of the Four-Function Theorem: 2/14/14

Proof of Theorem 16.5. The proof of the XYZ theorem will construct a distributive lattice $(L, \leq), \wedge, \vee$ and a μ such that $\mu(\underline{x})\mu(\underline{y}) \leq \mu(\underline{x}\wedge\underline{y})\mu(\underline{x}\vee\underline{y})$, such that the desired relation

$$\mathbb{P}(\sigma(a_1) \leq \sigma(a_2) \mid \sigma(a_1) \leq \sigma(a_3)) \geq \mathbb{P}(\sigma(a_1) \leq \sigma(a_2)).$$

Fix a large integer M (eventually, $M \to \infty$), and let L be the set of all ordered n-tuples $\underline{x} = (x_1,\dots,x_n)$, where $x_i \in M = \{1,\dots,m\}$, with a somewhat mysterious partial order where $\underline{x} \leq \underline{y}$ iff $x_1 \leq y_1$ and $x_i - x_1 \geq y_i - y_1$ for all $i = 2,\dots,n$. Then, one can explicitly calclate that $(\underline{x}\vee\underline{y})_i = \max(x_i - x_1, y_i - y_1) + \min(x_1, y_1)$, and similarly $(\underline{x}\wedge\underline{y})_i = \min(x_i - x_1, y_i - y_1) + \max(x_1, y_1)$.

Then, one can check that $(L, \leq)$ is distributive, or read he relevant section of Alon & Spencer (i.e. §6.4). Since the partial order is wacky, then it actually needs to be checked. Then, define a probability measure

$$\mu(\underline{x}) = \begin{cases} 1, & \text{if whenever } \underline{x} \text{ is such that } a_i \leq a_j \text{ in } (P,\leq), \text{ then } x_i \leq x_j, \\ 0, & \text{otherwise.} \end{cases}$$

Then, suppose $\mu(\underline{x}) = 1$ and $\mu(\underline{y}) = 1$. Since $(\underline{x}\wedge\underline{y})_i = \min(x_i - x_1, y_i - y_1) + \max(x_1, y_1)$, then if $a_i \leq a_j \in P$, then $x_i \leq x_j$ and $y_i \leq y_j$, so

$$(\underline{x}\wedge\underline{y})_i \leq \min(x_j - x_1, y_j - y_1) + \max(y_1, x_1) = (\underline{x}\wedge\underline{y})_j.$$

Now construct the functions

$$f(\underline{x}) = \begin{cases} 1, & \text{if } x_1 \leq x_2 \\ 0, & \text{otherwise.} \end{cases} \qquad \text{and} \qquad g(\underline{x}) = \begin{cases} 1, & \text{if } x_1 \leq x_3 \\ 0, & \text{otherwise.} \end{cases}$$

f and g are both increasing functions: if $f(\underline{x}) = 1$ and $\underline{y} \geq \underline{x}$, then $x_1 \geq y_1$ and $x_2 - x_1 \leq y_2 - y_1$, so $y_2 \geq y_1 + (x_2 - x_1) \geq y_1$, and therefore $f(\underline{y}) = 1$, so f is increasing. Then, g is increasing by nearly the same argument.

Define $Q(\underline{x}) = \mu(\underline{x})/\mu(1)$, so that $Q(x_1 \le x_2, x_1 \le x_3) \ge Q(x_1 \le x_2)Q(x_1 \le x_3)$. Then, as $m \to \infty$, the proportion of m-tuples with any $x_i = x_j$ goes to zero, and so $Q_m \to \mathbb{P}$, the probability measure of a uniform random distribution on the linear extensions. ⊠

Proof of Theorem 16.1. The proof of the Four-Function theorem will go in two parts.

The first step is to show that any finite distributive lattice $(L, \le)$ is isomorphic to a sub-lattice of $2^N, \subseteq, \cup, \cap$ (i.e. subsets with containment on some finite set of points).

An $x \in L$ is called *join-irreducible* (akin to a primeness condition) if whenever $x = y \vee z$, then $x = y$ or $x = z$. There is always a join-irreducible element, because L is finite.

Let $x_1, \dots, x_n$ be the join-irreducible elements of L. Then, every $x \in L$ is mapped to some $A(x) \subseteq N = \{1, \dots, n\}$ where $x = \bigvee_{i \in A} x_i$ and $\{x_i : i \in A\}$ are join-irreducible and less than x. Thus, it suffices to prove the Four-Function theorem in $(2^N, \subseteq)$. In this context, the statement of the theorem is that if $N = \{1, \dots, N\}$ and $\mathscr{P}(N) = 2^N$, then suppose $\varphi : \mathscr{P}(N) \to \mathbb{R}^+$.

If $\mathscr{A} \subseteq \mathscr{P}(N)$, then define $\varphi(\mathscr{A}) = \sum_{A \in \mathscr{A}} \varphi(A)$, and define $\mathscr{A} \cup \mathscr{B} = \{A \cup B \mid A \in \mathscr{A}, B \in \mathscr{B}\}$ and $\mathscr{A} \cap \mathscr{B} = \{A \cap B \mid A \in \mathscr{A}, B \in \mathscr{B}\}$. Then, if for all $A, B \subseteq N$ we have $\alpha(A)\beta(B) \le \gamma(A \cup B)\delta(A \cap B)$, then for all $\mathscr{A}, \mathscr{B} \subseteq \mathscr{P}(N)$, $\alpha(\mathscr{A})\beta(\mathscr{B}) \le \gamma(\mathscr{A} \cup \mathscr{B})\delta(\mathscr{A} \cap \mathscr{B})$.

Without loss of generality, one can assume that $\mathscr{A} = \mathscr{B} = \mathscr{A} \cap \mathscr{B} = \mathscr{A} \cup \mathscr{B} = \mathscr{P}(N)$, because one can set $\alpha(A) = 0$ when $A \notin \mathscr{A}$, $\beta(B) = 0$ for all $B \notin \mathscr{B}$, $\gamma = 0$ outside of $\mathscr{A} \cup \mathscr{B}$, and $\delta = 0$ outside of $\mathscr{A} \cap \mathscr{B}$. Then, these functions still satisfy the conditions of the theorem (since one side goes to zero iff the other does).

The third (of two) parts is to induct on n. Most of the work is somehow in the base case: if $n = 1$, then $\mathscr{P}(N) = \{\emptyset, N\}$, and let $\varphi_0 = \varphi(\emptyset)$ and $\varphi_1 = \varphi(N)$. These are all of the possible choices for $\alpha, \dots, \delta$, and in all possible cases life works: $\alpha_0\beta_0 \le \gamma_0\delta_0$, $\alpha_0\beta_1 \le \gamma_1\delta_0$, $\alpha_1\beta_0 \le \gamma_1\delta_0$, and $\alpha_1\beta_1 \le \gamma_1\delta_1$. Then, summing and redistributing, $(\alpha_0 + \alpha_1)(\beta_0 + \beta_1) \le (\gamma_0 + \gamma_1)(\delta_0 + \delta_1)$.

If $\gamma_1 = 0$ or $\delta_0 = 0$, then the required inequality follows, so assume they are nonzero. Then, $\gamma_0 \ge \alpha_0\beta_0/\delta_0$ and $\delta_1 \ge \alpha_1\beta_1/\gamma_1$. Then, it's enough to substitute in these lower bounds:

$$\left(\frac{\alpha_0\beta_0}{\delta_0} + \gamma_1\right)\left(\delta_0 + \frac{\alpha_1\beta_1}{\gamma_1}\right).$$

But multiplying this back out, a lot of the same terms are in common. Then, if $x = \alpha_0\beta_1$, $y = \alpha_1\beta_0$, and $z = \gamma_1\delta_0$, then $z \ge x$ and $z \ge y$, so after some laborious algebra, one can see that $x + y \le z + xy/2$.

And now for the inductive step! Assume it works for $n - 1$, and let $N' = \{1, \dots, n-1\}$ and $N = N' \cup \{n\}$. Then, given φ on subsets of N, define $\varphi'(A) = \varphi(A \cup \{n\}) + \varphi(A)$ on subsets of N'. Thus, $\varphi'(\mathscr{P}(N')) = \varphi(\mathscr{P}(N))$. Thus, the inductive hypothesis holds for $\alpha', \beta', \gamma', \delta' : \mathscr{P}(N) \to \mathbb{R}^+$, as long as they satisfy $\alpha'(A')\beta'(B') \le \gamma'(A' \cup B')\delta'(A' \cap B')$, but this is just the base case again, because one can define a one-element set T and functions $\overline{\alpha}(\emptyset) = \alpha(A')$ and $\overline{\alpha}(T) = \alpha(A' \cup \{n\})$ and similarly for the other three. But then, asking about any $S, R \subseteq T$ boils down to the base case, a statement about the empty sets and the $\alpha_0, \alpha_1, \dots, \delta_0, \delta_1$.Thus, $\overline{\alpha}(\mathscr{P}(T))\overline{\beta}(\mathscr{P}(T)) \le \overline{\gamma}(\mathscr{P}(T))\overline{\delta}(\mathscr{P}(T))$. ⊠

There are more general results, including an uncreatively named Six-Function theorem, and many applications of the FKG theorem to statictical physics and such.

18. The Behavior of $G(n, p)$ With Respect to Component Sizes: 2/19/14

Let $G(n, p)$ be the Erdős-Renyi random graph, and $p = c/n$. Then, let $C(v)$ denote the connected component of a vertex $v \in G$, so that $\{C(v)\}_{v \in G}$ denotes the set of connected components. Around $c \approx 1$ there is a "phase transition:" when $c < 1$, $\max|C(v)| \approx \log n \ll n$, and these connected components look approximately like trees. But when $c > 1$, there exists a unique component of size $O(n)$, which doesn't look like a tree.

Let $Z_i \sim \mathrm{Po}(c)$ be i.i.d. and let T_c denote the first time this crosses zero:

$$T_c = \inf\left\{t : 1 + \sum_{i=1}^{t} Z_i \le t\right\}.$$

This aproximates computing the size of the connected component, and is much easier to analyze.

Claim. $\Pr(T_c < \infty) = 1$ for $c < 1$ and $\mathbb{P}(T_c < \infty) < 1$ for $c > 1$.

This follows from a general large-deviation estimate for sums of i.i.d. random variables: compare the average and the mean. Then,

$$\mathbb{P}\left(\left|\frac{1}{t}\sum_{i=1}^{t} Z_i - c\right| \ge \varepsilon\right) \le K(c, \varepsilon)^{-t}$$

for $K(c, \varepsilon) > 1$. However, a different proof by way of generating functions is more enlightening.

Proof. For $0 < x < \infty$, let

$$\mathbb{R}(x) = \mathbb{E}\left[x^{Z_1}\right] = \sum_{k=0}^{\infty} x^k e^{-c}\frac{c^k}{k!} = e^{c(x-1)}$$

and $\mathbb{Q}(x) = \mathbb{E}\left[x^{T_c}\right]$.

Then, I claim that

$$\mathbb{Q}(x) = \sum_{s=0}^{\infty} \mathbb{P}(Z_1 = s)(x\mathbb{Q}(x))^s,$$

which then simplifies to $\mathbb{R}(x\mathbb{Q}(x)) = e^{c(x\mathbb{Q}(x)-1)|}$, so then $y = \mathbb{Q}(x)$ satisfies $y = e^{c(xy-1)}$. Thus,

$$\mathbb{Q}(1) = \sum_{i=0}^{\infty} \mathbb{P}(T_c = i) = \mathbb{P}(T_c < \infty) = y.$$

Thus, $y = 1$ is a solution, and if $c > 1$ there is an additional solution, as $-\ln y = c(y-1)$.

Conditioning on $Z_1 = s$, we have $T_c = 1 + \tau_1 + \tau_2 + \cdots + \tau_s$, where each τ_i corresponds to $\sum_{j=1}^{t} z_j^{(i)} - t$ first hitting level -1; these τ_i are independent. This should be thought of as a random walk on $\mathbb{Z}$ where the steps (increments) are nonnegative or -1; τ_i is the i^{th} time a step of -1 is taken, until 0 is reached. Then,

$$\mathbb{E}\left[x^{T_c} \mid Z_1 = s\right] = x\mathbb{E}\left[x^{\tau_1+\cdots+\tau_s}\right] = x\mathbb{E}\left[x^{\tau_1}\right]^s = x\mathbb{Q}(x)^s.$$

Then, there is a duality result: let $H = (z_1, \ldots, z_t)$ be the history of $Z_1, \ldots, Z_{T_c}$ given that $T_c = t$ is finite. Then, $\mathbb{P}_\lambda[H = (z_1, \ldots, z_t)] = e^{-\lambda}(\lambda e^{-\lambda})^{t-1}/\prod_{i=1}^{t} z_i!$ as long as this is a "legal history" (i.e. the probability is nonnegative) and $\sum_{i=1}^{t} z_i = t-1$.

Now, let $d < 1 < c$ be called a conjugate pair when $de^{-d} = ce^{-c}$. Since xe^{-x} is increasing on $[0,1)$ and decreasing on $[1,\infty)$, then any $c \neq 1$ has a conjugate. Since $y = e^{c(y-1)}$, then $cye^{-cy} = ce^{-c}$, so $d = cy$ is the conjugate of a $c > 1$. Then, the process with mean $c > 1$, conditioned to have $T = t$ finite, has the conjugate distribution of mean $d < 1$. In some sense, the correspondence says that if it succeeds, one ends up with a giant component (one large connected component of the graph), and if it fails, the random graph has many small disconnected components. The claim is shown by taking a istory $H = (z_1, \ldots, z_t)$; then, $\mathbb{P}_c(H \mid T_c < \infty) = \mathbb{P}_c(H)/\mathbb{P}(T < \infty) = e^{-c})ce^{-c})^{t-1}/y\prod_{i=1}^{t} z_i!$

Then, using the fact that $ce^{-c} = de^{-d}$, this probability also becomes $\mathbb{P}_d(H)$, which expands in the same way. ⊠

Now, one can use breadth-first search to relate this idea to connected components of random graphs. The algorithm for finding $C(v)$ given v looks like this: start with v live and all other nodes neutral (nodes can be live, which are in $C(v)$ but unchecked, dead, which have been checked, or neutral). Let $t = 0$ and $Y_0 = 1$. Then, at each time t, choose any live vertex w and check if $\{w, w'\} \in E$ for any neutral w' (i.e. things we haven't examined yet); if $\{w, w'\} \in E$, then make w' live, and after all such w',mark w as dead, and update Y_0 to be the number of live vertices. Then, if z_i is the number of extra edges found, then

$$Y_t = \sum_{i=1}^{t} z_i - t + 1.$$

The algorithm halts when $Y_t = 0$, with

$$|C(v)| = T = \inf\left\{t : 1 + \sum_{i=1}^{t} z_i - t + 1\right\}.$$

Then, $Z_t \sim \text{Binomial}(n-(t-1)-Y_{t-1}, p)$, and at tme t, there are $t-1$ dead and Y_{t-1} live, so there are $N_{t-1} \coloneqq n-(t-1)-Y_{t-1}$ neutral. As all trials are independent, one can check by induction that $Y_t = Y_{t-1} + z_t - 1 \sim \text{Binomial}(n-1, 1-(1-p)^t) + 1 - t$, and $N_t \sim = n - t - Y_t \sim \text{Binomial}(n-1, (1-p)^t)$.

Theorem 18.1 (Alon & Spencer, Theorem 11.5.1)**.** $\mathbb{P}(|C(v)| = t) \leq \mathbb{P}(\text{Binomial}(n-1, (1-p)^t) = n-t)$ *or is equal to* $\mathbb{P}(\text{Binomial}(n-1, 1-(1-p)^t) = t-1)$.

If $p = c/n$, then $Z_1 \sim \text{Binomial}(n-1.c/n) \to \text{Po}(c)$ as $n \to \infty$. The same holds as $n - N_{t-1} \sim O(n)$, so the number of dead and live vertices is about $o(n)$. Thus, $|C(v)| \approx T_c$, since they're from the same distribution. This comes from the following derivation.

Theorem 18.2. *For all fixed k,*

$$\lim_{h\to\infty} \mathbb{P}(|C(v)| = k \text{ in } G(n, c/n)) = \mathbb{P}(T_c = k).$$

Proof. First, see that

$$\mathbb{P}(T^{\text{po}} = k) = \sum_{\star} P\mathbb{P}(Z_i^{\text{po}} = z_i, 1 \le i \le k) \qquad \text{and} \qquad \mathbb{P}(T^{\text{gr}} = k) = \sum_{\star} P\mathbb{P}(Z_i^{\text{gr}} = z_i, 1 \le i \le k),$$

where T^{po}, etc., are from the Poisson model and T^{gr}, etc., are from the graph model. Here,

$$\star = \{\mathbf{z} = (z_1, \dots, z_k) \mid y_0 = 1, y_t = y_{t-1} + z_t - 1, y_t > 0, t < k, y_k = 0\}.$$

Then,

$$\mathbb{P}(Z_i^{\text{gr}} = z_i, 1 \le i \le k) = \prod_{i=1}^{k} P(\text{Binomial}(N_{i-1}^{\text{gr}}, p) = z_i),$$

but this goes to $\mathbb{P}(\text{Po}(c) = z_i)$, as $N_{i-1}^{\text{gr}} = n - o(n)$. ⊠

Next time, we will see how a coupling argument can be used to prove Theorem 18.2 and another, related theorem (in section 11.6 of [1]).

19. Branching Processes: 2/21/14

Recall that we are looking at the size $|C(v)|$ of the connected component of a vertex $v \in G(n,p)$, which can be written using $T^{\text{gr}} = \inf\{t : Y_t = 0\}$, where $Y_t = Y_{t-1} + Z_t - 1$ and $Y_0 = 1$. This comes form the breadth-first search algorithm (whihc is why it's sometimes called a branching process, but only when the Z_i introduced below are independent): Y_t is the number of live nodes at time t, $n - Y_t - N_t$ is the number of dead nodes, and N_t the number of neutral nodes. One can also define T_c^{Po}, which is the same with $Z_i \sim \text{Po}(c)$, and $T_{n,p}^{\text{Bin}}$, where $Z_i \sim \text{Binomial}(n,p)$. If Z_t is the number of neutral nodes made alive at step t, then $Z_t \sim \text{Binomial}(N_t, p)$.

We also had $y = \mathbb{P}(T_c^{\text{Po}} \text{ is finite})$, which is $y = ce^{c(y-1)}$. Thus, if $c > 1$, there exists a solution $y < 1$.

Theorem 19.1. $\mathbb{P}(T_c^{\text{Po}} = k) = -e^{ck}(ck)^{k-1}/k!$

Theorem 19.2. *For all u, $\mathbb{P}(T_{n-u,p}^{\text{Bin}} \ge u) \le \mathbb{P}(T_{n,p}^{\text{gr}} \ge u) \le \mathbb{P}(T_{n-1,p}^{\text{Bin}} \ge u)$.*

Then, it will be possible to say something about the subcritical case, where $p = c/n$, $c < 1$, or the barely critical graphs where $c = 1 - \varepsilon$, with $\varepsilon = \lambda n^{-1/3}$.

Proof of Theorem 19.2. For the latter inequality, modify the breadth-first search algoritm by replenishing neutral vertices so that there are always $n-1$ neutral nodes. This new algorithm produces possibly larger $|C(v)|$, but follow $T_{n,p}^{\text{Bin}}$.

For the first inequality, observe that all live and dead vertices are part of $C(v)$; hence, to compute $\mathbb{P}(T_{n,p}^{\text{gr}} \ge u)$, it's enough to run the breadth-first search algorithm until $N_t \le n - u$. But here, modify breadth-first search by keeping only $n - u$ vertices at a time. This produces a smaller probability for $|C(v)|$, and follows $T_{n-u,p}^{\text{Bin}}$. ⊠

Now, it's possible to analyze the subcritical case. Let $p = c/n$, with $c < 1$. Then,

$$\mathbb{P}(T_{n,p}^{\text{gr}} \ge u) \le \mathbb{P}(T_{n-1,p}^{\text{Bin}} \ge u) \simeq (1 + o(1))\mathbb{P}(T_c^{\text{Po}} \ge u) \le e^{-\alpha(u)(1+o(1))}$$

for some $\alpha > 0$, so write $u = K \ln n$ for some large, fixed K. But this probability is also $\mathbb{P}(|C(v)| \ge K \ln n) \le n^{-\alpha K(1+o(1))}$, so

$$\mathbb{P}(L_1 \ge K \ln n) \le \sum_{v=1}^{n} \mathbb{P}(|C(v)| \ge K \ln n) \to 0,$$

where $L_1 = \max_v |C(v)|$.

Then, the next step is to bound $\mathbb{P}(|C(v)| \ge u) \le (1 + o(1))\mathbb{P}(T_{1-\varepsilon}^{\text{Po}} \ge u)$, but $\mathbb{P}(T_{1-\varepsilon}^{\text{Po}} \ge A\varepsilon^{-2}) \le \varepsilon e^{-(1+o(1))A/2}$, so take $u = A\varepsilon^{-2}$. Using a Taylor expansion,

$$\mathbb{P}(|C(v)| \ge u) \le (1 + o(1))\mathbb{P}(T_{1-\varepsilon}^{\text{Po}} \ge u \simeq \varepsilon e^{-A/2} = (\lambda n^{-1/3}\lambda^{K/2}.$$

Let $X = \sum_v I_v$, where $I_v = \mathbb{1}_{\{|C(v)| \ge u\}}$, and let Γ be the number of components of size at least u (so we want to bound it above); then, $\Gamma \le X/u$.

Thus, $\mathbb{P}(\Gamma \ge 1) \le \mathbb{R}(\Gamma) \le (1/u)\mathbb{E}[X] = (n/u)\mathbb{P}(|C(v)| \ge u) = \lambda^3 \lambda^{-K/2}/(K \ln \lambda)$, which goes to 0 as $\lambda, n \to \infty$. Thus, the conclusion is that if $p = 1/n - \lambda/n^{4/3} = (1-\varepsilon)/n <$ then $n \to \infty$ and $\lambda \gg 1$ implies that

$$\mathbb{P}(L_1 \ge K\lambda^{-2} \ln \lambda n^{2/3}) = \mathbb{P}(\Gamma \ge 1) \to 0.$$

In other words, the relative size of the largest component decreases as n increases.

In some sense, there are several possible behaviors:

- Subcritical, with $c < 1$ and $p = c/n$. Here, L_1 obeys the relationship seen above: $L_1 \leq K \ln n$.
- Barely subcritical, where $p = (1-\varepsilon)/n$ and $\varepsilon = \lambda n^{-1/3}$, so $L_1 = O(\varepsilon^{-2} \ln \lambda) = O(n^{2/3})$, and all of the other components are of size $O(\ln n)$. The small components are simple (i.e. not far from trees: the number of edges and vertices are not far apart).
- Critical, where $\varepsilon \in [-\lambda n^{-1/3}, \lambda n^{1/3}]$. There are lots of relatively large connected components (relative, that is, to each other) $L_k = c_k n^{2/3}$ and $dk = (\#v - \#e)$ in L_k, so that there's a nontrivial relation between c_k and dk_k.
- Barely supercritical, where $p = (1+\varepsilon)/n$ and $\varepsilon = \lambda n^{-1/3}$. Here, $L_1 \sim 2\varepsilon n \sim O(n^{2/3})$, and $L_2 \sim O(\varepsilon^2 \ln \lambda) = O(n^{2/3} \lambda^{-2} \ln \lambda)$.
- Very supercritical, where $p = c/n$, and $c > 1$. Then, $L_1 = n(1-y)(1+o(1))$, and all of the other components are $O(\log n)$ and simple.

One can think of starting with c small and scaling it up over time. Then, one starts with many small components, and as more edges are added, the connected components grow and merge.

References

[1] Alon, Noga; Spencer, Joel H. *The probabilistic method*. New York: Wiley-Interscience: 2000.
[2] Barbour, A. D., Lars Holst, and Svante Janson. *Poisson Approximation*. Oxford University Press: 1992.
[3] Dembo, Amir and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer: 2009.
[4] Karlin, Samuel and William J. Studden. *Tchebycheff systems: with applications in analysis and statistics*. Interscience Publishers, 1966.