Math 295 - Topics in Discrete Probability: Random Structures and Algorithms

Taught by David Gamarnik
Notes by Dongryul Kim

Fall 2018

¡+instructor+¿ ¡+meetingtimes+¿ ¡+textbook+¿ ¡+enrolled+¿ ¡+grading+¿ ¡+courseassistants+¿

Contents

1 September 5, 2018

A graduate-level introductory course in probability will be enough. We will start with large deviations theory. Next we will talk about compressive sensing, using linear programming methods. Then we will dive into the theory of random graphs, discussing configuration models and component structures. We will introduce the differential equations method to analyze some combinatorial optimization problems. The last part of the course will be devoted to the Markov random field, which is a model with great universality. There are going to be five to six homeworks, and a final project done in teams.

1.1 Chernoff bound

If S_n is a random variable, we want to say something about the probability of S_n being unusually large:

$$\mathbb{P}(S_n \geq y) \approx \exp(-\Theta n).$$

Let $X_1, \dots, X_n$ be independent identically distributed random variables with $\mathbb{E}X_1 = \mu$. For

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

and $a > \mu$, we want to estimate $\mathbb{P}(S_n \geq a) \to 0$. By Chebyshev, we have

$$\mathbb{P}(S_n \geq a) \leq \frac{\mathrm{Var}(X_1)}{n(a-\mu)^2} = O(n^{-1}).$$

But this is not the correct speed.

Define the moment generating function

$$M(\theta) = \mathbb{E}e^{\theta X_1},$$

defined on the domain

$$D_X = \{\theta : M(\theta) < \infty\}.$$

This is a nonempty interval containing 0. This is because $\theta_1 < \theta_2 < \theta_3$ then $e^{\theta_2 X}$ can be bound by a linear combination of $e^{\theta_1 X}$ and $e^{\theta_3 X}$.

Let $\theta > 0$ be in the domain, so that $M(\theta) < \infty$. Fix $a > \mu = \mathbb{E}X_1$. Then

$$\mathbb{P}(S_n \geq a) = \mathbb{P}(e^{\theta n S_n} \geq e^{\theta n a}).$$

By Markov, we can bound

$$\mathbb{P}(S_n \geq a) \leq e^{n\theta a} M^n(\theta) = \Big(\frac{M(\theta)}{e^{\theta a}}\Big)^n.$$

Now the question is whether we can make the ratio smaller than 1. We know that $M(\theta)$ is differentiable in the interior of D_X. In fact, we have $\dot{M}(\theta) = \mathbb{E}Xe^{\theta X}$ by the dominated convergence theorem. Then we can compute

$$\frac{M(\theta)}{e^{\theta a}}\Big|_{\theta=0} = 1, \quad \frac{d}{d\theta}\frac{M(\theta)}{e^{\theta a}}\Big|_{\theta=0} = \mu - a < 0.$$

This means that there exists a θ' such that $M(\theta') < e^{\theta' a}$. So we get a bound of the form

$$\mathbb{P}(S_n \geq a) \leq \lambda^n, \quad \lambda < 1.$$

We can rewrite this bound as

$$\mathbb{P}(S_n \geq a) \leq \exp(-n(\theta a - \log M(\theta))).$$

One thing to observe is that we have the freedom to chose $\theta > 0$. So we should use the θ that gives us the best bound. This leads us to the concept of Legendre transformation.

Definition 1.1. We define the **Legendre transformation** as

$$I(a) = \sup(\theta a - \log M(\theta)), \quad I_+(a) = \sup_{\theta \geq 0}(\theta a - \log M(\theta)), \quad I_-(a) = \sup_{\theta \leq 0}(\theta a - \log M(\theta)).$$

This takes values in $\mathbb{R} \cup \{+\infty\}$.

Then the Chernoff bound states that

$$\mathbb{P}(S_n \geq a) \leq \exp(-nI_+(a)).$$

Actually, we have $I(a) = I_+(a)$ if $a > \mu$ and $I(a) = I_-(a)$ if $a < \mu$. So we just have

$$\mathbb{P}(S_n \geq a) \leq \exp(-nI(a)).$$

But the highlight of large deviations theory is that this is a tight bound. This requires a proof.

Example 1.2. Let us X is exponentially distributed, i.e., $\mathbb{P}(X > t) = e^{-\lambda t}$. Then the moment generating function is

$$D_X = (-\infty, \lambda), \quad M(\theta) = \frac{\lambda}{\lambda - \theta}.$$

Then we have

$$I(a) = a\lambda - 1 - \log \lambda - \log a$$

for all $a > \mu = \lambda^{-1}$.

Proposition 1.3. *(1) The Legendre transformation I is convex, nonnegative. Moreover, $I(\mu) = 0$ and it is non-decreasing on $[\mu, +\infty)$, non-increasing on $(-\infty, \mu]$. Finally, $I(a) = I_+(a)$ for all $a \geq \mu$ and $I(a) = I_0(a)$ for all $a \leq \mu$.*

(2) Assume $D_X = \mathbb{R}$, and also that the domain of X is $\mathbb{R}$, i.e., $\mathbb{P}(X \geq k) > 0$ and $\mathbb{P}(X \leq -k) > 0$ for all k. Then there exists a $\theta_0 \in \mathbb{R}$ such that

$$I(a) = \sup_{\theta}(\theta a - \log M(\theta)) = \theta_0 a - \log M(\theta_0).$$

Moreover, θ_0 satisfies the identity $a = \dot{M}(\theta_0)/M(\theta_0)$.

Proof. (1) Convexity can be proved directly. $I(\mu) = 0$ because Jensen's inequality gives $\log \mathbb{E}e^{\theta X} \geq \log e^{\theta \mathbb{E}X} = \theta\mu$. Monotonicity follows from convexity. □

Now we want to establish that

$$\mathbb{P}(S_n \geq a) \approx \exp(-nI(a))$$

for $a \geq \mu$. For a fixed $A \subseteq \mathbb{R}$, what is $\mathbb{P}(S_n \in A)$? Large deviations theory roughly says that we have

$$\mathbb{P}(S_n \in A) \approx \exp(-n \inf_{a \in A} I(a)).$$

It just happens to be that when A is an half-line, the a that minimizes $I(a)$ is the endpoint.

We want to prove that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in A) = -\inf_{a \in A} I(a).$$

This is not true, as formulated. Suppose $X_1 = \pm 1$ and $A = \{0\}$. Then we have $\mathbb{P}(S_n \in A) = 0$ for n odd, and $\mathbb{P}(S_n \in A) = \mathbb{P}(S_n = 0) \approx O(n^{-1/2})$ by the central limit theorem. So the sequence oscillates between 0 and $-\infty$. The topology of the set A matters.

Theorem 1.4 (Cramer's theorem)**.** *1. For a closed set $A \subseteq \mathbb{R}$, we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in A) \leq -\inf_{a \in A} I(a).$$

2. For an open set $A \subseteq \mathbb{R}$, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in A) \geq -\inf_{a \in A} I(a).$$

2 September 10, 2018

The setup is that we have independent identically distributed $X_1, \dots, X_n$ with common $M(\theta) = \mathbb{E}e^{\theta X}$. We wanted to look at

$$S_n = \frac{X_1 + \cdots + X_n}{n}$$

and $\mu = \mathbb{E}X_1$ with $a > \mu$. We showed that

$$\mathbb{P}(S_n \geq a) \leq \exp(-nI(a)), \quad I(a) = \sup_{\theta \in \mathbb{R}} (\theta a - \log M(\theta)).$$

2.1 Large deviations principle

But our goal is to establish more than this.

Theorem 2.1 (Cramer's theorem)**.** *(a) For every closed set $F \subseteq \mathbb{R}$,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \in F) \leq - \inf_{a \in F} I(a).$$

(b) For every open set $U \subseteq \mathbb{R}$,

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \in F) \geq - \inf_{a \in U} I(a).$$

Proof. (a) We can reduce it to the semi-interval case. Define

$$\alpha_+ = \min\{x \in F \cap (\mu, +\infty)\}, \quad \alpha_- = \max\{x \in F \cap (-\infty, \mu)\}.$$

Then we have

$$\mathbb{P}(S_n \in F) \leq \mathbb{P}(S_n \geq \alpha_+) + \mathbb{P}(S_n \leq \alpha_-) \leq \exp(-nI(\alpha_+)) + \exp(-nI(\alpha_-)).$$

(b) This is the clever part, and we introduce the idea of a change of measure, or exponential tilting. For simplicity, assume $M(\theta) < \infty$ for all θ, and also $\mathbb{P}(X \leq -k) > 0$ and $\mathbb{P}(X \geq k) > 0$ for all k. Recall that for all a, there exists a $\theta_0(a)$ such that

$$I(a) = \theta_0 a - \log M(\theta_0), \quad a = \frac{\dot{M}(\theta_0)}{M(\theta_0)}.$$

Fix any small $\epsilon > 0$. We can find $y \in U$ such that

$$I(y) \leq \inf_{a \in U} I(a) + \epsilon.$$

Then there exists $\delta > 0$ such that $(y - \delta, y + \delta) \subseteq U$, and we will only look at the probability that the average lands in this interval.

Find θ_0 such that $I(y) = \theta_0 y - \log M(\theta)$. Now we define a new random variable X_{θ_0} defined as

$$\mathbb{P}(X_{\theta_0} \leq z) = \frac{\mathbb{E}e^{\theta_0 X} 1_{\{X \leq z\}}}{M(\theta_0)}.$$

Intuitively, you're multiplying the measure by $\theta_0 X$ and renormalizing it. The fundamental fact is that the mean of this new distribution is y. This is because if $f_X(t)$ is the density of the original measure, then

$$\mathbb{E}X_{\theta_0} = \frac{1}{M(\theta_0)} \int z e^{\theta_0 z} f_X(z) = \frac{\dot{M}(\theta_0)}{M(\theta_0)} = y.$$

Now, our goal was to give a lower bound for $\mathbb{P}(S_n \in (y-\delta, y+\delta))$. We can do exponential tilting and get

$$\begin{aligned}
\mathbb{P}(S_n \in (y-\delta, y+\delta)) &= \int_{|\frac{x_1+\cdots+x_n}{n} - y| < \delta} d\mathbb{P}(x_1) \cdots d\mathbb{P}(x_n) \\
&= M^n(\theta_0) \int_{\cdots} \exp(-\theta_0(x_1 + \cdots + x_n)) \prod_{i=1}^n \frac{\exp(\theta_0 x_i)}{M(\theta_0)} d\mathbb{P}(x_i) \\
&\geq \exp(-\theta_0 yn - |\theta_0|dn) M^n(\theta_0) \int_{\cdots} \prod_{i=1}^n \frac{\exp(\theta_0 x_i) d\mathbb{P}(x_i)}{M(\theta_0)} \\
&= \exp(-\theta_0 yn - |\theta_0|dn) M^n(\theta_0) \mathbb{P}\left(\left| \frac{X_1^{\theta_0} + \cdots + X_n^{\theta_0}}{n} - y \right| < \delta \right).
\end{aligned}$$

This thing on the right hand side is covered by the law of large numbers. So

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \in (y-\delta, y+\delta)) \geq \theta_0 y + |\theta_0|\delta - \log M(\theta_0) = -I(y) - |\theta_0|\delta.$$

Now we take $\delta \to 0$. □

This is a multidimensional general theory, but let me describe one thing. Consider independent identically distributed $X_1, \ldots, X_n \in \mathbb{R}^d$, and consider

$$S_n = \frac{\sum_{i=1}^n X_i}{n}.$$

Then we should have

$$\mathbb{P}(S_n \in A) \approx \exp(-n \inf_{a \in A} I(a)),$$

where

$$I(a) = \sup_{\theta \in \mathbb{R}^d} (\langle \theta, a \rangle - \log \mathbb{E} e^{\langle \theta, X \rangle}).$$

2.2 Linear regression

Let us change gears and discuss applications. Suppose we have a model of the form

$$y = x\beta + w,$$

where y, w are scalars and $x, \beta \in \mathbb{R}^p$. We are given samples of x, y, and we want to extract β. In the context of regressions, "big data" has a specific meaning.

In the classical setting, p is small or $O(1)$. But in modern applications, p is very large and comparable to (or larger than) sample size n. (Medical history, cholesterol level one minutes ago, two minutes ago, so on.) In the classical case, we can take

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle)^2.$$

But if p is large, here is a whole space of possible β with zero difference. So it becomes a degenerate system and you don't learn anything.

So we need to something called regularization. This means restricting the possibility of β in some ways. So here is the setup we are interested in. Consider

$$Y = X\beta, \quad Y \in \mathbb{R}^n, \quad X \in \mathbb{R}^{n \times p},$$

where $p \gg n$. We have $\beta \in \mathbb{R}^p$ such that $\|\beta\|_0 \leq S$ with $s < p$, where $\|\beta\|_0$ is the cardinality of i such that $\beta_i \neq 0$. (We also have $\|\beta\|_0 = \lim_{q \to 0} \|\beta\|_q$.) The question is to find β given X and Y.

One possibility is to solve the following problem. Find the vector with the smallest $\|-\|_0$ solving the equation. Then "hope" that this solution $\hat{\beta}_0$ is what we are looking for. The caveat is that this is a very hard problem, because it is non-convex optimization. Indeed, it is NP-hard. So the idea of compressive sensing was instead solve the following:

Minimize $\|\beta\|_1$ subject to $Y = X\beta$.

This becomes a linear programming problem, and we hope that the answer is what we are looking for. If X satisfies the "Restrictive Isometry Property", then we are good.

3 September 12, 2018

We wanted to see when it is the case when we can recover β from

$$Y = X\beta, \quad X \in \mathbb{R}^{n\times p}, \quad Y \in \mathbb{R}^n.$$

We introduce some number $s < p$ called **sparsity**, and then we will look for solutions with sparsity s, i.e.,

$$\|\beta\|_0 \leq s.$$

Given $Y = X\beta$, we want to recover β from (X, Y). The norm $\|\beta\|_0$ is bad, so we instead consider the problem of:

Minimize $\|z\|_1$ subject to $Xz = Y$.

3.1 Restricted isometry property

Definition 3.1. We say that X satisfies the **restricted isometry property(RIP)** with $\delta \in (0, 1)$ and sparsity $s < p$, if for every s-sparse $\beta \in \mathbb{R}^p$,

$$\left|\|X\beta\|_2^2 - \|\beta\|_2^2\right| \leq \delta\|\beta\|_2^2.$$

This means that if you take any s columns in X, then this matrix is an isometry up to a factor.

So the main theorem is the following:

Theorem 3.2. *Suppose X is RIP with $\delta = \delta_{2s} < \frac{1}{3}$. Then for every s-sparse $\beta \in \mathbb{R}^p$, the solution to*

$$\textit{Minimize } \|z\|_1 \textit{ subject to } Xz = X\beta$$

is unique and is β.

We will prove this in two steps.

Lemma 3.3. *Suppose for all $0 \neq v$ such that $Xv = 0$, and for all $S \subseteq [p]$ with $|S| \leq s$, it holds that $\|v_S\|_1 < \frac{1}{2}\|v\|_1$. Then the unique solution of the above linear programming problem is $z = \beta$. (Here we are only assuming $\|\beta\|_0 \leq s$.)*

Proof. Fix an s-sparse $\beta \in \mathbb{R}^p$. Consider any $z \neq \beta$ such that $Xz = X\beta$. We want to show that

$$\|z\|_1 > \|\beta\|_1.$$

(There is no sparsity assumption on z.) Observe that $X(z - \beta) = 0$. Let S be the support of β (so that $|S| \leq s$), and consider its complement $S^c = [p] \setminus S$. Then

$$\|(z - \beta)_S\|_1 < \frac{1}{2}\|z - \beta\|_1$$

because $X(z - \beta) = 0$. Then we can write

$$\|z - \beta\|_1 = \|(z - \beta)_S\|_1 + \|(z - \beta)_{S^c}\|_1,$$

and this implies

$$\|z_S - \beta\|_1 = \|(z-\beta)_S\|_1 < \|(z-\beta)_{S^c}\|_1 = \|z_{S^c}\|_1.$$

Then by the triangle inequality,

$$\|\beta\|_1 \leq \|z_S - \beta\|_1 + \|z_S\| < \|z_{S^c}\|_1 + \|z_S\|_1 = \|z\|_1.$$

This finishes the proof. □

Lemma 3.4. *If X is RIP with $\delta_{2s} < \frac{1}{3}$, then any nonzero solution to $Xv = 0$ satisfies $\|v_S\|_1 < \frac{1}{2}\|v\|_1$ for any $|S| = s$.*

Proof. Let

$$\rho = \frac{2\delta_{2s}}{1-\delta_{2s}} < 1.$$

Therefore it suffices to show that

$$\|v\|_1 \leq \frac{\rho}{2}\|v\|_1$$

for all $Xv = 0$. Moreover, because $\|v_S\|_1 \leq \sqrt{s}\|v_S\|_2$, it suffices to show that

$$\|v_S\|_2 \leq \frac{\rho}{2\sqrt{s}}\|v\|_1.$$

Let $S_0 \subseteq [p]$ be the set of coordinates corresponding to the s largest absolute values, so that $\|S_0\| = s$. Likewise, let S_1, S_2, ... be the next largest s coordinates. Then we can always write

$$v = v_{S_0} + v_{S_1} + \cdots + v_{S_m}$$

and we want to show that

$$\|v_{S_0}\| \leq \frac{\rho}{2\sqrt{s}}\|v\|_1.$$

Because v is in the null space, we have

$$Xv_{S_0} = -\sum_{i=1}^{m} Xv_{S_i}.$$

By the RIP property, we have

$$\begin{aligned}\|v_{S_0}\|_2^2 &\leq \frac{1}{1-\delta_{2S}}\|Xv_{S_0}\|_2^2 \\ &= \frac{1}{1-\delta_{2s}}\sum_{i=1}^{m} -\langle Xv_{S_0}, Xv_{S_i}\rangle.\end{aligned}$$

Under the assumptions here, the following is true: for all $u, v \in \mathbb{R}^p$ which are disjoint and s-sparse,

$$|\langle Xu, Xv\rangle| \leq \delta_{2s}\|u\|_2\|v\|_2.$$

So we get

$$\|v_{S_0}\|_2^2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{i=1}^{m} \|Xv_{S_0}\|_2 \|Xv_{S_i}\|_2$$

and then

$$\|v_{S_0}\|_2 \leq \frac{\rho}{2} \sum_{i=1}^{m} \|v_{S_i}\|_2.$$

Observe that

$$\|v_{S_i}\|_2 \leq \frac{1}{\sqrt{s}} \|v_{S_{i-1}}\|_1$$

because all coordinates of v_{S_i} are smaller than coordinates of $v_{S_{i-1}}$. Now we get

$$\|v_{S_0}\|_2 \leq \frac{\rho}{2\sqrt{s}} \sum_{i=1}^{m} v_{S_{i-1}} \leq \frac{\rho}{2\sqrt{s}} \|v\|_1.$$

This finishes the proof. □

Next time, we will suppose that $X \in \mathbb{R}^{n \times p}$ consists of independent identical distributions $N(0,1)$. Then we will show that X is RIP provided that $n = \Theta(s \log(p/s))$. The key is the we can allow $n \ll p$. So you can do regression with very little samples.

4 September 17, 2018

Today we will show that random matrices have the restricted isometry property. Last time we showed that if some matrix has RIP then we can do linear programming.

4.1 Random matrices are RIP

Consider X be independent identically distributed in $\mathbb{R}^{n\times p}$ with $\mathcal{N}(0,1)$.

Theorem 4.1 (random matrices are RIP)**.** *For every $\delta > 0$, there exist $e_j = e_1(\delta), e_2(\delta)$ such that for all $\epsilon > 0$, if*

$$n \geq \frac{e_1 s \log(\frac{pe}{s}) + \log(\frac{1}{\epsilon})}{e_2}$$

then

$$\mathbb{P}\big(\tfrac{1}{\sqrt{n}}X \text{ is } (\delta, s)\text{-RIP}\big) \geq 1 - \epsilon.$$

(So this applies for $n = \Theta(s \log \frac{p}{s})$.)

Proof. The proof goes in three steps.

- We first prove this for a fixed β, using the Chernoff bound.
- We extend it to all β with the same support S, using the entropy bound, which is approximating the unit ball by some discrete things.
- We extend it to all s-sparse β, by the union bound.

We define

$$J(t) = \min\Big(\frac{t}{2} - \frac{1}{2}\log(1+t), -\frac{t}{2} - \frac{1}{2}\log(1-t)\Big).$$

Now we claim that for all $\beta \in \mathbb{R}^p$, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n}\|X\beta\|_2^2 - \|\beta\|_2^2\Big| \geq t\|\beta\|_2^2\Big) \leq 2\exp(-J(t)n).$$

This is just an application of the Chernoff bound.

Now we can look at a fixed support S with $|S| = s$. For all $\delta \in (0,1)$ and $\rho \in (0, \frac{1}{2})$, we claim that

$$\mathbb{P}\Big(\sup_{\|\beta\|_2 \leq 1}\Big|\frac{1}{n}\|X_S\beta\|_2^2 - \|\beta\|_2^2\Big| \leq \delta\Big) \geq 1 - 2\Big(1 + \frac{2}{\rho}\Big)^s \exp(-J((1-2\rho)\delta)n).$$

The idea is standard in statistics. We fix $\rho \in (0, \frac{1}{2})$. There exists a set

$$U \subseteq \{\beta \in \mathbb{R}^s : \|\beta\|_2 \leq 1\},$$

such that $|U| \leq (1 + \frac{2}{\rho})^s$ and for every β with $\|\beta\|_2 \leq 1$, there exists $\hat{\beta} \in U$ such that $\|\beta - \hat{\beta}\| \leq \rho$. (We call $\log|U|$ the metric entropy.)

Using this fact, let us do the second step. Let $A = \frac{1}{n} X_S^t X_S - I_S$ so that we are interested in estimating

$$\mathbb{P}\Big(\sup_{\beta \in \mathbb{R}^s} |\langle A\beta, \beta \rangle| \leq \delta \Big).$$

For every $\beta \in \mathbb{R}^s$ of unit length, we can find $\hat{\beta} \in U$ such that $|\beta - \hat{\beta}|_2 \leq \rho$. Then

$$|\langle A\beta, \beta \rangle - \langle A\hat{\beta}, \hat{\beta} \rangle| = |\langle A(\beta + \hat{\beta}), \beta - \hat{\beta} \rangle| \leq \|A\|_2 \|\beta + \hat{\beta}\| \|\beta - \hat{\beta}\|_2 \leq 2\rho \|A\|_2.$$

This implies that

$$\|A\|_2 \leq \max_{\hat{\beta} \in U} |\langle A\hat{\beta}, \hat{\beta} \rangle| + 2\rho \|A\|_2,$$

so we have

$$\|A\|_2 \leq \frac{1}{1 - 2\rho} \max_{\hat{\beta} \in U} |\langle A\hat{\beta}, \hat{\beta} \rangle|.$$

If we plug this back into the previous estimates, and use the union bound, we obtain the claim.

Now we get to the third step. We have

$$\mathbb{P}\left(\max_{|S|=s} \sup_{\|\beta\|_2 = 1} \left| \frac{1}{n} \|X_S \beta\|^2 - \|\beta\|^2 \right| \leq \delta \right) \geq 1 - \binom{p}{s} 2\Big(1 + \frac{2}{\rho}\Big)^s \exp(-J((1-2s)\delta)n).$$

We just estimate $\binom{p}{s} \leq p^s / s! \leq (pe/s)^s$. Fix $\rho = \frac{1}{4}$. Then we get the desired result. □

In compressive sensing, a region that is extensively studied is $s = \alpha p$. In that case, $n = \Theta(p)$ after fixing ϵ.

5 September 19, 2018

We are going to change topics and study random graphs.

5.1 Triangles in random graphs

A graph is $G = (V, E)$ where V are vertices and $E \subseteq \{(i, j) : i < j\} \subseteq V^2$ are edges. Naturally, $|E| \leq n(n-1)/2$. There is this famous **Erdös–Renyi** model $\mathbb{G}(n, p_n)$ of an random graph. Here, for all $i < j$, we have that (i, j) is an edge with probability p_n, independently. The set of vertices will just be $V = [n]$.

Nowadays, random graphs are used in many areas. Erdös, in 1947, was the first person to use this in extremal combinatorics. There are four main regimes we might want to consider.

- $p_n = p \in (0, 1)$; these are called dense Erdös–Renyi graphs
- $p_n = n^{-\alpha}$, $0 < \alpha < 1$; these are where certain small graphs, like K_4, emerge
- $p_n = \log n/n$; this is the threshold for connectivity
- $p_n = \frac{d}{n}$; this marks the emergence of giant components (if $d < 1$ then all components are of size $O(\log n)$, but for $d > 1$ one component has order $O(n)$)

Let us first look at the dense Erdös–Renyi graphs. Consider i, j, k vertices, and let

$$I_{ijk} = \begin{cases} 1 & (i, j, k) \text{ is a triangle} \\ 0 & \text{else.} \end{cases}$$

Then $\mathbb{P}(I_{ijk} = 1) = p^3$, so we see that the expectation value of N_3 is asymptotic to $n^3p^3/6$.

Theorem 5.1. *We have*

$$\frac{N_3}{n^3p^3/6} \to 1$$

in probability.

Proof. We use the second moment method. We see that

$$\mathbb{E}N_3 = \frac{n^3p^3}{6} + O(n^2).$$

If we look at the second moment,

$$\begin{aligned} \operatorname{Var}(N_3) &= \sum_{i<j<k} \operatorname{Var}(I_{ijk}) + \sum_{i,j,k,i',j',k'} \operatorname{Cov}(I_{ijk}, I_{i'j'k'}) \\ &= O(n^3) + \sum \operatorname{Cov}(I_{ijk}, I_{i'j'k'}). \end{aligned}$$

The covariance is zero if they share no edges. When they share an edge, the covariance becomes nonzero. So the sum of the covariances is $O(n^4)$. That is, $\mathrm{Var}(N_3) = O(n^4)$ and $\mathbb{E}(N_3) = n^3p^3/6$. □

We can also try to work our a large deviations theory on this distribution. We we do this, we get

$$\mathbb{P}(N_3 \geq (1+\epsilon)\mathbb{E}N_3) \leq \exp(-In^2),$$

and the correct rate I was worked out a few years ago by Chatterjee and Varadhan. But if you ask similar questions for $p_n = n^{-\alpha}$, this is an open problem.

5.2 Connectivity of random graphs

When is the random graph $\mathbb{G}(n, p_n)$ connected? As I alluded to earlier, the correct rate is $\log n/n$.

Theorem 5.2. *Let $p_n = c\log n/n$. Then the random graph $\mathbb{G}(n, p_n)$ is connected with high probability when $c > 1$ and disconnected with high probability when $c < 1$.*

It turns out that the bottleneck in connectivity is the set of isolated nodes. We say that a node u is isolated if it is not connected to the rest of the graph.

Proof. First suppose that $c > 1$. If the graph is not connected, there exists a subset $A \subseteq V$ such that there is no edge between A and $V \setminus A$. Set $\epsilon > 0$ so that $c > 1/(1-\epsilon)$. The expectation of the number of such subsets of size k is,

$$\mathbb{E}I_k = \binom{n}{k}(1-p_n)^{k(n-k)}.$$

Then we can add them and bound the sum.

Define I_u be the variable that is 1 if u is isolated and 0 otherwise. Then we have $\mathbb{E}I_n = (1-p_n)^{n-1}$. If we define $I_{\mathrm{iso}} = \sum_{u=1}^n I_u$, then we are going to show that this goes to infinity. The first moment is

$$\mathbb{E}I_{\mathrm{iso}} = n(1-p_n)^{n-1} = n^{1-c+o(1)}.$$

The second moment is

$$\mathbb{E}I_{\mathrm{iso}}^2 = n^{1-c+o(1)} + \sum_{u\neq v}\mathbb{E}I_uI_v = n(1-p_n)^{n-1} + 2\binom{n}{2}(1-p_n)^{2(n-2)+1} = (1+o(1))(\mathbb{E}I_{\mathrm{iso}})^2.$$

So you can see that the variance is small compared to the expectation. □

6 September 24, 2018

Today we are going to talk about the Azuma–Hoeffding bound.

6.1 Martingales

Let $Z_1, \dots, Z_n, \dots$ be random variables, and let $X_n = f(Z_1, \dots, Z_n)$ be measure with respect to $Z_1, \dots, Z_n$. Then X_n is a stochastic process.

Definition 6.1. We say that X_n is a **martingale** if

(1) $\mathbb{E}|X_n| < \infty$,

(2) $\mathbb{E}[X_{n+1}|Z_1, \dots, Z_n] = X_n$.

It follows that $\mathbb{E}X_n = \cdots = \mathbb{E}X_0$. An example is $Z_1, \dots, Z_n, \dots$ independent identically distributed with $\mathbb{E}Z_1 = 0$, and

$$X_n = \sum_{i=1}^{n} Z_i.$$

This is a random walk.

But this is not what we are interested in. We will look at **Doob's martingale**. Fix $f : \mathbb{R}^N \to \mathbb{R}$ and $Z_1, \dots, Z_N$ random variables. Now we are going to define

$$X_n = \mathbb{E}[f(Z_1, \dots, Z_N)|Z_1, \dots, Z_n]$$

for $1 \le n \le N$. Then X_n is a martingale for $0 \le n \le N$, provided that $\mathbb{E}|X_n| < \infty$. The intuition is that you are revealing more and more information and looking at the expected values. We can also formally prove

$$\begin{aligned}\mathbb{E}[X_{n+1}|Z_1, \dots, Z_n] &= \mathbb{E}[\mathbb{E}[f|Z_1, \dots, Z_{n+1}]|Z_1, \dots, Z_n] \\ &= \mathbb{E}[f|Z_1, \dots, Z_n] = X_n,\end{aligned}$$

using the tower property.

Now let us go back to $\mathbb{G}(n, p)$ and N_3 the number of triangles. Then we have the edges $Z_1, \dots, Z_N$ with $N = \binom{n}{2}$ the independent identically distributed as $\mathrm{Be}(p)$. We can look at the martingale

$$\mathbb{E}[N_3|Z_1, \dots, Z_k]$$

for $0 \le k \le N$. This is a martingale, and these are the types of martingales we will consider.

6.2 Azuma–Hoeffding bound

Theorem 6.2 (Azuma–Hoeffding)**.** *Let X_i for $0 \le i \le n$ be a martingale with $X_0 = 0$, such that $|X_n - X_{n-1}| \le d_i$ almost surely for all $1 \le i \le n$. Then for all $t \ge 0$,*

$$\mathbb{P}(X_n \ge t) \le 2\exp\left(-\frac{t^2}{2\sum d_i^2}\right).$$

If $d_i = d$ and we take $t = xn$, then

$$\mathbb{P}(X_n \geq nx) \leq 2\exp\left(-\frac{x^2n^2}{nd^2}\right) = \exp(-\Theta(n)).$$

So this is a kind of a large deviations bound where the variables might not be independent. The proof is very similar; use the Markov bound to a particular exponential.

Proof. Fix $|x| \leq d_i$, fix a $\lambda > 0$, and define

$$\begin{aligned} f(x) = \exp(\lambda x) &= f\left(\frac{1}{2}\left(\frac{x}{d_i}+1\right)d_i + \frac{1}{2}\left(1-\frac{x}{d_i}\right)(-d_i)\right) \\ &\leq \frac{1}{2}\left(\frac{x}{d_i}+1\right)f(d_i) + \frac{1}{2}\left(1-\frac{x}{d_i}\right)f(-d_i) \end{aligned}$$

by convexity. Here, note that $\frac{1}{2}(\exp(a)+\exp(-a)) \leq \exp(a^2/2)$. So for $|x| \leq d_i$, we get

$$\exp(\lambda x) \leq \exp\left(\frac{\lambda^2 d_i^2}{2}\right) + \frac{\exp(\lambda d_i) - \exp(-\lambda d_i)}{2}x.$$

Now we can use Markov to show

$$\begin{aligned} \mathbb{P}(X_n \geq t) = \mathbb{P}(e^{\lambda X_n} \geq e^{\lambda t}) &\leq \exp(-\lambda t)\mathbb{E}e^{\lambda X_n} \\ &= \exp(-\lambda t)\mathbb{E}\left(\lambda \sum_{i=1}^{n}(X_i - X_{i-1})\right). \end{aligned}$$

The key is the say something about the sum of the increments. We use the tower property.Then

$$\begin{aligned} &\mathbb{E}\exp(\lambda\textstyle\sum_{i=1}^{n}(X_i - X_{i-1})) \\ &\quad = \mathbb{E}[\exp(\lambda\textstyle\sum_{i=1}^{n}(X_i - X_{i-1}) \cdot \mathbb{E}[\exp(\lambda(X_n - X_{n-1}))|\mathcal{F}_{n-1}]]. \end{aligned}$$

Now it is this term $\mathbb{E}[\exp(\lambda(X_n - X_{n-1}))|\mathcal{F}_{n-1}]$ we have linearized. We are going to upper bound the exponential with the linearized term. Then

$$\begin{aligned} &\mathbb{E}[\exp(\lambda(X_n - X_{n-1}))|\mathcal{F}_{n-1}] \\ &\leq \exp\left(\frac{\lambda^2 d_n^2}{2}\right) + \frac{\exp(\lambda d_n) - \exp(-\lambda d_n)}{2d_n}\mathbb{E}[X_n - X_{n-1}|\mathcal{F}_n] = \exp\left(\frac{\lambda^2 d_n^2}{2}\right), \end{aligned}$$

by the martingale property. So we get

$$\mathbb{E}\exp(\lambda\textstyle\sum_{i=1}^{n}(X_i - X_{i-1})) \leq \exp\left(\frac{\lambda^2 d_n^2}{2}\right)\mathbb{E}\exp(\lambda\textstyle\sum_{i=1}^{n-1}(X_i - X_{i-1})).$$

If we iterate this, we get

$$\mathbb{P}(X_n \geq t) \leq \exp\left(-\lambda t + \frac{1}{2}\lambda^2 \sum_{i=1}^{n} d_i^2\right).$$

Now optimize $\lambda > 0$. □

6.3 McDiarmid inequality

This was motivated by the study of empirical processes. Let $Z_1, \dots, Z_n$ be independent identically distributed variables and let F be the cumulative distribution function. If we only have an empirical distribution, we would want to estimate it using

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\} \approx F(x).$$

Then does $\hat{F}_n \to F$ converge?

Theorem 6.3 (Glivenko–Cantelli)**.** *This converges in a strong sense: if we define*

$$L_n = \mathbb{E}\sup_x |\hat{F}_n(x) - F(x)|$$

then $L_n \to 0$.

But in fact, this convergence is stronger; we have $|L_n - \mathbb{E}L_n| \to 0$ almost surely as $n \to \infty$. We will show this using the following inequality.

Theorem 6.4 (McDiarmid)**.** *Let $g : \mathbb{R}^n \to \mathbb{R}$ a function and $d_1, \dots, d_n$ be such that for all $x, y \in \mathbb{R}^n$ we have*

$$|g(x) - g(y)| \le \sum_{i=1}^{n} d_i \mathbf{1}\{x_i \ne y_i\}.$$

If $X_1, \dots, X_n$ are independent identically distributed then

$$\mathbb{P}(|g(X_1, \dots, X_n) - \mathbb{E}g| \ge t) \le 2\exp\left(-\frac{t^2}{2\sum d_i^2}\right).$$

This condition can be thought of as some discrete version of Lipschitz continuity. Suppose $x = (x_1, \dots, x_n)$ with $|x_i| \le C_i$, and let us take g to be a K-Lipschitz continuous function with respect to the $\|-\|_1$ norm:

$$|g(x) - g(y)| \le K\|x - y\|_1.$$

Then this is Lipschitz in the discrete sense, if we set $d_i = 2KC_i$.

So let us apply this to the case of

$$L_n(Z_1, \dots, Z_n) = \sup_x |\hat{F}_n(x) - F(x)|.$$

Let us see how much change there can be, if we only change Z_i to $\hat{Z}_i$. Here, we see that

$$|\hat{F}_n(x) - F_n(x)| \le \frac{1}{n},$$

because we are changing one indicator function in an average. So by a direct application of McDiarmid, we get

$$\mathbb{P}(|L_n - \mathbb{E}L_n| > t) \le 2\exp\left(\frac{t^2}{nn^{-2}}\right) = 2\exp(-t^2 n).$$

So $|L_n - \mathbb{E}L_n| \to 0$ almost surely.

Index