9 Lecture 9, October 31, 2014

9.1 Gale-Berlekamp game

Let's remember a problem we saw in the first lecture (slightly retold):

- You are given an $n \times n$ grid of lightbulbs. For each bulb, at position (i, j), there is a switch b_{ij}; there is also a switch r_i on each row and a switch c_j on each column. The (i, j) bulb is lit if $b_{ij} + r_i + c_j$ is even. For a setting b, r, c of the switches, let $F(b, r, c)$ be the number of lit bulbs minus the number of unlit bulbs. Then $F(b, r, c) = \sum_{ij} (-1)^{b_{ij}+r_i+c_j}$.

  Let $F(b) = \max_{r,c} F(b, r, c)$.

  What is the greatest $f(n)$ such that for all b, $F(b) \geq f(n)$?

This is called the Gale-Berlekamp game after David Gale and Elwyn Berlekamp, who viewed it as a competitive game: the first player chooses b and then the second chooses r and c to maximize the number of lit bulbs. So $f(n)$ is the outcome of the game for perfect players. In the 1960s, at Bell Labs, Berlekamp even built a physical 10×10 grid of lightbulbs with b_{ij}, r_i and c_j switches. People have labored to determine the exact value of $f(n)$ for small n—see [22]. But the key issue is the asymptotics.

Theorem 39 $f(n) \in \Theta(n^{3/2})$.

Proof:

First, the upper bound $f(n) \in O(n^{3/2})$: We have to find a setting b that is favorable for the "minimizing f" player, who goes first. That is, we have to find a b with small $F(b)$.

We'll simply choose a setting u.a.r. and show that $\Pr_b(F(b)) > t) < 1$, so there is a b s.t. $F(b) \leq t$.

With this strategy, by a union bound,

$$\begin{aligned} \Pr_b(F(b)) > t) &\leq \sum_{r,c} \Pr_b(F(b, r, c) > t) \\ &= 2^{2n} \Pr_b(F(b, 0, 0)) > t) \\ &\leq 2^{2n - n^2 D_2(\frac{1+t/n^2}{2} \| \frac{1}{2})} \end{aligned}$$

We use inequality 17, $D(p\|1/2) \geq (2p-1)^2/2$, equivalently $D_2(p\|1/2) \geq (2p-1)^2/(2\log 2)$. Take $t = cn^{3/2}$ for constant c. Then

$$\ldots \leq 2^{2n - c^2 n/(2\log 2)}$$

For large $c > 2\sqrt{\log 2}$ this is < 1.

Next we show the lower bound. Here we must consider any setting b and show how to choose r, c favorably. Initially, set all $r_i = 0$ and pick c_j u.a.r. Then for any fixed i, the row sum

$$\sum_j (-1)^{b_{ij}+r_i+c_j} = \sum_j (-1)^{b_{ij}+c_j}$$

is binomially distributed, being an unbiased random walk of length n.

Now, unlike the Chernoff bound, we'd like to see not an *upper* but a *lower* tail bound on random walk. Let's derive this from the CLT:

Corollary 40 *For X the sum of m uniform iid ± 1 rvs, $E(|X|) = (1 + o(1))\sqrt{2m/\pi}$.*

Now for each row, flip r_i if the row sum is negative. So $E(\sum_j (-1)^{b_{ij}+r_i+c_j}) = (1 + o(1))\sqrt{2/\pi} n^{3/2}$.

This proves the theorem, subject to the CLT. We'll come back soon and finish the proof (with a weaker constant) from first principles. □

9.2 Chernoff bound for general distributions. Moment generating function

Now for a version of the Chernoff bound which we can apply to sums of independent real rvs with very general probability distributions.

After presenting the bound we'll see an application of it, with broad computational applications, in the theory of metric spaces.

Let X be a real-valued random variable with distribution μ: for measurable $S \subseteq \mathbb{R}$, $\Pr(X \in S) = \mu(S)$.

Definition 41 *The* moment generating function *(mgf) (or characteristic function) of X (or, more precisely but less commonly, of μ) is defined for $\beta \in \mathbb{R}$ by*

$$\begin{aligned} g_\mu(\beta) &= E\left[e^{\beta X}\right] \\ &= \sum_0^\infty \frac{\beta^k}{k!} E(X^k) \end{aligned}$$

Incidentally note that if instead of taking β to be real we take it to be imaginary, this gives the Fourier transform.

For any μ, $g_\mu(0) = E[1] = 1$.

Assume $E[X] = \theta$. We are interested in large deviation bounds for random walk with steps from μ. That is, if we sample $X_1, \ldots, X_n$ iid from μ and take $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, we want to know if the distribution of $\overline{X}$ is concentrated around θ. Without loss of generality, take $\theta = 0$, simply by shifting μ.

Perhaps not surprisingly, the quality of the large deviation bound that is possible, depends on how heavy the tails of μ are. What is interesting is that this is nicely measured by the smoothness of g_μ at the origin. Specifically, a moment-generating function that is differentiable at the origin guarantees exponential tails.

One way to think about this intuitively is to examine the Fourier transform (the imaginary axis), rather than the characteristic function, near the origin. If μ has light tails—as an extreme case suppose μ has bounded support—then near the origin, the Fourier coefficients are picking up only very long-wavelength information, and seeing almost no "cancellations"—negative contributions can come only from very far away and therefore be very small. So the Fourier coefficients near 0 are vanishingly different from the Fourier coefficient at 0, and so g_μ is differentiable at 0. This goes both ways—if μ has heavy tails, then even at very long wavelengths, the Fourier integral picks up substantial cancellation, and so the Fourier coefficients change a lot moving away from 0.

Theorem 42 (Chernoff) *If the mgf $g_\mu(\beta)$ is well-defined in a neighborhood of 0 and differentiable at 0, then $\forall \varepsilon \neq 0\ \exists c_\varepsilon < 1$ such that*

$$\Pr(\overline{X}/\varepsilon > 1) < c_\varepsilon^n.$$

Specifically

$$c_\varepsilon \leq \inf_\beta e^{-\beta\varepsilon} g_\mu(\beta) < 1. \tag{22}$$

Proof: Start with the case $\varepsilon > 0$.

$$\begin{aligned} \Pr(\overline{X} > \varepsilon) &= \Pr(e^{\beta\sum_i X_i} > e^{\beta n\varepsilon}) && \text{for any } \beta > 0 \qquad (23) \\ &< e^{-\beta n\varepsilon} E\left[e^{\beta\sum_i X_i}\right] && \text{Markov bound} \\ &= e^{-\beta n\varepsilon}\left(E\left[e^{\beta X_1}\right]\right)^n && X_i \text{ are independent} \\ &= \left(e^{-\beta\varepsilon} g_\mu(\beta)\right)^n && (24) \end{aligned}$$

We now need to show that there is a $\beta > 0$ such that $e^{-\beta\varepsilon} g_\mu(\beta) < 1$. At $\beta = 0$, $e^0 g_\mu(0) = 1$, so let's find the derivative of $e^{-\beta\varepsilon} g_\mu(\beta)$ at 0. Since g_μ is differentiable at 0 we have:

$$\begin{aligned}
\left.\frac{\partial g_\mu(\beta)}{\partial \beta}\right|_0 &= \left.\frac{\partial E\left[e^{\beta X}\right]}{\partial \beta}\right|_0 \\
&= E\left[\frac{\partial e^{\beta X}}{\partial \beta}\right]\Bigg|_0 \\
&= E\left[X e^{\beta X}\right]\Big|_0 \\
&= E[X] \;=\; \theta \;=\; 0
\end{aligned}$$

So, because we have shifted the mean to 0, the moment-generating function is flat at 0. Now we can differentiate the whole function:

$$\begin{aligned}
\left.\frac{\partial e^{-\beta\varepsilon} g_\mu(\beta)}{\partial \beta}\right|_0 &= e^{-\varepsilon\cdot 0} g'_\mu(0) - \varepsilon e^{-\varepsilon\cdot 0} g_\mu(0) && \text{product rule} \\
&= \underbrace{e^{-\varepsilon\cdot 0}}_{1}\underbrace{g'_\mu(0)}_{0} - \varepsilon\underbrace{e^{-\varepsilon\cdot 0}}_{1}\underbrace{g_\mu(0)}_{1} && \text{at } \beta = 0 \\
&= -\varepsilon && (25)
\end{aligned}$$

We have determined that $\exists \beta > 0$ such that $e^{-\beta\varepsilon} g_\mu(\beta) < 1$, and thus there is a $c_\varepsilon < 1$ as stated in the theorem.

The case $\varepsilon < 0$ is similar. All that changes is that for line 23 we substitute

$$\Pr(\overline{X} < \varepsilon) = \Pr(e^{\beta \sum_i X_i} > e^{\beta n \varepsilon}) \qquad \text{for any } \beta < 0 \qquad (26)$$

The rest of the derivation is identical up to and including line 25, which in this case shows that $\exists \beta < 0$ such that $e^{-\beta\varepsilon} g_\mu(\beta) < 1$, and thus there is a $c_\varepsilon < 1$ as stated in the theorem. □

This method also allows us, in some cases, to find the value of c_ε which gives the tightest Chernoff bound. (For general μ and ε this can be a complicated task and we may have to settle for bounds on the best c_ε.)

9.3 Johnson-Lindenstrauss embedding $L_2 \to L_2$

By a small sample we may judge the whole piece.
Cervantes

Today we'll see a geometric application of the Chernoff bound. At first glance the question we solve, which originates in analysis, appears to have nothing to do with probability. But actually it illustrates a shared geometric core between analysis and probability.

Definition 43 *A metric space (M, d_M) is a set M and a function $d_M : M \times M \to (\mathbb{R} \cup \infty)$ that is symmetric; nonnegative; 0 on, and only on, the diagonal; and obeys the triangle inequality, $d_M(x,y) \le d_M(x,z) + d_M(z,y)$.*

Examples:

1. A Euclidean space is a vector space $\mathbb{R}^n$ equipped with the metric $d(x,y) = \sqrt{\sum_1^n (x_i - y_i)^2}$.
2. The same vector space can be equipped with a different metric, for instance the ℓ_∞ metric, $\max_i |x_i - y_i|$.
3. ... or the ℓ_1 metric, $\sum_i |x_i - y_i|$.

4. ...or the same metric may be applied to only a portion of the vector space: let Δ_n denote the *probability simplex*, $\Delta_n = \{x \in \mathbb{R}^n : \sum_i x_i = 1, x_i \geq 0\}$. In this space (half of) the ℓ_1 distance is referred to as "total variation distance."

5. Many metric spaces have nothing to do with vector spaces. An important class of metrics are the *shortest path metrics*, derived from undirected graphs: If $G = (V, E)$ is a graph and $x, y \in V$, let $d(x, y)$ denote the length of (number of edges on) a shortest path between them.

Definition 44 *An embedding $f : M \to M'$ is a mapping of a metric space (M, d_M) into another metric space $(M', d_{M'})$. The distortion of the embedding is $\sup_{a,b,c,d \in M} \frac{d_{M'}(f(a),f(b))}{d_M(a,b)} \cdot \frac{d_M(c,d)}{d_{M'}(f(c),f(d))}$. The mapping is called isometric if it has distortion 1.*

A finite metric space is one in which the underlying set is finite. A finite ℓ_2 space is one that can be embedded isometrically into a Euclidean space of any dimension.

Exercise: The dimension need not be greater than $n - 1$.

Of course, n points span only at most an $(n-1)$-dimensional affine subspace.

Exercise: Generically, the dimension must be $n - 1$. (The distances between points in Euclidean space determine their coordinates up to a rotation, reflection and translation. Consider the volume of the simplex.)

What we'll see today is a method of embedding an n-point ℓ_2 metric into a very low-dimensional Euclidean space with only slight distortion. This is useful in the theory of computation because many algorithms for geometric problems have complexity that scales exponentially in the dimension of the input space. We'll have to skip giving example applications, but there are quite a few by now, and because of these, a variety of improvements and extensions of the embedding method have also been developed.

Our goal is to prove the following claim:

Theorem 45 (Johnson and Lindenstrauss [36]) *Given a set, A, of n points in a Euclidean space, there exists a map $f : A \to (\mathbb{R}^k, l_2)$ with $k = 8(1 + O(\varepsilon))\varepsilon^{-2} \log n$ that is of distortion e^ε on the metric restricted to A. Moreover, the map f can be taken to be linear and can be found with a simple randomized algorithm in expected time polynomial in n.*

Although the points of A generically span an $(n-1)$-dimensional affine space, and the map is linear, nonetheless observe that we are *not* embedding all of $\mathbb{R}^{n-1}$ with low distortion—that is impossible, as the map is many-one—we care only about the distances among our n input points.

9.4 Normed spaces

A real normed space is a vector space V equipped with a nonnegative real-valued "norm" $\|\cdot\|$ satisfying $\|cv\| = c\|v\|$ for $c \geq 0$, $\|v\| \neq 0$ for $v \neq 0$, and $\|v + w\| \leq \|v\| + \|w\|$. As is very familiar, norms immediately give rise to metrics, as in examples 1, 2, 3, by taking the distance between v and w to be $\|v - w\|$.

Let $\mathcal{S} = (S, \mu)$ be any measure space. For $p \geq 1$, the L_p normed space w.r.t. the measure μ, $L_p(\mathcal{S})$, is defined to be the vector space of functions

$$f : S \to \mathbb{R}$$

of finite "L_p norm," defined by

$$\|f\|_p = \left(\int_S \|f(x)\|^p d\mu(x) \right)^{1/p}$$

Exercise:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

So (like any normed space), $L_p(\mathcal{S})$ is also automatically a metric space.

This framework allows us to discuss the collection of all L_2 (Euclidean) spaces, all L_1 spaces, etc. The most commonly encountered cases are indeed L_1, L_2 and L_∞, which is defined to be the sup norm (so μ doesn't matter). Today we discuss embeddings $L_2 \to L_2$. Time permitting we may also discuss embeddings of general metrics into L_1.

We will use the shorthand L_p^k to refer to an L_p space on a set S of cardinality k, with the counting measure.