3.4 Lecture 16 (7/Nov): Application of CLT to Gale-Berlekamp. Khintchine-Kahane. Moment generating functions

3.4.1 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 $10 \times 10$ 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 [36]. But the key issue is the asymptotics.

**Theorem 52** $f(n) \in \Theta(n^{3/2})$.

**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 $F(b)$.

Fix any $r, c$. Then for $b$ selected u.a.r.,

$$\Pr(F(b, r, c) > kn^{3/2}) \leq 2^{-n^2D_2(\frac{1}{2} + \frac{1}{\sqrt{2\pi}})}$$

we’ll choose a value for $k$ shortly

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

using $D(p||1/2) \geq (2p - 1)^2/2$

Now take a union bound over all $r, c$.

$$\Pr(F(b) > kn^{3/2}) \leq 2^{2n-k^2n/(2\log 2)}$$

For $k > 2\sqrt{\log 2}$ this is $< 1$. So $\exists b$ s.t. $\forall r, c, F(b, r, c) \leq 2\sqrt{\log 2}n^{3/2}$.

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

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

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


(Proof sketch: for $X$ distributed as the unit-variance Gaussian $\mathcal{N}(0, 1)$, this value is exact; see [104]. The CLT shows this is a good enough approximation to our rv.)

Comment: Instead of using Corollary 53, we could alternatively have used the following result:

**Theorem 54 (Khintchine-Kahane)** Let $a = (a_1, \ldots, a_n), a_i \in \mathbb{R}$. Let $s_i \in \{\pm 1\}$ and set $S = \sum s_i a_i$.

Then $\frac{1}{\sqrt{2}} \|a\|_2 \leq \mathbb{E}(S) \leq \|a\|_2$.

The original result of this form is [62]; the above constant and generality are found in [68]; for an elegant one-page proof see [33]. Not coincidentally, both this result and the CLT are proven through Fourier analysis.

Comment: Since we haven’t provided proofs of either of these, and we are about to use them, let me mention that later in the course (Sec. 4.3.1) we’ll come back and finish the proof (with a weaker constant) through a more elementary argument, and with the added benefit that we will be able to give the player a deterministic poly-time strategy for choosing the row and column bits. (Here we gave the player only a randomized poly-time strategy.)

In any case we now continue, using the conclusion (with the largest constant, coming from the CLT): for every $i$, $\mathbb{E}(|X_i|) = (1 + o(1)) \sqrt{2n/\pi}$. Now for each row, flip $r_i$ if the row sum is negative. So $\mathbb{E}(\sum_i (-1)^{r_i} X_i) = \mathbb{E}(\sum_i |X_i|) = \sum_i \mathbb{E}(|X_i|) = (1 + o(1)) \sqrt{2/\pi n^{3/2}}$.

This shows (assuming the CLT) that for any $b$, $E_{\max} F(b, r, c)$ is $(1 + o(1)) \sqrt{2/\pi n^{3/2}}$. Consequently, for all $b$, $F(b) \geq (1 + o(1)) \sqrt{2/\pi n^{3/2}}$, which proves the theorem. \hfill \Box

Comment: It was convenient in this problem that the system of switches at your disposal was “bipartite”, that is, there are no interactions amongst the effects of the row switches, and likewise amongst the effects of the column switches. However, even when such effects are present it is possible to attain similar theorems. See [59].

### 3.4.2 Moment generating functions, Chernoff bound for general distributions

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 $\mu$: for measurable $S \subseteq \mathbb{R}$, $\Pr(X \in S) = \mu(S)$.

**Definition 55** The moment generating function (mgf) of $X$ (or, more precisely, of $\mu$) is defined for $\beta \in \mathbb{R}$ by

$$ g_\mu(\beta) = \mathbb{E}[e^{\beta X}] \quad \text{provided this converges in an open neighborhood of 0} $$

$$ = \sum_{k=0}^{\infty} \frac{\beta^k}{k!} \mathbb{E}(X^k) $$

Incidentally note that (a) if instead of taking $\beta$ to be real we take it to be imaginary, this gives the Fourier transform, (b) both are “slices” of the Laplace transform.

For any probability measure $\mu$, $g_\mu(0) = \mathbb{E}[1] = 1$.

We are interested in large deviation bounds for random walk with steps from $\mu$. That is, if we sample $X_1, \ldots, X_n$ iid from $\mu$ 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 $E[X]$. It will be convenient to re-center $\mu$, if necessary, so that $E[X] = 0$; clearly this just subtracts a known constant off each step of the rw, so it does not affect any probabilistic calculations. So without loss of generality we now take $E[X] = 0$.

Perhaps not surprisingly, the quality of the large deviation bound that is possible, depends on how heavy the tails of $\mu$ are. What is interesting is that this is nicely measured by the smoothness of $g_\mu$ 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 $\mu$ has light tails—as an extreme case suppose $\mu$ 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 $g_\mu$ is differentiable at 0. This goes both ways—if $\mu$ 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 56 (Chernoff)** If the mgf $g_\mu(\beta)$ is differentiable at 0, then $\forall \varepsilon \neq 0 \exists c_\varepsilon < 1$ such that

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

Specifically

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

**Proof:** Let $\mathcal{N}$ be a neighborhood of 0 in which the mgf converges. Start with the case $\varepsilon > 0$.

$$\Pr(\bar{X} > \varepsilon) = \Pr(e^{\beta \sum_i X_i} > e^{\beta \varepsilon n})$$

for any $\beta > 0$ \hspace{1cm} (3.7)

$$< e^{-\beta \varepsilon} E\left[ e^{\beta \sum_i X_i} \right]$$

Markov bound, for $\beta \in \mathcal{N}$

$$= e^{-\beta \varepsilon} \left( E\left[ e^{\beta X_1} \right] \right)^n$$

$X_i$ are independent

$$= \left( e^{-\beta \varepsilon} g_\mu(\beta) \right)^n$$  

(3.8)

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 $g_\mu$ is differentiable at 0 we have:

$$\frac{\partial g_\mu(\beta)}{\partial \beta} \bigg|_0 = \frac{\partial E\left[ e^{\beta X} \right]}{\partial \beta} \bigg|_0$$

$$= E\left[ X e^{\beta X} \right] \bigg|_0$$

$$= E\left[ X \right] = 0$$

So, because we have shifted the mean to 0, the moment-generating function is flat at 0.

Now we can differentiate the whole function:

$$\frac{\partial e^{-\beta \varepsilon} g_\mu(\beta)}{\partial \beta} \bigg|_0 = e^{-\varepsilon} g_\mu'(0) - \varepsilon e^{-\varepsilon} g_\mu(0)$$

product rule

$$= e^{-\varepsilon} g_\mu'(0) - \varepsilon e^{-\varepsilon} g_\mu(0)$$

at $\beta = 0$

$$= -\varepsilon$$  

(3.9)
We have determined that $\exists \beta > 0$ such that $e^{-\beta \epsilon g_\mu(\beta)} < 1$, and thus there is a $c_\epsilon < 1$ as stated in the theorem.

The case $\epsilon < 0$ is similar. All that changes is that for line 3.7 we substitute

$$\Pr(\bar{X} < \epsilon) = \Pr(e^{\beta \sum X_i} > e^{\beta n \epsilon})$$

for any $\beta < 0$ (3.10)

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

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

Exercise: What is the mgf of the uniform distribution on $\pm 1$? What is the best $c_\epsilon$?