Chapter 3

Concentration of Measure

3.1 Lecture 8 (25/Oct): Independent rvs, Chernoff bound, applications

3.1.1 Independent rvs

**Lemma 30** If $X_1, \ldots, X_n$ are independent real rvs with finite expectations (recall this assumption requires that the integrals converge absolutely), then

$$E(\prod X_i) = \prod E(X_i).$$

The proof is an exercise which we only suggest. It is enough to consider the case $n = 2$ and proceed by induction. Recall the definition of expectation from Eqn 1.4:

$$E(X) = \lim_{h \to 0} \sum_{\text{integer}} -\infty < j < \infty \, jh \Pr(jh \leq X < (j+1)h)$$

and apply

$$\Pr((jh \leq X < (j+1)h) \land (j'h \leq Y < (j'+1)h))$$

$$= \Pr(jh \leq X < (j+1)h) \cdot \Pr(j'h \leq Y < (j'+1)h)$$

for independent $X, Y$.

If you want to do the measure theory carefully, this boils down to the Fubini Theorem.

3.1.2 Chernoff bound for uniform Bernoulli rvs (symmetric random walk)

The Chernoff bound\(^1\) will be one of two ways in which we’ll display the concentration of measure phenomenon, the other being the central limit theorem. In the types of problems we’ll be looking at the Chernoff bound is the more frequently useful of the two but they’re closely related.

Let’s begin with the special case of iid fair coins, aka iid uniform Bernoulli rvs: $P(X_i = 1) = 1/2, P(X_i = 0) = 1/2$. Put another way, we have $n$ independent events, each of which occurs with probability 1/2.

We want an exponential tail bound on the probability that significantly more than half the events occur. This very short argument is the seed of more general or stronger bounds that we will see later.

It will be convenient to use the rvs $Y_i = 2X_i - 1$, where $X_i$ is the indicator rv of the $i$th event. This shift lets us work with mean-0 rvs. This leaves the $Y_i$ independent; that is a special case of the following lemma, which is an immediate consequence of the definitions in Sec 1.1.2.

\(^1\)First due to Bernstein \cite{12, 13, 11} but we follow the standard naming convention in Computer Science.
Lemma 31 If $f_1, \ldots$ are measurable functions and $X_1, \ldots$ are independent rvs then $f_1(X_1), \ldots$ are independent rvs.

(Proof omitted.)

Theorem 32 Let $Y_1, \ldots, Y_n$ be iid rvs, with $\Pr(Y_i = -1) = \Pr(Y_i = 1) = 1/2$. Let $Y = \sum_i^n Y_i$. Then $\Pr(Y > \lambda \sqrt{n}) < e^{-\lambda^2/2}$ for any $\lambda > 0$.

Proof: Fix any $\alpha > 0$. Exercise\footnote{Exercise 2: $E(e^{\alpha Y_i}) = \cosh \alpha \leq e^{\alpha^2/2}$.}

$$E(e^{\alpha Y_i}) = \cosh \alpha \leq e^{\alpha^2/2}.$$ By independence of the rvs $e^{\alpha Y_i}$,

$$E(e^{\alpha Y}) = \prod E(e^{\alpha X_i}) \leq e^{n \alpha^2/2}.$$ 

$$\Pr(Y > \lambda \sqrt{n}) = \Pr(e^{\alpha Y} > e^{n \lambda \sqrt{n}}) \leq \frac{E(e^{\alpha Y})}{e^{n \lambda \sqrt{n}}} \quad \text{Markov ineq.} \leq e^{n \alpha^2/2 - n \lambda \sqrt{n}}$$

We now optimize this bound by making the choice $\alpha = \lambda / \sqrt{n}$, and obtain:

$$\Pr(Y > \lambda \sqrt{n}) \leq e^{-\lambda^2/2}.$$ \hfill $\square$

3.1.3 Application: set discrepancy

For a function $\chi : \{1, \ldots, n\} \rightarrow \{1, -1\}$ and a subset $S$ of $\{1, \ldots, n\}$, let $\chi(S) = \sum_{i \in S} \chi(i)$. Define the discrepancy of $\chi$ on $S$ to be $|\chi(S)|$, and the discrepancy of $\chi$ on a collection of sets $S = \{S_1, \ldots, S_n\}$ to be $\text{Disc}(\chi) = \max_j |\chi(S_j)|$.

Theorem 33 (Spencer) With the definitions above, there is a function $\chi$ of discrepancy $\text{Disc}(\chi) \in O(\sqrt{n})$.

We won’t provide Spencer’s argument, but the starting point for it is the proof of the following weaker statement.

Theorem 34 With the definitions above, a function $\chi$ selected u.a.r. has $\text{Disc}(\chi) \in O(\sqrt{n \log n})$ with positive probability.

Proof: By Theorem 32, for any particular set $S_j$ (noting that $|S_j| \leq n$),

$$\Pr(|\chi(S_j)| > c \sqrt{n \log n}) = \Pr(|\chi(S_j)| > \frac{c \sqrt{n \log n}}{|S_j|} \sqrt{|S_j|})$$

$$\leq 2e^{-\frac{c^2 n \log n}{2|S_j|}} \leq 2e^{-\frac{c^2 \log n}{2}} = 2n - c^2/2,$$

\footnote{For $k \geq 0$, $(2k)! = \prod_{i=k}^{k+i} (i+1) \geq 2^k k!$, so for any real $x$, $e^{x^2/2} = \sum_{k \geq 0} x^{2k} / (2^k k!) \geq \sum_{k \geq 0} x^{2k} / (2k)! = \cosh x$.}
Now take a union bound over the sets.

\[
\Pr(\exists j: |\chi(S_j)| > c\sqrt{n \log n}) \leq n \Pr(|\chi(S_j)| > c\sqrt{n \log n}) < 2n^{1-c^2/2}.
\]

Plug in any \(c > \sqrt{2}\) to show the theorem for sufficiently large values of \(n\). \(\Box\)

### 3.1.4 Slightly stronger Chernoff bound; robustness of the definition of BPP

When we introduced BPP we specified that at the end of the poly-time computation, strings in the language should be accepted with probability \(\geq 2/3\), and strings not in the language should be accepted with probability \(\leq 1/3\). We also noted that these values were immaterial and did not even need to be constants—we need only that they be separated by some \(1/poly\). Here’s why. We start by defining two important functions.

**Definition 35** The entropy (base 2) of a probability distribution \(\{p_1, \ldots, p_n\}\) is

\[
h_2(p) = \sum p_i \log \frac{1}{p_i}.
\]

In natural units we use

\[
h(p) = \sum p_i \log \frac{1}{p_i}.
\]

**Definition 36** Let \(r = (r_1, \ldots, r_n)\) and \(s = (s_1, \ldots, s_n)\) be two probability distributions and suppose \(s_i > 0\) \(\forall i\). The (base 2) Kullback-Leibler divergence \(D_2(r\|s)\) “from \(r\) to \(s\),” or “of \(r\) w.r.t. \(s\),” is defined by

\[
D_2(r\|s) = \sum r_i \log \frac{r_i}{s_i}
\]

This is also known as information divergence, directed divergence or relative entropy\(^3\). In natural log units the divergence is \(D(r\|s) = \sum r_i \log \frac{r_i}{s_i}\), and we also use this notation when the base doesn’t matter. \(D(r\|s)\) is not a metric (it isn’t symmetric and doesn’t satisfy the triangle inequality) but it is nonnegative, and zero only if the distributions are the same.

**Exercise:**

(a) \(D(r\|s) \geq 0 \ \forall r, s\)

(b) \(D(r\|s) = 0 \Rightarrow r = s\)

(c) \(D(s + \epsilon\|s) = \sum \left(\frac{\epsilon_i^2}{2s_i} + O(\epsilon_i^3)\right)\)

The “\(\parallel\)” notation is strange but is the convention.

When \(s\) is the uniform distribution, we have:

\[
D(r\|\text{uniform}) = \sum r_i \log(nr_i) = \log n + \sum r_i \log r_i = \log n - h(r)
\]

So \(D(r\|\text{uniform})\) can be thought of as the entropy deficit of \(r\), compared to the uniform distribution.

In the case \(n = 2\) we will write \(p\) rather than \((p, 1-p)\), thus: \(h_2(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}\), \(D_2(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}\).

Let’s extend and improve the previous large deviation bound for symmetric random walk. The new bound is almost the same for relatively mild deviations (just a few standard deviations) but is much stronger at many (especially, \(\Omega(\sqrt{n})\)) standard deviations. It also does not depend on the coins being fair.

---

\(^3\)\(D\) is useful throughout information theory and statistics (and is closely related to “Fisher information”). See [17].
Theorem 37 If $X_1, \ldots, X_n$ are iid coins each with probability $q$ of being heads, the probability that the number of heads, $X = \sum X_i$, is $> pn$ (for $p \geq q$) or $< pn$ (for $p \leq q$), is $< 2^{-nD_2(p||q)} = \exp(-nD(p||q))$.

Exercise: Derive from the above one side of Stirling’s approximation for $\binom{n}{pn}$.

Note 1: this improves on Thm 32 even at $q = 1/2$ because the inequality $\cosh \alpha \leq \exp(\alpha^2/2)$ that we used before, though convenient, was wasteful. (But the two bounds converge for $p$ in the neighborhood of $q$.) Specifically we have (see Figure 3.1):

$$D(p\|1/2) \geq (2p - 1)^2/2 \quad (3.1)$$

Note 2: The divergence is the correct constant in the above inequality; and this remains the case even when we “reasonably” extend this inequality to alphabets larger than 2—that is, dice rather than coins; see Sanov’s Theorem [17, 69]. There are of course lower-order terms that are not captured by the inequality.

Proof: Consider the case $p \geq q$; the other case is similar. Set $Y_i = X_i - q$ and $Y = \sum Y_i$. Now for $\alpha > 0$,

$$\Pr(Y > n(p - q)) = \Pr(e^{\alpha Y} > e^{n(p-q)})$$

$$< E(e^{\alpha Y})/e^{n(p-q)} \quad \text{Markov}$$

$$= \left(\frac{(1-q)e^{-\alpha q} + qe^{\alpha(1-q)}}{e^{\alpha(p-q)}}\right)^n \quad \text{Independence}$$

Set $\alpha = \log \frac{p(1-q)}{(1-p)q}$. Continuing,

$$= \left(1 - q \right) \left( p \right) \left( \frac{1 - q}{1 - p} \right)^p q \left( \frac{1 - q}{1 - p} \right)^{1-p} \right)^n$$

$$= \left( \frac{q}{p} \right)^p \left( 1 - q \right)^{1-p} \right)^n$$

$$= e^{-nD(p||q)}$$

This is saying that the probability of a coin of bias $q$ empirically “masquerading” as one of bias at least $p > q$, drops off exponentially, with the coefficient in the exponent being the divergence.
Back to BPP

Suppose we start with a randomized polynomial-time decision algorithm for a language \( L \) which for \( x \in L \), reports “Yes” with probability at least \( p \), and for \( x \notin L \), reports “Yes” with probability at most \( q \), for \( p = q + 1 / f(n) \) for some \( f(n) \in n^{O(1)} \).

Exercise: \( D(q + \epsilon || q) \geq \Omega(\epsilon^2) \). (With the constant in the \( \Omega \) depending on \( q \).)

Also, \( D(q + \epsilon || q) \) is monotone in each of the regions \( \epsilon > 0, \epsilon < 0 \).

So if we perform \( O(n f^2(n)) \) repetitions of the original BPP algorithm, and accept \( x \) iff the fraction of “Yes” votes is above \( (p + q)/2 \), then the probability of error on any input is bounded by \( \exp(-n) \).

3.1.5 Balls and bins

Suppose you throw \( n \) balls, uniformly iid, into \( n \) bins. What is the highest bin occupancy?

Let \( A_i = \# \text{ balls in bin } i \). Claim: \( \forall c > 1, \Pr(\max A_i > c \log n / \log \log n) \in o(1) \).

Proof: by the union bound,

\[
\Pr(\max A_i > c \log n / \log \log n)
\leq n \Pr(A_i > c \log n / \log \log n)
\leq n \exp(-n D(\frac{c \log n}{n \log \log n} | 1/n))
= n \left( \frac{\log \log n}{c \log n} \right)^{\frac{c \log n}{n \log \log n}} \left( 1 - \frac{1 - 1/n}{1 - \frac{c \log n}{n \log \log n}} \right)^{\frac{1 - c \log n}{n \log \log n}}
\leq n \left( \frac{\log \log n}{c \log n} \right)^{\frac{c \log n}{n \log \log n}} \left( 1 - \frac{1 - c \log n}{1 - \frac{c \log n}{n \log \log n}} \right)^{\frac{1 - c \log n}{n \log \log n}}
\]

Exercise: for \( 0 \leq p < 1 \), \( \left( \frac{1}{1-p} \right)^{1-p} \leq e^p \)[4] So

\[
\ldots \leq n \left( \frac{\log \log n}{c \log n} \right)^{\frac{c \log n}{n \log \log n}} \exp \left( \frac{c \log n}{\log \log n} \right)
= \exp \left( (1 - c) \log n + \frac{c \cdot \log n \cdot (1 - \log c + \log \log \log n)}{\log \log n} \right)
\leq n^{1 - c + o(1)}
\]

Omitted: use Poisson approximation to show matching lower bound for suitable \( 0 < c < 1 \).

---

[4]I think actually \( \leq 1 + p \) but we don’t need this.
CHAPTER 3. CONCENTRATION OF MEASURE

3.2 Lecture 9 (27/Oct): Applications of large deviation bounds: Shannon’s coding theorem, Gale-Berlekamp

3.2.1 Preview of Shannon’s coding theorem

This is an exceptionally important application of large deviation bounds. Consider one party (Alice) who can send a bit per second to another party (Bob). She wants to send him a $k$-bit message. However, the channel between them is noisy, and each transmitted bit may be flipped, independently, with probability $p < 1/2$. What can Alice and Bob do? You can’t expect them to communicate reliably at 1 bit/second anymore, but can they achieve reliable communication at all? If so, how many bits/second can they achieve? This question turns out to have a beautiful answer that is the starting point of modern communication theory.

Before Shannon came along, the only answer to this question was, basically, the following naïve strategy: Alice repeats each bit some $\ell$ times. Bob takes the majority of his $\ell$ receptions as his best guess for the value of the bit.

We’ve already learned how to evaluate the quality of this method: Bob’s error probability on each bit is bounded above by, and roughly equal to, $\exp(-\ell D(1/2||p))$. In order for all bits to arrive correctly, then, Alice must use $\ell$ proportional to $\log k$. This means the rate of the communication, the number of message bits divided by elapsed time, is tending to 0 in the length of the message (scaling as $1/\log k$). And if Alice and Bob want to have exponentially small probability of error $\exp(-k)$, she would have to employ $\ell \sim k$, so the rate would be even worse, scaling as $1/k$.

Shannon showed that in actual fact one does not need to sacrifice rate for reliability. This was a great insight, and we will see next time how he did it. Roughly speaking—but not exactly—his argument uses a randomly chosen code. He achieves error probability $\exp(-\Omega(k))$ at a constant communication rate. What is more, the rate he achieves is arbitrarily close to the theoretical limit.

3.2.2 Shannon’s block coding theorem. A probabilistic existence argument.

In order to communicate reliably, Alice and Bob are going to agree in advance on a codebook, a set of codewords that are fairly distant from each other (in Hamming distance), with the idea that when a corrupted codeword is received, it will still be closer to the correct codeword than to all others. In this discussion we completely ignore a key computational issue: how are the encoding and decoding maps computed efficiently? In fact it will be enough for us, for a positive result, to demonstrate existence of an encoding map $E : \{0,1\}^k \rightarrow \{0,1\}^n$ and a decoding map $D : \{0,1\}^n \rightarrow \{0,1\}^k$ (we’ll call this an $(n,k)$ code) with the desired properties; we won’t even explicitly describe what the maps are, let alone specify how to efficiently compute them. We will call $k/n$ the rate of such a code. Shannon’s great achievement was to realize (and show) that you can simultaneously have positive rate and error probability tending to 0—in fact, exponentially fast.

Theorem 38 (Shannon [71]) Let $p < 1/2$. For any $\varepsilon > 0$, for all $k$ sufficiently large, there is an $(n,k)$ code with rate $\geq D_2(p||1/2) - \varepsilon$ and error probability $\exp(-\Omega(k))$ on every message. (The constant in the $\Omega$ depends on $p$ and $\varepsilon$.)

In this theorem statement, “Error” means that Bob decodes to anything different from $X$, and error probabilities are taken only with respect to the random bit-flips introduced by the channel.

Proof: Let

$$n = \frac{k}{D_2(p||1/2) - \varepsilon} \quad (3.2)$$
As a first try, let’s design \( \mathcal{E} \) by simply mapping each \( X \in \{0,1\}^k \) to a uniformly, independently chosen string in \( \{0,1\}^n \). (This won’t be good enough for the theorem.)

So (for now) when we speak of error probability, we have two sources of randomness: channel noise \( R \), and code design \( \mathcal{E} \).

To describe the decoding procedure we start with the notion of Hamming distance \( H \). The Hamming distance \( H(x,y) \) between two same-length strings over a common alphabet \( \Sigma \), is the number of indices in which the strings disagree: \( H(x,y) = |\{ i : x_i \neq y_i \}| \) for \( x,y \in \Sigma^n \).

Define the decoding \( \mathcal{D} \) to map \( Y \) to a closest codeword in Hamming distance.

In order to analyze how well this works, we pick \( \delta \) sufficiently small that

\[
p + \delta < 1/2
\]

and

\[
D_2(p + \delta \| 1/2) > D_2(p \| 1/2) - \epsilon/2.
\]  

Fix now a particular message \( X \) and analyze the probability that it is decoded incorrectly.

If Bob decodes \( X \) incorrectly then at least one of the following events has to have occurred:

**Bad\(_1\):** \( H(\mathcal{E}(X) + R, \mathcal{E}(X)) \geq (p + \delta)n \)

**Bad\(_2\):** \( \exists X' \neq X : H(\mathcal{E}(X) + R, \mathcal{E}(X')) \leq (p + \delta)n \)

So let’s analyze the probability of failure of either of these clauses.

For Bad\(_1\), we rely only on the randomness in \( R \). Applying Lemma 37, we have

\[
\Pr( H(\mathcal{E}(X) + R, \mathcal{E}(X)) \leq (p + \delta)n ) = \Pr(H(\mathcal{E}(X)) \geq (p + \delta)n) \leq 2^{-D_2(p+\delta\|1/2)n}.
\]

For Bad\(_2\), we rely only on the randomness in \( \mathcal{E} \). Consider that for every \( X' \neq X \),

\[
\Pr(H(\mathcal{E}(X) + R, \mathcal{E}(X')) \leq (p + \delta)n) \leq 2^{-nD_2(p+\delta\|1/2)} \text{ using Eqn 3.3} \]

Now we allow for both sources of error: that \( R \) is heavier than \( (p + \delta)n \), or (with a union bound) that one of the \( 2^k - 1 \) incorrect messages has its codeword land within \( (p + \delta)n \) of the corrupted codeword of \( X \).

\[
\Pr(\exists X' \neq X \text{ s.t. } \mathcal{D}(\mathcal{E}(X) + R) = X') \leq 2^{-nD_2(p+\delta\|p)} + 2^{k-n[D_2(p\|1/2)-\epsilon/2]}
\]

\[
= 2^{-nD_2(p+\delta\|p)} + 2^{n[D_2(p\|1/2)-\epsilon] - n[D_2(p\|1/2)-\epsilon/2]} \text{ by Eqn 3.2}
\]

\[
= 2^{-nD_2(p+\delta\|p)} + 2^{-n\epsilon/2}
\]

\[
\leq 2^{1-cn} \text{ where } c := \min\{D_2(p+\delta\|p), \epsilon/2\} \tag{3.5}
\]
Another way of stating this conclusion is by conditioning on the choice of $E$.

\[
2^{1-cn} \geq \Pr_{E,R}(\text{Error on } X) = E_E(\Pr_R(\text{Error on } X|E)) = E_E(M_X)
\]

where we have defined $M_X$ be the rv (which is a function of $E$) $M_X = \Pr_R(\text{Error on } X|E)$. So for any $E$, $0 \leq M_X \leq 1$.

Now let $Z$ be the rv which is the fraction of $X$'s for which $M_X \leq 2E(M_X)$. By the Markov inequality, $\exists E$ s.t. $Z \geq 1/2$. Let $E^*$ be a specific such code.

This isn’t quite what we want—we want $M_X$ to be small for all messages $X$.

But there is a simple solution. Choose a code $E^*$ as above for $k+1$ bits, then map the $k$-bit messages to the good half of the messages. Note that removal of some codewords from $E^*$ can only decrease any $M_X$. (Assuming we still use closest-codeword decoding.)

So now the bound $\Pr_R(\text{Error on } X) \leq 2E(M_X) \leq 2^{2-cn}$ applies to all $X$.

The asymptotic rate is unaffected by this trick; the error exponent is also unaffected.

To be explicit, using $E^*$ designed for $k+1$ bits and with $n = \frac{k+1}{D_2(p\|1/2)-\epsilon}$ we have for all $X \in \{0,1\}^k$

\[
\Pr_R(\text{Error on } X) \leq 2^{2-cn}
\]

Thus no matter what message Alice sends, Bob’s probability of error is exponentially small.

3.3 Central limit theorem

As I mentioned earlier in the course, there are two basic ways in which we express concentration of measure: large deviation bounds, and the central limit theorem. Roughly speaking the former is a weaker conclusion (only upper tail bounds) from weaker assumptions (we don’t need full independence—we’ll talk about this soon).

The proof of the basic CLT is not hard but relies on a little Fourier analysis and would take us too far out of our way this lecture, so I will just quote it:

Let $\mu$ be a probability distribution on $\mathbb{R}$, i.e., for $X$ distributed as $\mu$, measurable $S \subseteq \mathbb{R}$, $\Pr(X \in S) = \mu(S)$.

For $X_1, \ldots, X_n$ sampled independently from $\mu$ set $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

**Theorem 39** Suppose that $\mu$ possesses both first and second moments:

\[
\theta = E[X] = \int x \, d\mu \quad \text{mean}
\]

\[
\sigma^2 = E[(X - \theta)^2] = \int (x - \theta)^2 \, d\mu \quad \text{variance}
\]

Then for all $a < b$,

\[
\lim_{n} \Pr(\frac{a\sigma}{\sqrt{n}} < \bar{X} - \theta < \frac{b\sigma}{\sqrt{n}}) = \frac{1}{\sqrt{2\pi}} \int_{a}^{b} e^{-t^2/2} \, dt.
\]
3.3.1 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 [26]. But the key issue is the asymptotics.

**Theorem 40** $f(n) \in \Theta(n^{3/2})$.

**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,

$$
\Pr_b(F(b) > t) \leq \sum_{b,c} \Pr_b(F(b, r, c) > t) \\
= 2^{2n} \Pr_b(F(b, 0, 0) > t) \\
\leq 2^{2n - n^2 D_2(\|\cdot\|_1^2 / 2)}
$$

We use inequality [3.1] $D(p\|1/2) \geq (2p - 1)^2 / 2$, 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

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

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 41** For $X$ the sum of $m$ uniform iid $\pm 1$ rvs, $E(|X|) = (1 + o(1)) \sqrt{2m / \pi}$.
Consequently for every $i$, $E(|X_i|) = (1 + o(1))\sqrt{2n/\pi}$.

Now for each row, flip $r_i$ if the row sum is negative. So $E(\sum_i (\sum_i X_i)^i) = E(\sum_i |X_i|) = \sum_i E(|X_i|) = (1 + o(1))\sqrt{2/\pi n^{3/2}}$.

This shows (assuming the CLT) that for any $b$, $E_{\max, r} F(b, r, c)$ 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.

Later in the course (Sec. 5.2.2) 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.)