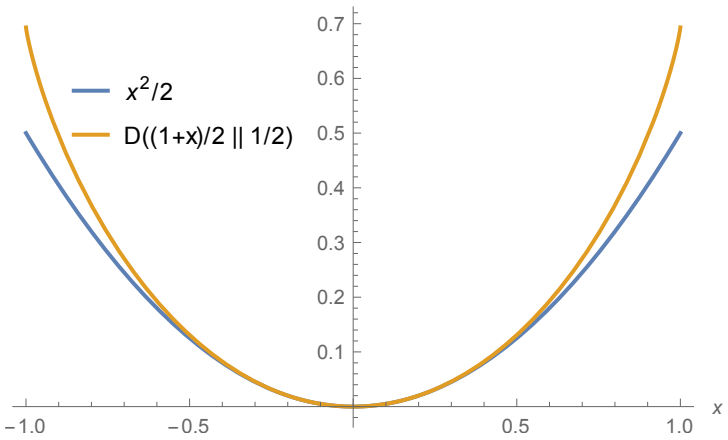

Figure 3.2: Comparing the two Chernoff bounds at $q = 1/2$

3.2 Lecture 14 (2/Nov): Stronger Chernoff bound, applications

3.2.1 Chernoff bound using divergence; robustness of BPP

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.

Theorem 48 *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 43 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 3.2):

$$D(p\|1/2) \geq (2p-1)^2/2 \tag{3.1}$$

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 [27, 92]. There are of course lower-order terms that are not captured by the inequality.

Note 3: Let's see what we mean by "concentration of measure". Clearly, the Chernoff bound is telling us that something, namely the rv X, is very tightly concentrated about a particular value. On the other hand, if you look at the full underlying rv, namely the vector $(X_1, \ldots, X_n)$, that is not concentrated at all; if say $q = 1/2$, then it is actually as smoothly distributed as it could be, being uniform on the hypercube! The concentration of measure phenomenon, then, is a statement about *low dimension representation of high dimensional objects*. In fact the "representation" does not have to be a nice linear function like $X = \sum X_i$. It is sufficient that $f(X_1, \ldots, X_n)$ be a *Lipschitz* function, namely that that there be some constant bound c s.t. flipping any one of the X_i's changes the function value by no more than c. From this simple information you can already get a large deviation bound on f for independent inputs X_i.

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 $\alpha > 0$,

$$\begin{aligned}
\Pr(Y > n(p-q)) &= \Pr(e^{\alpha Y} > e^{\alpha n(p-q)}) \\
&< E(e^{\alpha Y})/e^{\alpha n(p-q)} \quad \text{Markov} \\
&= \left(\frac{(1-q)e^{-\alpha q} + qe^{\alpha(1-q)}}{e^{\alpha(p-q)}}\right)^n \quad \text{Independence}
\end{aligned}$$

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

$$\begin{aligned}
&= \left(\left(\frac{q}{p}\right)^p \left(\frac{1-q}{1-p}\right)^{1-p}\right)^n \\
&= e^{-nD(p\|q)}
\end{aligned}$$

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 $f(n) \in n^{O(1)}$.

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

So if we perform $O(nf^2(n))$ 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 $\exp(-n)$.

3.2.2 Balls and bins

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

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

To avoid a morass of iterated logarithms, write $L = \log n, L_2 = \log\log n, L_3 = \log\log\log n$. So we wish to show $\Pr(\max A_i > cL/L_2) \in o(1)$.

Proof: by the union bound,

$$\begin{aligned}
&\Pr(\max A_i > cL/L_2) \\
\leq\ & n \Pr(A_i > cL/L_2) \\
\leq\ & n \exp(-nD(\frac{cL}{nL_2}\|\frac{1}{n})) \\
=\ & n \left(\frac{L_2}{cL}\right)^{\frac{cL}{L_2}} \left(\frac{1-1/n}{1-\frac{cL}{nL_2}}\right)^{(1-\frac{cL}{nL_2})n} \\
\leq\ & n \left(\frac{L_2}{cL}\right)^{\frac{cL}{L_2}} \left(\frac{1}{1-\frac{cL}{nL_2}}\right)^{(1-\frac{cL}{nL_2})n}
\end{aligned}$$

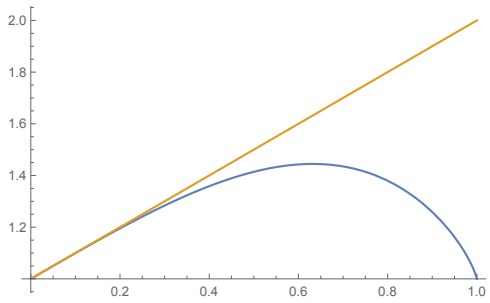

Figure 3.3: $(\frac{1}{1-p})^{1-p}$ vs. $1+p$

Expand the first term and apply the following inequality[4] to the second term: For $0 \leq p < 1$, $(\frac{1}{1-p})^{1-p} \leq e^p$.

$$\begin{aligned} \dots &\leq \exp\left(L + \frac{cL}{L_2}(L_3 - L_2 - \log c) + \frac{cL}{L_2}\right) \\ &= \exp((1-c)L + cL\frac{L_3 - \log c + 1}{L_2}) \\ &\leq \exp((1-c)L + o(1)) = n^{1-c+o(1)}. \end{aligned}$$

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

3.2.3 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 ℓ times. Bob takes the majority of his ℓ 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 ℓ 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.

[4] In fact we have the stronger $(\frac{1}{1-p})^{1-p} \leq 1+p$ (see Fig. 3.2.2) although we don't need this. Let $\alpha = \log \frac{1}{1-p}$, so $\alpha \geq 0$. Then $p = 1 - e^{-\alpha}$ and we are to show that $2 \geq e^{-\alpha} + e^{\alpha e^{-\alpha}} =: f(\alpha)$. $f(0) = 2$ and $f' = e^{-\alpha}(e^{\alpha e^{-\alpha}} - 1 - \alpha)$ so it suffices to show for $\alpha \geq 0$ that $g(\alpha) := e^{\alpha e^{-\alpha}} \leq 1 + \alpha$. At $\alpha = 0$ this is satisfied with equality so it suffices to show that $1 \geq g' = (1-\alpha)e^{-\alpha}g$. Since $1 - \alpha \leq e^{-\alpha}$, it suffices to show that $1 \geq e^{-2\alpha}g = e^{\alpha(e^{-\alpha}-2)}$, which holds (with room to spare) because $e^{-\alpha} \leq 1$.