CSE696 Week 6: Probabilistic Analysis and BPP.

The key concept is a **random variable** over a probability distribution D. D itself is formally a function on a set S whose values are nonnegative real numbers that sum to 1. When just S is given and S is finite, the default assumption is that D is the **uniform distribution** on S, that is, $D(x) = 1/|S|$ for all $x \in S$. The random variable is formally a function f from S to a range R that **induces** a distribution E on R by the rule

$$E(y) = \sum_{x \in S : f(x) = y} D(x).$$

For example, consider the number k of heads in n flips of a fair coin. For $n = 1$, we have $S = \{0, 1\}$ standing for the outcomes tails and heads, respectively, and D gives each equal probability $1/2$. For general n, we have the domain $S^n = \{0, 1\}^n$, and $f : S^n \to \{0, \ldots, n\}$. The induced distribution is the (unbiased) binomial distribution

$$B_n(k) = \frac{1}{2^n}\binom{n}{k}.$$

We have mapped uniform distribution to the binomial distribution. Often the task of finding a *fair* way to sample a target distribution E lies in defining a mapping from uniform to E.

The most immediate task is to make inferences about E given only partial information about D, f, and/or E itself. Some of this information involves the **mean** and **variance**, which are defined when the range is numeric:

$$\mu_E = \mathbf{E}_D[f] = \mathbf{E}_{x \leftarrow D}[f(x)] = \sum_{x \in S} f(x)D(x)$$

$$\sigma_E^2 = \mathbf{E}_{y \leftarrow E}\left[(y - \mu_E)^2\right] = \mathbf{E}_{x \leftarrow D}\left[(f(x) - E_{x \leftarrow D}[f(x)])^2\right]$$

We can also write these as μ_f and σ_f^2 when we want to emphasize the underlying distribution D rather than the top-level distribution E. Another form for the variance comes when you multiply the square out:

$$\sigma_E^2 = \mathbf{E}_{y \leftarrow E}\left[(y - \mu_E)^2\right] = \mathbf{E}_{y \leftarrow E}\left[y^2\right] - 2\mathbf{E}_{y \leftarrow E}[y\mu_E] + \mathbf{E}_{y \leftarrow E}\left[\mu_E^2\right] = \mathbf{E}_{y \leftarrow E}\left[y^2\right] - \mu_E^2 .$$

$$-2\mathbf{E}_{y \leftarrow E}[y\mu_E] = -2\mu_E\mathbf{E}[y] = -2\mu_E^2 .$$

The square root of the variance is called the **standard deviation** and is just σ_E or σ_f or just σ depending on context. Here are two bounds we get by using first μ and then σ as the given partial information:

Markov's Inequality: Let f be any random variable on D with nonnegative numerical values, and let $a \geq 0$. Then

$$\Pr_{x \leftarrow D}[f(x) \geq a] \leq \frac{\mu_f}{a}.$$

To sanity-check this, consider that $a = 0$ gives the inequality $1 \leq \infty$, but that is fine. The case $a = 1$ says that the probability of getting $f(x) \geq 1$ cannot exceed the mean. Of course, if the mean is ≥ 1 in that case then the conclusion is again trivial (since the probability is always ≤ 1). So there is "tension" only when the mean is < 1, in which case there must be values < 1 having nonzero probability, so $\Pr_{x \leftarrow D}[f(x) \geq 1] < 1$ as well. The proof throws away two terms.

Proof: Define $S_a = \{x \in S : f(x) \geq a\}$. Throwing away the complement of S_a from the sum of nonnegative values that defines μ_f as above, we have:

$$\mu_f \geq \sum_{x \in S_a} f(x)D(x) \geq \sum_{x \in S_a} aD(x) = a \Pr_{x \leftarrow D}[f(x) \geq a],$$

from which the result follows. ☒

So this looks like a very stupid theorem that throws away a lot. However, it can be used in ways that don't throw away so much, and thereby undergirds results that aren't as stupid. The first one might be called "semi-intelligent":

Chebyshev's Inequality: Let f be any numerical random variable (negative values now OK), and suppose μ_f and σ_f are both known. Then we know that for any $k > 0$,

$$\Pr_E[|y - \mu_f| \geq k\sigma_f] \leq \frac{1}{k^2}.$$

Proof: Define $g = (y - \mu_f)^2$. We could "ground" this on (S, D) by defining $g(x) = (f(x) - \mu_f)^2$, but we can reason without this "crutch." This is a random variable defined on E that takes non-negative values. We could "re-base" it on the original distribution D but part of the fun is that we can reason entirely at the level of E. First, note that $\mu_g = \mathbf{E}_{y \leftarrow E}\left[(y - \mu_f)^2\right]$ which is just the definition of σ_f^2. Markov's inequality tells us that for any $a > 0$,

$$\Pr_{y \leftarrow E}[g(y) \geq a] \leq \frac{\mu_g}{a} = \frac{\sigma_f^2}{a}.$$

Now choose $a = k^2\sigma_f^2$. Then the right-hand side becomes $\frac{1}{k^2}$, while the left-hand side becomes $\Pr_{y \leftarrow E}\left[g(y) \geq k^2\sigma_f^2\right]$, which is the same as saying $\Pr_{y \leftarrow E}[|y - \mu_f| \geq k\sigma_f]$. ☒

With $k = 2$, what this says is that *for any numerical distribution whatsoever*, the chance of getting a numerical result that is two standard deviations or more away from the mean is at most 0.25. There are distributions where this is tight. In my opinion, the most overt one is for the random variable X with

$$\Pr[X = -1] \;=\; 0.125,\;\; \Pr[X = 0] \;=\; 0.75, \text{ and } \Pr[X = +1] \;=\; 0.125.$$

Expressed as a function from a simpler distribution, this is $f(000) \;=\; -1$, $f(111) \;=\; +1$, $f(001) = f(010) = f(011) = f(100) = f(101) = f(110) = 0$. The mean is 0, while the standard deviation is $\sigma_f^2 \;=\; 0.125 \cdot (-1 \;-\; 0)^2 \;+\; 0.75 \cdot (0-0)^2 \;+\; 0.125 \cdot (1-0)^2 \;=\; 0.25$. This makes $\sigma = 0.5$. The values -1 and $+1$ are thus each 2 standard deviations away from the mean and have collective probability $0.25 \;=\; 1/2^2$, which is the limit by Chebychev's theorem.

It is amusing to try to make a tight case with a **rv.** that takes only 2 values. Suppose we insist the values be 0 and 1 with $p \;=\; \Pr(1)$ and $q \equiv (1-p) \;=\; \Pr(0)$. The mean is $\mu \;=\; p$ and the variance is

$$p(1-p)^2 \;+\; (1-p)(0-p)^2 \;=\; p \;-\; 2p^2 \;+\; p^3 \;+p^2 \;-p^3 \;=\; p - p^2 \;= p(1-p) \;= pq \;\;.$$

If the value 1 is to be $k = 2$ standard deviations above the mean, then we get the equation

$$1 = p \;+\; 2\sqrt{p(1-p)}, \;\; \text{so} \;\; 1 - p = 2\sqrt{p(1-p)}, \;\; \text{so} \;\; (1-p)^2 = 4p(1-p), \;\; \text{so} \;\; 1 - p = 4p.$$

Thus $p \;=\; 1/5$. This makes $\Pr[X = 1]$, i.e., $\Pr[X - \mu \geq 2\sigma]$, equal to 0.2, which is less than the Chebychev bound of 0.25. So it is not tight. Can we make it tight with numerical values x, y other than 0 and 1? Maybe with a different value of k? This is idle fun to work out.

Independence and Sums of RVs.

For less-idle application, let's see how close Chebychev comes for the case of the binomial distribution $B_{n,p}$ on $\{0, \ldots, n\}$. As a random variable, it is the **sum** of n **independent** random variables. Now I tend to insert extra layers of technicality, but what tends to get hidden in presentations is the concept of random variables being projected-out. Suppose we have a support set of the form $S \;=\; S_1 \times S_2$ with distribution D and a distribution E on $R \;=\; R_1 \times R_2$ induced by a function $F : S \to R$. For any $(y_1, y_2) \in R$ we have

$$E(y_1, y_2) \;=\; \sum_{(x_1,x_2):F(x_1,x_2)=(y_1,y_2)} D(x_1, x_2).$$

Now we define the projected random variables via their distributions as

$$E_1(y_1) \;=\; \sum_{(x_1,x_2):(\exists y)F(x_1,x_2)=(y_1,y)} D(x_1, x_2) \;\; \text{and} \;\; E_2(y_2) \;=\; \sum_{(x_1,x_2):(\exists y)F(x_1,x_2)=(y,y_2)} D(x_1, x_2)$$

It is important to note that these are induced by the functions $F_1(x_1, x_2) = y_1$ and $F_2(x_1, x_2) = y_2$, where for both cases $(y_1, y_2) = F(x_1, x_2)$, but are **not** necessarily induced by functions $f_1 : S_1 \to R_1$ and $f_2 : S_2 \to R_2$ defined on the separate components. This happens only when the projected random variables are independent, which we can now rigorously define as that for all $(y_1, y_2) \in R$,

$$E(y_1, y_2) \;=\; E_1(y_1) \cdot E_2(y_2).$$

It is high time for a revelatory example. Let us use $S_1 = S_2 = \{H, T\}$; I am not using $\{0, 1\}$ as the domain in order to ward off a potential confusion. Now define the function F by:

$$F(H,H) \;=\; F(H,T) \;=\; (0,0);\;\; F(T,H) \;=\; F(T,T) \;=\; (1,1)\,.$$

Under the presumed uniform distribution D on $\{H, T\}^2$, this gives the outcomes $(0, 0)$ and $(1, 1)$ equal probability 0.5, but gives the outcomes $(0, 1)$ and $(1, 0)$ zero probability. The projected rvs. are:

$$E_1(0) \;=\; D(H,H) \;+\; D(H,T) \;=\; 0.5\,,\;\; E_1(1) \;=\; D(T,H) + D(T,T) \;=\; 0.5\,,$$

and similarly,

$$E_2(0) \;=\; D(H,H) \;+\; D(H,T) \;=\; 0.5\,,\;\; E_2(1) \;=\; D(T,H) + D(T,T) \;=\; 0.5\,.$$

But we do not get independence, because e.g. $E_1(0) \cdot E_2(0) \;=\; 0.25$ but $E(0,0) \;=\; 0.5$. [In quantum mechanics, the putative domain will be two **qubits**, one held by "Alice" and the other by "Bob". The random variable $Y_A \;=\; y_1$ will give the outcome of a **one-qubit measurement** made by Alice and $Y_B \;=\; y_2$ the outcome for Bob. The lack of independence will translate to saying the the two qubits are **entangled**. This is quite against what you would get if Alice and Bob really flipped separate coins with outcomes H and T, as the notation may have suggested.]

Now consider the derived random variable giving the sum of the y_1 and y_2 outcomes. Formally it is induced by the function

$$G_F(x_1, x_2) \;=\; y_1 + y_2 \;\text{ where }\; (y_1, y_2) = F(x_1, x_2).$$

The possible values are $R_G = \{0, 1, 2\}$, but the induced distribution E_G has $E_G(0) \;=\; 0.5$, $E_G(1) \;=\; 0$, and $E_G(2) \;=\; 0.5$. The mean is 1, but the variance is $0.5(0-1)^2 + 0.5(2-1)^2 \;=\; 1\,.$

Now suppose we have random variables defined via functions f_1, f_2 on the separate components with distributions D_1, D_2. Then the Cartesian-product random variable is the function on $D_1 \times D_2$ defined by $f(x_1, x_2) \;=\; (f_1(x_1), f_2(x_2))$, and its induced distribution

$$E(y_1, y_2) \;\;=\;\; \sum_{(x_1,x_2):\, f(x_1,x_2)=(y_1,y_2)} D_1(x_1) \cdot D_2(x_2)$$

guarantees independence. Now if f_1 gives equal probability to the outcomes 0 and 1, and f_2 likewise, then each of the outcomes $(0, 0), (0, 1), (1, 0), (1, 1)$ has probability 0.25. Moreover, the sum, which is what we mean by the unbiased binomial distribution B_2 on $\{0, 1, 2\}$, has mean 1 but variance

$$\sigma_2^2 \;=\; 0.25(0-1)^2 \;+\; 0.5(1-1)^2 \;+\; 0.25(2-1)^2 \;\;=\;\; 0.5\,,$$

rather than variance 1 as in the entangled case. Now for some nomenclature:

- Even in the non-independent case, Y_A and Y_B are still considered **joint** random variables, since they are projected from the same random variable Y representing the outcome of the whole system.
- When there is independence in a Cartesian product of random variables, we can call them **independently joint**.
- When they are independent and all E_i are the same (not just that all D_i on the individual domain sets S_i are the same), they are **independently and identically distributed**, abbreviated **i.i.d.**
- We can apply similar terms---with care of context---to random variables that are really *derived* from the joint ones, such as the **sum**, that are not the joint rvs. themselves.

Now we return to top-level in Arora-Barak section 7.4 to discuss rules for sums of random variables $X_1, \ldots, X_n$:

1. The expectation $\mathbf{E}[X_1 + \cdots + X_n]$ of the sum is always the sum $\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]$ of the expectations, even if the variables are not independent.
2. If the variables are independent, then the variance is additive as well: $\boldsymbol{\sigma^2}[X_1 + \cdots + X_n] = \boldsymbol{\sigma^2}[X_1] + \cdots \boldsymbol{\sigma^2}[X_n]$.
3. In the case of possibly-biased binary distribution $B_{n,p}$ with $\Pr[X_i = 1] = p$ for all i, the variables are i.i.d. and the sum has mean pn and variance $n \cdot p(1-p) = npq$.

Now we can talk about how close B_n (taking $p = 0.5$) comes to Chebychev's inequality. We have $\sigma^2 = 0.25n$, so $\sigma = \frac{1}{2}\sqrt{n}$. If we are interested in $k = 2$ then this gives $2\sigma = \sqrt{n}$. Trying $n = 9$ gives $\mu = 4.5$ and $2\sigma = 3$. The only outcomes higher than $\mu + 2\sigma$ are 8 and 9. There are 9 combinations that produce the sum 8, with probability $1/2^9$ each. plus the one giving 9. Thus $\Pr[B_9 \geq \mu + 2\sigma] = \dfrac{10}{512} = 0.01953125$. Quite a bit less than 0.25.

OK, let's try $n = 100$. We have $\mu = 50$, $\sigma = 5$, so $2\sigma = 10$. We therefore want to know about $\Pr[B_{100} \geq 60]$. This looks like it has a lot more room to grow---after all, $60 \ldots 100$ is over 40% of the range $0 \ldots 100$. Using a widget such as https://stattrek.com/online-calculator/binomial.aspx, we get

$$B_{100}(\geq 60) = 0.02844396682$$

Bigger, but still nowhere near 0.25. A large chunk of this comes from the probability of getting exactly 60 heads in 100 tosses of a fair coin: $B_{100}(60) = 0.01084386671$, so $B_{100}(> 60) = 0.01760010011$. If we split the difference, we get

$$\frac{1}{2}B_{100}(60) + B_{100}(> 60) = 0.023022033465.$$

This is quite close to 2.3%. It is close to what we would get either way if we did $n = 10,000$, since it gives $\mu = 5000$, $\sigma^2 = 2500$, so $\sigma = 50$, so $\mu + 2\sigma = 5100$, and the individual value

$B_{10000}(5100) = 0.00107986433$, which leaves much less to "split". Putting it in or out, we have

$$B_{10000}(> 5100) = 0.02221289952, \qquad B_{10000}(\geq 5100) = 0.02329276385,$$

whose average is a little under 2.3%. The limit as $n \to \infty$ approaches $0.02275...$ Where did I get that from? From the convergence of the limit binomial distribution to the **Gaussian normal distribution**.

Thus the Chebychev bound is really weak for these distributions. It becomes far worse for $k = 3$ and $k = 4$ and higher. We can get far better bounds for these distributions. But ironically, the proof of these bounds comes from a judicious application of the "super-weak" Markov inequality...

Chernoff Bounds

The real essence of the binomial case is that whereas Chebychev's theorem gives only a $\sim \frac{1}{k^2}$ upper bound on the size of the "tail" of the distribution, for k standard deviations above the mean, the truth is exponential decrease with k. The binomial and normal distributions famously do not have closed-form formulas that involve only simple mathematical functions (counting exp and log as simple but factorial as not). The "art" is getting a simple upper bound that is strong *and* easy to manipulate. Our recent post

https://rjlipton.wordpress.com/2021/01/14/priming-random-restrictions/

shows a case where an inequality is weakened to make it "homomorphic" under certain transformations. Whether the original stronger formulas can be used to squeeze a useful drop of extra power might be food for "the next generation" but our intuition is often helped by the simpler version. The story of amplifying BPP shows several facets of this, but let's start with the general case of bounds on binomial/Gaussian tails that were popularized by Herman Chernoff but arguably discovered by someone else named Herman.

Theorem: Let $X_1, \ldots, X_N$ be 0-1 valued random variables that are independent but not neces-sarily identically distributed. Put $\mu = E[X_1] + \cdots + E[X_N]$ and $p = \frac{\mu}{N}$. Then for all $\delta > 0$:

$$\Pr[X_1 + \cdots + X_N \geq \mu + \delta\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} = \left(\frac{1}{e^{(p(\ln(1+\delta)+\delta\ln(1+\delta)-\delta))}}\right)^{N}$$

and

$$\Pr[X_1 + \cdots + X_N \leq \mu - \delta\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} = \left(\frac{1}{e^{(p(\ln(1-\delta)+\delta\ln(1-\delta)+\delta))}}\right)^{N}.$$

Before we prove this, some remarks:

1. The upper tail is expressed as $\mu + \delta\mu$ not $\mu + k\sigma$. This makes $\delta \leq \frac{(1-p)}{p}$ in the first statement, and of course $\delta \leq 1$ in the second statement.
2. In the i.i.d. case we get bounds on $B_{n,p}(\geq \mu + \delta\mu)$ and $B_{n,p}(\leq \mu - \delta\mu)$.
3. If we put $X_i' = 1 - X_i$ for each i, with mean $\mu' = N - \mu$, then $\left[\sum X_i \leq \mu - \delta\mu\right]$ is the same event as $\left[\sum X_i' \geq N - \mu + \delta\mu\right] \equiv \left[\sum X_i' \geq \mu' + \delta N - \delta\mu'\right]$. But this is not the same statement as the ' $\geq$ ' for above for X' and μ' unless $\mu = N - \mu = \frac{N}{2}$.
4. Thus the two statements above are not redundant by symmetry. Instead, there is a difference according to whether the deviation is "to the short side" of μ or "to the long side." But, each statement by itself has a "long side" and "short side" depending on whether $\mu < \frac{1}{2}$. So is it abstractly sufficient to prove just one of them? (??)
5. For the BPP amplification lemma, we will apply it to the long side crossing under $\frac{N}{2}$ in order to bound away the probability of a majority vote giving the wrong answer.
6. Using the identities $\ln(1+\delta) = \delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} + \cdots$ and $\ln(1-\delta) = -\delta + \frac{\delta^2}{2} - \frac{\delta^3}{3} + \cdots$ for $|\delta| < 1$ (and the former OK with $\delta = 1$ too), and bounding some constant times p by $\ln 2$, one can prove the following two-sided form:

Corollary: For every $\gamma > 0$, $\Pr[|X_1 + \cdots + X_N - \mu| \geq \gamma\mu] \leq \frac{1}{2^{0.5N\gamma^2}}$. $\boxtimes$

In any event, the asymptotic point is that the bound falls off exponentially with N. We have the freedom to make N as large a polynomial in the input size parameter n as we want. This will enable us to amplify strongly exponentially.

The proof of the theorem uses "only" Markov's inequality applied to the other kind of function besides a square that guarantees nonnegative values: an exponential-valued random variable. A key point is that if X and Y are independent, then the mean of their product is a product of means.

$$\mathbf{E}[X \cdot Y] = \sum_w w \cdot \Pr[X \cdot Y = w] = \sum_w \sum_{u,v:uv=w} w \cdot \Pr[X = u \wedge Y = v] .$$

Now applying independence, we get

$$= \sum_w \sum_{u,v:uv=w} u \cdot v \cdot \Pr[X = u] \cdot \Pr[Y = v] = \sum_{u,v} (u \cdot \Pr[X = u])(v \cdot \Pr[Y = v]) = \mathbf{E}[X] \cdot \mathbf{E}[Y] .$$

Proof: To prove the upper-tail inequality, let t be a positive parameter that we later optimize as $t = \ln(1+\delta)$,and consider the random variable $Y = e^{t(X_1 + \cdots + X_N)}$. The event $Y \geq e^{t(\mu+\delta\mu)}$ is the

same as the event $X_1 + \cdots + X_N \geq \mu + \delta\mu$ whose probability we wish to bound. By Markov's inequality, $\Pr\left[Y \geq e^{t(\mu+\delta\mu)}\right] \leq \dfrac{\mathbf{E}[Y]}{e^{t(\mu+\delta\mu)}}$. Now the random variables e^{tX_i} are likewise independent, so letting p_i be their probabilities of the value e^t, giving $(1-p_i)$ for value 1, we get

$$\mathbf{E}[Y] = \mathbf{E}\left[\prod_{i=1}^{N} e^{tX_i}\right] = \prod_{i=1}^{N} \mathbf{E}\left[e^{tX_i}\right] = \prod_{i=1}^{N} (1-p_i) + p_i e^t = \prod_{i=1}^{N} \left(1 + p_i\left(e^t - 1\right)\right) .$$

Now, IMHO incredibly, we throw away more pieces by using $1 + x \leq e^x$ to make a double-decker exponential, giving

$$\mathbf{E}[Y] \leq \prod_{i=1}^{N} e^{p_i\left(e^t-1\right)} = e^{\Sigma_{i=1}^{N} p_i\left(e^t-1\right)} = e^{\left(\Sigma_{i=1}^{N} p_i\right)\left(e^t-1\right)} = e^{\mu\left(e^t-1\right)} .$$

So for any t,

$$\Pr\left[\Sigma_{i=1}^{N} X_i \geq \mu + \delta\mu\right] = \Pr\left[Y \geq e^{t(\mu+\delta\mu)}\right] \leq \frac{e^{\mu\left(e^t-1\right)}}{e^{t(\mu+\delta\mu)}} = \left(e^{\mu\left(e^t-1\right)-\mu t(1+\delta)}\right) = \left(e^{e^t-t(1+\delta)-1}\right)^{\mu} .$$

The right-hand side is minimized when t is chosen to minimize $e^t - t(1+\delta)$, which is when $e^t - (1+\delta) = 0$, which is when $t = \ln(1+\delta)$. That choice makes $e^t - t(1+\delta) - 1$ become $1 + \delta - 1 - (1+\delta)\ln(1+\delta)$, so we get $\Pr[\cdots] \leq e^{\delta-(1+\delta)\ln(1+\delta)}$ and the bound follows. The other inequality for the "lower tail" is proved "similarly." ☒

Proof of the BPP Amplification Lemma

Amplification Lemma: If $A \in$ BPP with associated $R(x,y)$ and $p(n)$, then for any polynomial $q(n)$ we can build a polynomial-time decidable $R'(x,z)$ and associated polynomial $p'(n)$ such that for all x,

$$x \in A \implies \Pr_{|z|=p'(n)}[R'(x,z)] > 1 - 2^{-q(n)};$$
$$x \notin A \implies \Pr_{|z|=p'(n)}[R'(x,z)] < 2^{-q(n)}.$$

Moreover, we can achieve this even if the original R and p only give a "non-negligible" advantage, meaning that for some polynomial $r(n) \geq n$,

$$x \in A \implies \Pr_{|y|=p(n)}[R(x,y)] > \frac{1}{2} + \frac{1}{r(n)};$$
$$x \notin A \implies \Pr_{|y|=p(n)}[R(x,y)] < \frac{1}{2} - \frac{1}{r(n)}.$$

Proof: The algorithm, given any n and x of length n, is to do majority vote of N trials, where N depends on the polynomials $q(n)$ and $r(n)$ (and the overall time depends on $p(n)$ as well). For $1 \leq i \leq N$, let

$X_i = 1$ if the randomly-chosen string y_i in trial i makes $R(x, y_i) = A(x)$, $X_i = 0$ otherwise. Then the X_i are i.i.d. r.v.s all having $\mu_i = \frac{1}{2} + \frac{1}{r(n)}$, so μ for their sum X is just that multiplied by N. We want to bound $\Pr\left[X < \frac{N}{2}\right] = \Pr[X < \mu(1-\delta)]$ where we get δ by solving $\mu(1-\delta) = \frac{N}{2}$, so $\left(\frac{1}{2} + \frac{1}{r(n)}\right)(1-\delta) = \frac{1}{2}$, so $\frac{1}{2} - \delta\frac{1}{2} + \frac{1}{r(n)} - \delta\frac{1}{r(n)} = \frac{1}{2}$, so $\frac{1}{r(n)} = \delta\left(\frac{1}{2} + \frac{1}{r(n)}\right)$, so $\delta = \frac{2}{r(n)+2}$. It suffices to use the 2-sided estimate

$$\Pr[|X - \mu| \geq \delta\mu] \leq \frac{1}{2^{0.5N\delta^2}} = 2^{\frac{-2N}{(r(n)+2)^2}}.$$

We need to make $\frac{2N}{(r(n)+2)^2} \geq q(n)$, so take $N = q(n)(r(n)+2)^2/2$. This is a polynomial in n, so the majority vote condition gives a predicate R' decidable in $O\left(p(n)q(n)r(n)^2\right)$ time that meets the conditions of the Amplification lemma. ☒

The ability to take $q(n) \gg p(n)$ has some further ramifications that we will encounter. For immediate use, the lemma (and its more-obvious one-sided versions for RP and ZPP) simply allows us to be cavalier about the particular probability estimates involved for BPP. For example, we can show that the BP[·] operator is *idempotent* as follows:

Suppose L is in $\text{BP}_z[\text{BP}_y[R(x,y,z)]]$. Suppose we say that at least 9/10 of the z's give the property that at least 9/10 of the y's give the right answer $R(x,y,z) = L(x)$. The chance of being right both times is (**can we multiply?**) at least 81%, which is still bigger than the 75% in the original definition of BPP. So define $R'(x,w) =$ "w breaks in half into y, z such that $R(x,y,z)$ holds." Then R' gives a BPP predicate for L, so $L \in \text{BPP}$.

Chernoff Bounds in Practical Theory

How concretely good are the Chernoff bounds? Let's first revisit the case of unbiased binary distribution, where we had $B_{100}(60) = 0.01084386671$, $B_{100}(> 60) = 0.01760010011$, and

$$\frac{1}{2}B_{100}(60) + B_{100}(> 60) = 0.023022033465.$$

We have $\mu = 50$, $\delta = \frac{60-50}{50} = 0.2$, $e^\delta = 1.22140275816$, $1.2^{1.2} = 1.2445647472$, and finally

$$\left(\frac{1.22140275816...}{1.2445647472...}\right)^{50} = 0.98138948649...^{50} = 0.3909039476...$$

Not close! OK, let's try $N = 10,000$, $\mu = 5000$, $\mu + 2\sigma = 5100$, so $\delta = \frac{100}{5000} = 0.02$. We had:

$$B_{10000}(>5100) = 0.02221289952, \qquad B_{10000}(\geq 5100) = 0.02329276385,$$

Now $e^{0.02} = 1.0202013400$, $(1.02)^{1.02} = 1.02040405360$, ratio $r = 0.9998013398672$, and $r^{5000} = 0.37031571935$. Not much better! But if we keep δ the same, we get

$$B_{10000}(\geq 6000) \;=\; 8.7021582061 \times 10^{-90}$$

(from https://keisan.casio.com/exec/system/1180573199 this time) versus the Chernoff bound

$$0.98138948649...^{\,5000} \;=\; 1.61066648 \times 10^{-41}$$

Does this count as "close"? Well, for non-i.i.d. distributions, Chernoff bounds can be even worse. Suppose we have 5,000 variables X_i with the certain value 0 and 5,000 with the certain value 1. They count as "independent." We have $\mu = 5000$ again and that is all the Chernoff bounds care about. Yet the true probability of getting the sum to be 6000, or 5100, or even 5001, is zero.

Let's see a case where N is back down in the 100 to 200 range but δ is bigger and takes us up the "long side" across the middle from $\mu = 0.444N$ to almost $0.6N$. We take $N = 184$, so that we have on one hand the biased binomial distribution $B_{184,0.444}$. The tail is small:

$$B_{184,0.444}(\geq 110) \;=\; .0000196863761547 \;\approx\; \frac{1}{50796.55}$$

We have $\mu = 0.444 \times 184 = 81.696$, $1 + \delta \;=\; \dfrac{110}{81.696} \;= 1.3464551508$, $e^{0.3464551508} = 1.414046073577$, $1.3464551508^{1.3464551508} = 1.4926262961$, $r = 0.9473543895437$, and finally $r^{\mu} = 0.012054880$. Again, not close! Let's see how my chess program fares. This happens to be from an actual important cheating case this past week:

Name	ProjVal	St.Dev	Actual;	Proj% Actual%	2sigma range	z-score	BrierSc
AvgScaledDiffW	33.643	4.737:	18.148	0.1828: 0.0986	0.1314--0.2343,	z = +3.27, adj +3.27	0.072
ExpectationLossW	7.287	0.990:	3.882	0.0396: 0.0211	0.0288--0.0504,	z = +3.44, adj +3.04	0.003
MoveMatchWtd	81.699	6.108:	110.000	44.40%: 59.78%	37.76%--51.04%,	z = +4.63, adj +3.83	0.202
EqValueMatchW[4]	84.375	6.136:	111.000	45.86%: 60.33%	39.19%--52.53%,	z = +4.34, adj +3.74	0.198
						Combined: adj +3.92	

We will focus on the line for the move-match test. The random variable X_i for the i-th position among the 184 game turns that were tested is whether the player chose the move the computer listed first. Now chess positions are not i.i.d.: some positions have obvious recaptures where the player and computer will certainly agree, while others have many reasonable moves with each given a fairly close projected probability by my model. One issue is that empirically, the move decisions by the player are **not** independent, especially at consecutive turns, but this is a "sparse" dependence whose effect is reasonable to model by a constant-factor adjustment, with the constant determined empirically during model training. This is the "adj" column, but the original "z-score" column assumes independence, so it is what we should compare with $B_{184,0.444}(\geq 110)$.
So the number to compare with the binomial tail is $z = 4.63$. This is the "k" giving

$$\mu + \mu\delta \quad = \quad \mu + k\sigma$$

and the advantage is we can look up the corresponding probability directly under the approximation of the binomial distribution by the normal distribution, using an applet such as https://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html

Given z = 4.63

Calculate

The chance probability, Q, is: 0.000002 or about one in 547006

Whoa! My program gives odds over 10 times as sharp as the idealized distribution! The explanation is hinted by the presence of obvious-recapture moves. Consider if about half the projected agreements are obvious moves. Subtracting out 40 such cases leaves 70 agreements versus 41.7 projected out of 144 trials, for $B_{144,0.28958}(\geq 70) \approx 0.00000052069457 \approx \frac{1}{1920512}$. The actual distribution is somewhere between those cases. In any event, the empirically validated score that I reported on this test is $z = 3.83$, which gives

Given z = 3.83

Calculate

The chance probability, Q, is: 0.000064 or about one in 15608

This is in the same ballpark as the idealized binomial probability. The most important point is that from large tournaments played over-the-board, without a whiff of cheating, the distribution of these reported scores conforms (slightly conservatively on purpose) to the bell curve.

Thus, Chernoff bounds are not close enough to be relevant for my concrete statistical work. Whether one could make any real profit from closer bounds in asymptotic complexity is regarded as dubious, but who knows?