7 Lecture 7, October 15, 2014

7.1 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:

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

and apply

$$\begin{aligned} & \Pr((jh \le X < (j+1)h) \wedge (j'h \le Y < (j'+1)h)) \\ = \quad & \Pr(jh \le X < (j+1)h) \cdot \Pr(j'h \le Y < (j'+1)h) \end{aligned}$$

for independent X, Y.

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

The Chernoff bound[4] 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 will be the seed of more general and stronger bounds that we will see.

It will be convenient to use the rvs $Y_i = 2X_i - 1$, where X_i is the indicator rvs the ith event. This shift lets us work with mean-0 rvs. This (as any function that is applied to an rv) leaves the Y_i independent.

Theorem 31 *Let $Y_1, \ldots, Y_n$ be iid rvs, with $\Pr(Y_i = -1) = \Pr(Y_i = 1) = 1/2$. Let $Y = \sum_1^n Y_i$. Then $\Pr(Y > \lambda\sqrt{n}) < e^{-\lambda^2/2}$ for any $\lambda > 0$.*

Proof: Fix any $\alpha > 0$. *Exercise:*[5]

$$E(e^{\alpha Y_i}) = \cosh \alpha \le e^{\alpha^2/2}.$$

By independence of the rvs $e^{\alpha Y_i}$,

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

$$\begin{aligned} \Pr(S > \lambda\sqrt{n}) &= \Pr(e^{\alpha S} > e^{\alpha\lambda\sqrt{n}}) \\ &\le E(e^{\alpha S})/e^{\alpha\lambda\sqrt{n}} \qquad \text{Markov ineq.} \\ &\le e^{n\alpha^2/2 - \alpha\lambda\sqrt{n}} \end{aligned}$$

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

$$\Pr(S > \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

□

[4] First due to Bernstein [8, 9, 7] but we follow the standard naming convention in Computer Science.

[5] For $k \ge 0$, $(2k)! = \prod_1^k i(k+i) \ge 2^k k!$, so for any real x, $e^{x^2/2} = \sum_{k \ge 0} x^{2k}/(2^k k!) \ge \sum_{k \ge 0} x^{2k}/(2k)! = \cosh x$.

7.3 Application: set discrepancy

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

Theorem 32 (Spencer) *With the definitions above, there is a function χ of discrepancy $Disc(\chi) \in O(\sqrt{n})$.*

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

Theorem 33 *With the definitions above, a function χ selected u.a.r. has $Disc(\chi) \in O(\sqrt{n\log n})$ with positive probability.*

Proof: By Theorem (31), for any particular set S_j (noting that $|S_j| \le n$),

$$\begin{aligned}
\Pr(|\chi(S_j)| > c\sqrt{n\log n}) &= \Pr(|\chi(S_j)| > \frac{c\sqrt{n\log n}}{\sqrt{|S_j|}}\sqrt{|S_j|}) \\
&\le 2e^{\frac{-c^2 n\log n}{2|S_j|}} \\
&\le 2e^{-\frac{c^2\log n}{2}} \\
&= 2n^{-c^2/2}.
\end{aligned}$$

Now take a union bound over the sets.

$$\begin{aligned}
\Pr(\exists j : |\chi(S_j)| > c\sqrt{n\log n}) &\le n\Pr(|\chi(S_j)| > c\sqrt{n\log n}) \\
&< 2n^{1-c^2/2}.
\end{aligned}$$

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

7.4 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 $\ge 2/3$, and strings not in the language should be accepted with probability $\le 1/3$. 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 34 *The entropy (base 2) of a probability distribution $\{p_1,\ldots,p_n\}$ is $h_2(p) = \sum p_i \lg \frac{1}{p_i}$.*

In natural units we use $h(p) = \sum p_i \log \frac{1}{p_i}$.

Definition 35 *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 s to r," or "of r w.r.t. s," is defined by*

$$D_2(r\|s) = \sum_i r_i \lg \frac{r_i}{s_i}$$

This is also known as information divergence, directed divergence or relative entropy[6]. In natural log units the divergence is $D(r\|s) = \sum_i r_i \log \frac{r_i}{s_i}$, 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:

$$\begin{array}{ll} \text{(a)} & D(r\|s) \geq 0 \ \ \forall r, s \\ \text{(b)} & D(r\|s) = 0 \Rightarrow r = s \\ \text{(c)} & D(s+\varepsilon\|s) = \sum_i \left(\frac{\varepsilon_i^2}{2s_i} + O(\varepsilon_i^3) \right) \end{array}$$

The "$\|$" notation is strange but is the convention.

When s is the uniform distribution, we have:

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

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: $h_2(p) = p \lg \frac{1}{p} + (1-p) \lg \frac{1}{1-p}$, $D_2(p\|q) = p \lg \frac{p}{q} + (1-p) \lg \frac{1-p}{1-q}$.

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.

Theorem 36 *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 31 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 1):

$$D(p\|1/2) \geq (2p-1)^2/2 \tag{17}$$

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 [13, 58]. 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 $\alpha > 0$,

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

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

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

[6] D is useful throughout information theory and statistics (and is closely related to "Fisher information"). See [13].

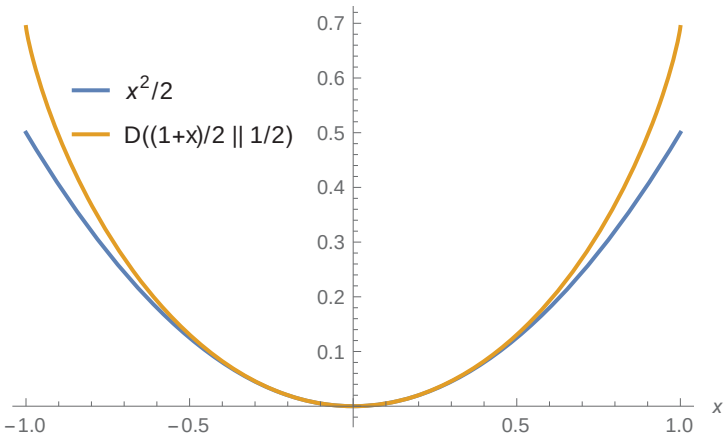

Figure 1: Comparing the two Chernoff bounds at $q = 1/2$

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

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 $p = q + 1/\text{poly}(n)$. We have *(Exercise)* $D((p+q)/2\|q), D((p+q)/2\|p) \geq 1/\text{poly}'(n)$ (for some other polynomial poly′). So if we perform $n \cdot \text{poly}'(n)$ repetitions of the original BPP algorithm, and accept x if the fraction of "Yes" votes is above $(p+q)/2$, then the probability of error on any input is bounded by $\exp(-n)$.

7.5 Shannon's coding theorem: a preview

This is an exceptionally important application. 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.