15-359: Probability and Computation

Assignment 7 Due: March 30, 2012

Problem 1: Spin me right round (20 pts.) Let X and Y be independent standard normals.

A. Write the joint PDF of (X, Y).

B. Show, via a change of variables, that the joint PDF's value at (x, y) depends only on the (Euclidean) distance r of (x, y) from the origin. (This property is called rotational symmetry, and it's really useful.)

C. Prove that $Z = X + Y$ has distribution Normal$(0, 2)$ via a geometric argument. Hint: $Z = (X, Y) \cdot (1, 1)$.

Problem 2: Unwieldy numbers (15 pts.) As of this morning, Twitter processes around 340 million tweets a day. Let's assume for simplicity that the tweets are independent, and each consists of a uniformly random number of characters between 10 and 140. Using the CLT, approximate the probability that Twitter processes between 25 billion and 26 billion characters per day.

Problem 3: Exp farming (15 pts.) Let $X \sim \text{Exp}(\lambda)$. Derive the kth moment of X.

Problem 4: Random amount of randomness (15 pts.)

1. Let $S = X_1 + X_2 + \cdots + X_{10}$ where the X_i's are i.i.d. continuous r.v.'s and $\widetilde{X}(s)$ is the Laplace transform of X. What is $\widetilde{S}(s)$?

2. Now let
$$S = X_1 + X_2 + \cdots + X_N$$
where the X_i's are i.i.d. continuous r.v.'s, and where N is a discrete random variable, where $N \perp X_i$, $\forall i$. Let $\widehat{N}(z)$ be the z-transform of N, and let $\widetilde{X}(s)$ be the Laplace transform of X_i. Prove that:
$$\widetilde{S}(s) = \widehat{N}\left(\widetilde{X}(s)\right)$$

3. Let $N \sim$ Geometric(p). Let $X_i \sim \text{Exp}(\mu)$. Let $S = \sum_{i=1}^{N} X_i$. What can you say about the distribution of S?

Problem 5: **Mouse in a maze** (20 pts.)

A mouse is trapped in a maze. The mouse is equally likely to turn left or right. If he turns left, then he will walk for Exp(1) time and return to his starting position. If he turns right, then with probability $\frac{1}{3}$, he will leave the maze after Exp(2) time and with probability $\frac{2}{3}$, he will walk for Exp(3) time before returning to his starting position. Let T be the time until the mouse leaves the maze.

1. What is $\mathrm{E}(T)$?
2. What is $\mathrm{Var}(T)$?
3. Derive $\widetilde{T}(s)$?
4. Check that your answer makes sense by deriving $\mathrm{E}(T)$ from the transform.

Problem 6: **Chernoff proof relents** (15 pts.)

It's possible to prove a Chernoff-like inequality without Markov's inequality or Z/Laplace transforms. Conceptually, the inequality says: the probability that a size n sample from distribution p looks like a sample from distribution q decreases exponentially in n and the distance between p and q.

Wait...distance between distributions? The KL-divergence (or 'relative entropy') measures distance (sort of) between two Bernoulli distributions with parameters $q, p \in (0, 1)$:

$$\mathrm{KL}(q, p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}$$

This function is a special case of a KL-divergence defined between two arbitrary distributions, which, in general, isn't symmetric in its arguments and doesn't satisfy the triangle inequality, hence the 'sort of'. The following plot shows $\mathrm{KL}(q, p)$ is reasonable.

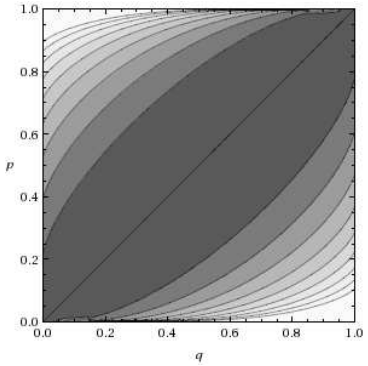

Figure 1: Lighter contours have higher values.

Suppose a coin of bias $p \in (0,1)$ is tossed n times. Prove the probability of qn heads or more, for $q > p$, is at most $\exp(-n\mathrm{KL}(q,p))$.

Problem 7: Polling is cheap. (Tell that to a systems programmer.) (10 pts.)

Suppose we ask n uniformly random chosen people if they approve or disapprove of the president. (We choose people with replacement, so it's possible, though very unlikely, that we ask the same person twice.) We publish the fraction d/n of people who disapprove, along with the guarantee "this poll is accurate to within $\pm 2\%$, 19 times out of 20." How big does n have to be to provide this guarantee?

Problem 8: Shifty columns (20 pts.)

Suppose A is an n by n binary matrix. Define the *max-col-sum of* A to be the maximum of all the column sums in A; in symbols $\mathsf{mcs}(A)$. For example, for

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

we have $\mathsf{mcs}(A) = 6$ because of the first column.

Now suppose we can rotate the rows of the matrix, independently of each other. Clearly, any such operation can be described by a shift vector $s = (s_1, s_2, \ldots, s_n)$ where $0 \le s_i < n$. Write $A[s]$ for the matrix obtained by rotating the rows according to shift vector s.

Our goal is to find a shift vector s such that $\mathsf{mcs}(A[s])$ is minimal; write $\mathsf{mmcs}(A)$ for this minimum value. For the example above we can achieve the minimal value of 3 by rotating according to $s = (0,0,0,1,4,2)$ (altogether there are 4416 shift vectors that work).

Obviously, we do not wish to try out all n^n shift vectors, so instead we use a randomized algorithm RandShift:

- Pick a shift vector $s \in [0, n-1]$ uniformly at random.
- Return $A[s]$.

That's it!

Task

Let A be an n by n binary matrix and $C = \mathsf{mmcs}(A)$ its min max col sum. Write R for the total number of 1's in A and call A *sparse* if $R \le n \ln n$. Define X to be the random variable: sum of the first column in the matrix B produced by RandShift.

A. Show that the exptected value of X is optimal.

B. Assume that A is sparse. Use a Chernoff bound to show that $\mathsf{mcs}(B) \geq C + 4\ln n$ with probability at most $1/n^2$.

C. Assume that A is not sparse. Use a Chernoff bound to show that $\mathsf{mcs}(B) \geq 5C$ with probability at most $1/n^2$.

In both cases the simplified bounds suffice.

Problem 9: BBB (15 pts.) Concentration of measure occurs not only for sums, but also for other smooth functions of independent random variables. Suppose:

- $X_1, \ldots, X_n$ are random variables taking values in some abstract space $\mathcal{X}$,
- $F : \mathcal{X}^n \to \mathbb{R}$ satisfies a 'Lipschitz' condition:

$$\sup_{x_1,\ldots,x_n,x_i'} |F(x_1,\ldots,x_n) - F(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)| \leq c_i$$

Then the 'bounded differences' inequalities holds:

$$P\left(F(X_1,\ldots,X_n) - E(F(X_1,\ldots,X_n)) > \epsilon\right) \leq \exp\left(\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$P\left(F(X_1,\ldots,X_n) - E(F(X_1,\ldots,X_n)) < \epsilon\right) \leq \exp\left(\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

Task

A. Show that the bounded differences inequality implies Hoeffding's inequality.

B. Remember balls and bins? (You may still be a little sore.) We threw n balls independently at random into m bins. For each $i \in [m]$, consider the indicator random variable

$$Z_i = \begin{cases} 1 & \text{if bin } i \text{ is empty,} \\ 0 & \text{otherwise} \end{cases}$$

So the number of empty bins is $Z = \sum_i Z_i$. Prove an upper bound on $P(|Z - E(Z)| > \epsilon)$.