Probability and Algorithms **Caltech CS150b, Fall 2019**
Leonard J. Schulman, schulman@caltech.edu **TA: Jenish Mehta, jenishc@gmail.com**
Problem set 1 **Out Friday Oct. 4. Due Friday Oct. 18 in Jenish's mailbox.**

1. In 2018-ps3-5 I assigned the following problem. If you were in that course feel free to resubmit your old solution, if it was correct.

   You are trying to count sheep that are walking past you in a very, very long line. You are a shepherd of very little brain: you don't even have a memory of size $\lg n$, which is what you would need to count n sheep. (Not to be handed in: argue that any deterministic counting algorithm requires this much space.)

   Instead, you come up with the following mechanism whose goal is to estimate the number of sheep within a constant factor, using memory only $O(\lg \lg n)$.

   Initialize $C := 0$.

   After a sheep walks by, flip a biased coin X, $\Pr(X = 1) = 2^{-C}$ (otherwise $X = 0$).

   Set $C := C + X$.

   Denote by $C(n)$ the random variable after n sheep have walked by. Show for any value of n, that $2^{C(n)}$ probably approximates n within a constant factor. More specifically, show that $\forall a > 0\ \exists b > 0$ s.t. with probability $\geq 1 - a$, $bn \leq 2^{C(n)} \leq n/b$.

   Also, suppose you do not have access to coins of arbitrary bias but only to a fair coin. Can you still solve the problem within the required memory limitation?

   Hint: One way to approach this is to turn things around and imagine there is an infinite list of sheep, and let $N(c)$ be the index of the first sheep to bring the register to c. E.g., for sure $N(1) = 1$. Show that $\forall a > 0\ \exists b > 0$ s.t. with probability $\geq 1 - a$, $b2^c \leq N(c) \leq 2^c/b$. Finally, get the quantification right: argue what we asked for any fixed number of sheep n. (You might have to pay a little in a.)

   Comment: For the upper bound on on $C(n)$ it helps if you show that $E(2^{C(n)}) = n + 1$. For the lower bound, there are two strategies: one is to calculate $\mathrm{Var}(2^{C(n)})$, and the other is to calculate $E(N(c))$.

2. Your job as a shepherd is not over. Yesterday you counted them so you know there are n. (For this exercise let's suppose you have the number exactly.) However, sheep are not all the same! – your flock is composed of sheep of up to n types. Being an expert shepherd, you can tell the type of a sheep at a glance. Let m_i ($1 \leq i \leq n$) be the (unknown) number of sheep of type i. Today, as the sheep walk past you in a long line, you want to use very little memory, yet when all the sheep have gone past, be able (with high probability) to produce a good estimate of

$$F_2 = \sum_1^n m_i^2.$$

   So the input to this "streaming" problem is a sequence $a_1, \ldots, a_n$ where each $a_j \in [n]$; here $m_i = |\{j : a_j = i\}|$.

   Give a randomized algorithm which for any given $\lambda, \delta > 0$ uses $O(\frac{\log 1/\delta}{\lambda^2} \log n)$ bits of memory, and outputs an estimate R such that $\Pr(|R - F_2| > \lambda F_2) \leq \delta$.

   Hint: To solve this problem you'll want to use four closely related things from CS150a. Feel free to look these up in the course notes (posted on my homepage) and to use them without proof. The first is the notion of k-wise (and in particular for this problem 4-wise) independent random variables. The second is that for sums of 4-wise independent random bits, we have (a) an upper tail bound (the probability that absolute value of the sum is large, is small) (actually this requires only 2-wise independence); (b) a lower tail bound (the probability that absolute value of the sum is small, is small). You can look up these bounds in the course notes. Third is that there exist, and we also can very efficiently construct, small 4-wise independent sample spaces: specifically, there is a set of $O(n^2)$ vectors in $(\pm 1)^n$ such that if we fix any four positions $b_1 < b_2 < b_3 < b_4$, and sample uniformly from the set a vector $z = (z_1, \ldots, z_n)$, then the probability distribution on the restriction of z to those four positions, is uniform. (I.e., uniform on the 16 possible vectors.) And fourth, having randomly chosen the $O(\log n)$-length bit string r which picks a particular z from this set of vectors, we can, given any $i \in [n]$, compute the bit z_i in space $O(\log n)$.

With all these ingredients in place, the basis of your method can be this: pick such a z. Now as the sheep go by, compute $Z = \sum_{i=1}^{n} z_{a_i}$. When they're all past, set $X = Z^2$. Show that $E(X) = F_2$. Bound $\mathrm{Var}(X)$.

For the full solution, compute several such estimators in parallel, and use the "averaging and then median" method. What this means is: generate enough independent samples $X_1, \ldots, X_m$ so that $Y = \frac{1}{m} X_i$ has at least $2/3$ of its probability mass in the desired interval. Then, take the median of ℓ such independent "Y" estimators, and apply a Chernoff bound to the event that the median of these Y's does not fall in the desired interval.

(N.B.: The previous problem was essentially that of estimating F_1, where $F_k = \sum m_i^k$. The question has been studied for all integer $k \geq 0$.)

3. The simplest and most universal method of upper bounding the probability of the union of n events is to use the sum of the individual probabilities.

   However, this can be a wasteful upper bound, particularly if the events have strong positive correlation. One typical case in which this is occurs is in time series. Here is an example.

   In this exercise we'll see that symmetric random walk is typically near its maximum value. To be specific: let $X_1, \ldots$ be iid, uniform in ± 1. Let $S_t = \sum_1^t X_i$. Let $M_t = \max\{S_1, \ldots, S_t\}$. For $a > 0$, clearly $\Pr(M_t > a) \geq \Pr(S_t > a)$. Show that $\Pr(M(t) \geq a) \leq 2 \Pr(S_t \geq a)$.

   Hint: define a *stopping time* of a time series such as $A_1, \ldots$ to be a variable T taking values in the positive integers, such that for all t, $A_1, \ldots, A_t$ determine whether $T = t$.

   Now show that after any stopping time of the series S_t, the distribution of the remainder of the walk (i.e., of $S_{T+1}, \ldots$) is again symmetric.

   Apply this to the stopping time $\tau_a = \min\{t : S_t = a\}$.

4. Prove the following *Theorem:* It is possible to place n points in the unit square in the plane in such a manner that all $\binom{n}{3}$ of the triangles whose corners are at three of the points, have area at least $\Omega(1/n^2)$.

   Hint: Use the deletion method.