CS 49 Lecture Notes

Sungil Ahn

2017 Nov. 1

1 Counting Probabilistically

If we wanted to count the number of items in a stream (i.e. F_0) of n items, we would need $c = \log_2 n$ bits to store it exactly. If we allow approximation to the count, we can count c instead to keep track of the number of bits that make up n. Then, we would need a total of $O(\log c) = O(\log \log n)$ bits. However, we cannot track c deterministically! Thus, we need a way to approximate c without using too much space.

1.1 Morris Counter

Morris Counter is one such way of counting c probabilistically. The algorithm is as follows:

1. Set $c = 0$
2. When an item arrives, increment c with probability 2^{-c}
3. Return $z = 2^c - 1$

Note that the mechanism to increment c can be implemented with $O(\log c)$ space plus a random bit (used as a coin), since we can toss the coin c times and increment c only when all of them land head.

1.1.1 Analysis

Let $z(n)$ be the state of the counter after n items arrived. Deterministically,

$$z(0) = 0 \text{ and } z(1) = 1 \tag{1}$$

Since we're incrementing c with a coin toss,

$$[z(n+1)|c(n) = j] = \begin{cases} 2^j & \text{with probability} 1 - \frac{1}{2^j} \\ 2^{j+1} & \text{with probability} \frac{1}{2^j} \end{cases}$$

Therefore, $\mathbb{E}[z(n+1)|c(n) = j] = 2^j \cdot (1 - 1/2^j) + 2^{j+1} \cdot (1/2^j) = 2^j + 1$ and

$$\begin{aligned} \mathbb{E}[z(n+1)] &= \sum_j P[c(n) = j]\mathbb{E}[z(n+1)|c(n) = j] \\ &= P[c(n) = j](2^j + 1) \\ &= 1 + \mathbb{E}[2^{c(n)}] \\ &= 1 + \mathbb{E}[z(n)] \end{aligned} \tag{2}$$

Then, iterating over the values of n,

$$\begin{aligned}\mathbb{E}[z(n)] &= 1 + \mathbb{E}[z(n-1)] \\ &= 1 + (1 + \mathbb{E}[z(n-2)]) \\ &\cdots \\ &= n - 1 + \mathbb{E}[z(1)] \\ &= n \ (\because (1))\end{aligned} \tag{3}$$

Therefore, the counter is an unbiased estimator of n! Similarly,

$$\begin{aligned}\mathbb{E}[z(n+1)^2] &= \sum_j P[c(n) = j]\mathbb{E}[z(n+1)^2 | c(n) = j] \\ &= P[c(n) = j](2^{2(j+1)}(1/2^j) + 2^{2j}(1 - 1/2^j)) \\ &= P[c(n) = j](2^{2j} - 2^j + 2^{j+2}) \\ &= P[c(n) = j](2^{2j} + 3 \cdot 2^j) \\ &= \mathbb{E}[2^{2c(n)}] + 3\mathbb{E}[2^{c(n)}] \\ &= \mathbb{E}[z(n)^2] + 3n\end{aligned} \tag{4}$$

Then, by iterating, we get

$$\begin{aligned}\mathbb{E}[z(n)^2] &= 3(n-1) + \mathbb{E}[z(n-1)^2] \\ &= 3(n-1) + 3(n-2) + \mathbb{E}[z(n-2)^2] \\ &= \cdots \\ &= 3(1 + 2 + \cdots + (n-1)) + \mathbb{E}[z(0)^2] \\ &= 3n(n-1)/2 + 1\end{aligned} \tag{5}$$

By Chebyshev's inequality, we can get the multiplicative error bound as follows:

$$\begin{aligned}P[|z(n)/n| > \epsilon] &\leq \frac{Var[z(n)]}{\epsilon^2 \mathbb{E}[z(n)]^2} \\ &\leq \frac{\mathbb{E}[z(n)^2]}{\epsilon^2 \mathbb{E}[z(n)]^2} \\ &= \frac{3n(n-1)/2 + 1}{\epsilon^2 n^2} \\ &= O(1/\epsilon^2)\end{aligned} \tag{6}$$

Question. Can we improve the error bound (sub-quadratic in ϵ^{-1})? How much will keeping around *multiple* counters of c help?

2 Counting Distinct Elements

Suppose we want to count $d =$ number of *distinct* items from a stream of m elements where $d >> 1$. The idea here is that we can use the maximum number of leading zeros (x) to approximate the cardinality ($\approx 2^x$).

Algorithm

1. Let $L = \lceil \log_2 m \rceil$, and create an array of counters $z[0 : L]$.
2. Randomly pick $h : [m]- > [N]$ from Pairwise Independent hash family, where $N = 2^x - 1$ for some $x >> 1$.
3. When an item $a \in 1, \ldots, m$ arrives, evaluate pos_a = number of trailing 0's in the binary representation of $h(a)$ (i.e. largest j such that $2^j | h(a)$).
4. Then, $z[pos_a] += 1$
5. After the stream has passed, return $\hat{d} = 2^k$, where k is the largest number with $z[k] > 0$.

Properties

1. Let $X_{a,j} = 1$ if 2^j is the largest power of 2 that divides $h(a)$, else $X_{a,j} = 0$. Then, $P[X_{a,j}] = 1] = 1/2^{j+1}$, since $h(a)$ would need to have $1000 \cdots 0$ (j 0's) as the last $j+1$ digits in its binary representation. Note that this is generally not true if N is small.
2. Let $Y_j = \sum_{a:distinct} X_{a,j}$ (i.e. the number of distinct elements that increment counter j). Then, $\mathbb{E}[Y_j] = d/2^{j+1}$ by linearity of expectation. This means that $Y_j \cdot 2^{j+1}$ is an unbiased of estimator of d $\forall j$! However, this does not mean that $Y_j = z[j]$ in general. In fact, since $z[j]$ can be incremented by duplicate items, $Y_j \leq z[j]$ $\forall j$.
3. $Y_j = 0$ iff $z[j] = 0$, since you can't have either of the values being greater than 0 without at least 1 (distinct) element incrementing them.
4. $Var[Y_j] \leq \mathbb{E}[Y_j]$ $\forall j$, since the variation of [sum of pairwise independent indicator variables] (in this case, Y_j) is always less than or equal to the expected value of the sum (this is true in general).

With these properties, we can bound the error for $\hat{d}$:

Theorem. With probability 5/8, $d/16 \leq \hat{d} \leq 16d$.

Proof. Let l be integer with $2^l < d \le 2^{l+1}$ and c some arbitrary constant. Then,

$$\begin{aligned}
P[\exists j \ge l+c : z[j] > 0] &\le \sum_{j \ge l+c} P[z[j] > 0](\because \text{union bound}) \\
&= \sum_{j \ge l+c} P[Y_j \ge 1](\because \text{property 3}) \\
&\le \sum_{j \ge l+c} \mathbb{E}[Y_j](\because \text{Markov's Inequality}) \\
&= \sum_{j \ge l+c} d/2^{j+1} \\
&\le \sum_{j \ge l+c} 2^{l+1}/2^{j+1}(\because d \le 2^{l+1}) \\
&= 1/2^c \sum_{j \ge l+c} 1/2^{j-(l+c)} \\
&\le 1/2^{c-1}
\end{aligned} \tag{7}$$

Thus, $P[\mathbb{E}j \ge l+4 : z[j] > 0] \le 1/8$ (we simply plugged in $c = 4$ to the above inequality). Hence, with probability at least 7/8, $\hat{d} = 2^k \le 2^{l+4} = 16d$. This satisfies the right side of the theorem's inequality. Similarly,

$$\begin{aligned}
P[z[l-c] = 0] &= P[Y_{l-c} = 0] \\
&\le P[|Y_{l-c} - \mathbb{E}[Y_{l-c}]| \ge \mathbb{E}[Y_{l-c}]] \\
&\le Var[Y_{l-c}]/\mathbb{E}[Y_{l-c}]^2(\because \text{Chebyshev}) \\
&\le \mathbb{E}[Y_{l-c}]/\mathbb{E}[Y_{l-c}]^2 = 1/\mathbb{E}[Y_{l-c}] \\
&= 2^{l-c+1}/d \\
&< 2^{l-c+1}/2^l(\because d > 2^l \text{ from our definition of } l) \\
&= 1/2^{c-1}
\end{aligned} \tag{8}$$

Then, plugging in $c = 3$ to the above inequality gives us $P[z[l-3] = 0] \le 1/4$. However, $z[l-3] = 0$ means $k < l-3$, since k is supposed to be the biggest number that has nonempty slot in z. Thus, the negation of the inequality asserts that with probability at least 3/4, $k \ge l-3$, i.e. $\hat{d} = 2^k \ge 2^{l-3} = 2^{l+1}/2^4 \ge d/16$. This satisfies the left side of the theorem's inequality. Thus, with probability $1 - 1/4 - 1/8 = 5/8$ (these probabilities are the probability that either of the theorem's inequalities fail), $d/16 \le \hat{d} \le 16d$. □

Question. Can we tighten the bound on $\hat{d}$ by keeping around *multiple* instances of z? If not, how else could we improve the bound?