CS5120: Probability & Computing **Jan-May 2020**

Lectures 13,14

AMS Algorithm for Counting Distinct Elements

Lecturer: N.R.Aravind *Scribe: N.R.Aravind*

1 The AMS algorithm

The following is the algorithm by Alon, Matias, Szegedy in 1996, to estimate the number of distinct elements.

1. Choose a random hash function $h : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, m^3\}$ from a pairwise independent hash family.
2. Initialize $z = 0$.
3. For each item x of the stream, update z as $z = Max(zeroes(h(x)), z)$, where $zeroes(y)$ denotes the number of trailing zeroes of y in the binary representation
4. Output 2^{z+c} for $c = 1/2$.

Analysis of space complexity: The hash function is of the form $h(x) = ax + b$, where $a, b \in O(m^3)$, thus the space needed to store the pair (a, b) is $O(\log m)$. Clearly the space needed to store the value of z and the final output are also in $O(\log m)$; thus the total space used is $O(\log m)$.

2 Analysis of correctness

2.1 Preserving distinctness

Let d denote the number of distinct elements in the stream. We first show that with probability at least $1 - 1/2m$, there are no collisions by the hash function, so that the number of distinct hashed values is also d.

The probability that $h(x) = h(y)$ for two distinct elements x, y is equal to $\frac{1}{m^3}$ since h is from a pairwise independent family. The number of pairs is

$\binom{m}{2} < m^2/2$, thus the probability that some pair collides, is by the union bound, at most $\frac{m^2}{2m^3} = \frac{1}{2m}$.

From now on, we condition on the event that the hashed values are all distinct.

2.2 Approximation and error guarantees

We now prove that the estimate is an approximation (although not a very good one).

Proposition 1 $Pr(2^{z+c} \geq 3d) \leq 0.472$ *and* $Pr(2^{z+c} \leq d/3) \leq 0.472$.

Proof of Proposition 1 We first show the following claim.

Claim 1

$$Pr(2^z \geq 2^r) \leq \frac{d}{2^r}.$$

and

$$Pr(2^z < 2^r) \leq \frac{2^r}{d}.$$

We first prove the proposition assuming the claim. We have:

$$Pr(2^{z+c} \geq 3d) = Pr(2^z \geq \frac{3d}{2^c}) = Pr(2^z \geq \frac{3d}{2^c}) \leq \frac{2^c}{3}$$

where we used Claim 1 for the inequality.

We have:

$$Pr(2^{z+c} \leq d/3) = Pr(2^{z+c} < 2d/3) = Pr(2^z < \frac{d}{3 \cdot 2^{c-1}}) \leq \frac{1}{3 \cdot 2^{c-1}}$$

where we used Claim 1 for the inequality.

Finally, substituting $c = 1/2$, we obtain the probability bounds to be $\frac{\sqrt{2}}{3} < 0.472$.

We now prove Claim 1. Let $L(r)$ denote the number of r-length trailing zeroes in the set $\{h(x)|x \text{ is in the stream}\}$. Then we have $E[L(r)] = \frac{d}{2^r}$.

Note that the event $2^z \geq 2^r$ is equivalent to $z \geq r$, which is equivalent to $L(r) \geq 1$. Thus we obtain the first part of the claim from Markov's inequality.

For the second part, we note that $2^z < 2^r$ is equivalent to $z < r$, which is equivalent to $L(r) = 0$. Now, by applying Proposition 2 (see last section), we obtain the second part of the claim.

This completes the proof of the proposition. ■

2.3 Reducing the error

Since the two probabilities of error (exceeding $3d$ and being less than $d/3$) are each less than $1/2$, they may be reduced by the median-of-means method each to less than $\delta/2$, by using $O\left(\log(\frac{1}{\delta})\right)$ copies of z in parallel. The total error would then be less than δ and the total space used is $O(\log(\frac{1}{\delta})\log m)$. We note however that the approximation guarantee remains unchanged, and is not an arbitarily-close approximation.

3 Chebyshev's Inequality: A useful special case

The following proposition follows from Chebyshev's inequality and the fact that if X is a sum of $0/1$ random variables, then $Var[X] \leq E[X]$.

Proposition 2 *If X is a sum of pairwise independent random variables taking values in $\{0,1\}$, with expectation μ, then:*

$$Pr(|X - \mu| \geq \varepsilon\mu) \leq \frac{1}{\varepsilon^2\mu}.$$

In particular, $Pr[X = 0] \leq \frac{1}{\mu}$.