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

Lecture 14:
The BJKST Algorithm for Counting Distinct Elements

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

1 The BJKST algorithm

The following is the algorithm by Bar-Yossef, Jayram, Kumar, Sivakumar and Trevisan, in 2002, to estimate the number of distinct elements.

1. Choose a random hash function $h : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, M = m^3\}$ from a pairwise independent hash family.
2. Maintain and update the k smallest elements of the hashed values of the stream seen so far.
3. Let z_k be the kth smallest element. Output $\frac{k(M+1)}{z_k}$ as the estimate.

The intuition behind this algorithm is the following: If we pick d random elements independently in $[0, M]$, and z_k is the kth smallest element, then $E[z_k] = \frac{kM}{d+1}$. Thus, $\frac{kM}{z}$ should be a good estimate for d. For technical reasons, seen in the analysis, there is a small modification in the value output.

We have replaced the set $[0, M]$ by the discrete set $\{1, 2, \ldots, M\}$, and instead of complete independence, we have only pairwise independence. However, we'll still be able to prove that the estimate is good.

2 Analysis of the BJKST algorithm

The space used by the algorithm is $O(k \log m)$. We will choose $k = \frac{96}{\varepsilon^2}$ so that the space used is $O(\frac{1}{\varepsilon^2} \log m)$.

We now argue about the approximation and error gurantee. Let $M = m^3$.

Claim 1 *We have:*

$$Pr\left(|\frac{k(M+1)}{z_k} - d| > \varepsilon d\right) \leq \frac{1}{3}.$$

Proof of Claim: The event $\frac{k(M+1)}{z_k} > d + \varepsilon d$ is equivalent to:
$z_k < \frac{k(M+1)}{(1+\varepsilon)d}$, that is: $z_k < \frac{k(M+1)}{d}(1 - \frac{\varepsilon}{1+\varepsilon})$, which by Corollary 2, has probability at most $\frac{8(1+\varepsilon)^2}{\varepsilon^2 k} \leq \frac{32}{\varepsilon^2 k}$. Substituting for k and noting that the bound is for the total deviation of z_k, we get the desired upper bound of $\frac{1}{3}$. ■

Thus, with probability at least 2/3, the algorithm returns an ε-approximation. The success probability can be boosted in the standard way.

3 The kth smallest element

Proposition 1 *Let $X_1, X_2, \ldots, X_d$ be uniformly random elements of $[0, M]$, that are pairwise independent Let $Y_1, Y_2, \ldots, Y_d$ be the elements in sorted order. Then, for $0 \leq \varepsilon \leq 1$, we have:*

$$Pr\left(|Y_k - \frac{kM}{d}| > \varepsilon \frac{kM}{d}\right) < \frac{2}{\varepsilon^2 k}.$$

We prove the proposition in the form of the following two claims.

Claim 2

$$Pr\left(Y_k < (1-\varepsilon)\frac{kM}{d}\right) < \frac{(1-\varepsilon)}{\varepsilon^2 k}.$$

Claim 3

$$Pr\left(Y_k > (1+\varepsilon)\frac{kM}{d}\right) < \frac{(1+\varepsilon)}{\varepsilon^2 k}.$$

Proof of Claim 1: Let $Z_i = 1$ if $X_i < (1-\varepsilon)\frac{kM}{d}$ and $Z_i = 0$ otherwise. Let $Z = Z_1 + \ldots + Z_d$. We have $E[Z_i] = \frac{(1-\varepsilon)k}{d}$ and hence $E[Z] = (1-\varepsilon)k$.

Thus, by Proposition 3,

$$Pr[Z \geq k] \leq Pr(|Z - E[Z]| \geq \varepsilon k) \leq (\frac{\varepsilon}{1-\varepsilon})^{-2}(1-\varepsilon)^{-1}k^{-1}.$$

Proof of Claim 2: Let $Z_i = 1$ if $X_i > (1+\varepsilon)\frac{kM}{d}$ and $Z_i = 0$ otherwise. Let $Z = Z_1 + \ldots + Z_d$. We have $E[Z_i] = \frac{(1+\varepsilon)k}{d}$ and hence $E[Z] = (1+\varepsilon)k$. Thus, by Proposition 3,

$$Pr[Z \leq k] \leq Pr(|Z - E[Z]| \geq \varepsilon k) \leq (\frac{\varepsilon}{1+\varepsilon})^{-2}(1+\varepsilon)^{-1}k^{-1}.$$

This completes the proofs of the two claims, and hence of Proposition 1.

Note that in proposition 1, we considered the uniform distribution on the real interval $[1, M]$. We can however use the above result to obtain a similar one for the discrete set $\{1, \ldots, M\}$.

Corollary 2 *Let $X_1, X_2, \ldots, X_d$ be uniformly random elements of $\{1, \ldots, M\}$, that are pairwise independent, with $M \geq d$. Let $Y_1, Y_2, \ldots, Y_d$ be the elements in sorted order, and let $\varepsilon \in [0, 1]$. Then*

$$Pr\left(|Y_k - \frac{k(M+1)}{d}| > \varepsilon\frac{k(M+1)}{d}\right) < \frac{8}{\varepsilon^2 k}.$$

To obtain the corollary, we can model the discrete variables as $\lfloor X \rfloor$, where X is u.a.r. from $[1, M+1]$. The difference in $E[Z_i]$ between the two cases is at most $\frac{1}{M+1}$, and we can argue that $E[Z]$ in Claim 2 is at least $(1 + \frac{\varepsilon}{2})k$ and at most $(1 + \frac{3\varepsilon}{2})k$. These approximations should comfortably give the bound in the corollary.

4 Chebyshev's Inequality: A useful special case

We restate the following inequality, also seen in the analysis of the AMS algorithm.

Proposition 3 *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, $Pr[X = 0] \leq \dfrac{1}{\mu}$.