1 Lecture 6

Frequency Moments and Counting Distinct Elements

Definition 1.1. Suppose that we receive m objects (not necessarily distinct) that are members of the set $[n]$. This set of m objects will, as they appear in a stream, be labeled σ; we have $B << m$ bits to process σ. Let g be a real valued non-negative function that operates on streams.

Definition 1.2. We say that a streaming algorithm $\mathcal{A}$ provides (ϵ, δ) additive approximation if

$$\mathbb{P}\left[|\mathcal{A} - g| > \epsilon\right] < \delta$$

J

Definition 1.3. We say that a streaming algorithm $\mathcal{A}$ provides (ϵ, δ) relative approximation if

$$\mathbb{P}\left[\frac{\mathcal{A}(\delta)}{g(\delta)} - 1 > \epsilon\right] < \delta$$

Remark 1.4. Many streaming problems are trivial to solve deterministically, if we have full access to the data that enters a stream. To solve them deterministically, however, there are lower bounds on the space and running time of working algorithms. These bounds are often unacceptable. To achieve useful space and running times, thus, streaming algorithms often end up using randomized algorithms.

Remark 1.5. The task of counting how many distinct elements appear in a stream has obvious, if costly, deterministic solutions: Put all objects in a binary search tree; when a new stream object appears, search the tree for the object and, if not found, add it to the tree.

This requires $O(k)$ space and $O(m \log k)$ running time where k is the true number of distinct objects. Even if you were told in advance that there would be k objects, a hashing solution would require $O(k)$ space and $O(m)$ running time.

Remark 1.6. A cute idea is to assume the existence of a hash function $h : [n] \to [n^3]$ that closely "resembles" the theoretical, non-existent hash function that maps all objects with perfect randomness to some real in $[0, 1]$. Then keep track of the random varible $X = \min_{e_i \in \text{ all objects}} h(e_i)$. The minimum statistic (for k uniform random variables over $[0, 1]$) has the property that $\mathbb{E}\left[[]\, X\right] = \frac{1}{k+1}$. Thus, at the end of streaming, take X, invert it and subtract 1 from it to obtain a statistical estimator of the number of distinct objects.

Convince yourself that the following integral represents the expectation.

$$\mathbb{E}\left[[]\, X\right] = \int_0^1 \binom{k}{1} (1-x)^{k-1}$$