11^{TH} TUTORIAL ON RANDOMIZED ALGORITHMS

Streaming algorithms: KMV and Count-Min sketches

1. *Even simpler count distinct: K Minimum Values (KMV) sketch.* We would like to count the number of distinct elements, i.e., estimate set cardinality. We look at a different approach (which is actually quite popular in practice): For a parameter k and a hash function $h : [N] \to (0, 1]$, store the k smallest hash values of the distinct stream elements, i.e., we store k pairs (item j, $h(j)$). When queried for cardinality, return $(k-1)/v_k$, where v_k is the k-th smallest hash value (the largest one stored).

a) Analyze the algorithm assuming h is fully random and prove that given $\varepsilon \in (0, 1)$, for $k \geq c/\varepsilon^2$ (where c is a large enough constant) the algorithm gives an ε-approximation of F_0 = the number of distinct elements with constant probability. Focus on bounding the probability of $(k-1)/v_k > (1+\varepsilon) \cdot F_0$; the other inequality is similar.

b) What is wrong with h being fully random? What kind of hash functions would be sufficient for the analysis?

2. *Count-Min sketch for frequency estimation.* We would like to estimate frequencies and find heavy hitters under both insertions and deletions (similarly as CountSketch but with a different guarantee). We will assume that all frequencies are non-negative at the end. We use the following sketch for estimating frequencies f_i (screenshot from lecture notes by A. Chakrabarti):

Algorithm 9 Count-Min Sketch

Initialize:

$C[1\dots t][1\dots k] \leftarrow \vec{0}$, where $k := 2/\varepsilon$ and $t := \lceil \log(1/\delta) \rceil$

Choose t independent hash functions $h_1, \dots h_t : [n] \to [k]$, each from a 2-universal family

Process (token (j,c)) **:**

for $i \leftarrow 1$ **to** t **do**

 $C[i][h_i(j)] \leftarrow C[i][h_i(j)] + c$

Output (query a) **:**

report $\hat{f}_a = \min_{1 \le i \le t} C[i][h_i(a)]$

a) Using the assumption that all frequencies are non-negative at the end, derive lower and upper bounds on the estimator of a single row. That is, for any $a \in [n]$ and row $i \in [t]$ show that

$$\Big| f_a - C[i][h_i(a)] \Big| \le \varepsilon \cdot \|\mathbf{f}\|_1 \,. \tag{1}$$

with a constant probability.

b) Show a high probability bound for the final estimator $\hat{a}_j$ for frequency f_a.

c) Compare CountSketch (from the lecture) and Count-Min sketch, both in terms of their description and their properties.

d) Can you derive a more refined bound on the error of Count-Min? That is, replace $\|\mathbf{f}\|_1$ by a smaller quantity in (1).

e) Count-Min is a linear sketch, that is, it can be viewed as a linear map of the frequency vector $\mathbf{f}$ to a much smaller dimension. What are the properties of the matrix of this linear map?