Weekplan: Streaming II.

Philip Bille Inge Li Gørtz Eva Rotenberg

References and Reading

[1] Amit Chakrabarti: *Data Stream Algorithms* 2011 (revised 2015) chapter 2.

[2] Kurt Mehlhorn and He Sun: *Streaming Algorithms* 2014.

[3] Jelani Nelson: *Algorithms for Big Data, lecture 3* 2015 section 2.1

[4] P. Flajolet: Approximate Counting: A Detailed Analysis

[5] J. S. Vitter: Random Sampling with a Reservoir

We recommend reading the specified chapters and sections of [1] and [3] in detail. The notes in [2] cover the same material as [1] but in other words.

Hash function cheat-sheet

The notation $[x]$ Throughout this sheet I will use the notation $[x]$ to denote the set $\{0, 1, 2, \ldots, x-1\}$.

Definition: Hash function A hash function $h : U \rightarrow [m]$ is a random variable in the class of all functions $U \rightarrow [m]$.

Definition: 2-independent Also known as *strongly universal* or *pairwise independent*.
A hash function $h : U \rightarrow [m]$ is 2-independent if for all $x \neq y \in U$ and $q, r \in [m]$: $P[h(x) = q \land h(y) = r] = \frac{1}{m^2}$.
Equivalently, the following two conditions hold:

- for any $x \in U$, $h(x)$ is uniform in $[m]$,
- for any $x \neq y \in U$, $h(x)$ and $h(y)$ are independent.

Exercises

The following exercises relate to chapter 2 in [1].

1 Sanity check Hash functions sometimes have collisions. Here, we choose our family of hash functions carefully to avoid collisions. Would collisions lead to overestimating or underestimating the number of distinct elements?

2 Suppose h is a 2-independent hash function from $[n]$ to $[n^3]$. Show that h is injective with probability at least $1 - \frac{1}{n}$.

3 Solve exercises 2-1 and 2-2 from the book.

4 We have seen an algorithm to estimate the number of distinct elements in astream. Equivalently it estimates the number of non-zero frequencies. Adapt the idea to estimate the number of frequencies that are odd

5 Analyse performance of the algorithm. The purpose of this exercise is to walk you through the proof in Section 2.3 of [1].

5.1 Describe the indicator variables $X_{r,j}$ and Y_r in your own words.

5.2 Calculate the expected value of $X_{r,j}$ and of Y_r. (How) Does the expected value of $X_{r,j}$ depend on j? (You can assume $h(j_1)$ and $h(j_2)$ are independent for any j_1, j_2.)

5.3 Bound the variance of Y_r.

5.4 Bound the probability of Y_r being > 0.

5.5 Bound the probability of Y_r being $= 0$.

5.6 Now, if $\hat{d} \geq 3d$, then our variable z must equal some value a with $2^{a+1/2} \geq 3d$.

Thus, we can rewrite $P[\hat{d} \geq 3d]$ to the form $P[Y_a > 0]$ for such an a.

Use this to bound $P[\hat{d} \geq 3d]$.

6 Similarly, ... In the exercise above we went through the proof of $P[\hat{d} \geq 3d] \leq \frac{\sqrt{2}}{3}$. Prove $P[\hat{d} \leq d/3] \leq \frac{\sqrt{2}}{3}$.