Weekplan: Streaming II.

Philip Bille Inge Li Gørtz Christian Wulff-Nilsen

References and Reading

[1] Amit Chakrabarti: *Data Stream Algorithms* 2011 (updated July 2020) Chapters 2.0–2.3 and 4.0–4.3.

Hash function cheat-sheet

The notation $[x]$ Throughout this sheet we let $[x] = \{1, 2, \ldots, x\}$.

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

Definition: 2-universal Also known as *strongly universal* or *pairwise independent*.
A hash function $h : U \to [n]$ is 2-*universal* if for all $x \neq y \in U$ and $q, r \in [n]$

$$\mathrm{P}[h(x) = q \wedge h(y) = r] = \frac{1}{n^2}.$$

Equivalently, the following two conditions hold:

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

Exercises

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

2 The Tidemark algorithm. The purpose of the following exercises is to walk you through part of the proof in Section 2.3 of [1]. The slides contain all solutions (and will be covered in the lecture) so try to avoid looking at them when solving the exercises.

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

2.2 Calculate $E[X_{r,j}]$ and $E[Y_r]$. You may use the fact that h is uniform. Does $E[X_{r,j}]$ depend on j, on r, or both?

2.3 Show that $\mathrm{P}[Y_r \geq 1] \leq \frac{d}{2^r}$.

2.4 Show that for any random variable X, $\mathrm{Var}[X] \leq E[X^2]$.

2.5 Show that $\mathrm{Var}[Y_r] \leq \frac{d}{2^r}$. You may use linearity of variance (this is applicable since the $X_{r,j}$-variables are 2-independent).

2.6 Use Chebyshev to show that $\mathrm{P}[Y_r = 0] \leq \frac{2^r}{d}$.

3 Counting rare elements[1] Paul goes fishing. There are u different fish species $U = \{1, \ldots, u\}$. Paul catches one fish at a time. Let a_t be the fish species he catches at time t. Let $c_t[j] = |\{a_i | a_i = j, i \leq t\}|$ be the number of times he catches a fish of species j up to time t. Species j is *rare* at time t if it appears precisely once in his catch up to time t. The rarity $\rho[t]$ of his catch at time t is defined as:

$$\rho(t) = \frac{\#\text{rare species}}{u} .$$

3.1 Explain how Paul can calculate $\rho(t)$ precisely, using $2u + \log m$ bits of space.

3.2 However, Paul wants to store only as many bits as will fit his tiny suitcase, i.e., $o(u)$, preferably $O(1)$ bits. Therefore, Paul picks k random fish species each independently, randomly with probability $1/u$ at the beginning and maintains the number of times each of these fish species appear in his bounty, as he catches fish one after another. Paul outputs the estimate

$$\hat{\rho}(t) = \frac{\#\text{rare species in the sample}}{k} .$$

Let $c_1(t), \ldots, c_k(t)$ be the value of the counters at time t. Show that $P[\hat{\rho}(t) \geq 3\rho] \leq 1/3$.

Hint: Calculate first $P[c_i(t) = 1]$.

4 Approximate Counting Solve exercise 4-1 from [1].

5 2-universal Hash Families Solve exercise 2-1 from [1].

[1] This exercise is from Muthukrishnan "Data Streams: Algorithms and Applications