Exercise Sheet for the Lecture
Intelligent Learning and Information Systems

Summer Term 2019

3 - Data Streams III

hand in: May 10, 2019 (2pm)
solutions will be discussed: May 10, 2019

Group I: Lukas Drexler (drexler@informatik.uni-bonn.de)
Group II: Tobias Elvermann (t.elvermann@gmail.com)
Group III: Maximilian Thiessen (s6mnthie@uni-bonn.de)

1. Count-Min Sketch for Heavy Hitters [25 points]

Using the method presented on Slides 7–11 of Lecture 2019–05–03, compute the ϕ-heavy hitters for $\phi = 1/3$ for the following data stream:

$$\sigma = \langle 2, 1, 2, 2, 3, 3, 6, 6, 2, 5 \rangle \ .$$

In your solution, set the parameters of the Count-Min Sketch algorithm to $d = 3$ and $w = 3$. Applying the method on Slide 13 of Lecture 2019–05–03 with $p = 3$ and $k = 2$, use the hash functions $h_{(1,2),1}$, $h_{(2,1),0}$, and $h_{(2,1),1}$.
Give the state of all count-min sketches (i.e., the matrices) after the processing of the 1st, 2nd, 5th, and the last item. Give also all steps of the evaluation of the binary tree for the entire data stream (cf. Slide 11 of Lecture 2019–05–03).

2. The Median Trick [25 points]

Prove Claim 4 on Slides 27–28 of Lecture 2019–05–03 for $s_1 = c \log \frac{1}{\delta}$ for some appropriate constant c.

Remark: The claim is stated for $c = 2$ on slides 28; the point is that $s_1 = O\left(\log \frac{1}{\delta}\right)$.

Hint: To prove the claim, you may use one of the Chernoff bounds stated below. In particular, one can show it for $c = 2$ by using (i).

Thm.: Let $X_1, \ldots, X_n$ be independent Poisson trials with $\Pr(X_i) = p_i$ (i.e., $\Pr(X_i = 1) = p_i$ and $\Pr(X_i = 0) = 1 - p_i$ for all $i \in [n]$). Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then the following bounds hold:

(i) For any $\gamma > 0$,

$$\Pr[X \geq (1+\gamma)\mu] < \left(\frac{e^{\gamma}}{(1+\gamma)^{(1+\gamma)}}\right)^{\mu}$$

(ii) For any $\gamma \in (0,1)$,

$$\Pr[|X-\mu| \geq \gamma\mu] \leq 2e^{-\mu\gamma^2/3}$$

3. The AMS Sketch [10 points]

Generalize the AMS Sketch given on Slide 29 of Lecture 2019–05–03 to the *cash register* model, i.e., in which each token in the data stream over $[n]$ is of the form (i, Δ), where $i \in [n]$ and $\Delta > 0$ is some integer. For a data stream $\sigma = \langle (a_1, \Delta_1), \ldots, (a_m, \Delta_m) \rangle$ over $[n]$ in the cash register model, we define the frequency f_i of i by

$$f_i = \sum_{l \in [m], a_l = i} \Delta_l$$

for all $i \in [n]$. Give the pseudocode of the generalized AMS Sketch for estimating F_k ($k \geq 1$). Sketch the proof of the correctness of your algorithm. In your proof, you may rely on the proof of the vanilla case given on Slides 21–30 of Lecture 2019–05–03, i.e., explain only what is new with respect to that proof.