Homework 9

Due: 27 November, 2020

Each problem is graded on the coarse scale of $\checkmark^+$, $\checkmark$, $\checkmark^-$ and no $\checkmark$. It is also assigned a multiplier, denoting the relative importance of the problem. Both correctness and presentation are grading criteria.

Please read and make sure you understand the collaboration policy on the course missive. The problems for 2000-level credit are clearly marked below: students who are registered for 1000-level credit are welcome to solve and submit these problems for extra credit (see course missive for details).

Remember to prove all your (non-elementary) mathematical claims, unless stated otherwise.

Each pair of students should submit only 1 pdf to the corresponding Canvas assignment.

Do 2 out of the 3 ✓ problems as well as Problem 4 for regular credit, and complete all problems for extra credit.

Problem 1

(3 ✓s)

Write out and analyse the "count-median" sketch that is an adaptation of the count-min sketch, but instead can work in the turnstile model, sacrificing the one-sidedness of the approximation error.

Problem 2

(3 ✓s)

Definition 1 (ℓ_1-Heavy Hitters)**.** Given a stream σ with frequency vector $\mathbf{f}$, the ℓ_1-heavy hitters problem with parameter k is to return a list of elements L such that

- If $f_i > ||\mathbf{f}||_1/k$, then $i \in L$.
- $|L| = O(k)$.

Note that there are at most $k-1$ heavy hitters, and we require that the returned list is no more than a constant factor longer than necessary.

State and analyse a streaming algorithm, which, by maintaining the count-min sketch of a stream (possibly with sub-constant failure probability), can answer a single ℓ_1-heavy hitters query with parameter k in $\tilde{O}(n)$ time (and polylogarithmic space in n and m), with a success probability of at least 2/3.

Problem 3

(3 ✓s)

Recall that for two relations (i.e. tables in a database) $r(A, B)$ and $s(A, C)$, with a common attribute A, we define the join $r \bowtie s$ to be a relation consisting of all tuples (a, b, c) such that $(a, b) \in r$ and $(a, c) \in s$. Therefore, if $f_{r,j}$ and $f_{s,j}$ denote the frequencies of j in the common attribute (i.e. A) of r and s respectively, then the size of the join is $\sum_j f_{r,j} f_{s,j}$.

As mentioned in class, the second frequency moment F_2 of a stream can be interpreted as the size of a self-join, if the stream consists of records in a database, and the frequency f_i denotes the number of occurrences of element i in the join attribute.

Show that, using the same tug-of-war sketching algorithm, we can also estimate the size of the join of two arbitrary databases represented by streams σ_r and σ_s. For simplicity, you may assume that the streams have 2nd frequency moment (namely their squared Euclidean norms are the same). Explain how to compute such an estimate, and state and prove a theorem capturing the accuracy and space complexity guarantees of your (1-pass) streaming algorithm.

Problem 4

(2 ✓s)

This is a non-technical problem. Throughout the semester, we have seen a number of models and algorithms, and even some lower bounds. What did you learn from the course? What were the biggest punchlines to you? Try to focus more on the big picture and less on each technical algorithm: how (if at all) has the course changed the way you think about computation? What kind of design/analysis/problem formulation principles have you learnt?

This problem is graded for completion (with reasonable effort) sonly. There is no right or wrong answer. It is a useful exercise to do though, for this course and really every course you take.

Also, feel free to discuss with your classmates (in addition to your homework partner)!