Homework 2
Due 2015-Oct-21 by 11:59pm

Prof. Amit Chakrabarti
Computer Science Department
Dartmouth College

General Instructions: Please write concisely, but rigorously, and show your calculations explicitly, as we do in class. Each problem is worth 7 points. You can find a detailed description of my grading principles under the 7-point scale on the course website.

Please submit your homework electronically (by email). For hand-written solutions, please scan and email. The big scanner/copier in the Sudikoff front office can be useful for this.

Honor Principle: You are allowed to discuss the problems and exchange solution ideas with your classmates. But when you write up any solutions for submission, you must work alone. You may refer to any textbook you like, including online ones. However, you may not refer to published or online solutions to the specific problems on the homework. *If in doubt, ask the professor for clarification!*

Start Early! This homework is harder than the previous one. Start early, or else you won't be able to do it justice.

---

5. Analyze the Count/Median variant of the Count-Min Sketch to prove that it estimates frequencies f_a by $\widehat{f_a}$, satisfying

$$\Pr\left[\widehat{f_a} - f_a \notin [-\varepsilon\|\mathbf{f}_{-a}\|_1, \varepsilon\|\mathbf{f}_{-a}\|_1]\right] \leq \delta\,,$$

using space $O\left(\frac{1}{\varepsilon}\log\frac{1}{\delta}\cdot(\log m + \log n)\right)$. The notation $\mathbf{f}_{-a}$ is explained in the lecture notes.

6. Consider a stream σ in the turnstile model, defining a frequency vector $\mathbf{f} \geq \mathbf{0}$. The Count-Min Sketch solves the problem of estimating f_j, given j, but does not directly give us a *quick* way to identify, e.g., the set of elements with frequency greater than some threshold. Fix this.

In greater detail: Let α be a constant with $0 < \alpha < 1$. We would like to maintain a suitable summary of the stream (some enhanced version of the Count-Min Sketch, perhaps?) so that we can, on demand, quickly produce a set $S \subseteq [n]$ satisfying the following properties w.h.p.: (1) S contains every j such that $f_j \geq \alpha F_1$; (2) S does not contain any j such that $f_j < (\alpha - \varepsilon)F_1$. Here, $F_1 = F_1(\sigma) = \|\mathbf{f}\|_1$. Design a data stream algorithm that achieves this. Your space usage, as well as the time taken to process each token and to produce the set S, should be polynomial in the usual parameters, $\log m$, $\log n$, and $1/\varepsilon$, and may depend arbitrarily on α.

Hint: Suppose you combined all tokens in $\{1, 2, \ldots, n/2\}$ into one "supertoken", and similarly for all the other tokens in $\{n/2+1, n/2+2, \ldots, n\}$. Now, if you estimated the frequency of one of these supertokens to be less than αF_1, then you have eliminated $n/2$ candidates from S in one shot.

7. To a scientist interested in Databases, the second frequency moment F_2 can be motivated by observing that it is the size of a self-join: the join of a relation in a database *with itself*. In fact, one can design a sketch that can scan a relation in one pass (i.e., in streaming fashion) such that, based on the sketches of two *different* relations, we can estimate the size of their join. Explain how.

Recall that for two relations (i.e., tables in a database) $r(A, B)$ and $s(A, C)$, with a common attribute (i.e., column) 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 first columns (i.e., "A"-columns) of r and s, respectively, and j can take values in $[n]$, then the size of the join is $\sum_{j=1}^{n} f_{r,j} f_{s,j}$.

Hint: You have already seen the sketch in class! But you have to prove that it solves the above problem, for some reasonable definition of a "good estimate."

8. Let σ be a data stream in the turnstile model, giving rise to a frequency vector $\mathbf{f} \in \mathbb{R}^n$. Recall that $F_k(\mathbf{f}) = \sum_{j=1}^{n} |f_j|^k$. By convention, $F_\infty(\mathbf{f}) = \max_{1 \leq j \leq n} |f_j|$. Consider the following outline of an algorithm for estimating F_k, where $k > 2$ is a constant:

- Obtain a good estimate $\widehat{F_2}$ of $F_2(\mathbf{f})$.
- In parallel, draw a good ℓ_2-sample $(j, \widehat{f_j})$ from $\mathbf{f}$.
- Output $\widehat{F_2}|\widehat{f_j}|^{k-2}$.

By adding details to this outline, and an analysis, prove that we can obtain an (ε, δ)-approximation to $F_k(\mathbf{f})$ using space

$$\widetilde{O}\left(\frac{1}{\varepsilon^2}\log\frac{1}{\delta}\cdot n^{1-2/k}\right),$$

where, as usual, the $\widetilde{O}$-notation suppresses $(\log n)^{O(1)}$ factors.

In greater detail: Let the random variable T denote the output of this algorithm. Work out $\mathbb{E}[T]$ and a good bound on $\mathrm{Var}[T]$. For the latter bound, after some algebra, you should have an expression involving F_2F_{2k-2}. Use either convexity (calculus!) or Hölder's inequality to prove that $F_2F_{2k-2} \le n^{1-2/k}F_k^2$. Then, based on the resulting bound on $\mathrm{Var}[T]$, work out (1) how good an estimate you need in the first two steps of the above description of the algorithm, and (2) with what parameters you need to use the median-of-means improvement to get an (ε, δ)-approximation.