Streams, Sketching and Big Data – Exercises

Graham Cormode
G.Cormode@warwick.ac.uk

July 11, 2014

1 Lecture 1: Sketches and Frequency Moments

1. Suppose we have arrival and departure streams where the frequencies of items are allowed to be negative. Extend the Count-Min sketch analysis to estimate these frequencies (note, the Markov argument no longer works)

2. The lectures showed that the inner product of two vectors, $x \cdot y$, can be approximated (up to additive error $\epsilon\|x\|_2\|y\|_2$) by sketch manipulations. A more direct way to do this is to compute the inner product of the sketches. Show that the AMS sketch yields an unbiased estimator for $x \cdot y$, and analyze the variance of the estimator to bound the additive error.

3. The hashing-based algorithms for F_0 estimation work for streams that consist of arrivals only. It is of interest to approximate F_0 for other models.
   (a) Design an algorithm to approximate F_0 over a stream of arrivals and departures.
   (b) Modify your algorithm algorithm to find the number of distinct elements among the most recent W arrivals

2 Lecture 2: Advanced Topics

4. (Graph sketching) Design a graph sketch to sketch a set of graph edges so that given a subset of nodes S we can approximate $\mathrm{cut}(S)$, the number of edges in $E \cap (S \times (V \setminus S))$.

5. (Linear Algebra) The method described for compressed matrix multiplication yields a sketch so that $(AB)_{ij}$ can be approximated with additive error $\epsilon\|AB\|_F^2$. Modify or build on the construction of this sketch to allow an efficient search for all entries of (AB) that are at least $\phi\|AB\|_F^2$ in magnitude.

6. (Verification) Suppose you are shown a stream that defines an $n \times n$ matrix A, and an n-dimensional vector x, followed by an n-dimensional vector y. Design a scheme to verify $Ax = y$. What is the memory needed by the verifier? Can you obtain a protocol where the space is, say, $O(\sqrt{n})$ if the prover provides a larger proof?

7. (Lower bounds) Use reductions to DISJ or INDEX to show the hardness of:
   (a) Frequent items: find all items in the stream whose frequency is greater than ϕN, for some $0 < \phi < 1$.
   (b) Sliding window: given a stream of binary (0/1) values, compute the sum of the last N values
   (c) Rank sum: Given a stream of (x, y) pairs and query (p, q) specified after stream, approximate $|(x, y)|x < p, y < q|$