CS 226 / 6.889 Sketching Algorithms for Big Data — Fall 2017

Problem Set 2

Due: 11:59pm, Monday, October 23rd

Submit to: `sketchingbigdata-f17-assignments@seas.harvard.edu`

See homework policy at `http://www.sketchingbigdata.org/fall17/syllabus/`

Problem 1: Point queries in insertions-only streams. (10 points)

Recall the algorithm for finding heavy hitters in insertions-only stream mentioned in Lecture 7. This algorithm, due to Misra and Gries, proceeds as follows. Let C be a parameter, defining the number of counters maintained by the algorithm. The algorithm maintains a set T of up to C elements, and for every $i \in T$ maintains a counter c_i. Initially $T = \emptyset$. Then, for each stream element i:

- If $i \in T$, then $c_i = c_i + 1$
- Else
  - If $|T| < C$, then $T = T \cup \{i\}$ and $c_i = 1$
  - Else for all $j \in T$ do $c_j = c_j - 1$
- Remove from T all items i such that $c_i = 0$

In Lecture 7, we have seen that some algorithms do more than just find heavy hitters. In particular, we have seen that Count-Min provides ℓ_1 point query guarantees, i.e., for each element i, it provides an estimate of the i-th frequency up to an error term of the form $\frac{1}{k}\|x_{tail(k)}\|_1$, where x is the frequency vector and $tail(k)$ denotes all coordinates of x except for the k largest ones. When the stream contains a small number of very frequent elements, $\|x_{tail(k)}\|_1$ can be much smaller than $\|x\|_1$.

Show that Misra-Gries algorithm also provides ℓ_1 point query guarantees, and specifically, that each c_i estimates x_i up to an additive error of

$$\frac{\|x_{tail(k)}\|_1}{C-k}$$

where we assume $c_i = 0$ for $i \notin T$.

Problem 2: RIP and incoherent matrices.

(a) (3 points) Let A be an arbitrary real, square matrix. Show that if $\lambda \in \mathbb{C}$ is an eigenvalue of A, then there exists an i so that $|\lambda - A_{i,i}| \leq \sum_{j \neq i} |a_{i,j}|$.

(b) (5 points) Show that any (ε/k)-incoherent matrix Π satisfies (ε, k)-RIP.

(c) (2 points) Conclude that for any $\varepsilon \in (0, 1/2)$ and integer $1 \le k \le n$ with n a prime, there is is an explicit (ε, k)-RIP matrix $\Pi \in \mathbb{R}^{m \times n}$ with $m = O((k^2/\varepsilon^2) \cdot poly(\lg n))$. By explicit, we mean there is a deterministic algorithm which, given ε, k, n for $1 \le k \le n$, returns Π as a 2-dimensional array in time $poly(n)$. You may assume that $\varepsilon > 1/\sqrt{n}$ since otherwise the $1/\varepsilon^2$ term in m is already at least n, and there's of course already an simple-to-describe $n \times n$ $(0, k)$-RIP matrix (the $n \times n$ identity matrix).

Problem 3: Chaining with limited independence. Recall in class the toy example of a random walk on a line: at time t we are at position $\langle \sigma, x^{(t)} \rangle$, where $x^{(t)} = \sum_{i=1}^{t} e_i$ and $\sigma \in \{-1, -1\}^n$ is chosen uniformly at random. We defined $v^{(t)} = x^{(t)}/\|x^{(n)}\|_2$ and showed via Dudley's inequality that

$$\mathbb{E} \sup_{1 \le t \le n} \langle \sigma, v^{(t)} \rangle = O(1). \tag{1}$$

Our goal is to show that this holds even if the σ_i are not fully independent, but rather simply p-wise independent for some $p = O(1)$.

(a) (2 points) Show that for integer $p > 1$, $\mathbb{E}_\sigma \langle \sigma, x \rangle^p$ is completely determined by p-wise independence.

(b) (2 points) In "bound 1" of lecture (the "union bound" argument), we showed that if σ has independent entries and T is an arbitrary set of vectors in $\mathbb{R}^n$, then $r(T) := \mathbb{E} \sup_{x \in T} \langle \sigma, x \rangle \lesssim (\lg^{1/2} |T|) \cdot \rho_{\ell_2}(T)$. Here $\rho_X(T)$ denotes $\sup_{x \in T} \|x\|_X$. It turns out that an equivalent form of Khintchine's inequality is that for all $p \ge 1$, $\|\langle \sigma, x \rangle\|_p \le \sqrt{p} \cdot \|x\|_2$ (you may use this fact without proof). Show that as long as the σ_i are p-wise independent for some $p \ge 2$, then $r(T) \lesssim \sqrt{p} \cdot |T|^{1/p} \cdot \rho_{\ell_2}(T)$ (note we can then recover the $\lg^{1/2} |T| \cdot \rho_{\ell_2}(T)$ bound under full independence by setting $p = \Theta(\lg |T|)$).

(c) (6 points) Conclude via a modified Dudley chaining argument that Eqn. (1) holds even if the σ_i are only p-wise independent for some constant $p = O(1)$. (We know how to show it for $p = 4$ — how small of a p can you get to work?)

(d) (4 points) Show that if $(x^{(t)})_t$ is not simply the sequence with $x^{(t)} = \sum_{i=1}^{t} e_i$, but rather is an arbitrary sequence of vectors generated by an insertion-only stream, then Eqn. (1) still holds.

Problem 4: Sparse Johnson-Lindenstrauss analysis is tight. (8 points)

Show that the analysis in class of the Sparse Johnson-Lindenstrauss Transform is tight. Specifically, consider the distribution over $\Pi \in \mathbb{R}^{m \times n}$ in which the columns are independent, and in each column there are exactly s non-zero entries chosen at uniformly random locations without replacement, and each one is independently $\pm 1/\sqrt{s}$. Show that for all $\varepsilon, \delta \in (0, 1/2)$ and constant $C > 0$ there exists constants $c, n_0 > 0$ such that for such Π with $m \le C\varepsilon^{-2} \log(1/\delta)$ and $s \le c\varepsilon^{-1} \log(1/\delta)$, for all $n > n_0$ there exists $x \in \mathbb{R}^n$ such that

$$\mathbb{P}_\Pi(|\|\Pi x\|_2^2 - 1| > \varepsilon) > \delta.$$

That is, there is a significant probability of failure unless s is large (larger than $c\varepsilon^{-1}\log(1/\delta)$). **Hint:** consider x having its first t coordinates equaling $1/\sqrt{t}$ and the rest 0 and choose t appropriately.

Bonus: (3 points — we do not know the answer!) can considering x being a random vector on the sphere also give a tight lower bound?

Problem 5: (1 point) How much time did you spend on this problem set? If you can remember the breakdown, please report this per problem. (sum of time spent solving problem and typing up your solution)