Exercise IV, Sublinear Algorithms for Big Data Analysis 2024-2025

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students, and solve as many problems as you can. Problems marked (*) are more difficult, but also more rewarding. These problems have been taken from various sources on the Internet, too numerous to cite individually.

1 Recall that the CountSketch algorithm discussed in class, given $x \in \mathbb{R}^n$ and a hash table with B columns and $O(\log n)$ rows, provides an estimate $y \in \mathbb{R}^n$ such that

$$||x - y||_\infty \leq O(1/\sqrt{B})||x_{(k+1,\ldots,n)}||_2$$

with probability at least $1 - 1/n$.

1a (30 pts) Prove that the vector $\tilde{x}$ of top k coefficients of y satisfies

$$||x - \tilde{x}||_2 \leq (1 + O(\epsilon))||x_{(k+1,\ldots,n)}||_2$$

if $B \geq k/\epsilon^2$.

Solution. Let S denote the top k coefficients of y. We have

$$||x - y_S||_2^2 = ||x_{[k]\setminus S}||_2^2 + ||(x-y)_S||_2^2 + ||x_{[n]\setminus([k]\cup S)}||_2^2 \tag{1}$$

For every $i \in [k] \setminus S$ and $j \in S \setminus [k]$ we have $|y_i| \le |y_j|$, so

$$x_i \le x_j + \frac{1}{\sqrt{B}}||x_T||_2.$$

Note that $|[k] \setminus S| = |S \setminus [k]|$, and let $\pi : [k] \setminus S \to S \setminus [k]$ denote an arbitrary bijection, so that for each $i \in [k] \setminus S$

$$x_i \le x_{\pi(i)} + \frac{1}{\sqrt{B}}||x_T||_2.$$

Summing over $i \in [k] \setminus S$, we get

$$\begin{aligned} ||x_{[k]\setminus S}||_2^2 &= \sum_{i\in[k]\setminus S} x_i^2 \\ &\le \sum_{i\in[k]\setminus S} (x_{\pi(i)} + \frac{1}{\sqrt{B}}||x_T||_2)^2 \\ &\le \sum_{i\in[k]\setminus S} \left(x_{\pi(i)}^2 + 2|x_{\pi(i)}|\frac{1}{\sqrt{B}}||x_T||_2 + \frac{1}{B}||x_T||_2^2\right) \qquad (2) \\ &\le ||x_{S\setminus[k]}||_2^2 + 2||x_{S\setminus[k]}||_1 \frac{1}{\sqrt{B}}||x_T||_2 + \frac{k}{B}||x_T||_2^2 \\ &\le ||x_{S\setminus[k]}||_2^2 + 2\sqrt{k/B}||x_T||_2^2 + \frac{k}{B}||x_T||_2^2 \qquad (\text{since } ||x_{S\setminus[k]}||_1 \le \sqrt{k}||x_{S\setminus[k]}||_2) \end{aligned}$$

We also have

$$||(x-y)_S||_2^2 \le k \cdot \left(\frac{1}{\sqrt{B}}||x_T||_2\right)^2. \tag{3}$$

Substituting (2) and (3) into (1), we get

$$\begin{aligned} ||x - y_S||_2^2 &= ||x_{[k]\setminus S}||_2^2 + ||(x-y)_S||_2^2 + ||x_{[n]\setminus([k]\cup S)}||_2^2 \\ &= \left(||x_{S\setminus[k]}||_2^2 + 2\sqrt{k/B}||x_T||_2^2 + \frac{k}{B}||x_T||_2^2\right) + k \cdot \left(\frac{1}{\sqrt{B}}||x_T||_2\right)^2 + ||x_{[n]\setminus([k]\cup S)}||_2^2 \\ &= ||x_T||_2^2 + 2\sqrt{k/B}||x_T||_2^2 + \frac{2k}{B}||x_T||_2^2 \\ &= (1 + 2\epsilon + 2\epsilon^2)||x_T||_2^2 \\ &\le (1 + O(\epsilon))||x_T||_2^2. \end{aligned}$$

□

1b Prove that the vector $\tilde{x}$ of top $2k$ coefficients of y satisfies

$$||x - \tilde{x}||_2 \le (1 + O(\epsilon))||x_{(k+1,\ldots,n)}||_2$$

if $B \ge k/\epsilon$.

Solution. Let S denote the top k coefficients of y. We have

$$||x - y_S||_2^2 = ||x_{[k]\setminus S}||_2^2 + ||(x-y)_S||_2^2 + ||x_{[n]\setminus([k]\cup S)}||_2^2 \tag{4}$$

For every $i \in [k] \setminus S$ and $j \in S \setminus [k]$ we have $|y_i| \leq |y_j|$, so

$$x_i \leq x_j + \frac{1}{\sqrt{B}}||x_T||_2.$$

Let $\Delta := |S \setminus [k]|$. Let $b := \min_{i \in S\setminus[k]} |x_i|$. Then

$$\begin{aligned}
||x_{[k]\setminus S}||_2^2 - ||x_{S\setminus[k]}||_2^2 &\leq \Delta \cdot \left(b + \frac{1}{\sqrt{B}}||x_T||_2\right)^2 - (k+\Delta)\cdot b^2 \\
&\leq \Delta \cdot \left(b^2 + \frac{2}{\sqrt{B}}||x_T||_2 \cdot b + \frac{1}{B}||x_T||_2^2\right) - (k+\Delta)\cdot b^2 \\
&\leq \Delta \cdot \left(\frac{2}{\sqrt{B}}||x_T||_2 \cdot b + \frac{1}{B}||x_T||_2^2\right) - k\cdot b^2 \\
&\leq \frac{2\Delta}{\sqrt{B}}||x_T||_2 \cdot b - k \cdot b^2 + \frac{\Delta}{B}||x_T||_2^2
\end{aligned}$$

The last term is bounded as

$$\frac{\Delta}{B}||x_T||_2^2 \leq \frac{2k}{B}||x_T||_2^2 \leq 2\epsilon||x_T||_2^2,$$

so it suffices to upper bound the sum of the first two, namely

$$\frac{2\Delta}{\sqrt{B}}||x_T||_2 \cdot b - k \cdot b^2.$$

This is a quadratic function of b (recall that $b = \min_{i\in S\setminus[k]} |x_i|$ by definition). The maximum over b is achieved at the solution to (setting the derivative to zero)

$$0 = \frac{2\Delta}{\sqrt{B}}||x_T||_2 - 2kb,$$

i.e.

$$b = \frac{\Delta}{k\sqrt{B}}||x_T||_2.$$

Thus the maximum over Δ is

$$\begin{aligned}
&\frac{2\Delta}{\sqrt{B}}||x_T||_2 \cdot \left(\frac{\Delta}{k\sqrt{B}}||x_T||_2\right) - k \cdot \left(\frac{\Delta}{k\sqrt{B}}||x_T||_2\right)^2 \\
&= 2k\left(\frac{\Delta}{k\sqrt{B}}||x_T||_2\right)^2 - k \cdot \left(\frac{\Delta}{k\sqrt{B}}||x_T||_2\right)^2 \\
&= k\left(\frac{\Delta}{k\sqrt{B}}||x_T||_2\right)^2 \\
&\leq k\left(\frac{2k}{k\sqrt{B}}||x_T||_2\right)^2 \\
&\leq \frac{4k}{B}||x_T||_2^2 \\
&\leq 4\epsilon||x_T||_2^2
\end{aligned}$$

To summarize, we showed that $||x_{[k]\setminus S}||_2^2 - ||x_{S\setminus[k]}||_2^2 \le 4\epsilon||x_T||_2^2$. We also have

$$||(x-y)_S||_2^2 \le |S| \cdot \left(\frac{1}{\sqrt{B}}||x_T||_2\right)^2 \le (2k/B)||x_T||_2^2 \le 2\epsilon||x_T||_2^2.$$

Putting this together with (4), we get

$$||x - y_S||_2^2 \le (1+6\epsilon)||x_T||_2^2,$$

as required. □

2 [Exact sparse recovery] Recall that the discrete Fourier transform for signals of length n is given by the matrix $F = (F_{jk}) = \exp(2\pi ijk/n)$. Show that every signal $x \in \mathbb{R}^n$ with at most s nonzero coordinates can be uniquely recovered from the first $2s$ rows of Fx, i.e. $(Fx)_i, i = 0, \ldots, 2s-1$. *Hint: your algorithm need not be stable to noise, nor efficient. You can assume infinite precision arithmetic.*