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

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

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

if $B \geq k/\epsilon$.

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.*