Exercise V, 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 in order to produce a list of heavy hitters in the previous lecture we used CountSketch to compute estimates for the number of times any element i occurred in the stream, and included those elements whose estimated count exceeded a certain fraction of the total Euclidean norm of the frequency vector x in the list. Thus, we need a way to maintain an approximation to the Euclidean norm of x. In this exercise you will show that the ℓ_2 norm of a single row of the matrix maintained by CountSketch is a good approximation to the norm.

Choose a pairwise independent hash function $h : [n] \to [m]$, and a four-wise independent hash function $\sigma : [n] \to \{-1, +1\}$. Define an $m \times n$ matrix Π by letting, for each $j \in [n] = \{1, 2, \ldots, n\}$

$$\Pi_{ij} = \begin{cases} \sigma(j) & \text{if h(j)=i} \\ 0 & \text{o.w.} \end{cases}$$

Note that this is the CountSketch matrix with m columns ($B = m$ buckets) and a single row.

Prove that if $m = C_2/\epsilon^2$ for a sufficiently large absolute constant $C_2 > 0$, then

$$(1-\epsilon)||x||_2^2 \le ||\Pi x||_2^2 \le (1+\epsilon)||\Pi x||_2^2$$

with probability at least 2/3 for every fixed $x \in \mathbb{R}^n$.