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

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

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

Solution. We first compute the mean:

$$\begin{aligned}
\mathbf{E}[||\Pi x||_2^2] &= \mathbf{E}\left[\sum_{i=1}^{m}\left(\sum_{j\in[n]}\Pi_{ij}x_j\right)^2\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m}\sum_{j\in[n]}\sum_{j'\in[n]}\Pi_{ij}\Pi_{ij'}x_jx_{j'}\right] \\
&= \sum_{i=1}^{m}\sum_{j\in[n]}\sum_{j'\in[n]}\mathbf{E}\left[\Pi_{ij}\Pi_{ij'}\right]x_jx_{j'} \\
&= \sum_{i=1}^{m}\sum_{j\in[n]}\mathbf{E}\left[\Pi_{ij}^2\right]x_j^2 \\
&= \sum_{i=1}^{m}\sum_{j\in[n]}\frac{1}{m}\cdot x_j^2 \\
&= ||x||_2^2
\end{aligned}$$

We now compute the variance:

$$\begin{aligned}
\mathbf{E}[||\Pi x||_2^4] &= \mathbf{E}\left[\left(\sum_{i=1}^{m}\left(\sum_{j\in[n]}\Pi_{ij}x_j\right)^2\right)^2\right] \\
&= \mathbf{E}\left[\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{a\in[n]}\sum_{b\in[n]}\sum_{a'\in[n]}\sum_{b'\in[n]}\Pi_{ia}\Pi_{ib}\Pi_{i'a'}\Pi_{i'b'}x_ax_bx_{a'}x_{b'}\right] \\
&= \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{a\in[n]}\sum_{b\in[n]}\sum_{a'\in[n]}\sum_{b'\in[n]}\mathbf{E}\left[\Pi_{ia}\Pi_{ib}\Pi_{i'a'}\Pi_{i'b'}\right]x_ax_bx_{a'}x_{b'}
\end{aligned}$$

We consider the following cases:

1. $i \neq i'$. Then we must have $a = b, a' = b'$. We get

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{i'=1,i\neq i'}^{m}\sum_{a\in[n]}\sum_{a'\in[n]}\mathbf{E}\left[\Pi_{ia}^2\Pi_{i'a'}^2\right]x_a^2x_{a'}^2 \\
=&\sum_{i=1}^{m}\sum_{i'=1,i\neq i'}^{m}\sum_{a\in[n]}\sum_{a'\in[n],a'\neq a}\mathbf{E}\left[\Pi_{ia}^2\Pi_{i'a'}^2\right]x_a^2x_{a'}^2 \quad (\text{since } \Pi_{ia}\Pi_{i'a} = 0 \text{ for } i \neq i') \\
=&\frac{m(m-1)}{m^2}\sum_{a\in[n]}\sum_{a'\in[n],a\neq a'}x_a^2x_{a'}^2 \\
\leq&\sum_{a\in[n]}\sum_{a'\in[n],a\neq a'}x_a^2x_{a'}^2
\end{aligned} \tag{1}$$

2. $i = i'$, $a = b, a' = b', a \neq a'$. Then we have

$$\begin{aligned}
&\sum_{i=1}^{m} \sum_{a\in[n]} \sum_{a'\in[n], a\neq a'} \mathbf{E}\left[\Pi_{ia}^2 \Pi_{ia'}^2\right] x_a^2 x_{a'}^2 \\
=&\sum_{i=1}^{m} \sum_{a,a'\in[n]:a\neq a'} \mathbf{E}\left[\Pi_{ia}^2 \Pi_{ia'}^2\right] x_a^2 x_{a'}^2 \\
=&\frac{1}{m^2} \sum_{i=1}^{m} \sum_{a,a'\in[n]:a\neq a'} x_a^2 x_{a'}^2 \\
\leq&\frac{1}{m} ||x||_2^4
\end{aligned}$$

3. $i = i'$, $a = a', b = b', a \neq b$. Similar to the above, gives no more than $\frac{1}{m}||x||_2^4$.

4. $i = i'$, $a = b', a' = b, a \neq a'$. Similar to the above, gives no more than $\frac{1}{m}||x||_2^4$.

5. $i = i'$, $a = b = a' = b'$. We get

$$\sum_{i=1}^{m} \sum_{a\in[n]} \mathbf{E}\left[\Pi_{ia}^4\right] x_a^4 = \frac{1}{m} \sum_{i=1}^{m} \sum_{a\in[n]} x_a^4 = \sum_{a\in[n]} x_a^4. \tag{2}$$

Putting the above bounds together, we get

$$\begin{aligned}
\mathbf{Var}(||\Pi x||_2^2) &= \mathbf{E}(||\Pi x||_2^4) - (\mathbf{E}(||\Pi x||_2^2))^2 \\
&\leq \sum_{a\in[n]} \sum_{a'\in[n], a\neq a'} x_a^2 x_{a'}^2 + \sum_{a\in[n]} x_a^4 \qquad \text{(by (1) and (2))} \\
&\quad + \frac{3}{m}||x||_2^4 \qquad \text{(by 2., 3., 4. above)} \\
&\quad - ||x||_2^4 \\
&= \frac{3}{m}||x||_2^4
\end{aligned}$$

By Chebyshev's inequality we have

$$\mathbf{Prob}\left[\left|||\Pi x||_2^2 - ||x||_2^2\right| > \epsilon||x||_2^2\right] \leq \frac{\mathbf{Var}(||\Pi x||_2^2)}{\epsilon^2||x||_2^4} \leq \frac{(3/m)||x||_2^4}{\epsilon^2||x||_2^4} \leq 3/(\epsilon^2 m) < 1/3$$

as long as $m > 9/\epsilon^2$, as required.

□