# Sublinear Time and Space Algorithms 2024A – Problem Set 2

Robert Krauthgamer

Due: January 17, 2024

**General instructions:** Please keep your answers short and easy to read. You can use results, calculations or notation seen in class without repeating them, unless asked explicitly to redo them.

1. Consider the frequency-vector model, where the stream contains additive updates to a vector $x \in \mathbb{R}^n$ whose coordinates are integers bounded by $\mathrm{poly}(n)$.

   Explain how to $(1+\epsilon)$-approximate $\sum_{i<j}(x_i + x_j)^2$ by a streaming algorithm with storage requirement $(\epsilon^{-1}\log n)^{O(1)}$ bits.

   Remark: As done in class, do not count storage of the algorithm's random coins.

2. Using the notation seen in class for Euclidean MST of $P \subset [\Delta]^d$, prove that $\mathrm{MST}_T(P) \geq \frac{1}{2}\,\mathrm{MST}(P)$. (Assume here that the quadtree is fixed and not randomly shifted.)

3. A matrix $A \in \mathbb{R}^{m\times n}$ is called $\varepsilon$-coherent if its columns $A^1, \ldots, A^n \in \mathbb{R}^m$ satisfy (1) all $i \in [n]$, $\|A^i\|_2 = 1$; and (2) for all $i \neq j \in [n]$, $|\langle A^i, A^j\rangle| \leq \varepsilon$.

   Show that for every $n$ and $\varepsilon \in (0, 1/2)$ there exists an $\varepsilon$-coherent matrix $A \in \mathbb{R}^{m\times n}$ with $m = O(\varepsilon^{-2}\log n)$.

   Hint: Consider a random matrix with independent random $\pm 1/\sqrt{m}$ entries.

4. Prove that Algorithm IncoherentSketch below (which uses the definition in the previous question) solves the $\ell_1$-point query problem, i.e., given a frequency vector $x \in \mathbb{R}^n$ it outputs $(A^T y)_i \in x_i \pm \varepsilon\|x\|_1$. What is its storage requirement (not including storing the matrix $A$)?

   **Algorithm IncoherentSketch**

   1. Init: Fix a matrix $A$ that is $\varepsilon$-coherent
   2. Update: Maintain a linear sketch $y = Ax$
   3. Output: to estimate $x_i$ report $(A^T y)_i$.

   How does this algorithm compare to CountMin+ seen in class (using $m$ from the previous question)? Try to find an advantage and a disadvantage.

## Extra credit:

5. Design a streaming algorithm for the bichromatic matching problem (aka earthmover distance), where the input is a set of colored points $P \subset [\Delta]^d$, half of them are blue and half are

red, i.e., $P = R \cup B$, and the goal is to compute a minimum-weight perfect matching between $R$ and $B$.

Hint: Use a randomly shifted quadtree (as seen in class), and for each level $i$ estimate $\|x^{(i)}\|_1$.

For simplicity, assume a randomized streaming algorithm that $(1+\varepsilon)$-approximates the $\ell_1$ norm (of the frequency vector $x$), for $\varepsilon = 0.1$, using storage $s(n) = \text{polylog}(n)$. Such algorithms are known, although we did not see it in class.