CSCE 689: Special Topics in Modern Algorithms for Data Science Fall 2023

Lecture 27 — November 3, 2023

Prof. Samson Zhou *Scribe: Ayesha Qamar*

1 Streaming Model

The input is elements of an underlying data set S, which arrive sequentially. We want to output the evaluation (or approximation) of a given function using only space sublinear in the size m of the input S.

So far we have assumed the data is fixed and therefore independent of the algorithm. This assumption doesn't always hold in the case where we need to interact with the algorithm multiple times.

Some use cases and if it is possible to handle them are given below.

- **Case Study #1** Suppose we run the same algorithm on multiple datasets.
  Yes, we can handle this case using union bound. Since the dataset is fixed, we already know the probability of failing on each dataset. By union bound the probability of success on all datasets is the complement of the sum of failure for all the datasets.
- **Case Study #2** Suppose we have a batch of queries for a randomized database algorithm.
  Yes, if the batch of queries is fixed.
- **Case Study #3** Suppose we ask a sequence of queries for a randomized database algorithm.
  Yes, if the batch of queries is fixed in advance. No, if each query can depend on the answer to previous queries. In the latter case, if the randomness of the algorithm remains fixed then it can cause problems.

2 Adversarially Robust Streaming

The input is elements of an underlying data set S, which arrive sequentially and *adversarially*. We want to output the evaluation (or approximation) of a given function using only space sublinear in the size m of the input S.

Here, *adversarially robust* means that the future queries may depend on previous queries.

2.1 Motivation

Having an adversarially robust model is essential for database queries and adversarial machine learning. For example, ML models can be adversarially trained on perturbed examples such that they output some malicious (wrong) answers whenever they are given inputs that contain the perturbations.

3 AMS F_2 Algorithm

Let $s \in \{-1, +1\}^n$ be a sign vector of length n and $Z = \langle s, f \rangle = s_1 f_1 + \ldots + s_n f_n$. Recall that we can take the mean of $O\left(\frac{1}{\varepsilon^2}\right)$ inner products for $(1+\varepsilon)$-approximation.

3.1 Attack on AMS

Can learn whether $s_i = s_j$ from $\langle s, e_i + e_j \rangle$.

Let $f_i = 1$ if $s_i = s_1$ and $f_i = -1$ if $s_i \neq s_1$. Since $Z = \rangle s, f \rangle = s_1 f_1 + \ldots + s_n f_n = m$, then we have $Z^2 = m^2$ deterministically.

3.2 Reconstruction Attack on Linear Sketches

Definition. Suppose stream S induces a frequency vector f. A **linear sketch** is an algorithmic framework that works by generating a random matrix A and maintaining $A \cdot f$. A post-processing function $g(A \cdot f)$ is applied to get the output.

Linear sketches are "not robust" to adversarial attacks and it has been shown that linear sketches must use $\Omega(n)$ space for certain problems. There exist attacks that can approximate the sketch matrix A, query in the kernel of A to map the output to the null space. i.e., input $f \in Ker(A)$ and $f \neq 0$ but $A \cdot f = 0$.

For example, if $A = \begin{bmatrix} 0 & 1 \end{bmatrix}$ and $f = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, then $A \cdot f = 0$.

4 Sketch Switching

Suppose we are trying to approximate some given function and we have a streaming algorithm for this function. Further, this function is monotonic and the stream is insertion-only. Sketch switching framework gives a robust algorithm for this function. It works by starting multiple instances of the same algorithm at the beginning. After using an instance of the algorithm, the output is "frozen". Each time the next instance has value $(1 + O(\varepsilon))$ more than the "frozen" output, use the next instance and "freeze" its output.

Sketch switching gives $\frac{1}{\varepsilon^3}$ dependency in space for many problems.