"Online" Algorithms (Data streams)

Wendy Rosettini

November 2023

1 Introduction

Summary of Online Algorithms for Data Streams:

1. Introduction

Efficient processing of massive data sets is crucial due to the proliferation of large-scale data in various applications. The focus is on real-time monitoring of rapidly changing data streams, necessitating algorithms that operate with limited memory and make a small number of passes over the data.

2. Challenges and Characteristics

- Limited Space: Algorithms for data streams use at most polylogarithmic space relative to the length of the input stream.
- Summary Structures: Summary data structures, also known as synopses, with small memory footprints are utilized for answering queries approximately.
- Accuracy Guarantees: Error bounds on query results are provided through user-specified parameters (ε and δ).

3. Progress in the Field

Significant progress has been made in designing streaming algorithms over the past decade. Various synopses have been proposed for fundamental problems such as frequency moments, norms, inner products, quantiles, histograms, and wavelets. Progress has also extended to computational geometry and graph problems.

4. Applications

- Network Management: Algorithms analyze network traffic patterns, estimate demands, and detect faults and congestion in real time.

- Database Monitoring: Statistical summaries are crucial for optimizing query plans in large databases undergoing transactions.
- Online Auctions: Statistical estimation aids in identifying economic trends and analyzing bid patterns in online auction systems.
- Sequential Disk Accesses: Streaming algorithms are effective for minimizing I/O operations when data reside on disks.

5. Overview of the Literature

The literature on online algorithms for data streams spans several decades. Notable contributions include works from the late 1970s to the 2000s, with seminal papers introducing techniques like randomized linear projections for approximating frequency moments.

6. Challenges and Future Directions

Ongoing challenges include addressing lower bounds, proving impossibility results, and tackling more complex graph problems. The field continues to evolve rapidly, and further research is needed to address emerging challenges and applications.

2 Simulations

I conducted a simulation[1] of an online algorithm applied to a data stream with the goal of incrementally calculating the mean and standard deviation as data is processed. More specifically:

- Online Algorithm: The algorithm processes data one at a time without revisiting previous elements. Its implementation simulates the online approach typical of algorithms operating on real-time data streams.
- Data Stream: I represented a data stream by inputting a sequence of numbers. The algorithm processes this data sequentially, reflecting the nature of real-time data arriving in a continuous stream.
- Memory Constraint: The algorithm maintains limited memory, using only a few state variables to calculate the mean and standard deviation. This limitation aligns with the online approach, which operates with constrained resources.
- Intermediate and Final Results: During the simulation,I displayed intermediate results at each step, showing how the algorithm updates statistical estimates as new data arrives. The final result reflects the mean and standard deviation of the entire data stream.

[1] http://wendy.altervista.org/data$_s tream.html$

In summary, the simulation illustrated how an online algorithm manages a real-time data stream, adapting to memory constraints and producing approximate results as data becomes available.

Input Data: 1,2,3,4,5 Process Data

```
Processing...
Step 1: Mean = 1.00, Standard Deviation = 0.00
Step 2: Mean = 1.50, Standard Deviation = 0.50
Step 3: Mean = 2.00, Standard Deviation = 0.82
Step 4: Mean = 2.50, Standard Deviation = 1.12
Step 5: Mean = 3.00, Standard Deviation = 1.41

Final Result: Mean = 3.00, Standard Deviation = 1.41
```