CMPSCI 711: More Advanced Algorithms

Course Overview

Andrew McGregor

Last Compiled: January 24, 2018

Goal and Focus of Class

- Overall Goal: Learn general techniques and gain experience designing and analyzing algorithms with randomization and approximation.
- Focus: Course will focus on problems processing large data sets.
  - Data Streams: You process a long sequence $\langle a_1, \ldots, a_m \rangle$ of items (numbers, edges of a graph, points in a geometric space etc.) sequentially. What can you compute without storing entire sequence?
  - Linear Sketches: Given a high-dimensional vector $x \in \mathbb{R}^n$, what can you compute about x given a projection $Ax \in \mathbb{R}^k$ where $k \ll n$ and A is a (random) $k \times n$ matrix of your choosing?
  - Sampling: Given a limited number of samples from some data set, what quantities can you accurately estimate with high probability?
  - Communication: Suppose data is stored on different machines in a cluster. How much communication does it require to compute a given function on the data.

Outline

Examples

Class Overview

Example 1: Simple Puzzles

- Puzzle 1: A bag contains n balls which are numbered 1 through n. Balls are removed one-by-one until only one remains. Can you find the number on the remaining ball?
- Answer: Missing value is $(\sum_{i=1}^{n} i)$-(sum of removed values).
- Puzzle 2: Suppose bag is initially empty. At each step, a numbered ball is either added or removed. Multiple balls can have same number. Can you determine if all balls in bag have same number?
- Answer: Let f_i be the number of balls with number i in the bag. It's easy to compute $m = \sum f_i$, $U = \sum i f_i / m$, and $V = \sum i^2 f_i / m$. Then balls in the bag have the same number iff $U^2 = V$.

Example 2: Randomization and Distinct Elements

- Puzzle: Suppose the bag is initially empty. At each step, a numbered ball is added to the bag. Multiple balls can have same number. Can you estimate how many distinct values are in the bag?
- Sorta Answer: For each $i \in [n]$, pick random $h_i \in_R [0, 1]$. Find

  $$X = \min\{h_i : \; i \text{ added to the bag}\}$$

  Can show $\mathbb{E}[X] = \frac{1}{F_0+1}$ where F_0 is the number of distinct values
- Problems:
  - How can we store all the h_i in limited space?
  - How can we find F_0 up to a factor $(1 + \epsilon)$ with high probability?
  - What if balls can be added and removed?

Example 3: Graphs

- Puzzle: Suppose your name is Mark Z. and you can see all the friendships that are forming in the social network that you invented. Can you tell if the network is connected with only $O(n \log n)$ bits of memory where n is the number of people in the network?
- Sorta Answer: Maintain spanning forest of growing network.
- Problems:
    - What can you do if people also de-friend each other?
    - What if you also want to approximate every cut in the graph?

Example 4: Clustering

- Suppose you are skimming through holiday photos and you want to find k photos that best summarize your trip. Given two photos p_i and p_j you define a distance metric $d(p_i, p_j)$ between the photos...
- Puzzle: Find k photos $\mathcal{C} = \{p_{i_1}, \ldots, p_{i_k}\}$ minimizing $\max_j d(p_j, \mathcal{C})$. Suppose we know that optimal solution has value t. Can we find a solution with value at most $2t$ while only remembering k photos.
- Answer: Add a new photo p to $\mathcal{C}$ if $d(p, \mathcal{C}) > 2t$.
- Problems:
  - What if you didn't know value of optimum solution?
  - What is you wanted to minimize the average distance rather than the maximum distance?

Outline

Examples

Class Overview

Approximate Schedule

- Part I: Graphs Connectivity, cuts, cliques, matchings, sparsification, correlation clustering, spanners, etc. (9 lectures)
- Part II: Vectors and Matrices Basic techniques and estimating numerical statistics such as quantiles, heavy-hitters, frequency moments, regression and randomized linear algebra, etc. (7 lectures)
- Part III: Lower Bounds: Basics of communication complexity and applications, information theory, etc. (3 lectures)
- Part IV: Additional Topics
  - Clustering and Geometry: Clustering, core-sets, ϵ-nets, convex hulls, minimum enclosing ball, etc.
  - Strings and Sequences: Longest increasing subsequences, formal language recognition, memory checking, etc.
  - Special Topics Sliding windows, stochastic streams, distributed monitoring, parallel algorithms, etc.

Things You'll Learn: Ways to make big data, small data. . .

- Dimensionality Reduction For any m vectors $v_1, \ldots, v_m \in \mathbb{R}^N$, there's a linear map $f : \mathbb{R}^N \to \mathbb{R}^{O(\epsilon^{-2} \lg m)}$ such that for all pairs of vectors v_i and v_j,

$$(1-\epsilon)\|v_i - v_j\|_2 \leq \|f(v_i) - f(v_j)\|_2 \leq (1+\epsilon)\|v_i - v_j\|_2$$

- Sparsifiers For any graph G on n vertices, there's a weighted subgraph graph G' on n vertices with only $O(\epsilon^{-2} n)$ edges such that the size of any cut in G' is within a $1+\epsilon$ factor of the size of the corresponding cut in G.
- Spanners For any graph G on n vertices, there's a graph subgraph G' on n vertices with $O(n^{1+1/t})$ edges such that the distance between any pair of nodes in G' is at most a factor $2t-1$ larger than it is in G.

Material

- All required material will be posted at:

  `https://www.cs.umass.edu/~mcgregor/courses/CS711S18/`
- No required textbook but you might find the following helpful:
  - Data Streams: Algorithms and Applications by Muthukrishnan
  - Randomized Algorithms by Motwani and Raghavan
  - Probability and Computing by Mitzenmacher and Upfal
- And there are some relevant surveys posted on the website.

Homework, Grading, and Stuff

- There will be four homework which you can do in small groups and an in-class midterm mainly related to the homework.
- There will be a few lectures for class presentations of a relevant paper of your choosing. I'll post a list of suggestions.
- Final mark will be based on a combination of homework marks, quality of presentation, and general class participation.