Exam

Algorithms for Data Science
Université Paris-Saclay, M2 Data Science

October 21st, 2022

This is the exam for the Algorithms for Data Science class, which will determine 50% of your grade for this class. The duration of the exam is *two hours.* This exam subject consists of 4 exercises and has 2 pages.

The exam is *strictly personal*: any communication or influence between students, or use of outside help, is prohibited. Any violation of the rules may result in a grade of 0 and/or disciplinary action.

Exercise I – Course Questions (2 points)

1. Explain two main challenges of designing algorithms for *data streams.*
2. What is the defintion of *competitive ratio* for online greedy algorithms?

Exercise II – Similarity and Min-hash (7 points)

Consider the following document matrix, where each document is a set of 11 possible shingles, denoted as $s_0, \ldots, s_6$:

	D_1	D_2	D_3
s_0	0	1	1
s_1	1	0	0
s_2	1	1	0
s_3	1	1	1
s_4	0	1	0
s_5	0	0	0
s_6	0	1	0

Questions:

1. Compute the Jaccard similarity for all 3 pairs of documents. Explain how you obtained it.
2. Consider the follow hash functions: $h_1(x) = 4x + 5 \bmod 7$, $h_2(x) = x + 3 \bmod 7$, $h_3(x) = 2x + 3 \bmod 7$. Compute min-hash signature for h_1 and all documents.
3. Compute the full min-hash matrix for the remaining two documents using the above functions, without first computing the permutations.

4. Give the resulting estimation of the Jaccard similarity of all 3 pairs of documents, using the min-hash signature. What do you notice?

Exercise III – Estimating Frequency Moments of a Stream (6 points)

Consider the following stream:

```
2 2 1 1 2 3 1 1 1
```

We will work in this exercise with estimations of the frequency moments of a stream and the Alon-Matias-Szegedy algorithm for estimating the second moment of a stream.

Questions:

1. Compute the surprise number (second frequency moment) for the above stream.
2. Suppose we apply the Alon-Matias-Szegedy algorithm on the above stream. For each possible value of the random choice $i \in \{1, \dots, 9\}$, have X_i the resulting variable (value) appearing at position i. Write the possible values of $X_i.c$ and $X_i.\mathrm{val}$
3. Exemplify how the AMS algorithm works (by using e.g., a random value of i), and how the final estimate of the moments is computed. Compare it with the one in Question III.1.
4. Explain and exemplify how the AMS algorithm's estimation can be improved.
5. Explain and exemplify how the AMS algorithm can be adapted to higher order moments, e.g., the third moment.

Exercise IV – Counting Ones in a Window (5 points)

Consider the following stream of bits, where the rightmost element is the most recent one:

```
...  1 1 1 0 1 0 1 1 0 0 1 1 0 0 1
```

We aim to estimate the number of 1s in the last k bits using the DCIM algorithm.

Questions:

1. Assume DCIM has created the following buckets, where the rectangles represent the buckets of increasing number of 1s inside:

   $$\ldots\ \ \boxed{1\ 1\ 1\ 0\ 1}\ 0\ \boxed{1\ 1}\ 0\ 0\ \boxed{1\ 1}\ 0\ 0\ \boxed{1}$$

   Explain how DCIM estimates the number of 1s from these buckets, for $k = 4$ and $k = 10$. If it is the case, explain why the estimation of the number of 1s is wrong.
2. Give one other way to divide the window into buckets, while respecting the DCIM restrictions on buckets. Explain how and why the estimations change (or do not change, as the case may be).
3. Assume the following bits arrive in the stream in order: 0, 1, 1, 0, 1, 0. Explain how the buckets are updated from the setting in Question IV.1 and what is the final result.