

Assignment 1 – Part A

In this exercise you will need to modify the Reservoir Sampling algorithm that we described in class, so that it samples K items uniformly at random from a stream of N items. Each item must have probability K/N of appearing in the sample. Your algorithm must work with a single pass over the data, reading the items one by one, without prior knowledge of the size of the stream (the size N), and must use $O(K)$ memory (assume that the size of each item is constant). This means that you cannot store the entire data stream in memory.

1. Describe the algorithm that selects a uniform sample of K items from a stream of N items. The description of the algorithm must not be in code or pseudocode, nor be a description of the code in natural language. In that case your answer will receive zero marks. The description should explain the logic of the algorithm in plain English.

For example, this is a description in plain English of the algorithm for sampling a single item that we saw in class:

The algorithm keeps space for one item and a counter with the number of items it has seen. It goes through the items one by one as they arrive from the stream. When it sees the n -th item, it selects it with probability $\frac{1}{n}$ and stores it, replacing the existing item (if it is not the first). It updates the counter. When the stream ends, it returns the item it has stored.

2. Prove that your algorithm produces a uniformly random sample, that is, for every i , $1 \leq i \leq N$, the i -th item has probability K/N of appearing in the sample.
3. Write a function **sample** in **Python** that implements your algorithm. Your function should take as arguments the name of a file and the number K , and return a list holding a sample of K random lines from the file. You should read the file line by line and not load it into memory. Create a file *input.txt* with 100 lines containing the numbers from 1 to 100, and use your function to sample 10 random lines. Print the lines of the sample.

Explain the correspondence between the description you gave in the 1st step and your code. For example, for the algorithm that selects only a single item, you could write something like: “In lines 2-3, the n -th item is selected with probability $\frac{1}{n}$ ”.

Create a Notebook with two code cells. One with the imports and the definition of the function `sample` and one in which you will use your function on the file `input.txt`, store the result, and print it. Also add two text cells, one with the description of the algorithm (Step 1), and one with the correspondence between the code and the description. You can add the proof (Step 2) in a separate cell, or write it separately and submit a pdf with the text (or photos if it is handwritten). Submit the Notebook, the file `input.txt` and the pdf (or photos) with your proof, if not in the Notebook.