CSC 284/484 - homework 2 (Streaming Algorithms)

http://www.cs.rochester.edu/~stefanko/Teaching/17CS484

Students that take the course as 484 are required to do **both** 284/484 and 484 parts of the homework. Students that take the course as 284 are only required to do 284/484 part of the homework (of course you are welcome to solve/turn-in the 484 part as well).

---

1 284/484 homework - solve and turn in

1.1 Theoretical/applied part

Exercise 1.1 (**due 3/7/2017**)(discussed in class) Given non-negative $p_1, \dots, p_n$ such that $\sum_i p_i = 1$ give a sampling algorithm that outputs X such that $P(X = i) = p_i$. The algorithm should spend $O(1)$ time per sample (worst-case) and use $O(n)$ time for preprocessing. Implement your algorithm. The algorithm should read input from `stddin`. The first line contains number n. The next line contains n non-negative numbers $p_1, \dots, p_n$ that sum to 1. The next line contains k—the number of samples to be output. Your algorithm should output k independent samples from the distribution on `stdout`.

1.2 Applied part

Exercise 1.2 (**due 3/7/2017**) Implement Tug-of-War Sketch and Count-Sketch algorithm. Your implementation should process a stream in the following format (read from `stdin`). The first line contains

- n (the elements in the stream will be from the set $[n] = \{1, \dots, n\}$),

Each of the next lines is of the following two types:

- line starting with `A` followed by an integer $x \in [n]$ adds x to the collection;
- line starting with `Q` followed by an integer $x \in [n]$ asks a query about element x.

For each line starting with `Q` output (to `stdout`): 1) the current estimate of F_2 (with precision and confidence) and 2) the estimate for the number of occurrences of x in the collection (the algorithm should also output an interval and a confidence value). The precision and confidence should be parameters in your program; default values are precision = 20%, confidence = 99%.

2 484 homework - solve and turn in

2.1 Theoretical/applied part

Exercise 2.1 (**due 3/7/2017**) The Cauchy distribution with parameter γ centered at x_0 has density

$$f(x) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x - x_0}{\gamma}\right)^2\right)}.$$

Let $X, X_1, \ldots, X_5$ be independent from Cauchy(x_0, γ). Let $Y = median(X_1, X_2, X_3)$. Let $Z = median(X_1, X_2, X_3, X_4, X_5)$. What is the squared coefficient of variation of X? What is the squared coefficient of variation of Y? What is the squared coefficient of variation of Z?

2.2 Applied part

Exercise 2.2 (**due 3/7/2017**) Implement BJKST algorithm (you can implement the simplified version that does not use the secondary hash function g). Your implementation should process a stream in the following format (read from `stdin`). The first line contains

- n (the elements in the stream will be from the set $[n] = \{1, \ldots, n\}$),

Each of the next lines is of the following two types:

- line starting with `A` followed by an integer $x \in [n]$ adds x to the collection;
- line starting with `Q`.

For each line starting with `Q` output (to `stdout`): 1) the current estimate of the number of distinct elements (with precision and confidence). The precision and confidence should be parameters in your program; default values are precision = 20%, confidence = 99%.