

## Algorithms for Big Data

Fall Semester 2019

### Exercise Set 4

#### Exercise 1:

Show that Cauchy distribution  $f(x) = \frac{1}{\pi(1+x^2)}$  is 1-stable, that is for  $X, Y \sim \text{Cauchy}$  and  $a, b \in \mathbb{R}$  we have  $a * X + b * Y \sim (|a| + |b|) \cdot \text{Cauchy}$ .

#### Exercise 2:

Modify the algorithm for computing  $F_p$  of a stream (the one working for any  $p$  in the sequence of values) to work in semi-turnstile streams (updates  $(x_i, c_i)$  where  $c_i \in \mathbb{R}^+$ ).

The goal of next few exercises is to show that  $p$ -stable distribution approach can be used for sketching of Hamming norm/distance. We follow the analysis from "*Comparing Data Streams Using Hamming Norms (How to Zero In)*" by Cormode, Datar, Indyk and Muthukrishnan (VLDB'02).<sup>1</sup>

#### Exercise 3:

Assume we operate in universe  $U$  of magnitude  $u$ , that is we have a promise that all our values we are ever going to see when sketching are integers from  $\{-u, \dots, 1, 0, 1, \dots, u\}$ . Show that for sufficiently small value  $p$ ,  $F_p$  is an  $1 \pm \varepsilon$  approximation of  $F_0$ . Show that  $p = \Theta(\varepsilon / \log u)$  is small enough.

Show that  $u$ -factor approximation of  $L_p$  approximation is enough to obtain  $1 \pm \varepsilon$  approximation of  $F_p$ , and thus  $1 \pm 2\varepsilon$  of  $F_0$ .

Useful fact: when  $p \rightarrow 0$ , we have  $\Pr(|X| > t) = \Theta(t^{-p})$  for  $X$  drawn from  $p$ -stable, 0 mean, normalized by median stable distribution.

#### Exercise 4:

(2 pts)

Take value of  $p$  from Exercise 3 and desired level of approximation. How many samples do we need to take from this  $p$ -stable distribution to reach this level of approximation for median estimation? What is the memory complexity of our algorithm? Take into account the actual bit-size of bignums needed in our algorithm.<sup>2</sup>

---

<sup>1</sup>With the only change that we actually want to do the correct analysis - the analysis in the paper has an error.

<sup>2</sup>And here is the omission of the authors...