

# Algorithms for Big Data

Fall Semester 2019

## Exercise Set 2

Pseudorandomness: emulating perfect randomness in a predictable manner. Recall a following measure of *quality*

**Definition 1** Consider a family of hash functions  $\mathcal{H} = \{h : [u] \rightarrow [m]\}^1$ . We say that  $\mathcal{H}$  is  $k$ -wise independent if for any distinct  $x_1, \dots, x_k \in [u]$  and any (not necessarily distinct)  $y_1, y_2, \dots, y_k \in [m]$  there is

$$\Pr_{h \in \mathcal{H}}(h(x_1) = y_1 \wedge \dots \wedge h(x_k) = y_k) = \Theta(m^{-k}).$$

Informally: those hash-functions are indistinguishable from perfectly random hashing when evaluated simultaneously at  $k$  values.

We claim that (i)  $k$ -wise independence is good enough to "fool" algorithms into behaving as if provided with perfect randomness and (ii) this type of pseudo-randomness can be stored using small space.

### Exercise 1:

Let  $p > u$  be prime number. Let  $\mathbf{a} = a_0, \dots, a_{k-1}$  be vector of coefficients. Let  $h_{\mathbf{a}} : [u] \rightarrow [m]$  be defined as  $h_{\mathbf{a}}(x) = [(\sum_{i=0}^{k-1} a_i x^i) \bmod p] \bmod m$ . Show that  $\mathcal{H} = \{h_{\mathbf{a}} \mid a_0, \dots, a_{k-1} \in [p]\}$  is  $k$ -wise independent.

**Hint:**

Polynomial of degree  $k - 1$  in  $\mathbb{Z}_p$  is uniquely defined by its value on  $k$  distinct points.

### Exercise 2:

Show that families of hash-functions from previous exercise are not  $(k + 1)$ -wise independent.

### Exercise 3:

Show a lower-bound of  $\Omega(k \log m)$  bits necessary to represent (store) a hash-function from  $k$ -wise independent hash-function family. How much space do we need to represent perfectly random hash-function?

### Exercise 4:

Let  $X_1, X_2, \dots, X_n$  be pairwise independent random variables. Show that  $\text{Var}[\sum_i X_i] = \sum_i \text{Var}[X_i]$ .

---

<sup>1</sup> $[u] = \{0, 1, \dots, u - 1\}$  is called an *universe*.

**Exercise 5:**

Missing part of Morris' algorithm analysis: show inductively that  $\mathbb{E} \left[ (2^{X_n})^2 \right] = \frac{3}{2}n^2 + \frac{3}{2}n + 1$ .

**Exercise 6:**

(2 pts)

Consider following idea for concentrating Flajolet-Martin approach. Let  $r_1, r_2, \dots, r_n \in [0, 1]$  be picked uniformly and independently at random, and let  $X_k$  be  $k$ -th smallest value among  $r_1, \dots, r_n$ . Find  $\mathbb{E}[X_k]$  and  $\text{Var}[X_k]$ . Use it to derive streaming algorithm for distinct elements (see Bar-Yossef et al. 2002).