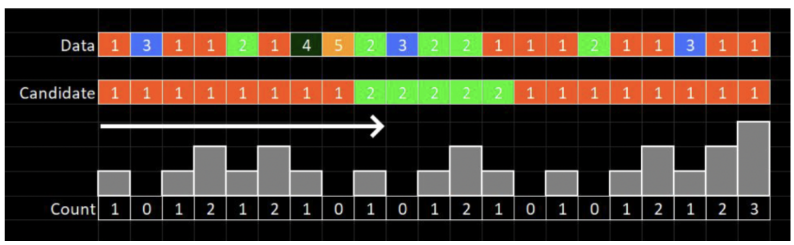
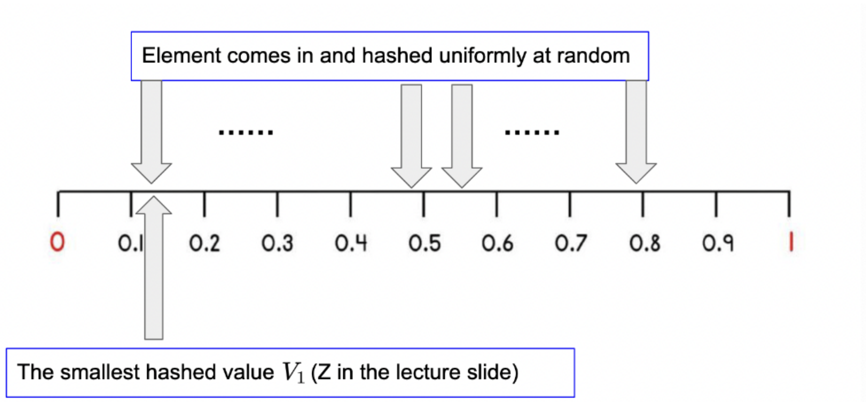
CAS CS 365
Lab 7

Streaming - Majority element and F0 estimation

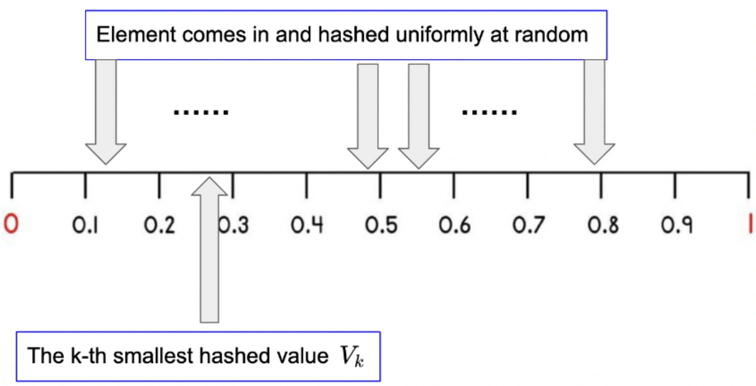
1. Majority element (heavy hitters algorithm)
    a. Assume the length of a data stream is n, how to find if there is a element that appears more than n/2 times with constant space? How many passes you need to make?
    b. Answer:
        i. Name key-value pair KV = (KV[0], KV[1])
        ii. For each element e in the stream:
            1. If the key-value pair is empty, set it to be (e,1)
            2. If KV not empty, and e = KV[0], then set KV[1] += 1
            3. If KV not empty, and e != KV[0], then KV[1] -= 1. Empty KV if KV[1] = 0
        iii. Go through the stream again to check the frequency of KV[0]
        iv. 

2. Lower bound on memory for exact deterministic algorithm
    a. Consider a sequence of m+1 elements
    b. There are 2^m - 1 possible subsets of elements for first m elements
    c. To determine the exact number of distinct elements in the sequence, we need at least m bits of memory
    d. If only m-1 bits are used, then the memory can only have 2^m - 1 states
        i. Two different subsets will share one state, which leads to incorrect answer
3. Can we use sampling to approximate F0?
    a. No
    b. Sampling cannot catch the minority with high probability, unless all elements appears with similar frequencies

4. Distinct element estimation using k-th min

a. 

b. $$\mathbb{E}[\mathbb{Z}] = \int_0^1 \Pr(Z > t)dt = \int_0^1 \Pr\left(X_1 > t\right)^n = \int_0^1 (1-t)^n dt = \frac{1}{n+1}$$

c. 

d. $$\Pr[V_k \leq x] = \Pr[\text{at least k observations are } \leq \text{ x}] = \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l}$$

e. $$\frac{d}{dx} \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l} = \sum_{l=k}^{n} \binom{n}{l} \left(l x^{l-1} (1-x)^{n-l} - x^l (n-l)(1-x)^{n-l-1} \right)$$

f. $$= n \binom{n-1}{k-1} x^{k-1} (1-x)^{(n-1)-(k-1)}$$

g. This is the pdf of beta distribution