Instructions

- The homework is due on **Friday 3/31 at 5pm ET**.
- No extension will be provided, unless for serious documented reasons.
- Start early!
- Study the material taught in class, and feel free to do so in small groups, but the solutions should be a product of your own work.
- This is not a multiple choice homework; reasoning, and mathematical proofs are required before giving your final answer.

1 Reservoir Sampling [15 points]

Design an algorithm that samples $k \geq 1$ elements uniformly at random from an insert-only stream, whose length is unknown. Present the pseudocode and prove the correctness of the proposed algorithm.

2 Median trick - a useful technique [15 points]

Prove the claim on slide 13. Be specific about the values of the constants C_1, C_2 you use in your proof, where $t = C_1 \log \frac{1}{\delta}$, $k = C_2 \frac{\mathrm{Var}[X]}{\epsilon^2 \mathbb{E}[X]^2}$.

3 Variance of Morris Counter [20 points]

Prove equation $\mathrm{Var}(Z) = \frac{m(m-1)}{2}$ on slide 47.

4 More on uniform RVs [10+10 points]

Let $X_1, \ldots, X_n$ be iid uniform random variables, $X_i \in U(0,1)$ for all i. (a) What is the pdf and (b) what is the expectation of the k-th smallest value among $X_1, \ldots, X_n$ for $k = 1, \ldots, n$?

5 Coding [40 points]

Check the Jupyter notebook on our Git repo.

1 Reservoir Sampling [15 points]

Design an algorithm that samples $k \geq 1$ elements uniformly at random from an insert-only stream, whose length is unknown. Present the pseudocode and prove the correctness of the proposed algorithm.

Algorithm: Reservoir (stream, k)

/* stream is the insert-only stream mentioned in Q1, k ≥ 1 */

```
res ← [0] * k   /* list with len=k */
for i in range(k):
        res[i] = stream.val
        stream = stream.next
while i < n :
        j = rand (0, i+1)  /* j = random selected
                               in range [0, i] */
        if (j < k):
            res[j] = stream.val
            stream = stream.next
        i += 1
return res
```

Description:

It stores first k elements in stream, and for elements later, if the element is selected then we put it in our return list. The loop ends after traversing every element in stream.

Correctness:

$\because$ each selection is pair-wise independent

$\Pr(\text{selection change}) = \frac{1}{i}$ $\quad \Pr(\text{unchange}) = 1 - \frac{1}{i} = \frac{i-1}{i}$

$$\therefore \Pr(\text{select old ele}) = \frac{1}{k} \prod_{i=k+1}^{N} \frac{i-1}{i}$$

$$= \frac{1}{k} \cdot \frac{k}{k+1} \cdot \frac{k+1}{k+2} \cdot \dots \cdot \frac{N-1}{N}$$

$$= \frac{1}{N}$$

Complexity: Time: $O(n)$ Space: $O(k)$

2 Median trick - a useful technique [15 points]

Prove the claim on slide 13. Be specific about the values of the constants C_1, C_2 you use in your proof, where $t = C_1 \log \frac{1}{\delta}$, $k = C_2 \frac{\mathrm{Var}[X]}{\epsilon^2 \mathbb{E}[X]^2}$.

$\because C_2 \frac{Var(X)}{\varepsilon^2 Q^2} = k \quad \therefore \frac{Var(X)}{\varepsilon^2 Q^2} = \frac{k}{C_2}$

$Pr(1 - \varepsilon Q \leq Y_i \leq 1 + \varepsilon Q) \leq \frac{Var(Y_i)}{\varepsilon^2 Q^2} = \frac{\frac{1}{k} Var(X)}{\varepsilon^2 Q^2} = \frac{1}{k} \cdot \frac{k}{C_2} = \frac{1}{C_2}$

$R_i \begin{cases} 1 & \text{fall in the range w.p. } \frac{1}{C_2} \\ 0 & , \text{o.w} \end{cases}$

$$\sum_{i=1}^{t} R_i = \frac{t}{C_2} = \mu$$

$$Pr\left(R \geq \frac{t}{2}\right) = Pr\left(\left|R - \frac{t}{C_2}\right| \geq \frac{t}{2} - \frac{t}{C_2}\right) = Pr\left(\left|R - \frac{t}{C_2}\right| \geq \left(\frac{C_2}{2} - 1\right)\frac{t}{C_2}\right)$$

$$\leq 2e^{-\frac{(\frac{C_2}{2}-1)^2}{3} \cdot \frac{t}{C_2}} = 2e^{-\frac{(\frac{C_2}{2}-1)^2}{3} \cdot \frac{C_1 \log \frac{1}{\delta}}{C_2}}$$

$\because 0 < \frac{C_2}{2} - 1 < 1$

$\therefore C_2 \in (2, 4)$

when C_2 not in $(2,4)$, C_1 $\boxed{}$?

the claim is true.

3 Variance of Morris Counter [20 points]

Prove equation $\mathrm{Var}(Z) = \frac{m(m-1)}{2}$ on slide 47.

$$Var = E(2^{2X_n}) - [E(2^{X_n})]^2$$

$$E(2^{2X_n}) = \sum_{j=0}^{\infty} 2^{2j} P(X_n = j)$$

$$= \sum_{j=0}^{\infty} 2^{2j} \left(\frac{1}{2^{j-1}} P(X_{n-1} = j-1) + \left(1 - \frac{1}{2^j}\right) \cdot P(X_{n-1} = j) \right)$$

$$= \sum_{j=0}^{\infty} 2^{j+1} P(X_{n-1} = j-1) + \sum_{j=0}^{\infty} 2^{2j} P(X_{n-1} = j) - \sum_{j=0}^{\infty} 2^{j} P(X_{n-1} = j)$$

$$= 4 \cdot E(2^{X_{n-1}}) + E(2^{2X_{n-1}}) - E(2^{X_{n-1}})$$

$$= 3E(2^{X_{n-1}}) + E(2^{2X_{n-1}})$$

$$= 3 \cdot (n+1-1) + E(2^{2X_{n-1}})$$

$$= 3n + E(2^{2X_{n-1}})$$

Base: $E(2^{X_0}) = 1$

n: $E(2^{2X_n}) = \sum_{i=1}^{n} 3i + 1 = \frac{3}{2} n(n+1) + 1$

$$Var = E(2^{2X_n}) - [E(2^{X_n})]^2 = \frac{3}{2} n(n+1) + 1 - (n-1)^2$$

$$= \frac{3}{2} n^2 + \frac{3}{2} n + 1 - n^2 - 2n - 1$$

$$= \frac{1}{2} n^2 - \frac{1}{2} n = \frac{n(n-1)}{2}$$

4 More on uniform RVs [10+10 points]

Let $X_1, \ldots, X_n$ be iid uniform random variables, $X_i \in U(0,1)$ for all i. (a) What is the pdf and (b) what is the expectation of the k-th smallest value among $X_1, \ldots, X_n$ for $k = 1, \ldots, n$?

a) $F(x) = Pr(\min X_1, \dots, X_n \le x) = 1 - Pr(X_1 > x, X_2 > x, \dots X_n > x)$

$= 1 - Pr(X_1 > x)\cdot Pr(X_2 > x) \cdots Pr(X_n > x)$ ∵ iid RV.

$= 1 - [1 - F_1(x)]\cdot[1 - F_2(x)] \cdots [1 - F_n(x)]$

$= 1 - [1 - F_1(x)]^n$ since $X_i \in U(0,1)$ for all i

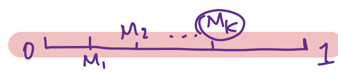
∵ uniform

$$1 - F_1(x) = \begin{cases} 1 & x < 1 \\ 1 - x & x \in (0,1) \\ 0 & x \ge 1 \end{cases}$$

$$F(x) = \begin{cases} 1 & x < 1 \\ (1-x)^n & x \in (0,1) \\ 0 & x \ge 1 \end{cases}$$

$$f(x) = \begin{cases} n(1-x)^{n-1} & x \in (0,1) \\ 0, & o.w. \end{cases}$$

b) $F_k(x) = Pr(V_k \le x) = \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{-l} = \Delta$



$f_k(x) = \Delta'$

$= \frac{d}{dx} \sum_{l=k}^{n} \binom{n}{l} \cdot x^l (1-x)^{n-l}$

$= \sum_{l=k}^{n} \binom{n}{l} \left[l \cdot x^{l-1} \cdot (1-x)^{n-l} - x^l \cdot (n-l)(1-x)^{n-l-1} \right]$

$= \sum_{l=k}^{n} \binom{n}{l} l \cdot x^{l-1} \cdot (1-x)^{n-l} - \sum_{l=k}^{n} \binom{n}{l} x^l (n-l)(1-x)^{n-l-1}$

$= \sum_{l=k}^{n} \binom{n}{l} l \cdot x^{l-1} \cdot (1-x)^{n-l} - \sum_{l=k}^{n-1} \binom{n}{l} (n-l)\, x^l (1-x)^{n-l-1}$

① $\binom{n}{l} l = \frac{n!}{(n-l)!\,l!} \cdot l = \frac{n \cdot (n-1)!}{(n-l)!\,(l-1)!} = n \cdot \binom{n-1}{l-1}$

② $(n-l)\binom{n}{l} = \frac{n!}{(n-l)!\,l!} \cdot (n-l) = \frac{n \cdot (n-1)!}{(n-l-1)!\,l!} = n \cdot \binom{n-1}{l}$

$= \sum_{l=k}^{n} n \cdot \binom{n-1}{l-1} x^{l-1} (1-x)^{n-l} - \sum_{l=k}^{n-} n \cdot \binom{n-1}{l} x^l (1-x)^{n-l-1}$

$= \sum_{i=k-1}^{n-1} n \binom{n-1}{i} x^i (1-x)^{n-i-1} - \sum_{l=k}^{n-1} n \binom{n-1}{l} x^l (1-x)^{n-l-1}$

$= n \cdot \binom{n-1}{k-1} x^{k-1} (1-x)^{n-(k-1)-1} + \sum_{i=k}^{n-1} n \binom{n-1}{i} x^i (1-x)^{n-i-1} - \sum_{l=k}^{n-1} n \binom{n-1}{l} x^l (1-x)^{n-l-1}$

$= n \binom{n-1}{k-1} x^{k-1} (1-x)^{(n-k+1)-1}$ ⇒ beta distribution

$\alpha = k, \ \beta = n+1-k$

$E_k(x) = \frac{\alpha}{\alpha+\beta} = \frac{k}{k+n+1-k} = \frac{k}{n+1}$