(1)

In this question, we have to present the pseudocode of an algorithm that samples $k \geq 1$ elements uniformly at random from an insert-only stream with unknown length.

Pseudocode:

```
lst = [ ]
index = 1    //Start from 1
while stream[index] exists :
    if index <= k :
        lst.append(stream[index])
    else:
        rf = random float from 0 to 1
        if rf < k/index:
            replace one element from lst with stream[index]
    index ++
```

Proof of Correctness

Base case: $\boxed{\text{index} <= k}$

The list can store k elements. Therefore, with probability of 1, it stores the first k elements from the stream. If the length of stream is less than the k value, we simply return every element in the stream.

Inductive Hypothesis:

We assume that the probability of replacing an element in the list with stream[index] is k/index for every element in the stream after k^{th} element

Induction:

The first index elements enter the list with probability k/index based on our hypothesis. This means that the next element (index) goes into the list with probability $k/(\text{index}+1)$. If the $(\text{index}+1)$th element enters the list, we need to take an element out of the list, randomly, which has length of k. Such probability is $1/k$. Therefore, the probability of the $(\text{index}+1)$th element entering the list is $1/k\,(k/(\text{index}+1)) = 1/(\text{index}+1)$.

The complementary, the probability of $(\text{index}+1)$th element not entering into the list, is therefore

$$1 - 1/(\text{index}+1) = \frac{\text{index}+1-1}{\text{index}+1} = \frac{\text{index}}{\text{index}+1}.$$

Finally, the probability that any element stays in the list is Pr(not replaced by stream)$\cdot$P(entered the list some point), which equals

$$\frac{k}{\text{index}}\left(\frac{\text{index}}{\text{index}+1}\right) = \frac{k}{\text{index}+1},$$

which matches with our assumption.

Furthermore, as we run through the end of stream, the value goes to

$$\frac{k}{\text{index}}\left(\frac{\text{index}}{\text{index}+1}\right)\left(\frac{\text{index}+1}{\text{index}+2}\right)\cdots\left(\frac{n-1}{n}\right)$$

$= k/n$, where n is the length of the stream.

(2)

Here, we are proving that $Pr(|Z-Q| \geq \varepsilon Q)$ by using Chernoff and Chebyshev bound where $Z = \text{median}_{i \in [t]} \frac{1}{k} \cdot \sum_{j=1}^{k} x_{ij}$ and

$t = C_1 \cdot \log\left(\frac{1}{\delta}\right)$, $k = \dfrac{C_2 \, Var[x]}{\varepsilon^2 E[x]^2}$

Let $A_i = \frac{1}{k} \cdot \sum_{j=1}^{k} x_{ij}$ for $1 \leq i \leq t$

$E[A_i] = Q$

$Var[A_i] = Var\left(\frac{1}{k}\sum_{j=1}^{k} x_{ij}\right) = \frac{1}{k^2} Var\left(\sum_{j=1}^{k} x_{ij}\right)$

- Since all X_s are independent RVs,

$Var[A_i] = \frac{1}{k^2} Var\left(\sum_{j=1}^{k} x_{ij}\right) = \frac{1}{k^2}\sum_{j=1}^{k} Var[x] = \frac{1}{k^2} \cdot (k \, Var[x]) = \dfrac{Var[x]}{k}$

When we apply Chebyshev inequality, we get

$$Pr(|A_i - Q| \geq \varepsilon Q) \leq \frac{Var[A_i]}{\varepsilon^2 Q^2} = \frac{Var[x]}{k \varepsilon^2 Q^2}$$

According to the question, $k = \dfrac{C_2 \, Var[x]}{\varepsilon^2 E[x]^2} = \dfrac{C_2 \, Var[x]}{\varepsilon^2 Q^2}$

When we plug in the values,

$$Pr(|A_i - Q| \geq \varepsilon Q) \leq \frac{Var[x]}{\frac{C_2 \, Var[x]}{\varepsilon^2 Q^2} \varepsilon^2 Q^2} = \frac{1}{C_2}$$

Now, let $C_2 = 4$

Note that Z refers to the median, meaning that Z is greater than half of the values.

↳ $Z > t/2$

Let $B_i \begin{cases} 1 & \text{if } |A_i - Q| \geq \varepsilon Q \text{ for } 1 \leq i \leq t \\ 0 & \text{otherwise} \end{cases}$

Z, therefore, is $\sum_{i=1}^{t} B_i$

The expected value of $\sum_{i=1}^{t} B_i$ is bounded to $t \cdot \frac{1}{c_2} = t/4$

Since for each B_i where $1 \leq i \leq t$, it is 1 with probability bounded to $1/c_2 = 1/4$

↳ $np = t\left(\frac{1}{c_2}\right) = t/4$

Now,

$Pr(Z > t/2)$ from the fact that Z is the median

According to Wikipedia and TA,

one version of the Chernoff bound is

$Pr(X > (1+\delta')\mu) \leq e^{\frac{-(\delta')^2 \cdot \mu}{2+\delta'}}$ and $\delta' \geq 0$

Therefore, we need to set $t/2 = (1+\delta')\mu$

$= (1+\delta') t/4$

$\delta' = 1$

Substituting the values,

$$Pr(Z > t/2)$$
$$= Pr(Z > (1+\delta')\, t/4) \leq e^{-\frac{(\delta')^2 (t/4)}{2+\delta'}} \quad \text{and } \delta' = 1$$
$$= e^{-\frac{(t/4)}{3}}$$
$$= e^{-t/12} = \delta$$

The question states $t = c_1 \cdot \log\left(\frac{1}{\delta}\right)$

Plugging in, we get

$$e^{-c_1 \log\left(\frac{1}{\delta}\right)/12} = \delta$$

In order for this equality to hold, $c_1 = 12$ since $e^{-\log\left(\frac{1}{\delta}\right)} = \delta$

Therefore, if $c_2 = 4$ and $c_1 = 12$, we proved that $Pr(|Z - Q| \geq \varepsilon Q) \leq \delta$.

(3) $Var(Z) = E(Z^2) - E(Z)^2$

From class, we proved that $E(Z) = m+1$

$Var(Z) = E(Z^2) - (m+1)^2$

$= E(Z^2) - (m^2+2m+1)$

Therefore, we need to find $E(Z^2)$

$Z = 2^{X_m}$

$Z^2 = (2^{X_m})^2$

$= 2^{2X_m}$

From the question, we are given that $Var(Z) = \frac{m(m-1)}{2}$

$E(Z^2) = Var(Z) + (m^2+2m+1)$

$= \frac{m^2-m}{2} + m^2+2m+1$

$= \frac{3m^2}{2} + \frac{3m}{2} + 1$

$= \frac{3m(m+1)}{2} + 1$

Therefore we need to prove that $E[2^{2X_m}] = \frac{3m(m+1)}{2} + 1$

Claim: Define X_m to be the value of the counter after m increments.

Then, $E[2^{2X_m}] = \frac{3m(m+1)}{2} + 1$

Proof: (induction)

Base case: If $m=0$, $X_m = 0$, thus the claim holds

Inductive step:

$$E[2^{2X_m}] = \sum_{j=0}^{\infty} P(X_{m-1} = j) E[2^{2X_m} \mid X_{m-1} = j]$$

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

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

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

$$= \sum_{j=0}^{\infty} P(X_{m-1}=j)\left(2^{2j}-2^{j}+4\cdot 2^{j}\right)$$

$$= \sum_{j=0}^{\infty} P(X_{m-1}=j)\left(2^{2j}+3\cdot 2^{j}\right)$$

$$= E[2^{2X_{m-1}}]+3E[2^{X_{m-1}}]$$

Notice that the next term depends on the first term

$$E[2^{2X_0}]=1$$
$$E[2^{2X_1}] = 3E[2^{X_{m0}}]+E[2^{2X_0}] = 3+1 = 4$$
$$E[2^{2X_2}] = 3E[2^{X_{m_1}}]+E[2^{2X_1}] = 3E[2^{X_{m_1}}]+3E[2^{X_{m0}}]+E[2^{2X_0}]$$
$$= 3(2)+3(1)+1$$
$$= 6+4 = 10$$

↳ Notice a pattern where

$$E[2^{2X_m}] = \sum_{j=1}^{m} 3j+1, \text{ which equals } 3\sum_{j=1}^{m} j+1.$$

This is $3\cdot\frac{(m)(m+1)}{2}+1$.

↳ $E[2^{2X_m}] = \frac{3(m)(m+1)}{2}+1$.

$$\boxed{\text{Therefore, } Var(Z) = \underbrace{\frac{3m(m+1)}{2}+1}_{E(Z^2)} - \underbrace{(m^2+2m+1)}_{E(Z)^2} = \frac{m(m-1)}{2}}$$

④ (a)

Here, we are solving for the pdf of the k-th smallest value among $X_1 \dots X_n$ for $k = 1 \dots n$ where $X_1 \dots X_n$ are iid uniform random variables and $X_i \in U(0,1)$ for all i.

To find the pdf, we must find the CDF and take the derivative of it.

First, we begin with the smallest hashed value.

Let $V_1 = \min(X_1, \dots X_n)$

$E[V_1] = \int_0^1 Pr(V_1 > t)\, dt = \int_0^1 Pr(X_1 > t)^n\, dt = \int_0^1 (1-t)^n\, dt = \frac{1}{n+1}$

In this question, we are concerned in the pdf of k-th smallest hashed value, V_k, where $k \in (1 \dots n)$.

By using the definition of CDF, we are looking for $Pr(V_k \leq x)$, which also means $Pr(\text{at least } k \text{ observations are } \leq x)$

If there are at least k observations, it means that the lower bound is k, which goes to n, the total number of observations.

↳ In addition, for each observation, it follows a binomial distribution with $p = x$.

$\therefore Pr(V_k \leq x) = \sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l} \rightarrow$ CDF

If we take the derivative, we get $\frac{d}{dx}\left(\sum_{l=k}^{n} \binom{n}{l} x^l (1-x)^{n-l}\right)$

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

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

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

$$= \sum_{l=K}^{n} \frac{n!}{l!(n-l)!} \cdot l \cdot (x)^{l-1} \cdot (1-x)^{n-l} - \sum_{l=K}^{n} \frac{n!}{l!(n-l)!} (x^l)(n-l)(1-x)^{n-l-1}$$

↳ when $l=n$, the value of $n-l$ is 0, which makes all of the values to 0 due to multiplication. Therefore, the last term when $n=l$ can be ignored.

$$= \sum_{l=K}^{n} \frac{n \cdot (n-1)!}{(l-1)!(n-l)!} (x)^{l-1} (1-x)^{n-l} - \sum_{l=K}^{n-1} \frac{n!}{l!(n-l)!} (n-l)(x)^l (1-x)^{n-l-1}$$

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

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

$$= n \binom{n-1}{K-1} x^{K-1} (1-x)^{(n-1)-(K-1)}$$

$$= \frac{n(n-1)!}{(K-1)!(n-K)!} x^{K-1} (1-x)^{(n-1)-(K-1)}$$

$$= \frac{n!}{(K-1)!(n-K)!} x^{K-1} (1-x)^{(n-1)-(K-1)}$$

(b)

From part a, we found that the pdf of k-th smallest value among $X_1 \dots X_n$ for $k = 1 \dots n$ is $\frac{n!}{(k-1)!(n-k)!} x^{k-1} (1-x)^{(n-1)-(k-1)}$

This is simply the pdf of the beta distribution where

$\alpha - 1 = k - 1$

$\beta - 1 = (n-1) - (k-1)$

in $B(\alpha, \beta)$

$\alpha = k$

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

$= n - k + 1$

Therefore, it is the beta distribution $B(k, n-k+1)$.

The expected value of beta distribution $B(\alpha, \beta) = \frac{\alpha}{\alpha + \beta}$

Plugging in, we get $\frac{k}{k + n - k + 1} = \frac{k}{n+1}$

$\therefore$ The expected value of the k-th smallest value among $X_1 \dots X_n$ for $k = 1 \dots n$ is $\frac{k}{n+1}$.