University of Edinburgh
INFR11156: Algorithmic Foundations of Data Science (2025)
Solution 3

Problem 1: Prove that the medium of the returned values from $\Theta(\log(1/\delta))$ independent copies of the BJKST algorithm gives an (ε, δ)-approximation of F_0.

***Solution*:** First, we will show that each instance of the algorithm outputs a good approximation of F_0, with constant probability. Let $X_{r,j}$ be a sequence of indicator random variables such that $X_{r,j} = 1$ if and only if $\rho(h(j)) \geq r$. Also define $Y_r := \sum_{j=1}^{n} X_{r,j}$ so that Y_r denotes the number of items j that reach level r. Smilarly to the analysis of the AMS algorithm, we have that

$$\mathbb{E}(Y_r) = \frac{F_0}{2^r} \quad \text{and} \quad \mathbb{V}(Y_r) \leq \frac{F_0}{2^r}.$$

Let $\bar{z}$ be the final value of z at the end of the algorithm and let Z be the output of the algorithm. It is easy to see that $Z = Y_{\bar{z}} \cdot 2^{\bar{z}}$. We further introduce a parameter s satisfying

$$\frac{\varepsilon^2 F_0}{10} \leq 2^s \leq \frac{\varepsilon^2 F_0}{5}.$$

Notice that such s always exists. Hence we have that

$$\begin{aligned}
\mathbb{P}\left(|Z - F_0| > \varepsilon F_0\right) &= \mathbb{P}\left(|Y_{\bar{z}} \cdot 2^{\bar{z}} - F_0| > \varepsilon F_0\right) \\
&= \mathbb{P}\left(\left|Y_{\bar{z}} - \frac{F_0}{2^{\bar{z}}}\right| > \frac{\varepsilon F_0}{2^{\bar{z}}}\right) \\
&= \mathbb{P}\left(|Y_{\bar{z}} - \mathbb{E}(Y_{\bar{z}})| > \frac{\varepsilon F_0}{2^{\bar{z}}}\right) \\
&= \sum_{z=1}^{\log n} \mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z} \wedge \bar{z} = z\right) \\
&= \sum_{z=1}^{s-1} \mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z} \wedge \bar{z} = z\right) + \sum_{z=s}^{\log n} \mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z} \wedge \bar{z} = z\right) \\
&\leq \sum_{z=1}^{s-1} \mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z}\right) + \sum_{z=s}^{\log n} \mathbb{P}\left(\bar{z} = z\right) \\
&= \sum_{z=1}^{s-1} \mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z}\right) + \mathbb{P}\left(\bar{z} \geq s\right)
\end{aligned}$$

By Chebyshev's inequality we have that

$$\mathbb{P}\left(|Y_z - \mathbb{E}(Y_z)| > \frac{\varepsilon F_0}{2^z}\right) \leq \frac{\mathbb{V}(Y_z)}{\left(\frac{\varepsilon F_0}{2^z}\right)^2} \leq \frac{2^z}{\varepsilon^2 F_0}.$$

Also by construction of the algorithm and Markov's inequality, we know that

$$\mathbb{P}\left(\bar{z} \geq s\right) = \mathbb{P}\left(Y_{s-1} > \frac{100}{\varepsilon^2}\right) \leq \mathbb{E}(Y_{s-1}) \cdot \frac{\varepsilon^2}{100} = \frac{\varepsilon^2 \cdot F_0}{100 \cdot 2^{s-1}}.$$

Therefore we can conclude that

$$\begin{aligned}
\mathbb{P}\left(|Z - F_0| > \varepsilon F_0\right) &\leq \sum_{z=1}^{s-1} \frac{2^z}{\varepsilon^2 F_0} + \frac{\varepsilon^2 \cdot F_0}{100 \cdot 2^{s-1}} \\
&\leq \frac{2^s}{\varepsilon^2 F_0} + \frac{\varepsilon^2 \cdot F_0}{100 \cdot 2^{s-1}} \\
&\leq 2/5,
\end{aligned}$$

where the last inequality holds by the choice of s. We can improve this δ by running $\Theta(\log(1/\delta))$ instances of the algorithm and returning the median of the returned values. Thus BJKST gives an (ε, δ)-approximation for F_0.

Problem 2: Let $Y_1, \ldots, Y_n$ be independent random variables with $\mathbb{P}[Y_i = 0] = \mathbb{P}[Y_i = 1] = 1/2$. Let $Y := \sum_{i=1}^{n} Y_i$ and $\mu := \mathbb{E}[Y] = n/2$. Apply the uniform Chernoff Bound to prove it holds for any $0 < \lambda < \mu$ that

$$\mathbb{P}[Y \geq \mu + \lambda] \leq \mathrm{e}^{-2\lambda^2/n}.$$

***<u>Solution</u>*:** Consider the substitution $X_i = 2(Y_i - \mathbb{E}[Y_i])$ and let $X = \sum_{i=1}^{n} X_i$. It is easy to see that $\mathbb{P}[X_i = -1] = [X_i = 1] = 1/2$. We have that

$$X = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} 2(Y_i - \mathbb{E}[Y_i]) = 2\sum_{i=1}^{n} Y_i - 2\mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = 2Y - 2\mathbb{E}[Y] = 2Y - 2\mu.$$

Therefore we see that $Y = \frac{1}{2}X + \mu$ and hence

$$\mathbb{P}[Y \geq \mu + \lambda] = \mathbb{P}\left[\frac{1}{2}X + \mu \geq \mu + \lambda\right] = \mathbb{P}[X \geq 2\lambda] \leq \mathrm{e}^{-(2\lambda)^2/2n} = \mathrm{e}^{-2\lambda^2/n},$$

where the inequality comes from applying the Chernoff Bound to the random variable X.

Problem 3: For any undirected graph $G = (V, E)$ with n vertices, we say three vertices u, v, w form a triangle if there are three edges connecting u, v, w respectively. This problem is to analyse a streaming algorithm for approximately computing the number of triangles in an undirected graph G. To describe the proposed algorithm, let $\mathcal{H}$ be a family of 12-wise independent hash functions, where every $h \in \mathcal{H}$ is of the form $h : V \to \{-1, 1\}$. Let Z be our estimator, which is set to be 0 initially. The algorithm is described as follows:

Algorithm 1 Approximate the number of triangles in G

```
Pick a function $h$ uniformly at random from $\mathcal{H}$;
$Z \leftarrow 0$;
while an edge $\{u, v\}$ arrives do
    $Z \leftarrow Z + h(u) \cdot h(v)$;
end while
Return $Z^3/6$.
```

You need to prove that the returned value $Z^3/6$ is an unbiased estimator of the number of triangles in G, i.e.,

$$\mathbb{E}\left(\frac{Z^3}{6}\right) = \text{number of triangles in } G.$$

Hence, the number of triangles can be approximately counted by running Algorithm 1 above multiple times in parallel and returning the medium of the returned values.

***<u>Solution</u>*:** We have that

$$\begin{aligned}
\mathbf{E}\left[\, Z^3 \,\right] &= \mathbf{E}\left[\left(\sum_{e=\{u,v\}} h(u)h(v)\right)^3\right] \\
&= \mathbf{E}\left[\sum_{e_1=\{u_1,v_1\}} \sum_{e_2=\{u_2,v_2\}} \sum_{e_3=\{u_3,v_3\}} \prod_{i=1}^{3} h(u_i)h(v_i)\right] \\
&= \sum_{e_1=\{u_1,v_1\}} \sum_{e_2=\{u_2,v_2\}} \sum_{e_3=\{u_3,v_3\}} \mathbf{E}\left[\prod_{i=1}^{3} h(u_i)h(v_i)\right],
\end{aligned}$$

where the last equality comes from the linearity of the expectation. We will now argue that the last formulation is exactly 6 times the number of triangles in G. Under expectation, only the terms with products of even powers of $h(u_i)$ and $h(v_i)$ survive. In a combination of three edges $e_i = \{u_i, v_i\}$, every vertex u_i or v_i is connected to at most three other vertices. Moreover, the power of each $h(x_i)$ is the number of times x_i appears in the combination. We see that only the terms where each vertex appears exactly twice survive, which can only happen if the three edges form a cycle. Since every triangle is counted 6 times (once for every permutation of its edges) wee see that $\mathbf{E}\left[Z^3\right]$ equals to 6 times the number of triangles in G.

Problem 4: We are given two independent streams of elements from $\{1, \ldots, n\}$, and we only consider the cash register model. Let $A[1, \ldots, n]$ and $B[1, \ldots, n]$ be the number of occurrences of item i in two streams, respectively. Design a streaming algorithm to estimate $X = \sum_{i=1}^{n} A[i]B[i]$ with additive error $\varepsilon \cdot \|A\|_1 \cdot \|B\|_1$. You need to analyse the space complexity of your proposed algorithm, and analyse the correctness of your algorithm.

***Solution*:** The algorithm follows the framework of the Count-Min sketch. We will make use of two tables C and D, each of size $d \times w$, where $d = \lceil \log(1/\delta) \rceil$ and $w = \mathrm{e}/\varepsilon$. The i-th row of each table corresponds to a hash function $h_i : [n] \to [w]$ chosen from a family of unievrsal hash functions. The two tables support two operations *Insert(x)* and *Query* as follows:

Algorithm 2 *Insert(x)*

```
Result: Inserts a new element $x$ from the stream
for $i = 1, d$ do
    Compute $h_i(x)$
    if $x$ is from the first stream then
        $C[i, h_i(x)] \leftarrow C[i, h_i(x)] + 1$
    else
        $D[i, h_i(x)] \leftarrow D[i, h_i(x)] + 1$
    end if
end for
```

Algorithm 3 *Querry*

```
Result: Provides the answer to the querry $X = \sum_{i=1}^{n} A[i]B[i]$
Return $X' := \min_{1 \le i \le d} C[i]D[i]$, where $C[i]D[i] = \sum_{j=1}^{w} C[i,j]D[i,j]$
```

By construction, for any $x \in [n]$ and any row i, x will be mapped to the same column $h_i(x)$ in the two tables. Thus, when computing the dot product $C[i]D[i]$, we are guaranteed to have the sum $A[x]B[x]$. By taking the minimum over all i's and taking into account the values in the two vectors are nonnegative, it follows that $X' \geq X$.

For the other direction, we will prove that with constant probability $1 - \delta$ we have that $X' \leq X + \varepsilon \|A\|_1 \|B\|_1$. Fix a row i and suppose $C[i]D[i] = \sum_{i=1}^{n} A[i]B[i] + Z_i$, where Z_i is the excess obtained from the dot product. Such an excess can occur if and only if we encounter collisions of the hash function. Namely, whenever two distinct $x, y \in [n]$ are such that $h_i(x) = h_i(y) = z$, computing $C[i, z]D[i, z]$ yields an excess of $A[x]B[y] + A[y]B[x]$. Hence, we conclude that

$$Z_i = \sum_{\substack{x \neq y \\ h_i(x) = h_i(y)}} A[x]B[y].$$

Since we used universal hash functions, it follows that $\forall x \neq y$,

$$\mathbf{P}[h_i(x) = h_i(y)] \leq \frac{1}{w} = \frac{\varepsilon}{\mathrm{e}}.$$

This, in turn, implies that

$$\mathbf{E}\left[\, Z_i \,\right] = \sum_{x \neq y} \mathbf{P}[h_i(x) = h_i(y)] A[x] B[y] \le \frac{\varepsilon}{\mathrm{e}} \left\| A \right\|_1 \left\| B \right\|_1 .$$

To complete the proof, observe that

$$\mathbf{P}\left[\, X' > X + \varepsilon \left\| A \right\|_1 \left\| B \right\|_1 \,\right] = \mathbf{P}[\forall i : Z_i > \varepsilon \left\| A \right\|_1 \left\| B \right\|_1] \le \mathbf{P}\left[\forall i : Z_i > \mathrm{e}\mathbf{E}\left[\, Z_i \,\right]\right] \le \mathrm{e}^{-d} \le \delta,$$

where the last inequality is obtained by applying Markov's inequality. The space used by the algorithm is essentially dominated by the two tables used to store the number of appearances of the elements in the two streams, which is $O(wd) = O\left(\frac{1}{\varepsilon} \log(1/\delta)\right)$.