University of Edinburgh
INFR11156: Algorithmic Foundations of Data Science (2019)
Solution 5

Problem 1: Let $H = \{h : [m] \to \{0,1\}^n\}$ be a family of pairwise independent hash functions. Let $I \subseteq [m]$ and $\mu := \frac{|I|}{2^n}$. Then, it holds for every $y \in \{0,1\}^n$ that

$$\mathbb{P}_{h\sim H}\Big[\big||\{i \in I : h(i) = y\}| - \mu\big| > \varepsilon\mu\Big] < \frac{1}{\varepsilon^2\mu},$$

where $h \sim H$ stands for the fact that h is chosen uniformly at random from H.

Solution: We fix an arbitrary $y \in \{0,1\}^n$, and for any $i \in I$ defined a random variable X_i, where $X_i = 1$ if $h(i) = y$, and $X_i = 0$ otherwise. Since H is a family of pairwise independent hash functions, we have that

$$\mathbb{P}[X_i = 1] = 1/2^n,$$

which implies that $\mathbb{E}[X_i] = 1/2^n$ and

$$\mathbb{V}[X_i] = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 \leq \mathbb{E}[X_i].$$

Moreover, we have that

$$\mathbb{E}\left[\sum_{i\in I} X_i\right] = \frac{|I|}{2^n} = \mu,$$

and

$$\mathbb{V}\left[\sum_{i\in I} X_i\right] = \sum_{i\in I} \mathbb{V}[X_i] \leq \sum_{i\in I} \mathbb{E}[X_i] = \mu.$$

By applying the Chebshev's inequality we have that

$$\begin{aligned}
&\mathbb{P}_{h\sim H}\Big[\big||\{i \in I : h(i) = y\}| - \mu\big| > \varepsilon\mu\Big] \\
&= \mathbb{P}_{h\sim H}\left[\left|\sum_{i\in I} X_i - \mathbb{E}\left[\sum_{i\in I} X_i\right]\right| > \varepsilon\mu\right] \\
&\leq \frac{1}{(\varepsilon\mu)^2} \cdot \mathbb{V}\left[\sum_{i\in I} X_i\right] \\
&\leq \frac{1}{(\varepsilon\mu)^2} \cdot \mu \\
&= \frac{1}{\varepsilon^2\mu},
\end{aligned}$$

which proves the statement.

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

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

Solution: 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: Prove that the median 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.