CS 498ABD Spring 2019 — Homework 2 Solutions

Exercise 2. This is mainly to make you work out a simple distinct elements analysis for yourself. Here is a variant of the algorithm we saw in lecture. Instead of using an ideal hash function h we choose a random hash function $h : [n] \to [n]$ from pairwise-independent hash family $\mathscr{H}$. Let Z be the minimum hash value seen in the stream. Suppose the number of distinct elements d is in the range $[2^i, 2^{i+1})$. Prove that $\mathrm{P}[Z \in [n/2^{i+3}, n/2^{i-2})] > c$ for some fixed constant c. Thus n/Z gives a constant factor estimate for d with probability at least c.

Solution: We make use of the observation that $Z \in [n/2^{i+3}, n/2^{i-2})$ if and only if nothing gets hashed into $[1, n/2^{i+3})$, and something gets hashed into $[1, n/2^{i-2})$. Specifically, we compute lower bounds on the probability of each event individually, and then use the union bound to lower bound $\mathrm{P}[Z \in [n/2^{i+3}, n/2^{i-2})]$:

$$\mathrm{P}[A \wedge B] = 1 - \mathrm{P}[\bar{A} \vee \bar{B}] \geq 1 - \mathrm{P}[\bar{A}] - \mathrm{P}[\bar{B}].$$

We can bound the probability that nothing is hashed into $[1, n/2^{i+3})$ using Markov's Inequality. Let X_j be indicator random variable for the event that the j-th *distinct* item is hashed into $[1, n/2^{i+3})$, and let $X = \sum_j X_j$ be the random variable for the number of items hashed into $[1, n/2^{i+3})$. Since $\mathrm{P}[X_j = 1] \leq 1/2^{i+3}$, $\mathrm{E}[X] \leq d/2^{i+3}$. By Markov's inequality,

$$\mathrm{P}[X \geq 1] \leq \frac{d}{2^{i+3}} \leq \frac{1}{4}.$$

We can bound the probability that something is hashed into $[1, n/2^{i-2})$ using Chebyshev's inequality. Let Y_j be the indicator random variable for the event that the j-th *distinct* item is hashed into $[1, n/2^{i-2})$, and $Y = \sum_j Y_j$ be the random variable for the number of items hashed into $[1, n/2^{i-2})$. Taking a very loose bound of $\mathrm{P}[Y_j = 1] \geq 1/2^{i-1}$, we get $\mathrm{E}[Y] \geq d/2^{i-1}$, and since the Y_j's are pairwise independent indicator random variables,

$$\mathrm{Var}[Y] = \sum_j \mathrm{Var}[Y_j] \leq \sum_j \mathrm{E}[Y_j^2] = \sum_j \mathrm{E}[Y_j] = \mathrm{E}[Y],$$

By Chebyshev's inequality,

$$\mathrm{P}[Y = 0] = \mathrm{P}[Y \leq \mathrm{E}[Y] - \mathrm{E}[Y]] \leq \mathrm{P}[|Y - \mathrm{E}[Y]| \geq \mathrm{E}[Y] \leq \frac{\mathrm{E}[Y]}{\mathrm{E}[Y]^2} = \frac{1}{\mathrm{E}[Y]} \leq \frac{2^{i-1}}{d} \leq \frac{1}{2}.$$

Putting the whole thing together gives

$$\mathrm{P}[Z \in [n/2^{i+2}, n/2^{i-1})] \geq 1 - \frac{1}{4} - \frac{1}{2} = \frac{1}{4}.$$

■

Solution (More Careful Analysis): Here we will use better analysis than in the previous version to bound the probability that Z falls into the smaller interval $[n/2^{i+2}, n/2^{i-1})$. This occurs if and only if nothing gets hashed into $[1, n/2^{i+2})$, and something gets hashed into $[1, n/2^{i-1})$. For brevity, we will omit details that are analagous to those in the previous analysis.

- If $n \leq 2^{i+2}$, the interval $[1, n/2^{i+2})$ is empty so $\mathrm{P}[X \geq 1] = 0$.
  Since $\mathrm{P}[Y_j = 1] \geq 1/2^{i-1} - 1/n$, $\mathrm{E}[Y] \geq d/2^{i-1} - d/n \geq d/2^{i-1} - 1$, so by Chebyshev's inequality,

  $$\mathrm{P}[Y = 0] \leq \frac{1}{\mathrm{E}[Y]} \leq \frac{2^{i-1}}{d - 2^{i-1}} \leq \frac{1}{2},$$

  and thus

  $$\mathrm{P}[Z \in [n/2^{i+2}, n/2^{i-1})] \geq \frac{1}{2}.$$

- Now suppose $n > 2^{i+2}$.

  Since $\mathrm{P}[X_j = 1] \le 1/2^{i+2}$, $\mathrm{E}[X] \le d/2^{i+2}$. By Markov's inequality,

  $$\mathrm{P}[X \ge 1] \le \frac{d}{2^{i+2}}.$$

  Since $\mathrm{P}[Y_j = 1] \ge 1/2^{i-1} - 1/n$, $\mathrm{E}[Y] \ge d/2^{i-1} - d/n \ge d/2^{i-1} - d/2^{i+2} = 7d/2^{i+2}$. By Chebyshev's inequality,

  $$\mathrm{P}[Y = 0] \le \frac{1}{\mathrm{E}[Y]} \le \frac{2^{i+2}}{7d}.$$

  By the union bound,

  $$\mathrm{P}[Z \in [n/2^{i+2}, n/2^{i-1})] \ge 1 - \frac{d}{2^{i+2}} - \frac{2^{i+2}}{7d}.$$

  Since $d \in [2^i, 2^{i+1})$, $d = \alpha 2^i$ for some $\alpha \in [1,2)$. Substituting this into the right hand side, we want to lower bound the function $f(\alpha) = 1 - \frac{\alpha}{4} - \frac{4}{7\alpha}$ on the interval $[1,2]$. f has negative second derivative on the interval $\alpha \in [1,2]$ and thus takes its minimum at one of the endpoints. $f(1) = \frac{5}{28}$ and $f(2) = \frac{3}{14}$, so we take the smaller of the two and obtain

  $$\mathrm{P}[Z \in [n/2^{i+2}, n/2^{i-1})] \ge \frac{5}{28}.$$

In either case, $\mathrm{P}[Z \in [n/2^{i+2}, n/2^{i-1})] \ge c$ for some $c > 0$, as desired. ∎

Exercise 3. We saw how to estimate the number of distinct elements from a stream. Now we consider the problem of sampling nearly uniformly from the set of distinct elements. Consider the BJKST algorithm we saw in lecture that used a random hash function $h : [n] \to [n^3]$ from a pairwise independent hash family $\mathscr{H}$. Now assume that we instead use a 3-wise independent hash family. The algorithm stores $t = 1/\epsilon^3$ elements associated with the smallest t hash values seen and at the end of the algorithm outputs one of them uniformly at random. Our goal is to show that each element $i \in [n]$ is output with probability at least $(1-10\epsilon)/d$ (i.e., nearly uniform). (You may assume that ϵ is smaller than some constant like $1/2$ if it makes the calculations easier.)

To this end, let $b_1, b_2, \dots, b_d$ are the distinct values in the stream. Assume $d > 1/\epsilon^3$ for otherwise we can store all of them and output a uniformly sampled element. Observe that an element b_i is among the t remaining elements if each of the following events all occur.

(a) $h(b_i) \le \lceil (1-\epsilon)tN/d \rceil$.

(b) The number of other elements b_j ($j \neq i$) such that $h(b_j) < \lceil (1-\epsilon)tN/d \rceil$ is at most $t-1$.

(c) $h(b_j) \neq h(b_i)$ for all $j \neq i$.

Show that the above events all occur with probability close to t/d via the following steps.

1. Let Z be an indicator for $h(b_i) \le \lceil (1-\epsilon)tN/d \rceil$. What is $\mathrm{P}[Z=1]$?
2. Conditioned on $Z=1$, show that the probability that more than $t-1$ of the remaining items (b_j where $j \neq i$) has value $< \lceil (1-\epsilon)tN/d \rceil$ is at most $(1+\epsilon)\epsilon$.
3. Show that the probability of a hash collision with b_i is at most $\epsilon t/d$.
4. Put the above together to show that b_i is one of the t selected elements with probability $\ge (1-10\epsilon)t/d$, hence output with probability $\ge \frac{1-10\epsilon}{d}$.
5. **Extra credit:** Show that the probability that b_i is output is at most $(1+10\epsilon)/d$.

Make note in particular of where you use the assumption that h is 3-wise independent.

Solution: Let $Z \in \{0,1\}$ indicate the event that $h(b_i) \le \lceil (1-\epsilon)tN/d \rceil$. For $j \neq i$, let $Y_j \in \{0,1\}$ indicated the event that $h(b_j) < \lceil (1-\epsilon)tN/d \rceil$. If $Z=1$, $\sum_{j\neq i} Y_j < t$, and $h(b_j) \neq h(b_i)$ for all $j \neq i$, then i will be included in the output. We claim that for $t = 1/\epsilon^3$, this probability is at least $(1-10\epsilon)/d$.

We first observe that

$$\mathrm{P}[Z=1] = \frac{\lceil (1-\epsilon)tN/d \rceil}{N} \ge (1-\epsilon)\frac{t}{d}. \tag{1}$$

Now consider the sum $\sum_{j\neq i} Y_j$. For each $j \neq i$, since $h(b_j)$ is independence of $h(b_i)$, we have

$$\begin{aligned} \mathrm{E}\big[Y_j \,\big|\, Z\big] &= \mathrm{P}\big[h(b_j) \le \lceil (1-\epsilon)tN/d \rceil - 1\big] \\ &= \frac{\lceil (1-\epsilon)tN/d \rceil - 1}{N} \le (1-\epsilon)\frac{t}{d}. \end{aligned} \tag{2}$$

The expected sum is

$$\mathrm{E}\left[\sum_{j\neq i} Y_j \,\middle|\, Z\right] \sum_{j\neq i} \mathrm{E}\big[Y_j \,\big|\, Z\big] \overset{(a)}{\le} (1-\epsilon)\frac{d-1}{d} t \le (1-\epsilon)t \tag{3}$$

by (a) inequality (2). The expected variance is

$$\mathrm{Var}\left[\sum_{j\neq i} Y_j \,\middle|\, Z\right] \overset{(b)}{=} \sum_{j\neq i} \mathrm{Var}\big[Y_j \,\big|\, Z\big] \overset{(c)}{\le} \sum_{j\neq i} \mathrm{E}[Y_i \,|\, Z] \overset{(d)}{\le} (1-\epsilon)t. \tag{4}$$

(b) since h is 3-wise independent, (c) $Y_j \in \{0,1\}$, and (d) inequality (3). Now

$$\begin{aligned}
\mathrm{P}\left[\sum_{j\neq i} Y_j > t-1 \,\middle|\, Z\right] &= \mathrm{P}\left[\sum_{j\neq i} Y_j \geq t \,\middle|\, Z\right] = \mathrm{P}\left[\sum_{j\neq i} Y_j - \mathrm{E}\left[\sum_{j\neq i} Y_j\right] \geq t - \mathrm{E}\left[\sum_{j\neq i} Y_j\right] \,\middle|\, Z\right] \\
&\overset{\text{(e)}}{\leq} \mathrm{P}\left[\sum_{j\neq i} Y_j - \mathrm{E}\left[\sum_{j\neq i} Y_j\right] \geq \epsilon t \,\middle|\, Z\right] \leq \mathrm{P}\left[\left|\sum_{j\neq i} Y_j - \mathrm{E}\left[\sum_{j\neq i} Y_j\right]\right| \geq \epsilon t \,\middle|\, Z\right] \\
&\overset{\text{(f)}}{\leq} \frac{\mathrm{Var}\left[\sum_{j\neq i} Y_j \,\middle|\, Z\right]}{\epsilon^2 t^2} \overset{\text{(g)}}{\leq} \frac{(1+\epsilon)t}{\epsilon^2 t^2} \overset{\text{(h)}}{=} (1+\epsilon)\epsilon
\end{aligned} \tag{5}$$

by (e) the upper bound on the mean in (3), (f) Chebyshev's inequality, (g) the upper bound on the variance (4) above, and (h) plugging in $t = 1/\epsilon^3$.

Now we have

$$\begin{aligned}
\mathrm{P}\left[Z = 1 \text{ and } \sum_{j\neq i} Y_j \leq t-1\right] &= \mathrm{P}[Z=1]\mathrm{P}\left[\sum_{j\neq i} Y_j \leq t-1 \,\middle|\, Z=1\right] \\
&= \mathrm{P}[Z=1]\left(1 - \mathrm{P}\left[\sum_{j\neq i} Y_j > t-1 \,\middle|\, Z=1\right]\right) \\
&\geq (1-\epsilon)\frac{t}{d}(1-(1+\epsilon)\epsilon).
\end{aligned}$$

The probability that $Z = 1$, $\sum_{j\neq i} Y_j \leq t-1$, and $h(b_j) \neq h(b_i)$ for all $j \neq i$, is

$$\begin{aligned}
&\mathrm{P}\left[Z = 1, \sum_{j\neq i} Y_j \leq t-1, h(b_j) \neq h(b_i) \text{ for all } j \neq i\right] \\
&\overset{\text{(i)}}{\geq} \mathrm{P}\left[Z = 1, \sum_{j\neq i} Y_j \leq t-1\right] - \mathrm{P}\left[h(b_j) = h(b_i) \text{ for some } j \neq i\right] \\
&\overset{\text{(j)}}{\geq} \mathrm{P}\left[Z = 1, \sum_{j\neq i} Y_k \leq t-1\right] - \sum_{j\neq i} \mathrm{P}\left[h(b_j) = h(b_i)\right] \\
&\overset{\text{(k)}}{\geq} (1-\epsilon)(1-(1+\epsilon)\epsilon)\frac{t}{d} - \frac{(d-1)}{N} \\
&\overset{\text{(l)}}{\geq} (1-\epsilon)(1-(1+\epsilon)\epsilon)\frac{t}{d} - \epsilon\frac{t}{d}.
\end{aligned}$$

by (i,j) the union bound, (k) h is pairwise independent, and (l) uses $(d-1)/N \leq 1/n^2 \leq 1/\epsilon^2 d = \epsilon t/d$ since $\epsilon^2 d \leq n^2$. One can now verify that for sufficiently small $\epsilon > 0$,

$$(1-\epsilon)(1-(1+\epsilon)\epsilon) - \epsilon \geq 1 - 10\epsilon.$$

(In fact, $(1-\epsilon)(1-(1+\epsilon)\epsilon) - \epsilon = (1-\epsilon)\left(1-\epsilon-\epsilon^2\right) - \epsilon = 1 - 3\epsilon + \epsilon^3 \geq 1 - 3\epsilon$.) ■

Problem 4. In class, we saw how to the AMS sampling procedure allows us to estimate the kth moment F^k in sublinear space for $k \geq 2$. Recall also that the AMS sampler still requires polynomial space because the variance of a single sample was polynomial. Here we will use AMS to estimate the *entropy* of a stream; in particular, we will show that we only need *logarithmic space* (modulo dependencies on ϵ and δ) to estimate the entropy.

Let $f \in \mathbb{Z}_{\geq 0}^n$ be frequency counts over n elements, and for each i, let $p_i = \frac{f_i}{m}$ be the corresponding probability distribution. The *entropy* of p is the quantity $\Phi = \sum_i p_i \ln \frac{1}{p_i}$, where $0 \ln\left(\frac{1}{0}\right) = 0$. Our high-level goal is to obtain an $(1 \pm \epsilon)$-multiplicative approximation to the entropy, but there is a technical issue because the entropy can be zero. We instead seek a $(1 \pm \epsilon)$-multiplicative approximation of $1 + \Phi$, which converts to a $(1 \pm 2\epsilon)$-multiplicative approximation of Φ if $\Phi \geq 1$ and a 2ϵ-additive approximation of Φ if $\Phi \leq 1$. You may assume m is larger than a fixed constant, say 42.

Let $g(\ell) = \frac{\ell}{em} \ln\left(\frac{me}{\ell}\right)$.

1. Show that $1 + \Phi = e \sum_{i \in [n]} g(f_i)$.
2. Show that $g(\ell) \geq g(\ell - 1)$ for $\ell \leq m$.
3. Show that for $\ell \leq m$, $g(\ell) - g(\ell - 1) \leq \frac{1 + \ln m}{me}$
4. Let Y be the AMS sample for the quantity $\frac{1}{e}(1 + \Phi) = \sum_i g(f_i)$. We know from class that $\mathrm{E}[Y] = \frac{1}{e}(1 + \Phi)$. Show that

$$\mathrm{Var}[Y] \leq c_1(1 + \ln(m))(1 + \Phi) = ec_1(1 + \ln(m))\,\mathrm{E}[Y]$$

   for some constant $c_1 > 0$.
5. Let $t = \frac{c(1+\ln m)}{\epsilon^2}$ for some (suitable) constant $c > 0$. For $i \in [t]$, let Y_i be an independent instance of the AMS sample for $\sum_i g(f_i)$, and let $Z = \frac{1}{t} \sum_{i=1}^t Y_i$ be the sum. Show that

$$\mathrm{P}[|eZ - (1 + \Phi)| \geq \epsilon(1 + \Phi)] \leq \frac{1}{4}$$

   for some constant $c > 0$.
6. Finally, explain how to arrange $O\left(\ln(1/\delta)\ln(m)/\epsilon^2\right)$ independent AMS samples to obtain a $(1 \pm \epsilon)$-approximation of $1 + \Phi$ with probability $\geq 1 - \delta$.

Solution:

1. We have

$$1 + \Phi = e\left(\frac{1}{e} + \frac{1}{e}\Phi\right) = e\left(\sum_{i=1}^n \frac{f_i}{em}\left(1 + \ln\left(\frac{m}{f_i}\right)\right)\right)$$
$$= e\sum_{i=1}^n \frac{f_i}{me} \ln\left(\frac{me}{f_i}\right) = e\sum_{i=1}^n g(f_i).$$

2. The derivative of g is

$$g'(x) = \frac{d}{dx}\left(\frac{\ell}{em}\ln\left(\frac{me}{\ell}\right)\right)$$
$$= \frac{\ln(me/\ell)}{me} - \frac{\ell}{me} \cdot \frac{1}{me/\ell} \cdot \frac{me}{\ell^2} = \frac{1}{me}\left(\ln\left(\frac{me}{\ell}\right) - 1\right),$$

   which is ≥ 0 iff $m \geq \ell$.

3. We have

$$g(\ell)-g(\ell-1)=\frac{\ell}{me}\ln\left(\frac{me}{\ell}\right)-\frac{\ell-1}{me}\ln\left(\frac{me}{\ell-1}\right)$$
$$=\frac{\ln(me/\ell)}{me}+\frac{\ell-1}{me}\ln\left(\frac{\ell-1}{\ell}\right)\overset{\text{(a)}}{\leq}\frac{\ln(me)}{me},$$

where (a) observes that $\ln(\ell-1)<\ln(\ell)$. Also

$$g(1)-g(0)=\frac{\ln(me)}{me}$$

4. AMS proof...

$$\begin{aligned}\mathrm{Var}[Z]&=\sum_i\frac{f_i}{m}\sum_{\ell=1}^{f_i}\frac{m^2}{f_i}(g(\ell)-g(\ell-1))^2\\&\leq\sum_i\frac{f_i}{m}\sum_{\ell=1}^{f_i}\frac{m^2}{f_i}\left(\frac{1+\ln(m)}{me}\right)(g(\ell)-g(\ell-1))\\&=\frac{1+\ln(m)}{e}\sum_i^n g(f_i)=\frac{1+\ln(m)}{e^2}(1+\Phi).\end{aligned}$$

We can take $c_1=1/e^2$.

5. For $c_1=1/e^2$, we can take $c=4$. Then we have

$$\mathrm{P}[|eZ-(1+\Phi)|\geq\epsilon(1+\Phi)]=\mathrm{P}\left[|Z-\mathrm{E}[Z]|\geq\frac{\epsilon}{e}(1+\Phi)\right]\overset{\text{(b)}}{\leq}\frac{e^2\,\mathrm{Var}[Z]}{\epsilon^2(1+\Phi)^2}$$
$$\overset{\text{(c)}}{\leq}\frac{(1+\ln(m))}{t\epsilon^2(1+\Phi)}\overset{\text{(d)}}{\leq}\frac{1}{4(1+\Phi)}\overset{\text{(e)}}{\leq}\frac{1}{4}$$

by (b) Chebyshev's inequality, (c) $\mathrm{Var}[Z]=\frac{1}{t}\mathrm{Var}[Y]$, (d) $t=4/\epsilon^2$, and (e) $\Phi\geq 0$. (For larger c_1, c needs to be a little bit larger too.)

6. We use the median trick. Above, we showed that the average of $O\left(1/\epsilon^2\right)$ counters get a $(1\pm\epsilon)$-multiplicative approximation with probability of error at most (say) $1/4$. To amplify the error down to δ, we use $O\left(\log(1/\delta)\right)$ averages each of $O\left(1/\epsilon^2\right)$ samples, and output the median. By the Chernoff inequality, we will be within a $(1\pm\epsilon)$-multiplicative factor with probability at most δ.

■