Algorithms for Big Data Fall 2020

Homework 2 Solutions

Problem 1: Approximate Counting, Remix

In Lectures 4 and 5, we analyzed Morris' algorithm, which approximated the number of updates n by using the following estimator. Initialize a counter X to 0, and for each update, increment X with probability $1/2^X$. Then, the algorithm output $\tilde{n} = 2^X - 1$. To obtain good bounds on the probability that $|\tilde{n} - n| < \varepsilon n$, we considered Morris+ and Morris++ that eventually took the median of many means, where each mean averaged many estimates. Consider a different algorithm, where we still initialize X to 0, but we increment it with probability $1/(1+a)^X$ for some parameter a.

(a) Determine an estimator $\tilde{n}$ as a function of X and a such that it is an unbiased estimator, that is, $\mathbb{E}[\tilde{n}] = n$ after n updates.

(b) How small must a be so that our estimate $\tilde{n}$ of n satisfies $|\tilde{n} - n| < \varepsilon n$ with at least 9/10 probability when we return the output of a single estimator (instead of averaging many estimators as in class)?

(c) Derive a bound S on the space (in bits) as a function of n, ε, a so that this algorithm uses at most S space with at least 9/10 probability after n increments (in addition to satisfying $|\tilde{n} - n| < \varepsilon n$ with probability at least 9/10).

Solution: (a) Analogous to the original Morris algorithm, let's try and calculate $\mathbb{E}[(1+a)^{X_{n+1}}]$ and see if we can come up with an unbiased estimator for $\tilde{n}$.

$$\begin{aligned}
\mathbb{E}[(1+a)^{X_{n+1}}] &= \sum_{j=0}^{\infty} \Pr(X_n = j) \cdot \mathbb{E}((1+a)^{X_n} | X_n = j) \\
&= \sum_{j=0}^{\infty} \Pr(X_n = j) \cdot ((1+a)^j (1 - \frac{1}{(1+a)^j}) + \frac{1}{(1+a)^j} \cdot (1+a)^{j+1}) \\
&= \mathbb{E}[(1+a)^{X_n}] + a.
\end{aligned}$$

Observe that this is a recursive definition,

$$\begin{aligned}
\mathbb{E}[(1+a)^{X_0}] &= 1 \\
\mathbb{E}[(1+a)^{X_1}] &= \mathbb{E}[(1+a)^{X_0}] + 1 = a + 1 \\
\mathbb{E}[(1+a)^{X_2}] &= \mathbb{E}[(1+a)^{X_1}] + 1 = 2a + 1
\end{aligned}$$

This can be written generally as $\mathbb{E}[(1+a)^{X_n}] = na + 1$. Therefore,

$$\tilde{n} = \frac{(1+a)^{X_n} - 1}{a}$$

is an unbiased estimator of n because

$$\mathbb{E}[\tilde{n}] = \mathbb{E}\left[\frac{(1+a)^{X_n} - 1}{a}\right] = \frac{1}{a}(\mathbb{E}[(1+a)^{X_n}] - 1) = n$$

(b) We need to find a bound on a such that for a given ε,

$$\Pr(|\tilde{n} - n| \geq \varepsilon n) \leq \frac{1}{10} \tag{1}$$

According to Chebyshev's Inequality,

$$\Pr(|\tilde{n} - n| \geq \varepsilon n) \leq \frac{\mathrm{Var}(\tilde{n})}{\varepsilon^2 n^2}$$

Now,

$$\mathrm{Var}(\tilde{n}) = \frac{1}{a^2}\,\mathrm{Var}((1+a)^{X_n} - 1) = \frac{1}{a^2}\,\mathrm{Var}((1+a)^{X_n}). \tag{2}$$

Observe,

$$\mathrm{Var}((1+a)^{X_n}) = \mathbb{E}[(1+a)^{2X_n}] - (\mathbb{E}[(1+a)^{X_n}])^2 \tag{3}$$

Similar to the original Morris algorithm ($a = 1$), $\mathbb{E}[(1+a)^{2X_n}]$ is quadratic.
To figure out the values of p, q and r, we can use polynomial interpolation, let's assume that

$$\mathbb{E}[(1+a)^{2X_n}] = pn^2 + qn + r.$$

We evaluate the expectation at three values, $n \in \{0, 1, 2\}$. This gives us three linear equations in p, q and r that can be solved to get the value of the coefficients. Doing this we get,

$$p = \frac{a^3}{2} + a^2 \quad \text{and} \quad q = 2a - \frac{a^3}{2} \quad \text{and} \quad r = 1.$$

(by setting $a = 1$, we get values for the original Morris algorithm, namely $p = q = 3/2$). Overall, we have

$$\mathrm{Var}[(1+a)^X_n] = \Theta(a^3 n^2).$$

We can figure out the constants with some effort, but for the sake of brevity, let k be a constant such that

$$\mathrm{Var}[(1+a)^X_n] \leq k a^3 n^2.$$

Using this in Eq. (2), we get,

$$\mathrm{Var}(\tilde{n}) \leq kan^2$$

Therefore Inequality 1 holds if,

$$\begin{aligned}
\frac{\mathrm{Var}(\tilde{n})}{\varepsilon^2 n^2} &\leq \frac{1}{10}, \\
\frac{kan^2}{\varepsilon^2 n^2} &\leq \frac{1}{10}, \\
a &\leq \frac{\varepsilon^2}{k10}, \\
a &= O(\varepsilon^2).
\end{aligned}$$

(c) In this part, we need to find a bound S on $\log_2(X)$ such that

$$\Pr(\log_2(X) \geq S) \leq \frac{1}{10} \tag{4}$$

Since, $\Pr(x > y) = \Pr(a^x > a^y)$ if $a > 1$

$$\Pr(\log_2(X) \geq S) = \Pr(X \geq 2^S) = \Pr((1+a)^X \geq (1+a)^{2^S})$$

Using Markov's Inequality,

$$\Pr((1+a)^X \geq (1+a)^{2^S}) \leq \frac{\mathbb{E}[(1+a)^X]}{(1+a)^{2^S}}$$

Therefore, Inequality 4 holds if,

$$\begin{aligned}
\frac{\mathbb{E}[(1+a)^X]}{(1+a)^{2^S}} &\leq \frac{1}{10}, \\
na+1 &\leq \frac{(1+a)^{2^S}}{10}, \\
(1+a)^{2^S} &\geq 10(na+1), \\
S &\geq \log_2 \log_{1+a}[10(na+1)] \\
S &= \Omega(\log \log_{1+a}(na))
\end{aligned}$$

In order to also satisfy the constraint in Part (b), a must be $O(\varepsilon^2)$.

Problem 2: Pairwise Independence

(a) Let q be a prime number. For integers $c, d \in \{0, 1, \ldots, q-1\}$, define the hash function $h_{c,d}$ as

$$h_{c,d}(x) = cx + d \bmod q.$$

Let $\mathcal{H}$ be the set of all such hash functions, defined as

$$\mathcal{H} = \big\{h_{c,d} \mid c, d \in \{0, 1, \ldots, q-1\}\big\}.$$

Prove that $\mathcal{H}$ is a pairwise independent hash family. That is, prove that for any distinct $i \neq i'$ and any j, j', we have that

$$\Pr_{h_{c,d} \in \mathcal{H}}[\ h_{c,d}(i) = j \text{ and } h_{c,d}(i') = j'\ ] = \frac{1}{q^2},$$

where $h_{c,d} \in \mathcal{H}$ is chosen uniformly by choosing c, d at random in $\{0, 1, \ldots, q-1\}$.
Hint: Start with $q = 2$ and $\{0, 1\}$ values; then, generalize to all prime $q \geq 2$. For the general case, you can use that $cx + d = j$ has a unique solution in terms of x if $c \neq 0$.

Solution 1: For distinct $i \neq i'$ and any j, j', we need to show that

$$\Pr_{h_{c,d} \in \mathcal{H}}[\ h_{c,d}(i) = j \text{ and } h_{c,d}(i') = j'\ ] = \frac{1}{q^2},$$

Observe that, $c \cdot i + d \bmod q = j$ and $c \cdot i' + d \bmod q = j'$ can also be written in the matrix form as:

$$\begin{bmatrix} i & 1 \\ i' & 1 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} j \\ j' \end{bmatrix}$$

where the addition and multiplication are under mod q. Therefore, c and d can be uniquely determined if the matrix

$$\begin{bmatrix} i & 1 \\ i' & 1 \end{bmatrix}$$

is invertible or non-singular as shown below:

$$\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} i & 1 \\ i' & 1 \end{bmatrix}^{-1} \begin{bmatrix} j \\ j' \end{bmatrix}$$

Luckily, inverse of a 2×2 matrix is easy to compute. Since the addition and multiplication in our universe are defined under mod q, the matrix $\begin{bmatrix} i & 1 \\ i' & 1 \end{bmatrix}$ is invertible as long as the following condition is satisfied:

$$i \neq i' \bmod q.$$

It's easy to see that one such case when this is always true is when $i, i' \in \{0, 1, 2, 3, \cdots, q-1\}$. Let the value of c and d determined using the above matrix equation be c_1 and d_1. Then,

$$\Pr_{h_{c,d} \in \mathcal{H}}[\ h_{c,d}(i) = j \text{ and } h_{c,d}(i') = j'\ ] = \Pr(c = c_1 \text{ and } d = d1).$$

Since c and d are uniformly and independently sampled from $\{0, 1, 2, \cdots, q-1\}$.

$$\begin{aligned} \Pr_{h_{c,d} \in \mathcal{H}}[\ h_{c,d}(i) = j \text{ and } h_{c,d}(i') = j'\ ] &= \Pr(c = c_1) \cdot \Pr(d = d1) \\ &= \frac{1}{q} \cdot \frac{1}{q} \\ &= \frac{1}{q^2}. \end{aligned}$$

When $q = 2$: We provide a direct proof for $q = 2$ and $\{0, 1\}$ values. For x = 1,

$$\Pr_{c \in \{0,1\}}[c \cdot x \bmod 2 = 0] = \frac{1}{2}.$$

For $x_1 \neq x_2 \in \{0, 1\}$ and $y_1, y_2 \in \{0, 1\}$, We need to prove :

$$\Pr_{c \in \{0,1\}, d \in \{0,1\}}[(c \cdot x_1 + d) \bmod\ 2 = y_1 \textit{ and } (c \cdot x_2 + d) \bmod 2 = y_2] = \frac{1}{4}.$$

If we randomize over c then for any y we get

$$\Pr_{c \in \{0,1\}}[c \cdot x_1 \oplus c \cdot x_2 = y] = P_{c \in \{0,1\}}[c \cdot (x_1 \oplus x_2) = y] = \frac{1}{2}.$$

Now, randomize over d

$$\Pr_{c \in \{0,1\}, d \in \{0,1\}}[(c \cdot x_1 + d) \bmod 2 = y_1 \textit{ and } (c \cdot x_2 + d) \bmod 2 = y_2] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

(b) Let $Y_1, \ldots, Y_n$ be pairwise independent random variables. Prove that $\mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Var}[Y_i]$.

Solution: We know by definition of variance and covariance that

$$\mathrm{Var}[x + y] = \mathrm{Var}[x] + \mathrm{Var}[y] + 2 \cdot \mathrm{Cov}[x, y].$$

Similarly, we see that

$$\mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Var}[Y_i] + 2 \cdot \sum_{i=1}^{n} \sum_{j > i} \mathrm{Cov}[Y_i, Y_j].$$

If x and y are independent then $\mathrm{Cov}[x, y] = 0$. Therefore,

$$\mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Var}[Y_i].$$

(c) **Extra Credit.** Let q be a prime, and let k be an integer with $q \geq k$. Consider the set $\mathcal{H}$ of degree $k-1$ polynomials over $\mathbb{F}_q$. More precisely, let $\mathcal{H}$ be the set of polynomials $h_{\vec{c}}$ defined by a vector $\vec{c}$ of k coefficients $c_0, c_1, \ldots, c_{k-1} \in \{0, 1, \ldots, q-1\}$ such that

$$h_{\vec{c}}(x) = c_{k-1}x^{k-1} + c_{k-2}x^{k-2} + c_1 x + c_0 \bmod q.$$

Prove that $\mathcal{H}$ is a k-wise independent hash family. That is, prove that for all distinct $i_1, i_2, \ldots, i_k$ and all $j_1, j_2, \ldots, j_k$, we have

$$\Pr_{\vec{c}}[\ h_{\vec{c}}(i_1) = j_1 \text{ and } h_{\vec{c}}(i_2) = j_2 \text{ and } \cdots \text{ and } h_{\vec{c}}(i_k) = j_k\ ] = \frac{1}{q^k},$$

where the probability is over uniformly random $c_0, c_1, \ldots, c_{k-1} \in \{0, 1, \ldots, q-1\}$.

Hint: Consider the $k \times k$ Vandermonde matrix, which is invertible.

Solution: Observe that proof for $k = 2$ case is essentially the solution to part a. In general, we can write the k hash function evaluations in the matrix form as:

$$\begin{bmatrix} 1 & i_1 & {i_1}^2 & \cdots & {i_1}^{k-1} \\ 1 & i_2 & {i_2}^2 & \cdots & {i_2}^{k-1} \\ 1 & i_3 & {i_3}^2 & \cdots & {i_3}^{k-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & i_1 & {i_1}^2 & \cdots & {i_1}^{k-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{k-1} \end{bmatrix} = \begin{bmatrix} j_1 \\ j_2 \\ j_3 \\ \vdots \\ j_k \end{bmatrix}$$

This is a square Vandermonde matrix which has a non-zero determinant and hence is invertible. Therefore, the coefficients $\{c_0, c_1, c_2, \cdots, c_{k-1}\}$ can be uniquely determined. Let $\vec{c}$ determined using the above matrix equation be $\vec{c_0}$. Since all c_is in $\vec{c}$ are uniformly and independently sampled from $\{0, 1, 2, \cdots, q-1\}$,

$$\Pr_{\vec{c}}[\ h_{\vec{c}}(i_1) = j_1 \text{ and } h_{\vec{c}}(i_2) = j_2 \text{ and } \cdots \text{ and } h_{\vec{c}}(i_k) = j_k\ ] = \Pr[\vec{c} = \vec{c_0}] = \frac{1}{q^k}$$

Problem 3: Streaming Sampling

Consider the following algorithm for sampling a random element in a stream. You see $x_1, x_2, \ldots, x_m$ one at a time. For the first element x_1, store it as $s = x_1$, and initialize a counter $i = 1$. Every time you see an new element x_{i+1}, increment the counter, and flip a biased coin that comes up heads with probability $1/(i+1)$. If you get heads, then replace the stored element s with x_{i+1}.

(a) Prove that if you have seen m elements in the stream so far, then the probability that you have stored any given element is exactly $1/m$. That is, show that $\Pr[s = x_i] = 1/m$ for all $i = 1, 2, \ldots, m$ after you have seen all m elements.

(b) For a parameter $k > 1$, generalize the algorithm to sample k elements *without* replacement from the stream. As a hint, you can store the first k elements, and then replace one of the stored elements with a new element using a random process. Prove that for any subset $S \subseteq \{x_1, x_2, \ldots, x_m\}$ of size $|S| = k$, the algorithm outputs S with probability $1/\binom{m}{k}$.

Solution: (a) We observe that $s = x_j$ if x_j is chosen when it is considered by the algorithm (which happens with probability $1/j$), and none of $x_{j+1}, ..., x_m$ are chosen to replace x_j . All the relevant events are independent and we can compute:

$$\Pr[s = x_j] = 1/j \cdot \prod_{i>j}(1 - 1/i) = 1/m.$$

(b) Let's generalize the algorithm to sample k elements without replacement. We observe $x_1, x_2, \cdots, x_m$ one at a time. We store the first k elements as they come and then for every x_t $(t > k)$, we decide to choose it for inclusion in S with probability k/t, and if it is chosen then we choose a uniform element from S to be replaced by x_t.
The output of the algorithm is the set S. We now prove that algorithm outputs a random sample of size k without replacement via induction. The base case $m = k$ is true, since the set S is just $\{x_1, x_2, \cdots, x_k\}$ and the $\Pr[S = \{x_1, x_2, \cdots, x_k\}] = 1$. Let's assume that the statement holds for $t = m - 1$. Therefore, after observing $m - 1$ elements, the probability of a random subset $S \subseteq \{x_1, x_2, \ldots, x_{m-1}\}$ of size $|S| = k$ is given by $1/\binom{m-1}{k}$.

Now for $t = m$, we divide all possible subsets of $\{x_1, x_2, \cdots, x_m\}$ in two cases:

(a) Case 1: When the subset does not contain x_m.
For this to happen, we need to discard x_m with probability $1 - k/m$. Therefore, the probability of any random subset S of size k can be written as:

$$\Pr_m[S] = (1 - k/m) \cdot \Pr_{m-1}[S] = (1 - k/m) \cdot \frac{1}{\binom{m-1}{k}} = \frac{1}{\binom{m}{k}}.$$

(b) Case 2: When the subset contains x_m.
For this to happen, we decide to keep x_m with probability k/m and choose a random element with probability $1/k$ in S to be replaced by x_m. Observe that, there are $m - k$ subsets of size k at $t = m - 1$ stage which only differ from S by a single element which can give us S at $t = m$. Therefore, the probability of the subset S can be written as:

$$\Pr_m[S] = (k/m) \cdot (1/k) \cdot (m - 1) \cdot \frac{1}{\binom{m-1}{k}} = \frac{1}{\binom{m}{k}}.$$