Data Mining - HW1

Nathan Geier
nathangeier@mail.tau.ac.il

1.

a. The arrangement would be ((x3,x4,x5)[10 times],x2[10 times],x1[50 times])
The count of x1 in this arrangement would be 50, and obviously it is the maximum count possible. Every time we see the triplet (x3,x4,x5), we create a count of 1 for x3, then create a count of 1 for x4, then see x5 and clear everything so the same will happen for the next triplet. After we finish with the ten triplets all counters are empty, and since $k = 2$ and we only have two kinds of keys inserted from now on, everything is counted - all 50 elements.

b. We learned in class that the estimate is smaller than the true count by at most $\frac{m-m'}{k+1} = \frac{90-60}{3} = 10$, so we know the actual frequency of x_1 is between 50 and 60.

c. The arrangement would be (x1[50 times],(x2,x3)[10 times],(x4,x5)[10 times]) and the count would be 30, because with every couple a new counter was created and then cleared with x_1 counter decreasing by 1. Now we show that this is indeed the minimum: we know that if the count is x then the actual frequency (50) is at most $x + \frac{m-m'}{k+1} \leq x + \frac{90-x}{3}$, we used $m' \geq x$ because obviously the sum of counters is at least the size of x_1 counter. We get that $50 \leq 30 + \frac{2x}{3}$ giving that $30 \leq x$.

d. The estimate is smaller than the true count by at most $\frac{m-m'}{k+1} = \frac{90-30}{3} = 20$, so we know the actual frequency of x_1 is between 30 and 50. $m' = 30$ because only x_1 counter exists in the end with our arrangement.

2.

a.

$$\mathbb{E}(S) = \mathbb{E}(\sum_{i=1}^{n} I_{[u_i \le p]}) = \sum_{i=1}^{n} \mathbb{E}\left(I_{[u_i \le p]}\right) = \sum_{i=1}^{n} p = np$$

Therefore, the estimator $\hat{n} = \frac{S}{p}$ is unbiased, because $\mathbb{E}(\hat{n}) = \mathbb{E}(\frac{S}{p}) = \frac{\mathbb{E}(S)}{p} = n$.

b.

$$\mathrm{Var}(\hat{n}) = \mathrm{Var}(\frac{S}{p}) = \frac{\mathrm{Var}(\sum_{i=1}^{n} I_{[u_i \le p]})}{p^2} = \frac{\sum_{i=1}^{n} \mathrm{Var}(I_{[u_i \le p]})}{p^2} = \frac{\sum_{i=1}^{n} p(1-p)}{p^2} = \frac{n(1-p)}{p}$$

$$\mathrm{CV}(\hat{n}) = \frac{\sigma(\hat{n})}{\mu(\hat{n})} = \frac{\sqrt{\frac{n(1-p)}{p}}}{n} = \sqrt{\frac{1-p}{np}}$$

c.

$$\Pr_n(S = s) = \Pr_n(\sum_{i=1}^{n} I_{[u_i \le p]} = s) = \binom{n}{s} p^s (1-p)^{n-s}$$

d. $\Pr_n(S = 1) = np(1-p)^{n-1}$, the derivative (by n) is $p((1-p)^{n-1} + n(1-p)^{n-1}\ln(1-p))$. Trying to find zero points, we get that

$$(1-p)^{n-1} = -n(1-p)^{n-1}\ln(1-p)$$
$$1 = n\ln(\frac{1}{1-p})$$
$$n = \frac{1}{\ln(\frac{1}{1-p})}$$

(It is easy to see that as $n \to \infty$ the deravitave is negative, and for $n \to -\infty$ the derivative is positive, therefore in this point we went from pos to neg and it is a maximum point)

e. Since $n = 1$, we either have $s = 1$ w.p. p or $s = 0$ w.p. $1 - p$.
We already know that for $s = 1$, $\hat{n}_{\texttt{MLE}} = \frac{1}{\ln(\frac{1}{1-p})}$.
For $s = 0$, $\Pr_n(S = 0) = (1-p)^n$, so in order to maximize it we choose $\hat{n}_{\texttt{MLE}} = 0$.

$$\mathbb{E}\left[\hat{n}_{\texttt{MLE}}\right] - n = 0 * (1-p) + \frac{1}{\ln(\frac{1}{1-p})} * p - 1 = \frac{p + \ln(1-p)}{\ln(\frac{1}{1-p})}$$

$$\mathrm{Var}(\hat{n}_{\texttt{MLE}}) = \mathbb{E}\left[\hat{n}^2_{\texttt{MLE}}\right] - \mathbb{E}^2\left[\hat{n}_{\texttt{MLE}}\right] = 0^2 * (1-p) + \left(\frac{1}{\ln(\frac{1}{1-p})}\right)^2 * p - \left(\frac{p}{\ln(\frac{1}{1-p})}\right)^2 = \frac{p(1-p)}{\left(\ln(\frac{1}{1-p})\right)^2}$$

$$\mathrm{MSE} = \mathrm{Var}(\hat{n}_{\texttt{MLE}}) + \mathrm{Bias}^2(\hat{n}_{\texttt{MLE}}) = \frac{p(1-p)}{\left(\ln(\frac{1}{1-p})\right)^2} + \left(\frac{p}{\ln(\frac{1}{1-p})} - 1\right)^2 =$$

$$\frac{p}{\left(\ln(\frac{1}{1-p})\right)^2} - \frac{2p}{\ln(\frac{1}{1-p})} + 1 = \frac{p - 2p\ln(\frac{1}{1-p}) + \left(\ln(\frac{1}{1-p})\right)^2}{\left(\ln(\frac{1}{1-p})\right)^2}$$

f. Yes, two Flipper counters that have the same parameter p are composable.
We merge two counters by simply summing them.
Since our processing condition only depends on p and not the rest of the stream, creating a single Flipper counter and going over all elements at once or creating flipper counters for each element individually and then merging them, would produce the same result.

g. Morris has expected size $\log \log n$ while Flipper counter has expected size $\log(\frac{n}{p}) = \log n - \log p$ which is only better than trivial counter by some constant.
Morris has CV $\approx \frac{1}{\sqrt{2}}$ while Flipper has CV $= \sqrt{\frac{1-p}{np}}$ (goes to 0 as $n \to \infty$) so Flipper is more accurate for large enough streams.
Overall, Morris is better space-wise, and Flipper is better accuracy-wise.

3.

a. If both S_1 and S_2 are full:
For $1 \leq i \leq k$, choose $S[i] = S_1[i]$ with probability $\frac{n_1}{n_1+n_2}$, otherwise $S[i] = S_2[i]$.
Otherwise, say S_1 is not full, then we go over all S_1 elements and process them with S_2 as if we are seeing them at the end of our stream after processing all E_2 elements. Then we return $S := S_2'$.

b. If both are full, then $k \leq n_1 + n_2$, and the probabiliy of element from E_1 of getting included in S is

$$\frac{k}{n_1} * \frac{n_1}{n_1 + n_2} = \frac{k}{n_1 + n_2}$$

because with probability $\frac{k}{n_1}$ it gets into E_1, and then with probability $\frac{n_1}{n_1+n_2}$ from E_1 into S. The probability of element from E_2 of getting included in S is

$$\frac{k}{n_2} * \left(1 - \frac{n_1}{n_1 + n_2}\right) = \frac{k}{n_2} * \frac{n_1 + n_2 - n_1}{n_1 + n_2} = \frac{k}{n_1 + n_2}$$

because with probability $\frac{k}{n_2}$ it gets into E_2, and then with probability $1 - \frac{n_1}{n_1+n_2}$ the corresponding element from E_1 does not get chosen, and this element is added into S.
Overall, every element from $E_1 \cup E_2$ gets into S with probability $\frac{k}{n_1+n_2} = min\{1, \frac{k}{n_1+n_2}\}$.
If at least one is not full, then the analysis is the same as processing all elements continuously in one go, and so the probability of each element of getting included is $min\{1, \frac{k}{n_1+n_2}\}$ just as learned in class.

4.

a. Initialize: Declare integer array S of size m; For $1 \leq i \leq m$, $S[i] \leftarrow 0$.
Insert(a): For $1 \leq i \leq k$, $S[h_i(a)] = S[h_i(a)] + 1$.
Delete(a): For $1 \leq i \leq k$, $S[h_i(a)] = S[h_i(a)] - 1$.
Membership(a): Return ($S[h_1(a)] \geq 1$ and $S[h_2(a)] \geq 1$ and ... $S[h_k(a)] \geq 1$).
Merge(S_1, S_2): Two structures of same size and same set of hash functions. For $1 \leq i \leq k$, $S[i] = S_1[i] + S_2[i]$.
It is easy to see that according to our rules (cannot delete something that is not there, cannot insert something that is already there), if a was inserted and was not deleted then other elements can only increment our relevant counters, and so our membership query will always return true for a.
If a was never inserted or was inserted and then deleted, the analysis stays the same as the original Bloom Filters, because the probability of a cell having 0 in our algorithm is the same as having F is the original one: it just means that no existing element hit it.

b. Let c be any cell. Define $I_{a,i}$ to be the indicator of $h_i(a) = c$, for element a and $1 \leq i \leq k$. We know that $\Pr[I_{a,i} = 1] = \frac{1}{n}$. Define $S = \sum_{a,i} I_{a,i}$.

$$\mathbb{E}(S) = \sum_{a,i} \mathbb{E}(I_{a,i}) = \frac{mk}{n}$$

$$\Pr[S \geq \frac{(1+\epsilon)mk}{n}] \leq \exp(-\frac{\epsilon^2 \frac{mk}{n}}{2})$$

$$\Pr[\text{OVERFLOW}] = \Pr[\bigcup_i \{\texttt{cell i overflow}\}] \leq \sum_{i=1}^{n} \exp(-\frac{\epsilon^2 \frac{mk}{n}}{2}) =$$

$$n \exp(-\frac{\epsilon^2 \frac{mk}{n}}{2}) := \delta$$

$$\exp(-\frac{\epsilon^2 \frac{mk}{n}}{2}) = \frac{\delta}{n}$$

$$-\frac{\epsilon^2 \frac{mk}{n}}{2} = \ln\left(\frac{\delta}{n}\right)$$

$$\epsilon^2 = -\frac{2n \ln\left(\frac{\delta}{n}\right)}{mk}$$

$$\epsilon = \sqrt{\frac{2n \ln\left(\frac{n}{\delta}\right)}{mk}}$$

So the bound on the length of each counter is

$$\lceil \log_2(\frac{(1+\epsilon)mk}{n}) \rceil = \lceil \log_2(\frac{mk}{n} + \sqrt{\frac{2mk \ln\left(\frac{n}{\delta}\right)}{n}}) \rceil = \lceil \log_2(\frac{mk + \sqrt{2mkn \ln\left(\frac{n}{\delta}\right)}}{n}) \rceil$$

5.

a.

$$f(\{s_i\};n) = \prod_{i=1}^{k} ne^{-ns_i} = n^k e^{-n\sum_{i=1}^{k} s_i}$$

$$\ell(\{s_i\};n) = k\ln(n) - n\sum_{i=1}^{k} s_i$$

$$a := \frac{1}{n}, \quad b := n^2$$

$$\ell(\{s_i\};a) = -k\ln(a) - \frac{\sum_{i=1}^{k} s_i}{a}$$

$$\ell(\{s_i\};b) = \frac{k\ln(b)}{2} - \sqrt{b}\sum_{i=1}^{k} s_i$$

$$\frac{\partial \ell(\{s_i\};a)}{\partial a} = -\frac{k}{a} + \frac{\sum_{i=1}^{k} s_i}{a^2} = 0$$

$$\frac{\partial \ell(\{s_i\};b)}{\partial b} = \frac{k}{2b} - \frac{\sum_{i=1}^{k} s_i}{2\sqrt{b}} = 0$$

$$\hat{a}_{\texttt{MLE}} = \frac{\sum_{i=1}^{k} s_i}{k} = \frac{1}{\hat{n}_{\texttt{MLE}}}$$

$$\hat{b}_{\texttt{MLE}} = \left(\frac{k}{\sum_{i=1}^{k} s_i}\right)^2 = \hat{n}^2_{\texttt{MLE}}$$

b. The sufficient statistic for all 3 likelihood functions $(n, \frac{1}{n}, n^2)$ is $\sum_{i=1}^{k} s_i$, because they do not depend on $\{s_i\}$ any other way.