Data Mining - HW2

Nathan Geier
nathangeier@mail.tau.ac.il

1.

(a) b contains exactly one non-zero entry iff

$$s_0 \neq 0 \text{ and } s_2 \cdot s_0 = s_1^2$$

The following are equivalent:

$$s_2 \cdot s_0 = s_1^2$$

$$\sum_{i=1}^{n} i^2 b_i \cdot \sum_{i=1}^{n} b_i = \sum_{i=1}^{n} i b_i \cdot \sum_{i=1}^{n} i b_i$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} i^2 b_i b_j = \sum_{i=1}^{n} \sum_{j=1}^{n} i j b_i b_j$$

$$\sum_{i<j} \left(i^2 + j^2\right) b_i b_j + \sum_{i=1}^{n} i^2 b_i b_i = \sum_{i<j} 2ij b_i b_j + \sum_{i=1}^{n} i^2 b_i b_i$$

$$\sum_{i<j} (i-j)^2 b_i b_j = 0$$

$$\forall i \neq j : \quad b_i = 0 \text{ or } b_j = 0$$

At most one element is non-zero

Where we used that $\forall i : \ b_i \geq 0$ and that $\forall i \neq j \ (i-j)^2 > 0$.
We know that $s_0 \neq 0$ iff there is at least one non-zero element, therefore both hold iff there is at least one non-zero element but at most one, meaning that there must be exactly one non-zero element. In this case, we have that:

$$\frac{s_1}{s_0} = \frac{\sum_{i=1}^{n} i b_i}{\sum_{i=1}^{n} b_i} = \frac{k b_k}{b_k} = k$$

So we can return $\frac{s_1}{s_0}$.

(b) A simple sketch that samples a random subset of 2^j elements from b in expectation for $j = 0, \ldots, \log n$: Let $h_0, \ldots, h_{\log n}$ be "fully random" hash functions mapping $\{1, 2, \ldots, n\}$ to $\{0, 1\}$ where $h_j(i) = 1$ w.p. $\frac{2^j}{n}$. The matrix A_j of size $n \times n$ with

$$A_j[x, y] = \begin{cases} 1 & x = y \text{ and } h_j(x) = 1 \\ 0 & \textit{otherwise} \end{cases}$$

defines a linear sketch giving a vector a_j of size n containing a random subset of size 2^j from b in expectation (other coordinates are zero). The matrix of size $3\log n \times n$ defining our linear sketch is:

$$P = \begin{bmatrix} M & M & \cdots & M \end{bmatrix} \times \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_{\log n} \end{bmatrix}$$

that just means that for all j we maintain the composition of M over A_j. What we actually do when processing an element (i, x) is:

Algorithm 1 Process Element

```
function ProcessElement(i, x)
    for j ← 0 to log n do
        if h_j(i) = 1 then
            for k ← 0 to 2 do
                s[j, k] ← s[j, k] + i^k · x
            end for
        end if
    end for
end function
```

Algorithm 2 Query

```
function Query()
    for j ← 0 to log n do
        if s[j] detects exactly one non-zero entry then
            return s[j]
        end if
    end for
    return FAIL
end function
```

Now, lets give a lower bound for the success probability. If at least one $s[j]$ succeeds, we succeed. Say there are $k > 1$ non-zero elements, and let $2^i \leq \frac{n}{k} < 2^{i+1}$. The probability for $s[i]$ to contain exactly one non-zero element is

$$k\frac{2^i}{n}\left(1 - \frac{2^i}{n}\right)^{k-1} \geq \frac{1}{2}\left(1 - \frac{1}{k}\right)^{k-1} \geq \frac{1}{2}\left(1 - \frac{1}{k}\right)^{k} \geq \frac{1}{8}$$

The last inequality because $(1 - \frac{1}{k})^k$ is an increasing function (calculus 1). Actually, a tighter lower bound for $s[i]$ is $\frac{1}{2e}$, but we will not prove it, we just want it to be lower bounded by some constant. If $k = 0$ we always return FAIL, and if $k = 1$ the set $s[\log n]$ will contain exactly one non-zero element with probability $\frac{2^{\log n}}{n} = 1$.

We keep $\log n + 1$ hash functions and an array of size $3\log n$, therefore the size of our sketch is $O(\log n) \cdot (size(hash) + O(1))$.

Our sketch must return a positive element **uniformly** because for all j, $h_j[i]$ does not depend on i, so the set $s[j]$ contains any positive element with the same probability.

2.

(a) M is of size $(d \cdot w) \times n$, and therefore s is of size $d \cdot w$.

$$M\left[(i,j),k\right] = \begin{cases} 1 & h_i[k] = j \\ 0 & otherwise \end{cases}$$

In the classic definition, when processing (k,x) we go over all (i,j) and do $s_{(i,j)} \leftarrow s_{(i,j)} + M_{(i,j),k} \cdot x$. But from the definition of M, for every i the only non-zero $M_{(i,j),k}$ is where $j = h_i[k]$ where it has the value 1, therefore it is completely equivalent to just go over all i and do $s_{(i,h_i[k])} \leftarrow s_{(i,h_i[k])} + x$.

(b) It is obvious that $b_i \leq \hat{b_i}$ because every time that b_i is incremented by x, $count[j,h_j[i]]$ is also incremented by x for all j, and since updates from other elements can only increase the counts, it holds that every $count[j,h_j[i]]$ (therefore also $\hat{b_i}$) must be at least b_i.

$$\Pr\left[\hat{b_i} > b_i + \varepsilon\|b\|_1\right] = \Pr\left[\forall j : \ \ count[j,h_j[i]] > b_i + \varepsilon\|b\|_1\right] =$$

$$\left(\Pr\left[count[1,h_1[i]] - b_i > \varepsilon\|b\|_1\right]\right)^d \leq \left(\frac{\mathbb{E}[count[1,h_1[i]] - b_i]}{\varepsilon\|b\|_1}\right)^d \leq$$

$$\left(\frac{\frac{1}{w} \cdot \|b\|_1}{\varepsilon\|b\|_1}\right)^d \leq \left(\frac{\frac{\varepsilon}{e} \cdot \|b\|_1}{\varepsilon\|b\|_1}\right)^d = e^{-d} \leq e^{-\ln(1/\delta)} = \delta$$

(c)

$$\Pr\left[b_i - 2\varepsilon\|b\|_1 > count[j,h_j[i]] \text{ or } count[j,h_j[i]] > b_i + 2\varepsilon\|b\|_1\right] =$$

$$\Pr\left[|count[j,h_j[i]] - b_i| > 2\varepsilon\|b\|_1\right] \leq \frac{\mathbb{E}\left[|count[j,h_j[i]] - b_i|\right]}{2\varepsilon\|b\|_1} \leq \frac{\frac{1}{w}\|b\|_1}{2\varepsilon\|b\|_1} \leq \frac{\frac{\varepsilon}{e}\|b\|_1}{2\varepsilon\|b\|_1} = \frac{1}{2e}$$

The median is bad $\Rightarrow$ at least half of our d' counters are bad

$$\Pr\left[\sum_{j=1}^{d'} I_{count[j,h_j[i]] \text{ is bad}} \geq \frac{d'}{2}\right] = \Pr\left[\sum_{j=1}^{d'} I_{count[j,h_j[i]] \text{ is bad}} \geq (1 + (e-1))\frac{d'}{2e}\right] \leq$$

$$e^{-\frac{(e-1)\frac{d'}{2e}}{3}} \leq \delta$$

$$d' \geq \frac{6e\ln(1/\delta)}{e-1}$$

3.

(a) Return the sum over

$$a_i = \begin{cases} \frac{\min\{u_i, v_i\}}{p_i} & i \text{ sampled by } S(U) \\ 0 & i \text{ not sampled by } S(U) \end{cases}$$

for every i such that $(i, u_i) \in S(U)$, where p_i is the inclusion probability of (i, u_i) in $S(U)$. (the same inclusion probability we used when making subset weight estimates using $S(U)$, as learned in class)

$$\mathbb{E}\left[\hat{est}\right] = \sum_{i=1}^{n} \mathbb{E}\left[a_i\right] = \sum_{i=1}^{n} \min\{u_i, v_i\} = \|min\{U, V\}\|_1$$

(b) We can create $S(\max\{U, V\})$ from $S(U)$ and $S(V)$. For every i, let:

$$r'(i) = \begin{cases} \frac{h(i)}{\max\{u_i, v_i\}} & i \in S(U) \cap S(V) \\ \frac{h(i)}{u_i} & i \in S(U) \setminus S(V) \\ \frac{h(i)}{v_i} & i \in S(V) \setminus S(U) \\ \infty & otherwise \end{cases}$$

then we take the k $(i, h(i)/r'(i))$'s with the lowest $r'(i)$. We don't always have that $r'(i) = r(i, \max\{u_i, v_i\})$, but we always have that $r'(i) \geq r(i, \max\{u_i, v_i\})$, and we claim that for every i in $S(\max\{U, V\})$ we do have that $r'(i) = r(i, \max\{u_i, v_i\})$, and therefore the k bottom $r'(i)$ are the same as the k bottom $r(i, \max\{u_i, v_i\})$, that are $S(\max\{U, V\})$.
$r'(i) \geq r(i, \max\{u_i, v_i\})$ because $\infty, \frac{h(i)}{v_i}, \frac{h(i)}{u_i} \geq \frac{h(i)}{\max\{u_i, v_i\}}$.
Say $i \in S(\max\{U, V\})$, and assume w.l.o.g that $u_i = \max\{u_i, v_i\}$. It holds that $r(i, \max\{u_i, v_i\}) \leq r(j, \max\{u_j, v_j\}) \Rightarrow r(i, u_i) \leq r(j, u_j)$, because

$$r(i, u_i) = r(i, \max\{u_i, v_i\}) \leq r(j, \max\{u_j, v_j\}) = \frac{h(j)}{\max\{u_j, v_j\}} \leq \frac{h(j)}{u_j} = r(j, u_j)$$

and that means that $r(i, u_i)$'s rank in $\{r(j, u_j)\}_j$ is at most as large as $r(i, \max\{u_i, v_i\})$'s rank in $\{r(j, \max\{u_j, v_j\})\}_j$, which is at most k. Therefore, $(i, u_i) \in S(U)$, and so it must be that $r'(i) = \frac{h(i)}{u_i} = \frac{h(i)}{\max\{u_i, v_i\}} = r(i, \max\{u_i, v_i\})$.
After creating $S(\max\{U, V\})$ we can estimate $\|\max\{U, V\}\|_1$ using the inclusion probabilities, the same way we learned in class for estimating subset weights.

(c) Our estimate for $\|\max\{U, V\}\|_1$ is completely equivalent to an estimate obtained directly from a bottom-k sample of $\max\{U, V\}$. Actually, we could have even done a little better than that in (b). Let $max_U = \max_{i \in S(U)} \{r(i, u_i)\}$ and $max_V = \max_{i \in S(V)} \{r(i, v_i)\}$. If instead of collecting the bottom-k $r'(i)$'s we take all $i's$ such that $r'(i) \leq \min\{max_U, max_V\}$, we get the bottom-$k'$ i's of $S(\max\{U, V\})$ for $k \leq k' < 2k$. (k' is a random variable, it may be different for different hash function and samples). This holds because $|r'(i) \leq \min\{max_U, max_V\}| \geq k$, and if $r(i, \max\{u_i, v_i\}) \leq \min\{max_U, max_V\}$ then either $(r(i, u_i) \leq max_U$ and $u_i \geq v_i)$ or $(r(i, v_i) \leq max_V$ and $v_i \geq u_i)$ and in any case $r'(i) = r(i, \max\{u_i, v_i\}) \leq \min\{max_U, max_V\}$ so it gets taken.

4.

(a) Denote by $p_i := \min\{1, \alpha w_i\}$, and $w_{1,2} = w_1 - w_2$. Our estimator:

$$\hat{w}_{1,2} = \begin{cases} \frac{w_1 - w_2}{p_1 p_2} & w_1, w_2 \in S \\ 0 & otherwise \end{cases}$$

$$\mathbb{E}\left[\hat{w}_{1,2}\right] = \frac{w_1 - w_2}{p_1 p_2} \cdot \Pr\left[w_1 \in S \cap w_2 \in S\right] = \frac{w_1 - w_2}{p_1 p_2} \cdot p_1 \cdot p_2 = w_1 - w_2$$

(b)

$$\begin{aligned} &\mathrm{Var}\left[\hat{w}_{1,2}\right] = \mathbb{E}\left[\hat{w}_{1,2}^2\right] - \mathbb{E}^2\left[\hat{w}_{1,2}\right] = \\ &\left(\frac{w_1 - w_2}{p_1 p_2}\right)^2 \cdot \Pr\left[w_1 \in S \cap w_2 \in S\right] - (w_1 - w_2)^2 = \\ &\frac{(w_1 - w_2)^2}{p_1 p_2} - (w_1 - w_2)^2 = (w_1 - w_2)^2 \cdot \left(\frac{1}{\min\{1, \alpha w_1\} \min\{1, \alpha w_2\}} - 1\right) \end{aligned}$$

(c) If the sample does not include the i'th entry, we know that $\alpha w_i < 1$ and therefore that $0 \leq w_i < \frac{1}{\alpha}$

(d)

$$\hat{w}_{1,2} = \begin{cases} \frac{-w_1 p_2 - w_2}{p_1 p_2} & w_1, w_2 \in S \\ \frac{w_1}{p_1} & w_1 \in S, w_2 \notin S \\ 0 & otherwise \end{cases}$$

First, note that our estimator is well defined, because each denominator cannot be zero in case the corresponding case happened. (For example, if we got a sample with $w_1 \in S, w_2 \notin S$, we know that $p_1 \neq 0$)

$$\begin{aligned} &\text{If } w_2 > 0: \\ &\mathbb{E}\left[\hat{w}_{1,2}\right] = \frac{-w_1 p_2 - w_2}{p_1 p_2} \cdot p_1 \cdot p_2 + \frac{w_1}{p_1} \cdot p_1 \cdot (1 - p_2) + 0 = \\ &- w_1 p_2 - w_2 + w_1 \cdot (1 - p_2) = w_1 - w_2 \\ &\text{If } w_2 = 0, w_1 > 0: \\ &\mathbb{E}\left[\hat{w}_{1,2}\right] = 0 + \frac{w_1}{p_1} \cdot p_1 \cdot 1 + 0 = w_1 = w_1 - w_2 \\ &\text{If } w_1 = 0: \\ &\mathbb{E}\left[\hat{w}_{1,2}\right] = 0 = w_1 - w_2, \text{ because we are always in the third case.} \end{aligned}$$

Now, we will show that there cannot be an unbiased non-negative estimator:
Let $0 < w_2 < w_1 < \frac{1}{\alpha}$.

$$\begin{aligned} &\mathbb{E}\left[\hat{w}_{1,2}\right] = \mathbb{E}\left[\hat{w}_{1,2} | w_2 \in S\right] \cdot \Pr\left[w_2 \in S\right] + \mathbb{E}\left[\hat{w}_{1,2} | w_2 \notin S\right] \cdot \Pr\left[w_2 \notin S\right] \geq \\ &0 + (w_1 - 0) \cdot (1 - \alpha w_2) = w_1 - \alpha w_1 w_2 > w_1 - \alpha \frac{1}{\alpha} w_2 = w_1 - w_2 \end{aligned}$$

Where we used that $\mathbb{E}\left[\hat{w}_{1,2} | w_2 \notin S\right] = w_1 - 0$ because the distribution of our sample given that $w_2 \notin S$ is the same as setting $w_2 = 0$, and we "assume towards contradiction" that the estimator is unbiased and therefore the expectation of our estimator over that distribution gives $w_1 - 0$.

5.

(a) If the sample does not include the i'th entry (we get h with the sample), we know that $h(i) > \alpha w_i$ and therefore that $0 \le w_i < \frac{h(i)}{\alpha}$.

(b)

$$w_1 - w_2 \in \begin{cases} \{w_1 - w_2\} & w_1, w_2 \in S \\ \left(\max\{0, w_1 - \frac{h(2)}{\alpha}\}, w_1\right] & w_1 \in S, w_2 \notin S \\ \left[0, \frac{h(1)}{\alpha} - w_2\right) & w_1 \notin S, w_2 \in S \\ \left[0, \frac{h(1)}{\alpha}\right) & w_1, w_2 \notin S \end{cases}$$

(here $[a, a) = (a, a] = \{a\}$)

(c)

$$\hat{w}_{1,2} = \begin{cases} \frac{(w_1 - w_2)(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2))}{\alpha w_1 \alpha w_2} & w_1, w_2 \in S \\ \max\{0, w_1 - \frac{h(2)}{\alpha}\} & w_1 \in S, w_2 \notin S \\ 0 & otherwise \end{cases}$$

$\hat{w}_{1,2} \ge 0$ because $\max\{0, w_1 - \frac{h(2)}{\alpha}\} \ge 0$ and $(w_1 - w_2) \ge 0, \alpha w_1 \alpha w_2 \ge 0$, $(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2)) \ge 0$ because $\alpha^2 w_1 (w_1 - w_2) \le (\alpha w_1)^2 \le 1$, therefore the first case is also non-negative.

$$\mathbb{E}\left[\hat{w}_{1,2}\right] = \frac{(w_1 - w_2)(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2))}{\alpha w_1 \alpha w_2} \Pr\left[w_1, w_2 \in S\right] +$$
$$\Pr\left[w_1 \in S\right] \int_{\alpha w_2}^{1} \max\{0, w_1 - \frac{h(2)}{\alpha}\} dh(2) = (w_1 - w_2)(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2)) +$$
$$\alpha w_1 \int_{\alpha w_2}^{\alpha w_1} w_1 - \frac{h(2)}{\alpha} dh(2) = (w_1 - w_2)(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2)) +$$
$$\alpha w_1 \left[w_1 \alpha w_1 - \frac{\alpha^2 w_1^2}{2\alpha} - w_1 \alpha w_2 + \frac{\alpha^2 w_2^2}{2\alpha} \right] = (w_1 - w_2)(1 - \frac{\alpha^2}{2} w_1 (w_1 - w_2)) +$$
$$\alpha^2 w_1 \frac{1}{2} (w_1 - w_2)^2 = w_1 - w_2$$

$$\mathrm{Var}\left[\hat{w}_{1,2}\right] = \mathbb{E}\left[\hat{w}_{1,2}^2\right] - \mathbb{E}^2\left[\hat{w}_{1,2}\right] = \left(\frac{(w_1-w_2)(1-\frac{\alpha^2}{2}w_1(w_1-w_2))}{\alpha w_1 \alpha w_2}\right)^2 \Pr\left[w_1, w_2 \in S\right] +$$

$$\Pr\left[w_1 \in S\right] \int\limits_{\alpha w_2}^{1} \left(\max\{0, w_1 - \frac{h(2)}{\alpha}\}\right)^2 dh(2) - (w_1-w_2)^2 = \frac{(w_1-w_2)^2(1-\frac{\alpha^2}{2}w_1(w_1-w_2))^2}{\alpha w_1 \alpha w_2} +$$

$$\alpha w_1 \int\limits_{\alpha w_2}^{\alpha w_1} w_1^2 - 2w_1\frac{h(2)}{\alpha} + \frac{h(2)^2}{\alpha^2} dh(2) - (w_1-w_2)^2$$

$$\int\limits_{\alpha w_2}^{\alpha w_1} w_1^2 - 2w_1\frac{h(2)}{\alpha} + \frac{h(2)^2}{\alpha^2} dh(2) = \left[w_1^2 h(2) - 2w_1\frac{h(2)^2}{2\alpha} + \frac{h(2)^3}{3\alpha^2}\right]_{\alpha w_2}^{\alpha w_1} =$$

$$w_1^2\alpha w_1 - 2w_1\frac{(\alpha w_1)^2}{2\alpha} + \frac{(\alpha w_1)^3}{3\alpha^2} - w_1^2\alpha w_2 + 2w_1\frac{(\alpha w_2)^2}{2\alpha} - \frac{(\alpha w_2)^3}{3\alpha^2} =$$

$$\alpha w_1^3 - \alpha w_1^3 + \alpha\frac{w_1^3}{3} - \alpha w_1^2 w_2 + \alpha w_1 w_2^2 - \alpha\frac{w_2^3}{3} = \frac{\alpha}{3}\left[w_1^3 - 3w_1^2 w_2 + 3w_1 w_2^2 - w_2^3\right] =$$

$$\frac{\alpha}{3}(w_1-w_2)^3$$

$$\mathrm{Var}\left[\hat{w}_{1,2}\right] = (w_1-w_2)^2\left[\frac{(1-\frac{\alpha^2}{2}w_1(w_1-w_2))^2}{\alpha w_1 \alpha w_2} + \frac{\alpha}{3}(w_1-w_2) - 1\right] =$$

$$(w_1-w_2)^2\left[\frac{1 - 2\frac{\alpha^2}{2}w_1(w_1-w_2) + \frac{\alpha^4}{4}w_1^2(w_1-w_2)^2 + \frac{\alpha}{3}(w_1-w_2)\alpha w_1 \alpha w_2}{\alpha w_1 \alpha w_2} - 1\right] =$$

$$(w_1-w_2)^2\left[\frac{1 - \alpha^2 w_1(w_1-w_2)\left[1 - \frac{\alpha^2}{4}w_1(w_1-w_2) - \frac{\alpha}{3}w_2\right]}{\alpha w_1 \alpha w_2} - 1\right]$$

Recall that the variance in the previous question was

$$(w_1-w_2)^2 \cdot \left(\frac{1}{\alpha w_1 \alpha w_2} - 1\right)$$

so this one is better, because

$$\alpha^2 w_1(w_1-w_2)\left[1 - \frac{\alpha^2}{4}w_1(w_1-w_2) - \frac{\alpha}{3}w_2\right]$$

is non-negative, since $\frac{\alpha^2}{4}w_1(w_1-w_2) \leq \frac{(\alpha w_1)^2}{4} \leq \frac{1}{4}$ and $\frac{\alpha w_2}{3} \leq \frac{1}{3}$, and therefore

$$1 - \frac{\alpha^2}{4}w_1(w_1-w_2) - \frac{\alpha}{3}w_2 \geq 1 - \frac{1}{4} - \frac{1}{3} \geq \frac{5}{12} > 0$$

(If $w_1 = w_2$ both variances are zero, but in any other case our current variance is better).

(d) Let $p_2 = \min\{1, \alpha w_2\}$. Since $w_1 \geq 1/\alpha$, we always have that $w_1 \in S$.

$$\hat{w}_{1,2} = \begin{cases} \frac{w_1 p_2 - w_2}{p_2} & w_2 \in S \\ w_1 & w_2 \notin S \end{cases}$$

Note that the estimator is well-defined because if $x_2 \in S$ then $p_2 \neq 0$. The estimator is non-negative because if $p_2 = 1$ then $w_1 p_2 - w_2 = w_1 - w_2 \geq 0$, and if $p_2 = \alpha w_2$ then $w_1 p_2 - w_2 = w_1 \alpha w_2 - w_2 \geq \frac{1}{\alpha} \alpha w_2 - w_2 = w_2 - w_2 = 0$.

If $w_2 > 0$:

$$\mathbb{E}\left[\hat{w}_{1,2}\right] = \frac{w_1 p_2 - w_2}{p_2} p_2 + w_1(1 - p_2) = w_1 p_2 - w_2 + w_1 - w_1 p_2 = w_1 - w_2$$

If $w_2 = 0$:

$\mathbb{E}\left[\hat{w}_{1,2}\right] = w_1 = w_1 - w_2$, because we are always in the second case.