Intelligent Learning and Analysis Systems SS19, Exercise Sheet 2

by Andreas Hene, Niklas Mertens, Richard Palme.
Tutor: Maximilian Thiessen, Group 3

May 3, 2019

Exercise 3

Lemma 1

Each bucket has exactly size $w = \lceil \frac{1}{\varepsilon} \rceil$. $\Rightarrow$ total number of buckets is

$$\frac{m}{w} = \frac{m}{\lceil 1/\varepsilon \rceil} \leq \varepsilon m$$

So the current bucket id is at most εm.

Lemma 2

Proof by induction.
Base Case: Let $b_{\text{current}} = 1$, let $(e, f, \Delta) \in \mathcal{D}$. Since there have been no deletions and no decrementations of f until now, we know that f is the true frequency up until this point, i.e. $f = f_e$.

Also, (e, f, Δ) is only deleted when $f \leq 1$, in other words (e, f, Δ) is only deleted when $f_e = f \leq 1 = b_{\text{current}}$ which proves the base case.

Step Case: Let $k > 1$. Suppose that for $b_{\text{current}} < k$ we know that $f_e \leq b_{\text{current}}$ when (e, f, Δ) gets deleted.

Now let $b_{\text{current}} = k$ and $(e, f, \Delta) \in \mathcal{D}$. The true frequency of e in the buckets with ids $\Delta + 1, \ldots, b_{\text{current}}$ is $\leq f + (b_{\text{current}} - 1 - \Delta)$, because f gets decremented by 1 at most $b_{\text{current}} - 1 - \Delta$ times.
Let b' be the bucket id where e was deleted the last time, if it exists. Else set $b' = 0$. By the induction hypothesis, the true frequency of e in buckets $1, \ldots, b'$ is $\leq b'$, if b' exists. The true frequency in buckets $b' + 1, \ldots, \Delta$ is 0. So the true frequency f_e at the current time (bucket id b_{current}) is at most $f_e \leq f + b_{\text{current}} - 1 - \Delta + b'$.
Since $b' \leq \Delta$, we get: $f_e \leq f + b_{\text{current}} - 1$.

If (e, f, Δ) gets deleted, we have $f \leq 1$, so $f_e \leq f + b_{\text{current}} - 1 \leq b_{\text{current}}$.

Lemma 3

If there is no entry for e in $\mathcal{D}$, then there are 2 cases:
Case 1: There has never been an entry for e in $\mathcal{D}$.
Then $f_e = 0 \leq \varepsilon m$.
Case 2: There has been an entry for e in $\mathcal{D}$ before.
Let b be the bucket id when e was deleted the last time. Then $f_e \leq b$ by Lemma 2, and $f_e \leq b \leq \varepsilon m$ by Lemma 1.

Lemma 4

$f \leq f_e$, because f gets incremented by 1 at most f_e times.
If the last time e was deleted was after processing bucket b, then by Lemma 3 the true frequency of e in buckets $1, \ldots, b$ is at most εm. The true frequency of e in buckets $b+1, \ldots, \Delta$ is 0, and the true frequency in buckets $\Delta+1, \ldots, b_{\text{current}}$ is at most f. So:

$$f_e \leq \varepsilon m + 0 + f = f + \varepsilon m$$

Exercise 4

d_i is the number of elements of $\mathcal{D}$ that were last added to $\mathcal{D}$ during processing of bucket $B - i + 1$. Call the i-th summand the contribution of bucket $B - i + 1$ to the sum. We want that the contribution of all summands does not exceed the total size of buckets $B - j + 1, \ldots, B$ (which is jw).

So if element e was created during processing of $B - i + 1$, then e survives the deletion/decrementing process $i - 1$ times. This can only happen if e occurs at least i times in buckets $B - i + 1, \ldots B$. So e is allowed a contribution of i. So the contribution of all the elements last added to $\mathcal{D}$ during processing of bucket $B - i + 1$ is allowed to be id_i, because this way the contribution of all summands can't exceed the total size of $B - j + 1, \ldots B$, i.e.

$$\sum_{i=1}^{B} id_i \leq jw$$

Exercise 2

Let $m_1 := |\sigma_1|, m_2 := |\sigma_2|, m := m_1 + m_2$
$m_1' := \sum_{l=1}^{k-1} A_1[l], m_2' := \sum_{l=1}^{k-1} A_2[l], m' := \sum_{l=1}^{k-1} A[l]$.
$c_k :=$ counter of k-th most frequent item in $A_1 \oplus A_2$
Let $i \in \text{keys}(A)$.
The estimated frequency of i in σ_1 is at most $\frac{m_1 - m_1'}{k}$ smaller than the true frequency of i in σ_1.
The estimated frequency of i in σ_2 is at most $\frac{m_2 - m_2'}{k}$ smaller than the true frequency of i in σ_2.
This the estimated frequency of i in $\sigma_1 \otimes \sigma_2$ is at most $\frac{m_1 - m_1'}{k} + \frac{m_2 - m_2'}{k} + c_k$ smaller than the true frequency of i in $\sigma_1 \otimes \sigma_2$.

$$\Rightarrow \hat{f}_i \geq f_i - \frac{m - (m_1' + m_2' - kc_k)}{k}$$

<u>Claim:</u> $m_1' + m_2' - kc_k \geq m'$
<u>Proof:</u> Case 1: $|\text{keys}(A_1) \cup \text{keys}(A_2)| \leq k - 1$
Then $c_k = 0$ and $m' = m_1' + m_2'$.
Case 2: $|\text{keys}(A_1) \cup \text{keys}(A_2)| \geq k$
In line 5 of Misra-Gries synopses algorithm we have $m_1' + m_2' \geq \widetilde{m}' + c_k$, since the k-th most frequent item is not in A ($\widetilde{m}'$ denotes m' at this point of the algorithm).
In line 7 of the algorithm, $\widetilde{m}'$ is reduced by $(k-1)c_k$, i.e.

$$m' = \widetilde{m}' - (k-1)c_k \Rightarrow m_1' + m_2' \geq m' + kc_k$$

■

So

$$\hat{f}_i \geq f_i - \frac{m - m'}{k}$$

$\hat{f}_i \leq f_i$, because $A_1[i]$ and $A_2[i]$ are underestimations of the true frequencies of i in σ_1, σ_2. So $A_1[i] + A_2[i]$ is an underestimation of f_i in $\sigma_1 \otimes \sigma_2$, so $A[i]$ is an underestimation of f_i as well.