SEEM5020 2026 Spring: Homework 3 Solution

- **Q1. Second Moment of MinHash**

  Proof: Let $Z = X_{min}^2 \in [0,1]$. We have

$$\begin{aligned}\mathbb{E}[Z] &= \int_0^{+\infty} (1 - F_Z(x))dx = \int_0^1 (1-\sqrt{x})^{\mathsf{NDE}} dx = \int_0^1 (1-t)^{\mathsf{NDE}} 2t dt \\ &= \int_0^1 (1-t)^{\mathsf{NDE}} 2t dt = \frac{-(1-t)^{\mathsf{NDE}+1} 2t}{\mathsf{NDE}+1}\Big|_0^1 + 2\int_0^1 \frac{(1-t)^{\mathsf{NDE}+1}}{\mathsf{NDE}+1} dt \\ &= \frac{2}{\mathsf{NDE}+1} \cdot \frac{-(1-t)^{\mathsf{NDE}+2}}{\mathsf{NDE}+2}\Big|_0^1 = \frac{2}{(\mathsf{NDE}+1)(\mathsf{NDE}+2)}.\end{aligned}$$

- **Q2. Bottom-k Sketch (Lower Tail)**

  Proof: Define $X_i = 1$ if the i-th element has a hash value no larger than $\frac{bk}{(1-\epsilon)\mathsf{NDE}}$ and $X = \sum_{i=1}^{\mathsf{NDE}} X_i$. Then we have $\mathbb{E}[X_i] = \frac{1}{b}\lfloor \frac{bk}{(1-\epsilon)} \rfloor$ and

$$\mathrm{Var}[X_i] \le \mathbb{E}[X_i^2] = \frac{1}{b}\lfloor \frac{bk}{(1-\epsilon)\mathsf{NDE}} \rfloor \le \lfloor \frac{k}{(1-\epsilon)\mathsf{NDE}} \rfloor,$$

  leading to

$$\mathbb{E}[X] = \sum_{i=1}^{\mathsf{NDE}} \mathbb{E}[X_i] = \frac{\mathsf{NDE}}{b}\lfloor \frac{bk}{(1-\epsilon)\mathsf{NDE}} \rfloor > \frac{\mathsf{NDE}}{b}(\frac{bk}{(1-\epsilon)\mathsf{NDE}} - 1) = \frac{k}{1-\epsilon} - \frac{\mathsf{NDE}}{b} > \frac{k}{1-\epsilon} - \frac{\epsilon k}{4}$$

  and

$$\mathrm{Var}[X] = \sum_{i=1}^{\mathsf{NDE}} \mathrm{Var}[X_i] + \sum_{1 \le i < j \le \mathsf{NDE}} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{\mathsf{NDE}} \mathrm{Var}[X_i] \le \lfloor \frac{k}{1-\epsilon} \rfloor \le \frac{k}{1-\epsilon}.$$

  So we have

$$\begin{aligned}\mathbb{P}\left[\frac{bk}{h_{kth}} \le (1-\epsilon)\mathsf{NDE}\right] &= \mathbb{P}\left[h_{kth} > \frac{bk}{(1-\epsilon)\mathsf{NDE}}\right] = \mathbb{P}[X \le k-1] \\ &= \mathbb{P}[\mathbb{E}[X] - X \ge \mathbb{E}[X] - k + 1] \le \mathbb{P}[|X - \mathbb{E}[X]| \ge \mathbb{E}[X] - k + 1] \\ &\le \mathbb{P}\left[|X - \mathbb{E}[X]| \ge \frac{k}{1-\epsilon} - \frac{\epsilon k}{4} - k\right] = \mathbb{P}\left[|X - \mathbb{E}[X]| \ge \frac{\epsilon k(3+\epsilon)}{4(1-\epsilon)}\right] \\ &\le \frac{\mathrm{Var}[X]}{\frac{\epsilon^2 k^2 (3+\epsilon)^2}{16(1-\epsilon)^2}} \le \frac{\frac{k}{1-\epsilon}}{\frac{\epsilon^2 k^2 (3+\epsilon)^2}{16(1-\epsilon)^2}} = \frac{16(1-\epsilon)}{\epsilon^2 k (3+\epsilon)^2}.\end{aligned}$$

  If $k = \lceil \frac{24}{\epsilon^2} \rceil$ where $0 < \epsilon < 0.5$, we have $\frac{16(1-\epsilon)}{\epsilon^2 k(3+\epsilon)^2} \le \frac{2(1-\epsilon)}{3(3+\epsilon)^2} \le \frac{2}{3} \cdot \frac{1}{9} = \frac{2}{27} \le \frac{1}{6}$.

- **Q3. Tight Error Bound for Misra-Gries (MG)**

  Proof: Let f_i be the true frequency of item i, and c_i be the counter value in MG. The MG algorithm decrements counters only when the summary is full (size m) and a new item arrives. This operation decrements m existing counters and the virtual counter of the new item, totaling $m+1$ decrements.

  Let d be the total number of such decrement rounds. The total count removed from the stream is $d(m+1)$. The sum of counters in the summary is $\|C\|_1 = N - d(m+1)$. For any item i, $c_i \ge f_i - d$, so the error $\delta_i = f_i - c_i \le d$.

Consider the k most frequent items. Summing their counters:

$$\sum_{j=1}^{k} c_j \geq \sum_{j=1}^{k} (f_j - d) = \sum_{j=1}^{k} f_j - k \cdot d.$$

The sum of the top k counters cannot exceed the total counter sum:

$$\sum_{j=1}^{k} f_j - kd \leq \|C\|_1 = N - d(m+1).$$

Rearranging for d:

$$d(m+1) - kd \leq N - \sum_{j=1}^{k} f_j.$$

$$d(m-k+1) \leq \sum_{j=k+1}^{n} f_j = F_1^{res(k)}.$$

$$d \leq \frac{F_1^{res(k)}}{m-k+1}.$$

Since $\delta_i \leq d$, the error is bounded by $\frac{F_1^{res(k)}}{m-k+1}$.

- **Q4. Equivalence of MG and SpaceSaving**

  Proof: Proof: We prove the hypothesis $H(n) : \hat{f}^{SS}(x) - \hat{f}^{MG}(x) = min^{SS}$ after n items by induction.

  Base Case ($n = 0$): Both summaries are empty. $\hat{f}^{SS}(x) = 0, \hat{f}^{MG}(x) = 0, min^{SS} = 0$. $0 - 0 = 0$. Holds.

  Inductive Step: Assume $H(n)$ holds. Consider the arrival of the $(n+1)$-th item y.

  Case 1: y is maintained in MG (counter > 0). MG increments: $\hat{f}^{MG}_{new}(y) = \hat{f}^{MG}_{old}(y) + 1$. By hypothesis, $\hat{f}^{SS}_{old}(y) = \hat{f}^{MG}_{old}(y) + min^{SS}_{old} > min^{SS}_{old}$. Thus y is also in SS and is not the minimum element. SS increments: $\hat{f}^{SS}_{new}(y) = \hat{f}^{SS}_{old}(y) + 1$. The minimum does not change: $min^{SS}_{new} = min^{SS}_{old}$. New difference:

  $$\hat{f}^{SS}_{new}(y) - \hat{f}^{MG}_{new}(y) = (\hat{f}^{SS}_{old}(y) + 1) - (\hat{f}^{MG}_{old}(y) + 1) = min^{SS}_{old} = min^{SS}_{new}.$$

  Case 2: y is not in MG, and MG is not full ($< k$ items). MG adds y: $\hat{f}^{MG}_{new}(y) = 1$. By hypothesis, $\hat{f}^{SS}_{old}(y) = 0 + min^{SS}_{old} = min^{SS}_{old}$. This means y is either not in SS or has the minimum counter value. SS increments the counter for y (or inserts it): $\hat{f}^{SS}_{new}(y) = min^{SS}_{old} + 1$. Note that $\hat{f}^{SS}_{old}(y)$ <u>cannot</u> be the unique minimum counter in SS. If it were the unique minimum, then all other k counters in SS would be strictly greater than min^{SS}_{old}. By the induction hypothesis, this would imply there are k items x with $\hat{f}^{MG}_{old}(x) > 0$, which would mean the MG summary is full. This contradicts the assumption of Case 2. Therefore, there is at least one other counter in SS with value min^{SS}_{old}. Thus, incrementing $\hat{f}^{SS}(y)$ does not increase the global minimum: $min^{SS}_{new} = min^{SS}_{old}$. New difference:

  $$\hat{f}^{SS}_{new}(y) - \hat{f}^{MG}_{new}(y) = (min^{SS}_{old} + 1) - 1 = min^{SS}_{old} = min^{SS}_{new}.$$

  Case 3: y is not in MG, and MG is full (k items). MG decrements all k counters by 1 (removing those that become 0). Effectively $\hat{f}^{MG}_{new}(x) = \hat{f}^{MG}_{old}(x) - 1$ for maintained items.

For y, $\hat{f}^{MG}_{new}(y) = 0$. By hypothesis, $\hat{f}^{SS}_{old}(y) = min^{SS}_{old}$. This means y either is not present in SpaceSaving or has the minimum counter. Since MG is full, there are k items x with $\hat{f}^{MG}_{old}(x) > 0$. By the hypothesis, for these k items, $\hat{f}^{SS}_{old}(x) = \hat{f}^{MG}_{old}(x) + min^{SS}_{old} > min^{SS}_{old}$. SS has exactly $k+1$ counters. Since k of them are strictly greater than min^{SS}_{old}, the remaining one (corresponding to y) must be the unique minimum. SpaceSaving will then increase $\hat{f}^{SS}(y)$, as well as min^{SS}, by 1. $min^{SS}_{new} = min^{SS}_{old} + 1$.

Check for y:

$$\hat{f}^{SS}_{new}(y) - \hat{f}^{MG}_{new}(y) = (min^{SS}_{old} + 1) - 0 = min^{SS}_{new}.$$

Check for $x \neq y$ (maintained in MG):

$$\hat{f}^{SS}_{new}(x) - \hat{f}^{MG}_{new}(x) = \hat{f}^{SS}_{old}(x) - (\hat{f}^{MG}_{old}(x) - 1) = (\hat{f}^{SS}_{old}(x) - \hat{f}^{MG}_{old}(x)) + 1 = min^{SS}_{old} + 1 = min^{SS}_{new}.$$

The relationship holds for all cases.