SEEM5020 2026 Spring: Homework 4 Solution

- **Q1. Mergeability of MG-Sketch**

  Solution: (1) **Algorithm:**

  1. **Combine:** Let S_{temp} be the union of counters from S_1 and S_2. For common items, sum their counts: $c_{temp}(e) = c_1(e) + c_2(e)$.
  2. **Prune:** Sort counters in descending order. Let C_{k+1} be the $(k+1)$-th largest counter value (or 0 if $|S_{temp}| \leq k$).
  3. **Update:** Subtract C_{k+1} from all counters in S_{temp} and remove non-positive entries. The result is S.

  (2) **Proof:** Let $N = n_1 + n_2$. We want to show $f(e) - \frac{N}{k+1} \leq \hat{f}(e) \leq f(e)$.

  Since MG only increments on arrival and decrements during pruning, $\hat{f}(e)$ never exceeds the true count. Thus $\hat{f}(e) \leq f(e)$.

  Let D_{total} be the total mass subtracted across all steps (processing streams and merging). In MG logic, a decrement of value δ is applied to at least $k+1$ distinct items simultaneously (either k counters + 1 arrival, or the $k+1$ largest counters during merge). Thus, for every unit of error introduced to a specific item e, at least $k+1$ units of mass are removed from the total stream count.

  $$(k+1)(f(e) - \hat{f}(e)) \leq D_{total}$$

  Since the remaining mass is non-negative, the total removed mass cannot exceed the input size: $D_{total} \leq N$.

  $$(k+1)(f(e) - \hat{f}(e)) \leq N \implies f(e) - \hat{f}(e) \leq \frac{N}{k+1}$$

- **Q2. Count-Sketch Analysis**

  Solution: (1) **Unbiasedness:**

  $$X_i = g_i(e)\left(f(e)g_i(e) + \sum_{x \neq e} f(x)g_i(x)\mathbf{1}_{h_i(x)=h_i(e)}\right) = f(e)g_i(e)^2 + \sum_{x \neq e} f(x)g_i(e)g_i(x)\mathbf{1}_{h_i(x)=h_i(e)}$$

  Since $g_i(e) \in \{+1, -1\}$, $g_i(e)^2 = 1$. Thus $X_i = f(e) + \sum_{x \neq e} f(x)g_i(e)g_i(x)\mathbf{1}_{h_i(x)=h_i(e)}$. Taking expectation:

  $$\mathbb{E}[X_i] = f(e) + \sum_{x \neq e} f(x)\mathbb{E}[g_i(e)g_i(x)]\mathbb{P}[h_i(x) = h_i(e)]$$

  Since g_i is pairwise independent with mean 0, for $x \neq e$, $\mathbb{E}[g_i(e)g_i(x)] = \mathbb{E}[g_i(e)]\mathbb{E}[g_i(x)] = 0 \cdot 0 = 0$. Thus, $\mathbb{E}[X_i] = f(e)$.

  (2) **Variance:**

  $$\mathrm{Var}[X_i] = \mathbb{E}[(X_i - f(e))^2] = \mathbb{E}\left[\left(\sum_{x \neq e} f(x)g_i(e)g_i(x)\mathbf{1}_{h_i(x)=h_i(e)}\right)^2\right]$$

  Let $Y_x = f(x)g_i(e)g_i(x)\mathbf{1}_{h_i(x)=h_i(e)}$. Then $\mathrm{Var}[X_i] = \mathbb{E}[(\sum_{x \neq e} Y_x)^2] = \sum_{x \neq e} \mathbb{E}[Y_x^2] + \sum_{x,y \neq e, x \neq y} \mathbb{E}[Y_x Y_y]$. Cross terms $(x, y \neq e, x \neq y)$: $\mathbb{E}[Y_x Y_y]$ involves $\mathbb{E}[g_i(e)^2 g_i(x)g_i(y)\dots] =$

$\mathbb{E}[g_i(x)g_i(y)\dots]$. Since g_i is independent for distinct elements, $\mathbb{E}[g_i(x)g_i(y)] = 0$. Thus cross terms are 0. Squared terms:

$$\mathbb{E}[Y_x^2] = f(x)^2\mathbb{E}[g_i(e)^2 g_i(x)^2 \mathbf{1}^2_{h_i(x)=h_i(e)}] = f(x)^2 \cdot 1 \cdot 1 \cdot \mathbb{P}[h_i(x) = h_i(e)] = \frac{f(x)^2}{w}$$

Summing over all $x \neq e$:

$$\mathrm{Var}[X_i] = \sum_{x\neq e} \frac{f(x)^2}{w} = \frac{1}{w}\|f_{-e}\|_2^2.$$

- **Q3. Sliding Window Cardinality Estimation**

  Solution: (1) **Bit-Decomposition Approach:** Each integer $x_t \in \{0, 1, \dots, W\}$ can be represented in binary form as $x_t = \sum_{j=0}^{L-1} b_{t,j}2^j$, where $b_{t,j} \in \{0, 1\}$ is the j-th bit of x_t and $L = \lceil \log_2(W + 1) \rceil$. We maintain L independent DGIM structures $D_0, D_1, \dots, D_{L-1}$. For each arriving integer x_t, we update D_j with the bit $b_{t,j}$. To estimate the total sum S, we let $\hat{C}_j$ be the estimated count of 1s from D_j in the window. The total sum estimate is:

$$\hat{S} = \sum_{j=0}^{L-1} 2^j \hat{C}_j$$

  (2) **Error Bound Proof:** Let $C_j = \sum_{t=c-N+1}^{c} b_{t,j}$ be the true count of 1s for the j-th bit in the window. The true total sum is $S = \sum_{j=0}^{L-1} 2^j C_j$. By the property of DGIM, each $\hat{C}_j$ satisfies:

$$(1 - \epsilon)C_j \leq \hat{C}_j \leq (1 + \epsilon)C_j$$

  which implies $|\hat{C}_j - C_j| \leq \epsilon C_j$. The absolute error of the combined estimator $\hat{S}$ is:

$$|\hat{S} - S| = \left| \sum_{j=0}^{L-1} 2^j \hat{C}_j - \sum_{j=0}^{L-1} 2^j C_j \right| = \left| \sum_{j=0}^{L-1} 2^j (\hat{C}_j - C_j) \right|$$

  Applying the triangle inequality:

$$|\hat{S} - S| \leq \sum_{j=0}^{L-1} 2^j |\hat{C}_j - C_j| \leq \sum_{j=0}^{L-1} 2^j \epsilon C_j = \epsilon \sum_{j=0}^{L-1} 2^j C_j = \epsilon S$$

  The combined estimator provides a $(1 \pm \epsilon)$-approximation of the total sum.

- **Q4. Sliding Window Frequency Estimation**

  Solution: (a) W replace the simple bit-count in each bucket with a Misra-Gries summary.

  - **Bucket Content:** Each bucket B stores a Misra-Gries sketch initialized with $m = \lceil 2/\epsilon \rceil$ counters. It also stores the timestamp of the most recent item in that bucket.
  - **Merge Logic:** When merging two buckets of the same size (capacity), we merge their MG sketches using the algorithm described in Q1.
  - **Granularity Modification:** We refine the DGIM structure to maintain $k = \lceil 4/\epsilon \rceil$ buckets of the same size level. This ensures that the size of the oldest bucket (the one overlapping the window boundary) is sufficiently small relative to N.

(b) **Proof of the ϵN Bound:** Let W be the set of items in the current sliding window of size N. Let S_{in} be the set of buckets fully contained within the window. Let B_{last} be the oldest bucket that partially overlaps the window boundary. We ignore B_{last} in our estimation to avoid overcounting items outside the window.

The total error consists of two parts:

1. **Sketch Error (E_{sketch}):** The error inherent in the MG summaries. Since MG is mergeable and linear, the total sketch error is bounded by the sum of errors of individual buckets. With $m = 2/\epsilon$ counters, the error in bucket B is $\frac{1}{m} n_B = \frac{\epsilon}{2} n_B$.

$$E_{sketch} \leq \sum_{B \in S_{in}} \frac{\epsilon}{2} n_B = \frac{\epsilon}{2} \sum_{B \in S_{in}} n_B \leq \frac{\epsilon}{2} N$$

2. **Boundary Error ($E_{boundary}$):** By ignoring B_{last}, we miss the occurrences of e that are inside the window but belong to B_{last}. The maximum possible error is the total size of B_{last}.

   Let the size of the last bucket B_{last} be 2^j. The total count of items N_{in} strictly inside the window satisfies:

$$N_{in} \geq k \cdot 2^{j-1}$$

   Since $N > N_{in}$, we have $N > k \cdot 2^{j-1} \geq 4 \cdot 2^{j-1}/\epsilon$. Thus $2^j \leq \epsilon N/2$ and $j = O(\log \epsilon N)$.

Total Error:

$$|f(e) - \hat{f}(e)| \leq E_{sketch} + E_{boundary} \leq \frac{\epsilon}{2} N + \frac{\epsilon}{2} N = \epsilon N$$

(c) **Memory Complexity Trade-off:**

- **Standard MG:** Requires $O(1/\epsilon)$ space but cannot handle sliding windows.
- **Hybrid Sliding Window:** We maintain $O(\log \epsilon N)$ levels. At each level, we maintain $k = O(1/\epsilon)$ buckets. Total number of buckets is $O(\frac{1}{\epsilon} \log \epsilon N)$. Each bucket stores an MG sketch of size $O(1/\epsilon)$. Total space is $O(\frac{1}{\epsilon^2} \log \epsilon N)$.
- **Trade-off:** We pay a multiplicative factor of $\frac{1}{\epsilon} \log \epsilon N$ in memory to support sliding window queries with the same error guarantee.