SEEM5020 2026 Spring: Homework 4

This assignment is due at 1:30 pm, February 6, 2026. Please bring your solution and submit it at the beginning of the next class. Total weight of Homework 4: 4%.

- **Q1. [10 Point] Mergeability of MG-Sketch** The Misra-Gries (MG) algorithm is a deterministic algorithm for finding frequent items. One important property of MG is its mergeability, which allows for parallel processing of data streams.

  (1) Suppose we have two MG summaries S_1 and S_2, each with k counters, representing streams D_1 and D_2 of lengths n_1 and n_2 respectively. Describe an efficient algorithm to merge S_1 and S_2 into a single summary S with k counters.

  (2) Let $f(e)$ be the true frequency of element e in the combined stream $D_1 \cup D_2$, and $\hat{f}(e)$ be its estimated frequency from the merged summary S. Prove that $f(e) - \frac{n_1+n_2}{k+1} \leq \hat{f}(e) \leq f(e)$.

- **Q2. [10 Point] Count-Sketch Analysis.** The Count-Sketch algorithm (Slide 21) uses hash functions $h_i : [m] \to [w]$ and $g_i : [m] \to \{+1, -1\}$. The update rule is $A[i][h_i(e)] \leftarrow A[i][h_i(e)] + g_i(e) \cdot c_e$. The estimate is the median of $g_i(e)A[i][h_i(e)]$ over d rows.

  Focus on a single row i. Let $X_i = g_i(e) \cdot A[i][h_i(e)]$ be the estimator from the i-th row for the frequency $f(e)$. Assume h_i is pairwise independent and g_i is pairwise independent with $\mathbb{P}[g_i(x) = 1] = \mathbb{P}[g_i(x) = -1] = 0.5$.

  (1) Prove that $\mathbb{E}[X_i] = f(e)$. (Unbiasedness)

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

- **Q3. [8 Point] Sliding Window Cardinality Estimation.** The standard DGIM algorithm is designed to estimate the number of 1s in a sliding window of size N. Suppose we want to extend this to handle a stream of non-negative integers $x_1, x_2, \ldots$, where each $x_t \in \{0, 1, \ldots, W\}$, and we want to estimate $S = \sum_{t=c-N+1}^{c} x_t$ in the sliding window.

  (1) Describe an approach to solve this problem by using $L = \lceil \log_2(W + 1) \rceil$ independent DGIM structures.

  (2) Prove that if each DGIM structure provides a $(1 \pm \epsilon)$-approximation for the count of bits at its corresponding position, the combined estimator $\hat{S}$ also provides a $(1 \pm \epsilon)$-approximation for the total sum S.

- **Q4. [12 Point] Sliding Window Frequency Estimation.**

  Suppose you are given a stream of items where you need to estimate the frequency of any element e within a sliding window of the last N elements. You must ensure that your estimate $\hat{f}(e)$ satisfies $f(e) - \hat{f}(e) \leq \epsilon N$. Please design a hybrid algorithm that combines the "Bucket" logic of DGIM with the "Counter" logic of the Frequent Algorithm (MG). Notice that in Q1 we established the mergeability of MG. You can take advantage of that.

  (a) Describe how you would modify the DGIM bucket structure to store frequency information instead of just bit counts.

  (b) Using the fact that Misra-Gries with $m = 1/\epsilon$ counters has an error bound of $\epsilon \times$ (number of elements processed), prove that your hybrid structure maintains the ϵN bound across the entire sliding window.

  (c) Briefly explain the trade-off in memory complexity when moving from a standard Misra-Gries sketch to your sliding window version.