SEEM5020 2026 Spring: Homework 3

This assignment is due at 1:30 pm, January 30, 2026. Please bring your solution and submit it at the beginning of the next class. Total weight of Homework 3: 4%.

- **Q1. [10 Point] Second Moment of MinHash.** In the lecture (Slide 27 of Lecture 2), we stated Lemma 5 regarding the second moment of the minimum hash value X_{min} used in the MinHash algorithm.

  Prove that:
  $$\mathbb{E}[X_{min}^2] = \frac{2}{(\mathsf{NDE}+1)(\mathsf{NDE}+2)}$$
  You may use the fact that for a non-negative random variable Y, $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > t)dt$.

- **Q2. [10 Point] Bottom-k Sketch (Lower Tail).** In the lecture, we analyzed the upper tail of the Bottom-k sketch. Now, prove the lower tail bound. Specifically, let $\hat{n} = \frac{b \cdot k}{h_{kth}}$ be the estimator for NDE. Prove that:
  $$\mathbb{P}[\hat{n} < (1-\epsilon)\mathsf{NDE}] < \frac{1}{6}$$
  You may assume $k = \lceil \frac{24}{\epsilon^2} \rceil$, $\epsilon < 0.5$, and $\frac{1}{b} = \frac{1}{n^3} < \frac{\epsilon \cdot k}{4\mathsf{NDE}}$.

- **Q3. [10 Point] Tight Error Bound for Misra-Gries (MG).** The standard analysis of the Misra-Gries algorithm with m counters guarantees an error bound of $N/(m+1)$. However, more careful analysis gives a stronger "tail" guarantee.

  Prove that for the Misra-Gries algorithm using m counters, the error δ_i in the estimated frequency of any item i satisfies:
  $$\delta_i \leq \frac{F_1^{res(k)}}{m-k+1}$$
  for any $1 \leq k \leq m$, where $F_1^{res(k)}$ is the sum of the frequencies of all items except the k most frequent ones.

  Remark: Notice that $\frac{F_1^{res(k)}}{m-k+1} \leq \frac{N}{m+1}$, just imagine that the right "cup" has more "sugar".

- **Q4. [10 Point] Equivalence of MG and SpaceSaving.** Please prove Theorem 3 of Lecture 3 with the following rephrased statement. Consider an MG summary with k counters and a SpaceSaving summary with $k+1$ counters processing the same stream. Let $\hat{f}^{MG}(x)$ be the counter for item x in MG (0 if not maintained) and $\hat{f}^{SS}(x)$ be the counter for item x in SS (0 if not maintained). Let min^{SS} be the minimum counter value in the SS summary.

  Prove by induction that after processing any sequence of items:
  $$\hat{f}^{SS}(x) - \hat{f}^{MG}(x) = min^{SS}$$
  for all items x.

  Remark: Denote $\|SS\|_1$ (resp. $\|MG\|_1$) as the sum of the counters in SS (resp., MG). The result implies that $min^{SS} = (\|SS\|_1 - \|MG\|_1)/(k+1)$.