Math 126 Problem 3: Count-min Works Super Well

Spring 2027

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Monday, February 1, 2027 (Pacific time).**

- You are encouraged to discuss ideas and work with your classmates. However, you *must acknowledge* your collaborators at the top of each solution on which you collaborated with others and you *must write* your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown.
- I recommend that you write your solutions in a Python notebook using Google Colab.
- You should submit your solutions as a *single PDF* via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by *not* grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.
Original notebook: Google Colab.

Count-min Works Super Well

We showed that Count-Min can estimate the frequency of any item in a stream of n items up to additive error $\frac{1}{m}n$ using $O(m)$ space. In practice it is often observed that this bound is pessimistic: the algorithm performs better than expected. In this problem, you will establish one reason why.

For any positive integer m, let $f_1, \ldots, f_m$ be the frequencies of the m most frequent items in our stream. Let $C = n - \sum_{i=1}^{m} f_i$. In general, we can have that $C \ll n$. For example, it has been observed that up to 95% of YouTube video views come from just 1% of videos. Prove that using $O(m)$ space, Count-Min actually returns an estimate $\tilde{f}(v)$ to $f(v)$ for any item v satisfying:

$$f(v) \leq \tilde{f}(v) \leq f(v) + \frac{2}{m}C$$

with 9/10 probability. This is strictly better than the $\frac{1}{m}n$ error bound shown in class.

Part A

Explain why it suffices to show the equation holds with any constant probability $c > 0$.

Then write the estimate $\tilde{f}(v) = A[h(v)]$ as the true frequency $f(v)$ and two error terms; the first error term is for the m most frequent items that could collide with v and the second error term is for the remaining items that could collide with v..

Part B

Prove that the first error term is 0 with constant probability; that is, there is a constant probability that none of the m most frequent items collide with v.

Hint: Use one of these inequalities:

$$\frac{1}{e} \geq \left(1 - \frac{1}{m}\right)^m \geq \frac{1}{2e}.$$

You can check these inequalities for yourself on Desmos.

Following the analysis in class, prove that the second error term is at most $\frac{2}{m}C$ with constant probability.

Show how both results together imply that the equation holds with a constant probability.