Problems 1, 3, 4 require you to have read the lecture notes or watched the lecture, but you should be able to attempt them on your own after that. They are important to attempt, in order to build a good understanding about the algorithms seen in class.

Problem 2 does not require having seen the lecture, and can be skipped. It is a useful fact, good to know, but that you can take for granted.

Problem 5 is good to go over if you have time, but also can be skipped. In that can, go over the solution afterwards.

Problem 6 is on the more difficult side, but worth doing to understand why the MG algorithms is a sketching algorithm.

Problem 7 is interesting if you have time, but not necessary.

Warm-up

Problem 1. Discussion: what are the parallels between Bloom filters and CountMinSketch?

Problem 2. Prove the following fact about "monotonicity of ℓ_p norms": if $x \in \mathbb{R}^d$, then $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1$. Show, in addition, that $\|x\|_2 \geq \|x\|_1/\sqrt{d}$. When are these inequalities tight?
(($\star$) More generally: if $1 \leq p \leq q \leq \infty$, then $\|x\|_q \leq \|x\|_p$.)

Problem 3. Discuss the advantages and disadvantages of Misra-Gries versus CountMinSketch when used in the cash register model: speed, memory, approximation. Can you think of a situation where having an overestimate (CountMinSketch) is better than an underestimate (Misra-Gries)?

Problem solving

Problem 4. For the same space budget s (ignoring the constants in the $O(\cdot)$'s), are the theoretical guarantees provided by CountMinSketch better, worse, or incomparable to those of CountSketch?

Problem 5. Generalise the analysis of the CountMinSketch algorithm to show it works in the *strict* turnstile model, where updates of the stream are of the form $(j, c) \in [n] \times \{-B, \ldots, B\}$ (can be negative) but one must have $f_j \geq 0$ at every time. Check the guarantees you can provide on the output $\widehat{f}$. Does the analysis extend to the general turnstile model, where f_j can become negative?

Problem 6. ($\star$) Show that the Misra-Gries algorithm is a sketching algorithm: namely, suppose we run Misra-Gries (with the same parameter $k = \lceil 1/\varepsilon \rceil$) on two streams σ_1, σ_2, getting output vectors $\widehat{f}^{(1)}, \widehat{f}^{(2)}$. Combine then as follows:

1. Set $\widehat{f} \leftarrow \widehat{f}^{(1)} + \widehat{f}^{(2)}$
2. If $\widehat{f}$ has more than k non-zero entries, let $v > 0$ be the value of the $(k+1)$-th, in non-increasing order.
3. Set $\widehat{f}_j \leftarrow \max(\widehat{f}_j - v, 0)$ for all j

a) Argue that $\widehat{f}$ has at most k non-zero entries.

b) Show that the sketch $\widehat{f}$ provides the original MISRA-GRIES estimation guarantees, for the combined stream $\sigma_1 \circ \sigma_2$.

c) Is this sketch a linear sketch?

Advanced

Problem 7. Modify the COUNTMINSKETCH algorithm so that it outputs a *list* of the ℓ_1 Heavy Hitters in the strict turnstile model: that is (similarly to an exercise in Tutorial 8), given parameter $\varepsilon \in (0, 1]$, it should output a set $H \subseteq [n]$ such that $H_\varepsilon(\sigma) \subseteq H \subseteq H_{\varepsilon/2}(\sigma)$, where

$$H_\varepsilon(\sigma) = \{j \in [n] : f_j \geq \varepsilon \cdot \|f\|_1\}$$