Problem: Given a stream of numbers, want to track the majority if it exists.

\[ n \text{ numbers: } x_1, x_2, \ldots, x_n \]

\[ m = \text{majority}(x_1, \ldots, x_n) \]

\[ \text{if } \# \{ i : x_i = m \} > n/2 \]

\[ O(\log n) \text{ bits of memory?} \]

- Initialize counter := 0, current := NULL.
  [current stores the frontrunner for the majority element]
- For \( i = 1 \) to \( n \):
  - If counter == 0:
    [In this case, there is no frontrunner.]
      * current := A[i]
      * counter++
  - else if A[i] == current:
    [In this case, our confidence in the current frontrunner goes up.]
      * counter++
  - else
    [In this case, our confidence in the current frontrunner goes down.]
      * counter--
- Return current

**E-Hit (Heavy hitters):** Given \( x_1, x_2, \ldots, x_n \).
say that $y$ is an $\epsilon$-HH if

$$\# \{ i : x_i = y \} \geq \epsilon n$$

Problem: Output the most popular element in the stream.

Claim: Requires $\Omega(n \log n)$ bits of memory to solve.

Given a stream $x_1, x_2, x_3, \ldots, x_m$.

$$S = \{ x_1, x_2, \ldots, x_m \}$$

$$\# \text{ of such } S \text{ is } 2^n$$

$$\Rightarrow 2^n \text{ bits of memory required}$$

Problem: Given a stream $x_1, x_2, \ldots, x_n$ and parameters $\epsilon > 0$, $k$.

Goal:

(i) Output all $x$ that occur $\geq \frac{n}{k}$ times
(ii) If we output $x$, then $x$ occurs $\geq \frac{n}{k} - \epsilon n$

Space usage: $O(\frac{1}{\epsilon^2})$

- $e = \frac{1}{2k}$
- Output all $x$ that occur $\geq \frac{n}{k}$ times, and only $x$ that occur $\geq \frac{n}{2k}$ times.

Space: $O(k)$ "words"

Count-min sketch: $Inc(x), Count(x)$

- $b = \#buckets$
- $l = \#hash functions$

$Inc(x): CMS[i, h_0(x)]++$ for $i = 1, \ldots, l$
$\text{Count}(x)$: Suppose $x$ occurs $f_x$ times

\[
\min (\text{CMS}[L, h(x)], \text{CMS}[L, h(x)]).
\]

\[
\text{CMS}[L, h(x)]
\]

\[
 Z = \text{CMS} [L, h(x)] = f_x + \sum_{y: \ y \neq x} \frac{n - f_x}{b} \quad \text{if } x \neq y
\]

\[
\frac{n - f_x}{b} \leq \frac{n}{b}
\]

\[
E[Z] \leq f_x + \frac{n}{b} \leq f_x + \frac{en}{2}
\]

Markov's ine. $Y \geq 0$ is a r.v.,

\[
\Pr [Y \geq x] \leq \frac{E[Y]}{x}
\]

\[
\Rightarrow \Pr [Z - f_x \geq en] \leq \frac{1}{2}
\]

\[
\Rightarrow \Pr [Z \leq f_x + en] \geq \frac{1}{2}
\]
\[ \Pr \left[ \min \{ \text{CMS}_{1}, h(x) \}, \ldots, \text{CMS}_{m}, h(x) \right] \]
\[
\geq f_x + \epsilon \] \[ \leq \left[ 2^{-\ell} \right]^{1/8} \]
\[ \ell := \log_2 (1/\delta) \]

\( \epsilon, \delta > 0 \)

Count \( x \) to be accurate within \( \epsilon \)
with prob \( \geq 1 - \delta \)

\[
\Rightarrow \text{space} = O(\beta \ell) 
\]
\[
= O\left( \frac{1}{\epsilon} \log \frac{1}{\delta} \right) 
\]

---

Problem: Given a stream \( X_1, X_2, \ldots, X_n \)
and params \( \epsilon, \delta, k \)

Goal:
(i) Output all \( x \) that occur \( \geq \frac{n}{k} \) times
(ii) If we output \( x \), then \( x \) occurs \( \geq \frac{n}{k} - \epsilon \)
$$b := \frac{2}{6} \quad \epsilon = \frac{1}{2k} \quad \Theta(k)$$