Problem: Maintain heavy hitters using Count-Min Sketch

i is a heavy hitter if $f_i \geq \alpha \|f\|_1$ for given $0<\alpha<1$

approximation: set H s.t.

$$HH_\alpha \subseteq H \subseteq HH_{\alpha-\varepsilon} \quad \text{for some } 0<\varepsilon<\alpha$$

Count-Min-Sketch gave us the following guarantee (see Theorem)

For every fixed $i \in [n]$: $\cdot\ f_i \leq \tilde{f}_i$ ← estimate $\tilde{f}_i$ that can be queried

$\cdot\ \tilde{f}_i < f_i + \varepsilon \|f\|_1$ with probability $\geq 1-\delta$

using space $O(\frac{1}{\varepsilon} \log \frac{1}{\delta} \cdot (\log n + \log \|f\|_1))$

modification:

- Maintain a set H using a binary search tree
- $H = \{i \mid \tilde{f}_i \geq \alpha \|f\|_1\}$ Let's assume we maintain $\|f\|_1$ separately ($O(\log \|f\|_1)$ additional space)

- Maintain an approximation of H: Every time a frequency is changed in the stream for some item i (i.e., item i is read from the stream), we check whether i is in H or not
  ↳ We don't prove this formally here because it's a bit tedious

<u>Claim</u>: $HH_\alpha \subseteq H \subseteq HH_{\alpha-\varepsilon}$ with probability $\geq 1 - n\delta$

<u>Proof</u>: Let $i \in HH_\alpha$

$$\Rightarrow f_i \geq \alpha \cdot \|f\|_1$$

By guarantee of Count-Min-Sketch:

$$\tilde{f}_i \geq f_i$$

$$\Rightarrow \tilde{f}_i \geq \alpha \cdot \|f\|_1 \Rightarrow i \in H$$

Thus: $HH_\alpha \subseteq H$

Let $i \in H$

$$\Rightarrow \tilde{f}_i \geq \alpha \cdot \|f\|_1$$

With probability $\geq 1-\delta$, the following inequality holds:

$$\tilde{f}_i \leq f_i + \varepsilon \|f\|_1$$

$$\Rightarrow \alpha \|f\|_1 \leq f_i + \varepsilon \|f_1\|$$

$$\Rightarrow f_i \geq (\alpha - \varepsilon) \cdot \|f\|_1$$

$$\Rightarrow i \in HH_{\alpha-\varepsilon}$$

$$\Rightarrow H \subseteq HH_{\alpha-\varepsilon} \text{ with probability } \geq 1 - n\delta \qquad \square$$

Running Count-Min Sketch with failure probability $\delta' = \frac{\delta}{n}$ gives use success probability of $\geq 1 - \delta$ requiring space

$$O\left(\frac{1}{\varepsilon} \cdot \log\left(\frac{n}{\delta}\right) \cdot (\log(n) + \log(\|f\|_1)\right)$$

Additional space for storing H $\leq |HH_{\alpha-\varepsilon}| \cdot O(\log n)$ with prob $1-\delta$

$$\leq \frac{\|f\|_1}{(\alpha - \varepsilon)\|f\|_1} \cdot O(\log n) \stackrel{?}{=} O((\alpha - \varepsilon) \log(n))$$

$$= O(\alpha \log(n))$$