Input stream \( a_1, ... a_n \).
Each \( a_i \) belongs to \( \{0, 1\}^b \) (set of "tokens")
\( m = 2^b = \# \) of possible tokens.

Algorithm with working space \( s = \text{poly} (\log n, \log m) \) observes \( a_1, ... a_n \) one by one, and afterward must estimate \( \# \) distinct tokens appearing in the sequence.

**Goal**: \((e, \delta)\)-PAC meaning: if true number of distinct tokens is \( d \), then
\[
\Pr(\text{ALG}'s \text{ output is not in } [\left[ (1-e)d, (1+e)d \right] ]) < \delta.
\]

Use hash function \( h : \{0, 1\}^b \rightarrow [M] \) for some large integer \( M \). (Typically \( M \geq 2^b \).)

Suppose for now that \( a_1, ... a_d \) are the \( d \) distinct tokens that appear in the stream and that \( h(a_1), h(a_2), ..., h(a_d) \) are independent uniformly random elements of \([M]\).

\[
\Pr( h(a_1), ... h(a_d) > \frac{M}{2d} )
= \left( 1 - \frac{1}{d} \right)^d \approx e^{-1}.
\]

\[
\Pr( \min_i h(a_i) \leq \frac{M}{2d} )
\leq \sum_{i=1}^{d} \Pr( h(a_i) \leq \frac{M}{2d} )
= \frac{1}{d}.
\]

\[
\mathbb{E} \left[ \min \{ h(a_1), ..., h(a_d) \} \right] \approx \frac{M}{2d}.
\]
Alg. 1. Initialize $Z = M$

2. for each ad in succession
   - compute $h(a_i)$
   - if $h(a_i) < Z$, $Z \leftarrow h(a_i)$
   // invariant: after $t$ iterations
   \[ Z = \min \{ h(a_j) \mid 1 \leq j \leq t \} \]

3. output $M/Z$ as our estimate of $d$.

Analysis. For $k \in [M]$ let $X_{ik} = \begin{cases} 1 & \text{if } h(a_i) \leq k \\ \emptyset & \text{if } h(a_i) > k \end{cases}$

\[ Y_k = \sum_{i=1}^{d} X_{ik} = \text{# distinct tokens in stream whose hash value is } k. \]

Fact 1. $E[X_{ik}] = \frac{k}{M}$

Fact 2. $E[Y_k] = \frac{dk}{M}$

Fact 3. $\text{Var}(Y_k) < \frac{dk}{M}$ for all $i$

\[ \begin{align*}
\text{Var}(Y_k) &= E(Y_k^2) - E(Y_k)^2 \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} E[X_{ik}X_{jk}] - \left( \frac{dk}{M} \right)^2 \\
&= \sum_{i=1}^{d} \sum_{j=1}^{d} E[X_{ik}^2] + 2 \sum_{1 \leq i < j \leq d} E[X_{ik}X_{jk}] - \left( \frac{dk}{M} \right)^2
\end{align*} \]
\[
= \frac{dk}{M} + 2 \left( \frac{d}{2} \right) \left( \frac{k}{M} \right)^2 - \left( \frac{dk}{M} \right)^2
\]
\[
= \frac{dk}{M} + \left( \frac{d^2-d}{2} \right) \left( \frac{k}{M} \right)^2 - \left( \frac{dk}{M} \right)^2
\]
\[
= \frac{dk}{M} - \frac{dk^2}{M^2}
\]
\[
< \frac{dk}{M}.
\]

**Def.** A 2-universal hash function (UHF) is a probability distribution over hash functions that satisfies:

1. \( h(a_i) \) is uniform distribution over its range for all \( i \).
2. \( h(a_i), h(a_j) \) independent \( \forall i \neq j \).

"Pairwise independence."

**Example.** Say \( M \) is prime, and identify \( \{0,1,\ldots,m-1\} \) with \([m]_M\), and assume \( M \geq m \).

Define \( h \) as follows: \( h_{ab}(x) = ax+b \mod M \)

Sample \( a,b \) as independent, uniform random \( \mod M \), set \( h = h_{ab} \).
Storing a description of $h$ takes $2 \log(M)$ bits to store $a,b$.

To verify 2-universality, must show

$$\forall a_i \neq a_j \Pr_{a_i,b} \left( h(a_i) = x, \ h(a_j) = y \right) = \frac{1}{M^2} \quad \forall x,y$$

$$h_{ab} = x, \ h_{ab} = y$$

means

$$\begin{pmatrix} a_i & 1 \\ a_j & 1 \end{pmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix} \quad \pmod{M}$$

The linear system has a unique solution \( \pmod{M} \) because \( \mathbb{Z}/(M) \) is a field and

$$\det \begin{pmatrix} a_i & 1 \\ a_j & 1 \end{pmatrix} = a_i - a_j \not\equiv 0 \pmod{M}$$