1 Alternate Method for Number of Distinct Elements

We want an algorithm that takes space $O(\log n)$. Suppose we have a sequence $x_1, x_2, ..., x_n$ where $x_i \in \{1, 2, ..., m\}$.

Let’s select a subset $S \subseteq \{1, 2, ..., m\}$ uniformly at random where $|S| \leq \sqrt{m}$. Let $\min$ be the smallest element in $S$. Because we selected uniformly at random, $\min \approx \frac{m}{|S| + 1}$.

$$(\min)|S| + \min = m$$

$$|S| = \frac{m}{\min} - 1$$

We have a problem, however, because our $S$ really won’t be selected uniformly at random. What’s the solution? Just pick a function $h$ such that $h(i)$ is selected uniformly at random from $\{1, 2, ..., m\}$. The difficulty now is having to actually store all the values.

2 Universal Hash Functions

The set of hash functions $H = \{h| h : \{1, 2, ..., m\} \rightarrow \{1, 2, ..., m\}\}$ is 2-universal if for all $x \neq y$ in $\{1, 2, ..., m\}$ where $x \neq y$ and for all $z$ & $w$:

$$\Pr[h(x) = z \text{ and } h(y) = w] = \frac{1}{m^2}$$

This must hold for any randomly selected hash function $h$ in $H$.

An example that is not 2-universal: $H = \{h_i|h_i(x) = i\}$

An example that is 2-universal: for each pair of integers $a$ & $b$, let $h_{ab}(x) = ax + b \mod m$. To see that this family of hash functions is 2-universal, note that $h(x) = z$ and $h(y) = w$ when:

$$\begin{pmatrix} x & 1 \\ y & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} z \\ w \end{pmatrix} \mod m$$

If $x \neq y$, matrix is invertible so there must be a unique solution for $a$ and $b$. 