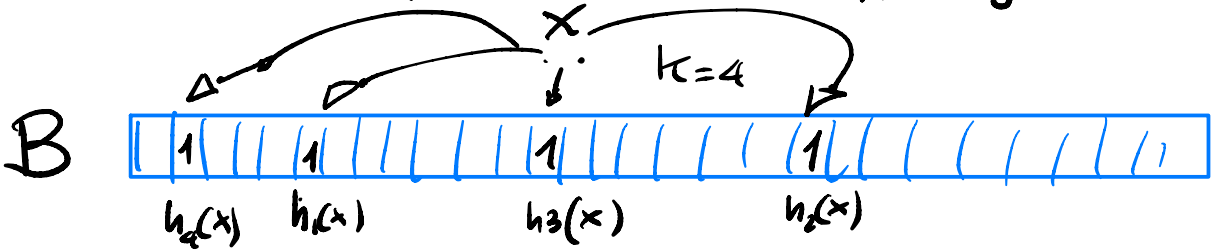
BLOOM FILTERS : it does not store the keys (Montecarlo algs.)

- set $S \subseteq U$ of keys (but we do not want to store them in the BF.!)
- universal hash family $\mathcal{H}$
- k rnd $h_1, h_2, \ldots, h_k \in \mathcal{H}$ randomly and uniformly chosen
  $h_i : U \to [m]$

BUILDING

- B = bitvector = compact array of m bits (like C++ bitset)
  (initialized to all 0s)
- Idea: $x \in S \xrightarrow{BF}$ for $i = 1, 2, \ldots, k$: $B[h_i(x)] = 1$



What we store:

- B — m bits + $O(1)$ words
- $h_1, h_2, \ldots, h_k$ — $2k$ words

(or e.g. $m, k, p, n = |S|$, etc.)

Space cost is dominated by m. Q. How to choose m, k?

Implict condition:

$$x \in S \iff B[h_1(x)] = 1 \wedge B[h_2(x)] = 1 \wedge \ldots \wedge B[h_k(x)] = 1 \quad (*)$$

Error due to the implicit condition:

1-side error:

- if $x \in S$ really $\Rightarrow$ $(*)$ is true
- if $x \notin S$ $\Rightarrow$ $(*)$ is false

: however, $(*)$ = true it is possible that $x \notin S$

f = FAILURE PROBABILITY

Query: $x \rightsquigarrow$ test $x \in S$ in $O(k)$ time

Insertion: $x \rightsquigarrow$ set $x \in S$ in $O(k)$ time

Deletion: it cannot be executed in this version (use counters)

Q: ① error probability f

② choose "best" parameters

- Universal hashing: $Pr[h_i(x) = y] = \frac{1}{m}$
- $f = Pr(\text{ERROR}) = Pr(x \notin S \text{ but } \circledast = \text{true})$

Small steps: suppose we built B, take any position $q \in [m]$

$\underline{Q1}$: $Pr[B[q] = 1] = 1 - Pr[B[q] = 0]$

Q_1': fix h_i: $Pr[B[q]=0 \mid h_i \text{ is chosen}] = 1 - \frac{1}{m}$

as $Pr[h_i(x) = q] = \frac{1}{m}$

Q_1: $Pr[B[q]=0] = (1 - \frac{1}{m})^k$ as all h_i's must give values $\neq q$

Repeating the above argument for all keys in S

$Pr[B[q]=0] = ((1 - \frac{1}{m})^k)^n$: this is what we have after building B

$P' = \left(1 - \frac{1}{m}\right)^{nk}$ $\rightsquigarrow$ $B[q] = 0$

$1 - p' \rightsquigarrow B[q] = 1$ Q1

$f = Pr[\otimes] = (1-p')^k$ where $p' = \left(1 - \frac{1}{m}\right)^{nk}$

FAILURE PROBABILITY that we want to minimize

Taylor expansion $(1+y) \sim e^y$ for small y

$$p' = \left(1 - \underbrace{\frac{1}{m}}_{y}\right)^{nk} \sim \left(e^{-\frac{1}{m}}\right)^{nk} = e^{-\frac{nk}{m}} = p$$

we minimize $(1-p)^k$

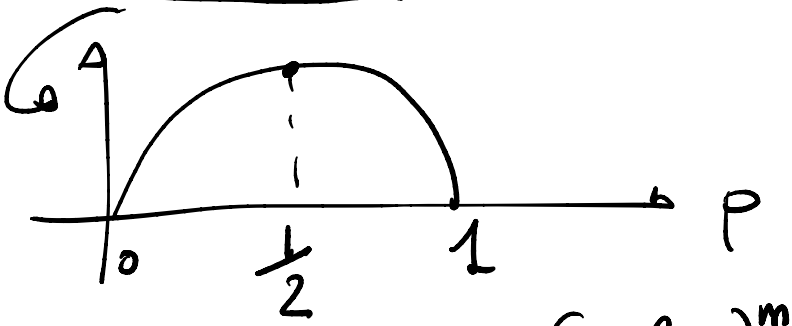
minimize $f \sim (1-p)^k$ where $p = e^{-\frac{nk}{m}}$

take logs. $\ln (1-p)^k = k \ln(1-p)$

$\ln p = -\frac{nk}{m} \underbrace{\ln e}_{1} \Rightarrow k = -\frac{m}{n} \ln p$

replace k

$k \ln(1-p) = -\frac{m}{n} \cdot \ln p \cdot \ln(1-p)$ ← minimize over p



Best choice is $p = \frac{1}{2}$

$f \sim (1-p)^k = \frac{1}{2^k} = \frac{1}{2^{-\frac{m}{n} \ln \frac{1}{2}}} = \frac{1}{2^{\frac{m}{n} \ln 2}} = \underbrace{\left(2^{-\ln 2}\right)^{m/n}}_{\sim 0.618} < 0.618^{m/n}$

$k = -\frac{m}{n} \ln p = \frac{m}{n} \ln 2$

you have to choose m accordingly