Advanced Algorithms

南京大学

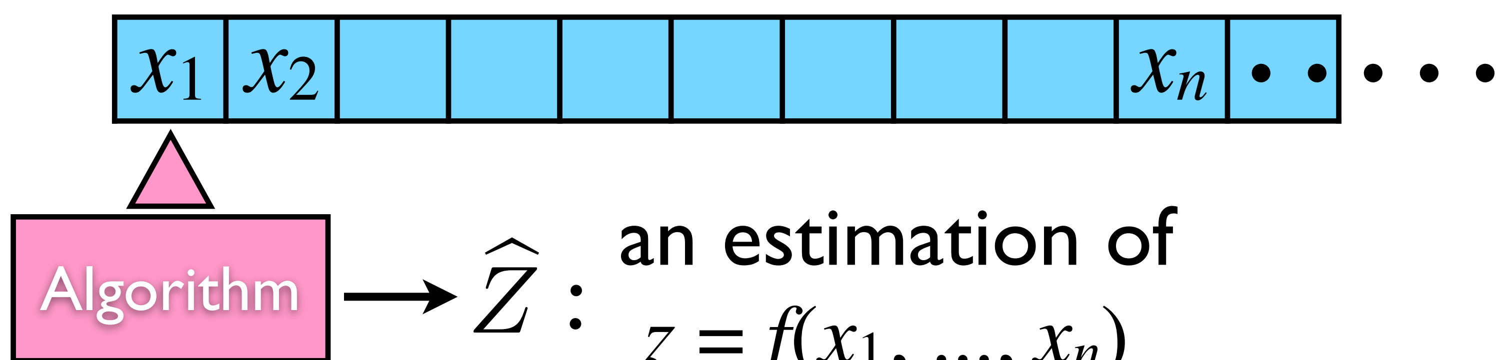
尹一通

Count Distinct Elements

Input: a sequence $x_1, x_2, \ldots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \ldots, x_n\}|$

- data stream: input comes one at a time
- naive algorithm: store everything with $O(n)$ space



- (ε,δ)-estimator: $\Pr\left[(1-\epsilon)z \le \widehat{Z} \le (1+\epsilon)z\right] \ge 1-\delta$

Using only memory equivalent to 5 lines of printed text, you can estimate with a typical accuracy of 5% and in a single pass the total vocabulary of Shakespeare. -----Flajolet

Input: a sequence $x_1, x_2, \ldots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \ldots, x_n\}|$

- (ε,δ)-estimator: $\Pr\left[(1-\epsilon)z \le \widehat{Z} \le (1+\epsilon)z\right] \ge 1-\delta$

uniform hash function $h: \Omega \to [0,1]$

$h(x_1), \ldots, h(x_n)$: z uniform independent values in $[0,1]$

(partition $[0,1]$ into z+1 subintervals)

$$\mathbb{E}\left[\min_{1\le i\le n} h(x_i)\right] = \mathbb{E}[\text{ length of a subinterval}] = \frac{1}{z+1}$$

(by symmetry)

estimator: $\widehat{Z} = \frac{1}{\min_i h(x_i)} - 1$?

But $\mathrm{Var}[\min_i h(x_i)]$ is too large!

(think of $z = 1$)

Markov's Inequality

Markov's Inequality:

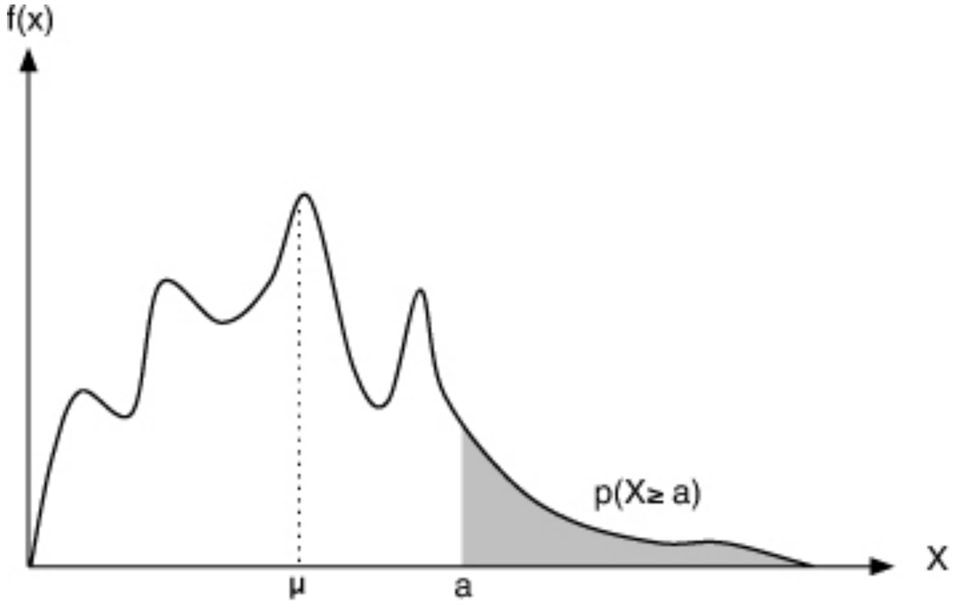
For *nonnegative* X, for any $t > 0$,

$$\Pr[X \geq t] \leq \frac{\mathbf{E}[X]}{t}.$$

Proof:

Let $Y = \begin{cases} 1 & \text{if } X \geq t, \\ 0 & \text{otherwise.} \end{cases} \Rightarrow Y \leq \left\lfloor \frac{X}{t} \right\rfloor \leq \frac{X}{t},$

$$\Pr[X \geq t] = \mathbf{E}[Y] \leq \mathbf{E}\left[\frac{X}{t}\right] = \frac{\mathbf{E}[X]}{t}.$$



tight if we only know the expectation of X

A Generalization of Markov's Inequality

Theorem:

For any X, for $h: X \mapsto \mathbb{R}^+$, for any $t > 0$,

$$\Pr[h(X) \geq t] \leq \frac{\mathbf{E}[h(X)]}{t}.$$

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr\left[|X - \mathbf{E}[X]| \geq t\right] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Variance:

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

$$\mathbf{Var}[cX] = c^2\mathbf{Var}[X]$$

$\mathbf{Var}\left[\sum_i X_i\right] = \sum_i \mathbf{Var}[X_i]$ for pairwise independent X_i

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr\left[|X - \mathbf{E}[X]| \geq t\right] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Proof:

Apply Markov's inequality to $(X - \mathbf{E}[X])^2$

$$\Pr\left[(X - \mathbf{E}[X])^2 \geq t^2\right] \leq \frac{\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]}{t^2}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \ldots, x_n\}|$

- (ε,δ)-estimator: $\Pr\left[(1-\epsilon)z \le \widehat{Z} \le (1+\epsilon)z\right] \ge 1-\delta$

uniform independent hash functions:

$h_1, h_2, \ldots, h_k : \Omega \to [0,1]$ $\qquad Y_j = \min_{1\le i\le n} h_j(x_i)$

average-min: $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$

Flajolet-Martin estimator: $\widehat{Z} = \frac{1}{\overline{Y}} - 1$

UHA: Uniform Hash Assumption

unbiased estimator: $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$

- Deviation: $\Pr\left[\widehat{Z} < (1-\epsilon)z \text{ or } \widehat{Z} > (1+\epsilon)z\right] <?$

$$z = |\{x_1, x_2, \ldots, x_n\}|$$

For j=1, 2, …, k, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \ldots, X_{jz} \in [0, 1]$

$$\left.\begin{array}{l} Y_j = \min_{1\le i\le n} X_{ji} \\ \overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j \end{array}\right\} \xrightarrow{\text{symmetry}} \boxed{\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\boxed{\text{let } \widehat{Z} = \frac{1}{\overline{Y}} - 1}$

goal: $\Pr\left[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z\right] < \delta$

for $\varepsilon \le 1/2$

$$\boxed{\left|\overline{Y} - \frac{1}{z+1}\right| > \frac{\epsilon/2}{z+1}}$$

$z = |\{x_1, x_2, \ldots, x_n\}|$

For j=1, 2, …, k, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \ldots, X_{jz} \in [0, 1]$

$$\left.\begin{array}{l} Y_j = \min_{1 \le i \le n} X_{ji} \\ \overline{Y} = \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Y_j \end{array}\right\} \xrightarrow{\text{symmetry}} \boxed{\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\boxed{\text{let } \widehat{Z} = \dfrac{1}{\overline{Y}} - 1}$

goal: $\Pr\left[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z\right] < \delta$

$\Downarrow$ for $\varepsilon \le 1/2$

$$\boxed{\left|\overline{Y} - \mathbb{E}[\overline{Y}]\right| > \frac{\epsilon/2}{z+1}}$$

$$z = |\{x_1, x_2, \ldots, x_n\}|$$

For $j=1, 2, \ldots, k$, hash values of h_j:

uniform independent $X_{j1},\ X_{j2}, \ldots, X_{jz} \in [0, 1]$

$$\left.\begin{aligned} Y_j &= \min_{1\le i\le n} X_{ji} \\ \overline{Y} &= \frac{1}{k}\sum_{j=1}^{k} Y_j \end{aligned}\right\} \xrightarrow{\text{symmetry}} \boxed{\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\boxed{\text{let } \widehat{Z} = \frac{1}{\overline{Y}} - 1}$

$$\Pr\left[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z\right]$$

(for $\varepsilon \le 1/2$) $\le \Pr\left[\left|\overline{Y} - \mathbb{E}[\overline{Y}]\right| > \frac{\epsilon/2}{z+1}\right]$

Chebyshev: $\le \frac{4}{\epsilon^2}(z+1)^2\mathbf{Var}[\overline{Y}]$

$z = |\{x_1, x_2, \ldots, x_n\}|$

For $j=1, 2, \ldots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \ldots, X_{jz} \in [0, 1]$

$$\left.\begin{aligned} Y_j &= \min_{1\le i\le n} X_{ji} \\ \overline{Y} &= \frac{1}{k}\sum_{j=1}^{k} Y_j \end{aligned}\right\} \xrightarrow{\text{symmetry}} \boxed{\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

geometry

probability

$$\Rightarrow \Pr[Y_j \ge y] = (1-y)^z \Rightarrow \text{pdf} = z(1-y)^{z-1}$$

$$\mathbb{E}[Y_j^2] = \int_0^1 y^2 z(1-y)^{z-1}\,\mathrm{d}y = \frac{2}{(z+1)(z+2)}$$

$$\mathbf{Var}[Y_j] = \mathbb{E}[Y_j^2] - \mathbb{E}[Y_j]^2 \le \frac{1}{(z+1)^2}$$

$$\mathbf{Var}[\overline{Y}] = \frac{1}{k^2}\sum_{j=1}^{k}\mathbf{Var}[Y_j] = \frac{1}{k}\mathbf{Var}[Y_j] \le \frac{1}{k(z+1)^2}$$

2-wise independence

$$z = |\{x_1, x_2, \ldots, x_n\}|$$

For j=1, 2, …, k, hash values of h_j :

uniform independent $X_{j1},\ X_{j2}, \ldots, X_{jz} \in [0, 1]$

$$\left.\begin{aligned} Y_j &= \min_{1\le i\le n} X_{ji} \\ \overline{Y} &= \frac{1}{k}\sum_{j=1}^{k} Y_j \end{aligned}\right\} \xRightarrow{\text{symmetry}} \boxed{\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\boxed{\text{let } \widehat{Z} = \frac{1}{\overline{Y}} - 1}$

$$\Pr\left[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z\right] \le \frac{4}{\epsilon^2 k}$$

(for $\varepsilon \le 1/2$) $\le \Pr\left[\left|\overline{Y} - \mathbb{E}[\overline{Y}]\right| > \frac{\epsilon/2}{z+1}\right]$

Chebyshev: $\le \frac{4}{\epsilon^2}(z+1)^2 \mathbf{Var}[\overline{Y}]$

$$\boxed{\mathbf{Var}[\overline{Y}] \le \frac{1}{k(z+1)^2}}$$

Input: a sequence $x_1, x_2, \ldots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \ldots, x_n\}|$

uniform independent hash functions:

$h_1, h_2, \ldots, h_k : \Omega \to [0,1]$ $\qquad Y_j = \min_{1\le i\le n} h_j(x_i)$

average-min: $\overline{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$

Flajolet-Martin estimator: $\widehat{Z} = \frac{1}{\overline{Y}} - 1$

UHA: Uniform Hash Assumption

$$\Pr\left[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z\right] \le \frac{4}{\epsilon^2 k} \le \delta$$

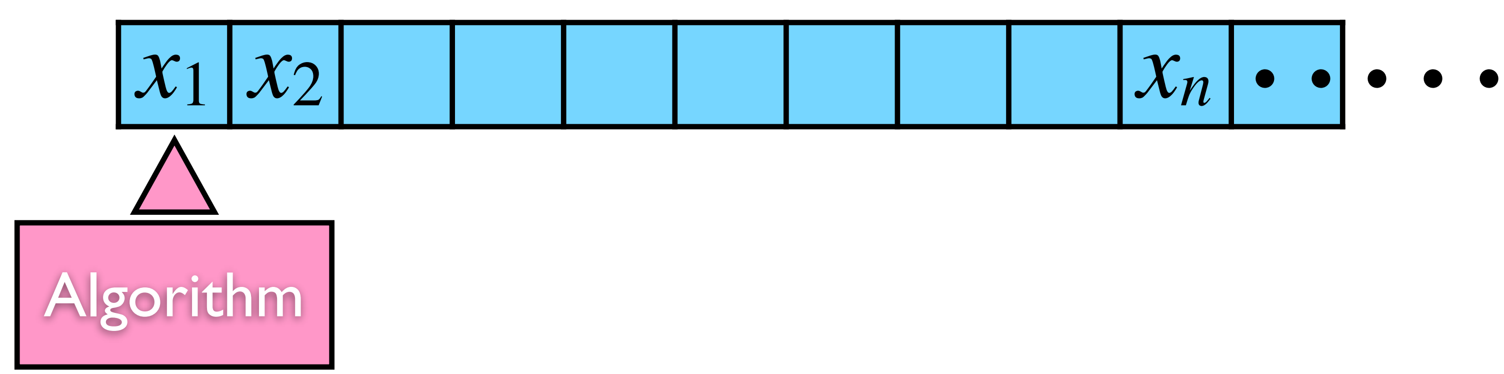
choose $k = \frac{4}{\epsilon^2\delta}$

Frequency Estimation

Data: a sequence $x_1, x_2, \ldots, x_n \in \Omega$
Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x within additive error ϵn.

- data stream: input comes one at a time

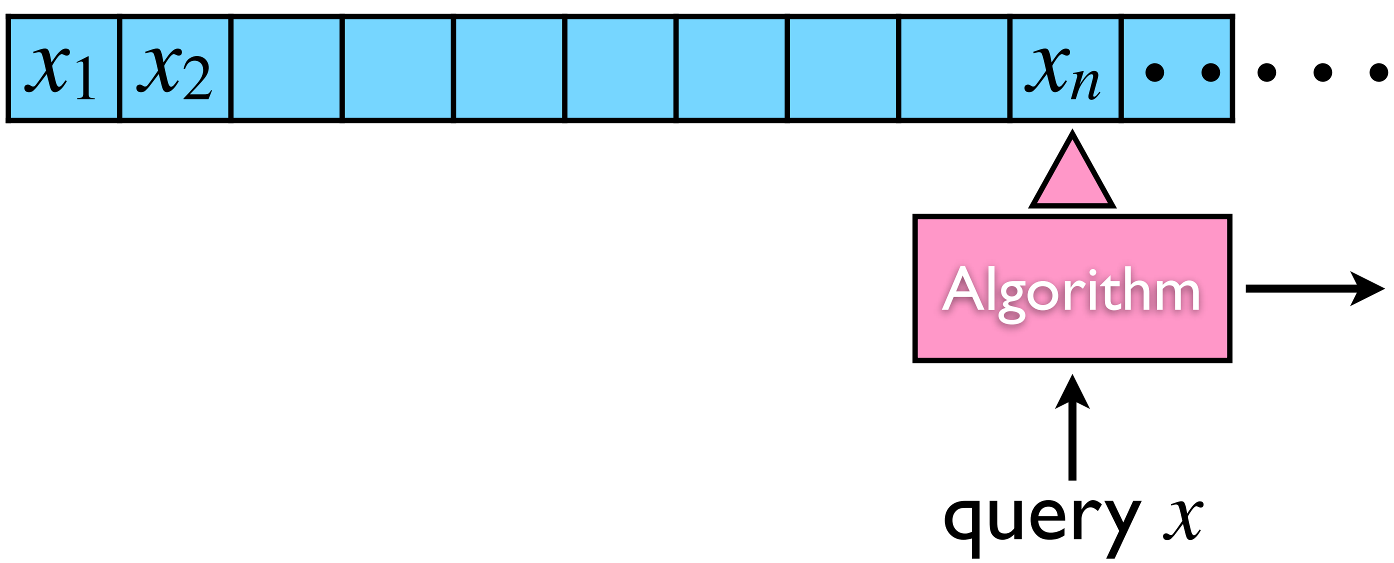


Frequency Estimation

Data: a sequence $x_1, x_2, \ldots, x_n \in \Omega$
Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x within additive error εn.

- data stream: input comes one at a time



$\hat{f}_x$: estimation of frequency f_x

$$\Pr[|\hat{f}_x - f_x| \geq \epsilon n] \leq \delta$$

- heavy hitters: items that appears $> \varepsilon n$ times

Data Structure for Set

Data: a set S of n items $x_1, x_2, \ldots, x_n \in \Omega$
Query: an item $x \in \Omega$

Determine whether $x \in S$.

- space cost: size of data structure (in bits)
  - entropy of a set: $O(n \log|\Omega|)$ bits
- time cost: time to answer a query
- balanced tree: $O(n \log|\Omega|)$ space, $O(\log n)$ time
- perfect hashing: $O(n \log|\Omega|)$ space, $O(1)$ time
- using < entropy space ? a sketch of the set

(approximate representation)

Approximate a Set

Data: a set S of n items $x_1, x_2, \ldots, x_n \in \Omega$

Query: an item $x \in \Omega$

Determine whether $x \in S$.

uniform hash function $h: \Omega \to [m]$

data structure: an m-bit vector $v \in \{0, 1\}^m$

initially v is all-0;

set $v[h(x_i)]=1$ for each $x_i \in S$;

query x: answer "yes" if $v[h(x)]=1$;

$x \in S$: always correct

$x \notin S$: false positive $\Pr[\ v[h(x)]=1\ ] = 1-(1-1/m)^n = 1-e^{-n/m}$

Bloom Filters

(Bloom 1970)

Data: a set S of n items $x_1, x_2, \ldots, x_n \in \Omega$

Query: an item $x \in \Omega$

Determine whether $x \in S$.

uniform independent hash functions

$$h_1, h_2, \ldots, h_k \colon \Omega \to [m]$$

data structure: an m-bit vector $v \in \{0, 1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)]=1$ for all $j=1,\ldots,k$;

query x: "yes" if $v[h_j(x)]=1$ for all $j=1,\ldots,k$;

Bloom Filters

uniform independent hash functions

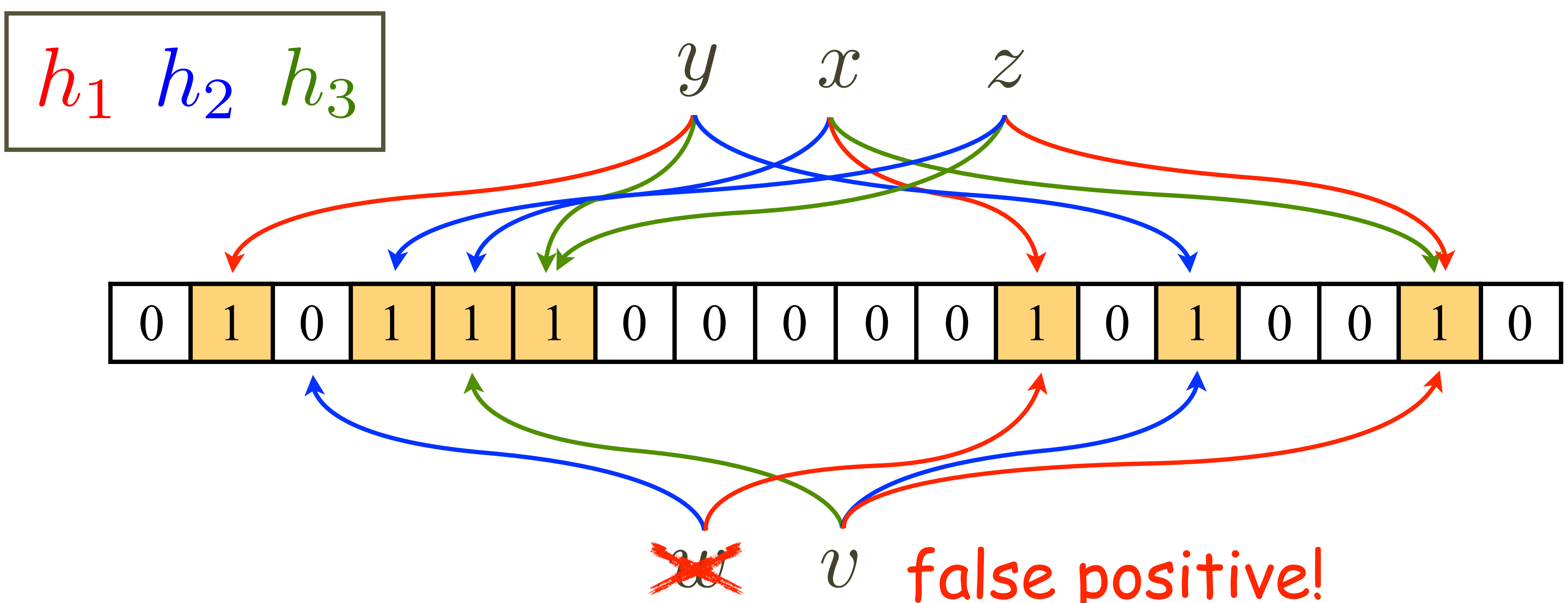
$$h_1, h_2, \ldots, h_k : \Omega \to [m]$$

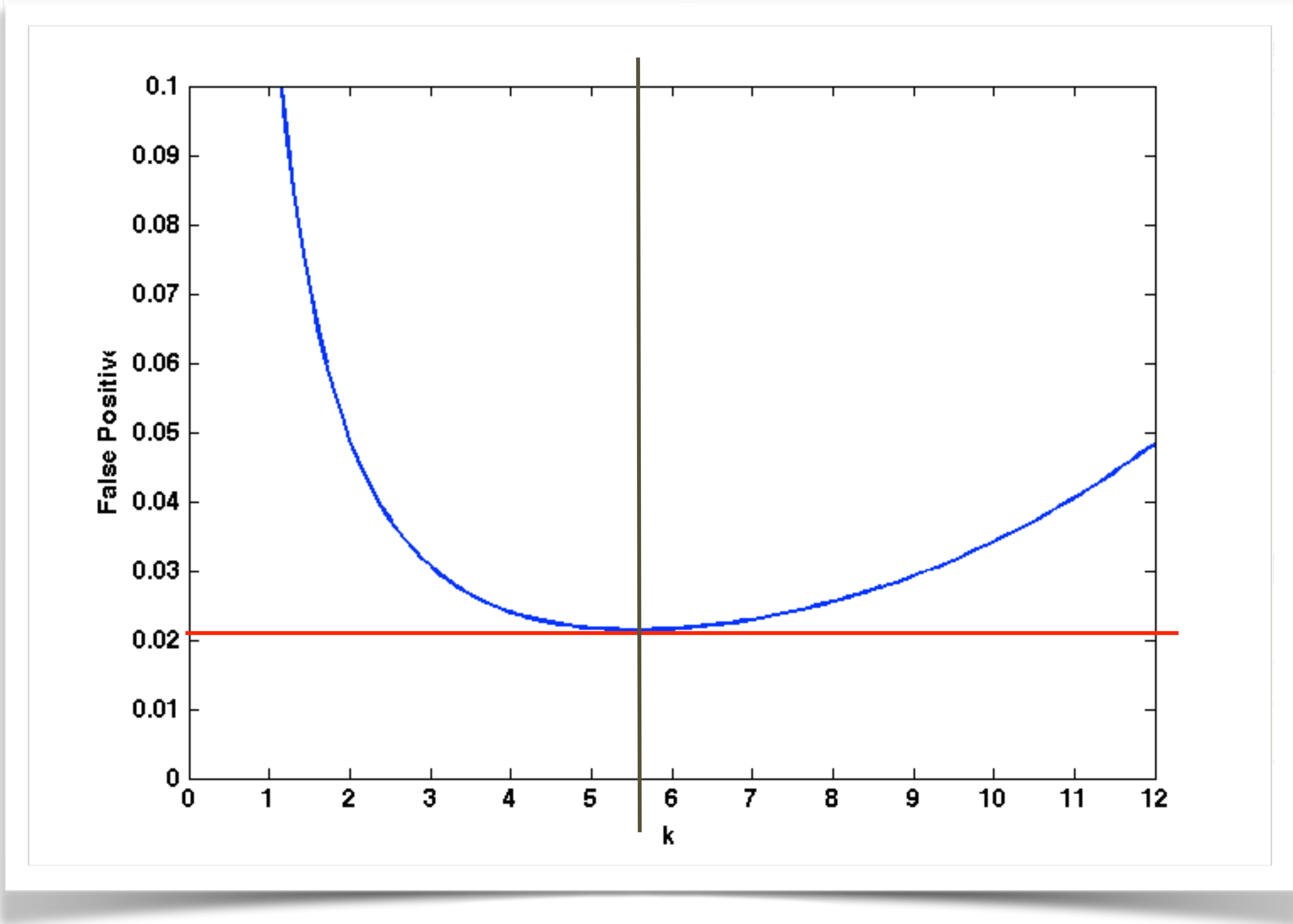
data structure: an m-bit vector $v \in \{0,1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)]=1$ for all $j=1,\ldots,k$;

query x: "yes" if $v[h_j(x)]=1$ for all $j=1,\ldots,k$;





ry: $x \in \Omega$

$0, 1\}^m$

| j=1,...,k;

,...,k;

$x \notin S$: false positive

choose $k = \dfrac{m \ln 2}{n}$

$m = cn$

$$\Pr[\forall 1 \leq j \leq k: \ v[h_j(x)] = 1]$$

$$= (\Pr[v[h_j(x)] = 1])^k \ = (1 - \Pr[v[h_j(x)] = 0])^k$$

$$\leq (1 - (1 - 1/m)^{kn})^k = (1 - \mathrm{e}^{-kn/m})^k \approx (0.6185)^c$$

Bloom Filters

data: set $S \subseteq \Omega$ of size $|S|=n$ **query**: $x \in \Omega$

uniform independent hash functions

$$h_1, h_2, \ldots, h_k: \Omega \to [m]$$

data structure: an m-bit vector $v \in \{0, 1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)]=1$ for all $j=1,\ldots,k$;

query x: "yes" if $v[h_j(x)]=1$ for all $j=1,\ldots,k$;

choose $m = cn$ $\quad k = \dfrac{m \ln 2}{n} = c \ln 2$

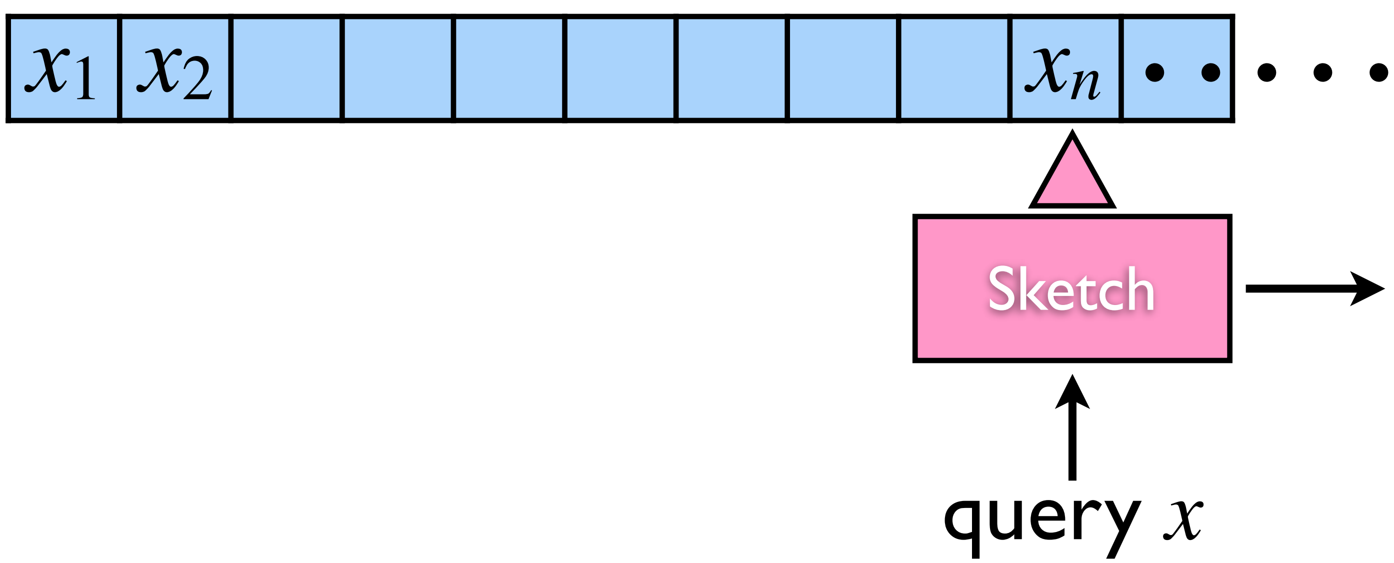
- space cost: cn bits; time cost: $c \ln 2$
- false positive: $< (0.6185)^c$

Heavy Hitters

Data: a sequence $x_1, x_2, \ldots, x_n \in \Omega$
Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x within additive error εn.

- data stream: input comes one at a time



$\hat{f}_x$: estimation of frequency f_x

$$\Pr[|\hat{f}_x - f_x| \geq \epsilon n] \leq \delta$$

- heavy hitters: items that appears $> \varepsilon n$ times

Count-Min Sketch

Data: a sequence $x_1, x_2, \ldots, x_n \in \Omega$

Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x within additive error εn.

uniform independent hash functions

$$h_1, h_2, \ldots, h_k : \Omega \to [m]$$

count-min sketch: $\mathrm{CMS}[k][m]$

initially $\mathrm{CMS}[][]$ is all-0;

for each x_i and each h_j: $\mathrm{CMS}[j][h_j(x_i)]$ ++;

query x: return $\hat{f}_x = \min_{1 \le j \le k} \mathrm{CMS}[j][h_j(x)]$

obviously $\mathrm{CMS}[j][h_j(x)] \ge f_x$ for all $j = 1, 2, \ldots, k$

data: $x_1, x_2, \ldots, x_n \in \Omega$ **query**: $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$$h_1, h_2, \ldots, h_k : \Omega \to [m]$$

count-min sketch: $\mathrm{CMS}[k][m]$

initially $\mathrm{CMS}[\,][\,]$ is all-0;

for each x_i and each h_j: $\mathrm{CMS}[j][h_j(x_i)]$ ++;

query x: return $\hat{f}_x = \min_{1\le j\le k} \mathrm{CMS}[j][h_j(x)]$

for any $x \in \Omega$, for any j :

$$\mathrm{CMS}[j][h_j(x)] = f_x + \sum_{\substack{y\in\{x_1,\ldots,x_n\}\setminus\{x\}\\ h_j(y)=h_j(x)}} f_y$$

$$\mathbb{E}\left[\,\mathrm{CMS}[j][h_j(x)]\,\right] = f_x + \sum_{y\in\{x_1,\ldots,x_n\}\setminus\{x\}} f_y \Pr[h_j(y) = h_j(x)]$$

data: $x_1, x_2, \ldots, x_n \in \Omega$ **query**: $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$$h_1, h_2, \ldots, h_k : \Omega \to [m]$$

count-min sketch: $\mathrm{CMS}[k][m]$

initially $\mathrm{CMS}[][]$ is all-0;

for each x_i and each h_j: $\mathrm{CMS}[j][h_j(x_i)]$ ++;

query x: return $\hat{f}_x = \min_{1 \le j \le k} \mathrm{CMS}[j][h_j(x)]$

for any $x \in \Omega$, for any j :

$$\mathbb{E}\left[\,\mathrm{CMS}[j][h_j(x)]\,\right] = f_x + \sum_{y \in \{x_1, \ldots, x_n\} \setminus \{x\}} f_y \Pr[h_j(y) = h_j(x)]$$

$$= f_x + \frac{1}{m} \sum_{y \in \{x_1, \ldots, x_n\} \setminus \{x\}} f_y \le f_x + \frac{1}{m} \sum_{y \in \{x_1, \ldots, x_n\}} f_y = f_x + \frac{n}{m}$$

biased estimator

data: $x_1, x_2, \ldots, x_n \in \Omega$ **query**: $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$$h_1, h_2, \ldots, h_k : \Omega \to [m]$$

count-min sketch: $\mathrm{CMS}[k][m]$

initially $\mathrm{CMS}[\,][\,]$ is all-0;

for each x_i and each h_j: $\mathrm{CMS}[j][h_j(x_i)]$ ++;

query x: return $\hat{f}_x = \min_{1\le j\le k} \mathrm{CMS}[j][h_j(x)]$

$$\forall x, \forall j: \quad \mathrm{CMS}[j][h_j(x)] \ge f_x$$

$$\mathbb{E}\left[\,\mathrm{CMS}[j][h_j(x)]\,\right] \le f_x + \frac{n}{m}$$

Markov's inequality: $\Pr[\ \mathrm{CMS}[j][h_j(x)] - f_x \ge \varepsilon n\ ] \le 1/(\varepsilon m)$

$\Pr\left[\left|\hat{f}_x - f_x\right| \ge \epsilon n\right] = \Pr[\ \forall j: \mathrm{CMS}[j][h_j(x)] - f_x \ge \varepsilon n\ ] \le 1/(\varepsilon m)^k$

data: $x_1, x_2, \ldots, x_n \in \Omega$ **query**: $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$$h_1, h_2, \ldots, h_k: \Omega \to [m]$$

count-min sketch: $\mathrm{CMS}[k][m]$

initially $\mathrm{CMS}[\,][\,]$ is all-0;

for each x_i and each h_j: $\mathrm{CMS}[j][h_j(x_i)]$ ++;

query x: return $\hat{f}_x = \min_{1 \le j \le k} \mathrm{CMS}[j][h_j(x)]$

$$\Pr\left[\left|\hat{f}_x - f_x\right| \ge \epsilon n\right] \le 1/(\varepsilon m)^k \le \delta$$

choose $m = \left\lceil \frac{\mathrm{e}}{\epsilon} \right\rceil$ $k = \left\lceil \ln \frac{1}{\delta} \right\rceil$

- space cost: $km = O\left(\frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$
- time cost for each query: $k = O\left(\ln \frac{1}{\delta}\right)$