Estimating Number of Distinct Elements

Fahad Panolan

Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad, India

Streaming Model

- The input consists of m objects/items/tokens $e_1, e_2, \ldots, e_m$ that are seen one by one by the algorithm.
- The algorithm has "limited" memory say for B tokens where $B < m$ (often $B << m$) and hence cannot store all the input
- Want to compute interesting functions over input

Distinct Elements

How many distinct items in the stream of integers? Here we know that each token is a postive integer from $[n] = \{1, 2, \dots, n\}$.

- Input stream: $e_1, \dots, e_m$.
- We associate a frequence vector $f = (f_1, \dots, f_n)$.
- f_i is the frequency of the element i in the input stream.
- We want to estimate $|\{f_i > 0 : i \in [n]\}|$

$n = 7$

2 4 2 3 1 6

$$\begin{array}{ccccccc} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 \\ (0 & 2 & 1 & 1 & 0 & 1 & 0) \end{array} = f.$$

$e_1 \quad e_2 \quad \ldots \quad e_m$

$0^0 \to 0$

$x^0 = 1$

$f = (f_1 \; \ldots \; f_n)$

$$\|f\|_1 = \sum_{i=1}^{n} |f_i| = m \leftarrow \text{length of the stream}$$

$O(\log \log m)$

ε, δ

$$\|f\|_0 = \sum_{i=1}^{n} f_i^0 \leftarrow \#\ \text{distinct elements}$$

$$\|f\|_2 = \left(\sum_{i=1}^{n} f_i^2\right)^{1/2}$$

$$\|f\|_\infty = \max_i |f_i|$$

Distinct Elements

How many distinct items in the stream of integers? Here we know that each token is a postive integer from $[n] = \{1, 2, \dots, n\}$.

- Input stream: $e_1, \dots, e_m$.
- We associate a frequence vector $f = (f_1, \dots, f_n)$.
- f_i is the frequency of the element i in the input stream.
- We want to estimate $|\{f_i > 0 \ : \ i \in [n]\}|$

Obvious: counter for each $i \in \{1, \dots, n\}$.
Space complexity $O(n \log m)$. — bits.

$O(\log n + \log m)$

Distinct Elements: Our objective

$d = \|f\|_0$

- We will discuss a $(O(1), \delta)$-estimate.

$$\Pr\left[\frac{3}{4} d \le \hat{d} \le 4d\right] \ge 1 - \delta$$

Distinct Elements: Our objective

- We will discuss a $(O(1), \delta)$-estimate.
- Let d be the no. of distinct elements.
- Then the algorithm will output $\widehat{d}$ with the following guarantee.

$$\Pr[\frac{d}{3} \leq \widehat{d} \leq 3d] \geq 1 - \delta.$$

[Flajolet and Martin' 85], [Alon, Matias and Szegedy' 99]

Distinct Elements: Our objective

- We will discuss a $(O(1), \delta)$-estimate.
- Let d be the no. of distinct elements.
- Then the algorithm will output $\widehat{d}$ with the following guarantee.

$$\Pr[\frac{d}{3} \leq \widehat{d} \leq 3d] \geq 1 - \delta.$$

[Flajolet and Martin' 85], [Alon, Matias and Szegedy' 99]

- In this algorithm, we use
  - Pairwise Independent Hash family ✓
  - Median Trick and Chernoff bound ✓

Probabilistic Inequalities
and
Pairwise Independent Hash Family

Markov's Inequality

Let X be a **non-negative** random variable over a probability space $(\Omega, \Pr)$. For any $a > 0$,

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

In other words, for any $t > 0$,

$$\Pr[X \geq t\mathbf{E}[X]] \leq \frac{1}{t}.$$

Chebyshev's Inequality: Variance

Variance

Given a random variable X over probability space $(\Omega, \Pr)$, variance of X is the measure of how much does it deviate from its mean value. Formally,

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

Chebyshev's Inequality: Variance

Variance

Given a random variable X over probability space $(\Omega, \Pr)$, variance of X is the measure of how much does it deviate from its mean value. Formally,

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

Independence

Random variables X and Y are called mutually independent if

$$\forall x, y \in \mathbb{R},\ \Pr[X = x \wedge Y = y] = \Pr[X = x] \Pr[Y = y]$$

Chebyshev's Inequality: Variance

Variance

Given a random variable X over probability space $(\Omega, \Pr)$, variance of X is the measure of how much does it deviate from its mean value. Formally,

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

Independence

Random variables X and Y are called mutually independent if

$$\forall x, y \in \mathbb{R},\ \Pr[X = x \wedge Y = y] = \Pr[X = x] \Pr[Y = y]$$

Lemma

If X and Y are independent random variables then $Var(X + Y) = Var(X) + Var(Y)$.

Chebyshev's Inequality

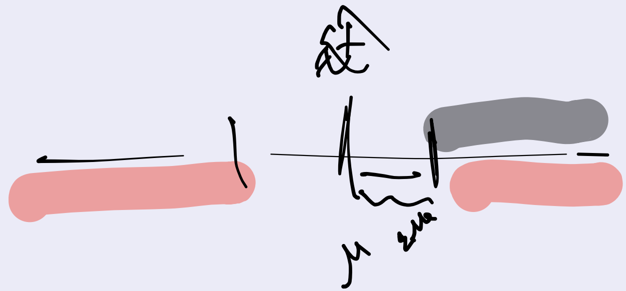
If $Var(X) < \infty$, then for any $a \geq 0$,

$$\Pr[|X - \mathbf{E}[X]| \geq a] \leq \frac{Var(X)}{a^2}.$$

Recap: Chernoff bound

Let $X_1, \ldots, X_k$ be k independent random variables such that, for each $i \in \{1, \ldots, k\}$, X_i equals 1 with probability p_i, and 0 with probability $(1 - p_i)$. Let $X = \sum_{i=1}^{k} X_i$ and $\mu = \mathbf{E}[X] = \sum_i p_i$. For any $0 < \varepsilon < 1$, it holds that:

- $\Pr[X \geq (1+\varepsilon)\mu] \leq e^{\frac{-\varepsilon^2 \mu}{3}}$
- $\Pr[X \leq (1-\varepsilon)\mu] \leq e^{\frac{-\varepsilon^2 \mu}{2}}$



For $0 < \varepsilon < 1$ and $\mu_{min} < \mu < \mu_{max}$,

- $\Pr[X \geq (1+\varepsilon)\mu_{max}] \leq e^{\frac{-\varepsilon^2 \mu_{max}}{3}}$
- $\Pr[X \leq (1-\varepsilon)\mu_{min}] \leq e^{\frac{-\varepsilon^2 \mu_{min}}{2}}$

Pairwise Independent Hash Family

- Hash functions are used in various fields in the CS.

Pairwise Independent Hash Family

- Hash functions are used in various fields in the CS.
- Want the hash function to behave like a “random function” and has a compact representation.

Let $Z_1, Z_2 \ldots Z_n$ are n random variables over (Ω, Pr). We say that $Z_1, Z_2, \ldots, Z_n$ are pairwise independent if for any two distinct i and j and any two values a and b.

$$Pr[Z_i = a \wedge Z_j = b] = Pr[Z_i = a] \cdot Pr[Z_j = b]$$

Pairwise Independent Hash Family

- Hash functions are used in various fields in the CS.
- Want the hash function to behave like a "random function" and has a compact representation.
- A family of hash functions $\mathcal{H} \subseteq \{f : X \to Y\}$, is a Pairwise Independent Hash Family if the following two conditions hold.

$\Omega = \mathcal{H}$.

$\forall$ any $h \in \mathcal{H}$, $\Pr[h] = \frac{1}{|\mathcal{H}|}$.

$(\Omega, \Pr)$

Pairwise Independent Hash Family

- Hash functions are used in various fields in the CS.
- Want the hash function to behave like a "random function" and has a compact representation.

p 1

- A family of hash functions $\mathcal{H} \subseteq \{f : X \to Y\}$, is a Pairwise Independent Hash Family if the following two conditions hold.
  - Uniformly distributed: for any $x \in X$ and $y \in Y$,

$$\Pr_{h \sim \mathcal{H}}[h(x) = y] = \frac{1}{|Y|}.$$

$\mathcal{H} = \{g_1, g_2 \dots g_t\}$

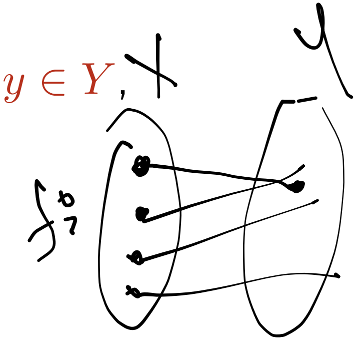
$g_i : X \to Y$

$|Y|$

Pairwise Independent Hash Family

- Hash functions are used in various fields in the CS.
- Want the hash function to behave like a "random function" and has a compact representation.
- A family of hash functions $\mathcal{H} \subseteq \{f : X \to Y\}$, is a Pairwise Independent Hash Family if the following two conditions hold.
    - Uniformly distributed: for a any $x \in X$ and $y \in Y$,

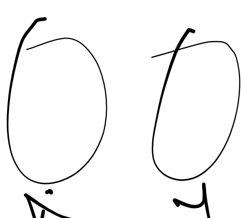
$$\Pr_{h\sim\mathcal{H}}[h(x) = y] = \frac{1}{|Y|}.$$



    - For any $x, x' \in X$ and $y, y' \in Y$ s.t $x \neq x'$,

$$\Pr_{h\sim\mathcal{H}}[h(x) = y \wedge h(x') = y'] = \frac{1}{|Y|^2}.$$

$= \Pr_{h\sim\mathcal{H}}[h(x)=y] \cdot \Pr[h(x')=y']$



Example: Pairwise Independent Hash Family

Let $X = \{0,1\}^N$ and $Y = \{0,1\}^K$ where $K \leq N$.

- For a matrix $A \in \{0,1\}^{K \times N}$ and vector $b \in \{0,1\}^K$, define $h_{A,b} \colon X \to Y$ as follows:

$$h_{A,b}(x) = (Ax + b) \mod 2.$$

$N = 3$

000
001
010
011
100
101
110
111

$K = 2$

00
01
10
11

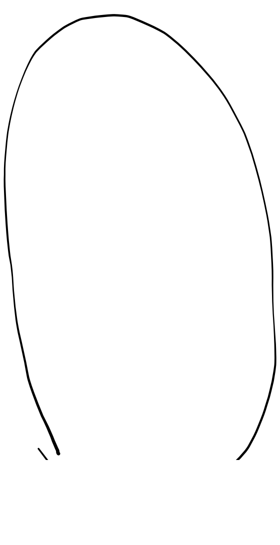
$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

$$(Ax + b) \bmod 2$$

$h_{A,b}$:

$(0\ 0\ 0)^T$	$(0\ 0)^T$
$(0\ 0\ 1)^T$	$(0\ 1)^T$
$(0\ 1\ 0)^T$	$1\ 0$
$(0\ 1\ 1)$	$1\ 1$
$1\ 0\ 0$	
$1\ 0\ 1$	
$1\ 1\ 0$	
$1\ 1\ 1$	

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \bmod 2$$



Example: Pairwise Independent Hash Family

$l = \log n$ $\qquad$ $l = \log n.$

Let $X = \{0,1\}^N$ and $Y = \{0,1\}^K$ where $K \leq N$.

- For a matrix $A \in \{0,1\}^{K \times N}$ and vector $b \in \{0,1\}^K$, define $h_{A,b}\colon X \to Y$ as follows:

$$h_{A,b}(x) = (Ax + b) \mod 2.$$

A, h

$\log n \times \log n$ $\qquad$ $\log n$ $\qquad$ $\log n \times \log n$

- $\mathcal{H} = \{h_{A,b}\colon A \in \{0,1\}^{\cancel{K \times N}}, b \in \{0,1\}^{\cancel{K}}\}$ is a pairwise independent hash family.

$$|\mathcal{H}| = 2^{K \times N} \cdot 2^{K}$$

$\mathcal{H}$ $(\Omega, \Pr)$

$\Omega = \mathcal{H}$, $\Pr[h_{A,b}] = \frac{1}{|\mathcal{H}|}$

Tidemark Algorithm

Notation

p - 10101110010000

zeros 4.

For an integer $p > 0$, $zeros(p)$ is the number of zeros that the binary representation of p ends with. That is,

$$zeros(p) = \max\{i \ : \ 2^i \text{ divides } p\}.$$

$p = 5$, 101. $zeros(5) = 0$.

$p = 6$ 110. $zeros(6) = 1$

$p = 8$ 1000 $zeros(8) = 3$

$\{0,1\}^\ell \to \{0,1\}^\ell$. $n = 2^\ell$.

Algorithm 1: Tidemark Algorithm

$\mathcal{H}$ is a pairwise independent hash family from $[n]$ to $[n]$;

choose h at random from $\mathcal{H}$; $(a, b) = \log n + \log n$.

$z \leftarrow 0$; $\{1, 2 \ldots n\}$

while a new token e_j arrives **do**

    if $zeros(h(e_j)) > 0$ **then** 2 $\{1, 2 \ldots n\}$.

        $z \leftarrow zeros(h(e_j))$

    end

end $ax + b =$

return $2^{z+\frac{1}{2}}$

$h(e_j)$ $e_1 \ldots e_m$

0, $\{1, 2 \ldots n\}$.

Intuition

$e_1 \quad e_2 \quad e_3 \quad e_4 \quad e_5 - e_6$

h: 1 2 1 2 3 4

3 3 3 3 2 4

0 0 0 0 1 2

zeros 0 0 0 1 2

0

$z=0, \quad z=0$

e_7: 1

$h(1)$ 3

0

$2 \quad 2L \quad 2$

$2+1/2$

$2^{2+1/2} = 4\sqrt{2}$

Space complexity

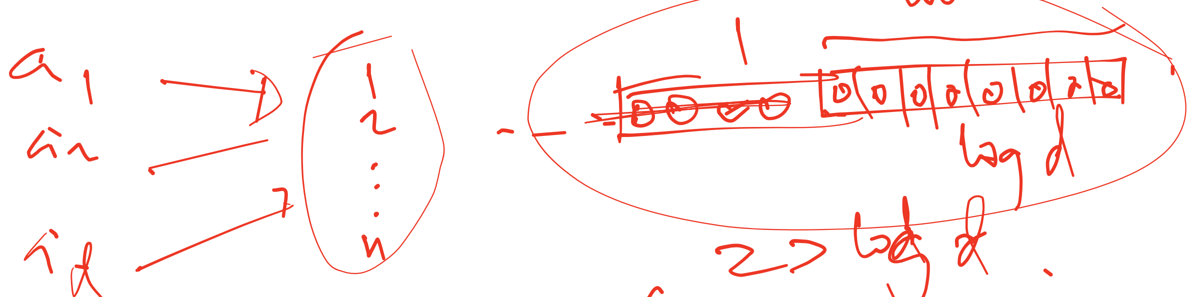
$h_{A,b}$ $\quad A \in \{0,1\}^{d \times d}$ $\quad d = \log n$

Algorithm 2: Tidemark Algorithm

$\mathcal{H}$ is a pairwise independent hash family from $[n]$ to $[n]$;
choose h at random from $\mathcal{H}$;
$z \leftarrow 0$;
while a new token e_i arrives **do**
  if $zeros(h(e_j)) > 0$ **then**
    $z \leftarrow zeros(h(e_j))$
  end
end
return $2^{z+\frac{1}{2}}$

$\log n \times \log n$ $\quad O(\log^2 n)$ $\quad 2^{\log d}$ $\quad d \cdot 2$

\# distinct elements in d.



$z > \log d$.

Space complexity $= O(\log^2 n)$.

Analysis: $(O(1), \frac{1}{\sqrt{2}})$-Estimate

- For each integer $t \in [n]$ and each integer $r \geq 0$, $X_{r,t}$ be the indicator random variable s.t.

$$X_{r,t} = \begin{cases} 1 & \text{if } zeros(h(t)) \geq r \\ 0 & \text{Otherwise} \end{cases}$$

Analysis: $(O(1), \frac{1}{\sqrt{2}})$-Estimate

- For each integer $t \in [n]$ and each integer $r \geq 0$, $X_{r,t}$ be the indicator random variable s.t.

$$X_{r,t} = \begin{cases} 1 & \text{if } zeros(h(t)) \geq r \\ 0 & \text{Otherwise} \end{cases}$$

1, 5, 8, 9

- $Y_r = \sum_{t: f_t > 0} X_{r,t}$.

$$Y_r = X_{r,1} + X_{r,5} + X_{r,8} + X_{r,9}$$

Analysis: $(O(1), \frac{1}{\sqrt{2}})$-Estimate

- For each integer $t \in [n]$ and each integer $r \geq 0$, $X_{r,t}$ be the indicator random variable s.t.

$$X_{r,t} = \begin{cases} 1 & \text{if } zeros(h(t)) \geq r \\ 0 & \text{Otherwise} \end{cases}$$

- $Y_r = \sum_{t: f_t > 0} X_{r,t}$.

$2^{T+1/2}$

- Let T be the value of z at the end of the algorithm.

Analysis: $(O(1), \frac{1}{\sqrt{2}})$-Estimate

- For each integer $t \in [n]$ and each integer $r \geq 0$, $X_{r,t}$ be the indicator random variable s.t.

$$X_{r,t} = \begin{cases} 1 & \text{if } zeros(h(t)) \geq r \\ 0 & \text{Otherwise} \end{cases}$$

$r = 3$

$X_{1,8} = 1$

$X_{3,1} = 0$

$X_{3,4} = 1$

$X_{1,1} = 1$

- $Y_r = \sum_{t: f_t > 0} X_{r,t}$.
- Let T be the value of z at the end of the algorithm.
- Then, $Y_r \geq 0$ iff $T \geq r$.
- Equivalently, $Y_r = 0$ iff $T \leq r - 1$.

stream: 1, 8, 2, 1, 2

h(): 2 1 3 2 3

zero: 1 0 0 1 0

4

$h(4) = 9$

Fix an r:

Suppose $Y_r = 0$

Then all the numbers you see in the stream is mapped to numbers that has less than or equal to $r-1$ # of zeros

$\Rightarrow$ The value in Z will be at most $r-1$

3

Expectation and Variance of Y_r

- $\mathbf{E}[X_{r,t}] = \Pr[zeros(h(t)) \geq r] = \Pr[2^r \text{ divides } h(t)] = \frac{1}{2^r}$

$0\;0\;0\;0\;0\;0$

$= \frac{1}{2^r}$

Expectation and Variance of Y_r

- $\mathbf{E}[X_{r,t}] = \Pr[zeros(h(t)) \geq r] = \Pr[2^r \text{ divides } h(t)] = \frac{1}{2^r}$
- $\mathbf{E}[Y_r] = \sum_{t:f_t>0} \mathbf{E}[X_{r,t}] = \frac{d}{2^r}$ (Here d is the no. of distinct elements)

Expectation and Variance of Y_r

- $\mathbf{E}[X_{r,t}] = \Pr[zeros(h(t)) \geq r] = \Pr[2^r \text{ divides } h(t)] = \frac{1}{2^r}$
- $\mathbf{E}[Y_r] = \sum_{t:f_t>0} \mathbf{E}[X_{r,t}] = \frac{d}{2^r}$ (Here d is the no. of distinct elements)
- $Var[Y_r] = \sum_{t:f_t>0} Var[X_{r,t}] \leq \sum_{t:f_t>0} \mathbf{E}[X_{r,t}^2] = \frac{d}{2^r}$

$\forall t$, $X_{r,t}$ are pairwise independent.

Applying Markov's and Chebyshevs inequalities

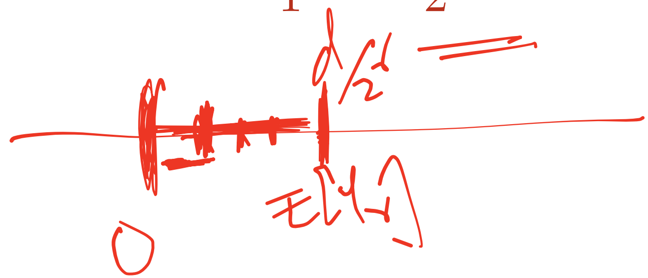
- By Markov's inequality,

$$\Pr[Y_r > 0] = \Pr[Y_r \geq 1] \leq \frac{\mathbf{E}[Y_r]}{1} = \frac{d}{2^r}.$$

Applying Markov's and Chebyshevs inequalities

- By Markov's inequality,

$$\Pr[Y_r > 0] = \Pr[Y_r \geq 1] \leq \frac{\mathbf{E}[Y_r]}{1} = \frac{d}{2^r}.$$



- By Chebyshevs inequality,

$$\begin{aligned}
\Pr[Y_r = 0] &\leq \Pr[|Y_r - \mathbf{E}[Y_r]| \geq \frac{d}{2^r}] \\
&\leq \frac{Var[Y_r]}{(d/2^r)^2} \\
&= \frac{2^r}{d}
\end{aligned}$$

Quality of the estimate $\widehat{d}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.

Quality of the estimate $\widehat{d}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.
- Let a be the smallest integer such that $2^{a+\frac{1}{2}} \geq 4d$.

Quality of the estimate $\widehat{d}$

$\cancel{\frac{1}{2} + \frac{1}{2} > \frac{3}{4}}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.
- Let a be the smallest integer such that $2^{a+\frac{1}{2}} \geq 4d$.
- Then, $\Pr[\widehat{d} \geq \cancel{3}4d] \leq \Pr[T \geq a] = \Pr[Y_a > 0] \leq \frac{d}{2^a} \leq \frac{\sqrt{2}}{4}$.

Quality of the estimate $\widehat{d}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.
- Let a be the smallest integer such that $2^{a+\frac{1}{2}} \geq 4d$.
- Then, $\Pr[\widehat{d} \geq \cancel{3}4d] \leq \Pr[T \geq a] = \Pr[Y_a > 0] \leq \frac{d}{2^a} \leq \frac{\sqrt{2}}{4}$.
- Let b be the largest integer such that $2^{b+\frac{1}{2}} \leq d/4$.

Quality of the estimate $\widehat{d}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.
- Let a be the smallest integer such that $2^{a+\frac{1}{2}} \geq 4d$.
- Then, $\Pr[\widehat{d} \geq \cancel{3}4d] \leq \Pr[T \geq a] = \Pr[Y_a > 0] \leq \frac{d}{2^a} \leq \frac{\sqrt{2}}{4}$.
- Let b be the largest integer such that $2^{b+\frac{1}{2}} \leq d/4$.
- Then, $\Pr[\widehat{d} \leq d/4] \leq \Pr[T \leq b] = \Pr[Y_{b+1} = 0] \leq \frac{2^{b+1}}{d} \leq \frac{\sqrt{2}}{4}$

Quality of the estimate $\widehat{d}$

- $\widehat{d} = 2^{T+\frac{1}{2}}$.
- Let a be the smallest integer such that $2^{a+\frac{1}{2}} \geq 4d$.
- Then, $\Pr[\widehat{d} \geq 3d] \leq \Pr[T \geq a] = \Pr[Y_a > 0] \leq \frac{d}{2^a} \leq \frac{\sqrt{2}}{4}$.
- Let b be the largest integer such that $2^{b+\frac{1}{2}} \leq d/4$.
- Then, $\Pr[\widehat{d} \leq d/4] \leq \Pr[T \leq b] = \Pr[Y_{b+1} = 0] \leq \frac{2^{b+1}}{d} \leq \frac{\sqrt{2}}{4}$
- Then, by union bound,

$$\Pr[d/4 \leq \widehat{d} \leq 4d] \geq 1 - \frac{1}{\sqrt{2}}.$$

Error reduction via median trick

We have:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \frac{1}{\sqrt{2}}$$

Want:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \delta$$

for some given parameter δ.

Error reduction via median trick

We have:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \frac{1}{\sqrt{2}}$$

Want:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \delta$$

for some given parameter δ.

Idea: Repeat independently $\ell = 12 \log(2/\delta)$ times.

Error reduction via median trick

We have:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \frac{1}{\sqrt{2}}$$

Want:

$$\Pr[\widehat{d} \geq 4d \text{ or } \widehat{d} \geq d/4] \leq \delta$$

for some given parameter δ.

Idea: Repeat independently $\ell = 12 \log(2/\delta)$ times.

Algorithm: Output median of the estimates $Q^{(1)}, Q^{(2)}, \ldots, Q^{(\ell)}$.

Error reduction via median trick

Let Z be median of the $\ell = 12\log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z > 4d] \leq \delta/2$.

Error reduction via median trick

Let Z be median of the $\ell = 12\log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z > 4d] \leq \delta/2$.

- Let A_i be event that estimate $Q^{(i)}$ is bad: that is, $Q^{(i)} > 4d$. Then, $\Pr[A_i] < \frac{\sqrt{2}}{4}$. Hence expected number of bad estimates is at most $\ell \cdot \frac{\sqrt{2}}{4}$.

Error reduction via median trick

Let Z be median of the $\ell = 12 \log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z > 4d] \leq \delta/2$.

- Let A_i be event that estimate $Q^{(i)}$ is <u>bad</u>: that is, $Q^{(i)} > 4d$. Then, $\Pr[A_i] < \frac{\sqrt{2}}{4}$. Hence expected number of bad estimates is at most $\ell \cdot \frac{\sqrt{2}}{4}$.
- For median estimate to be bad, more than half of A_i's have to be bad.

Error reduction via median trick

Let Z be median of the $\ell = 12\log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z > 4d] \leq \delta/2$.

- Let A_i be event that estimate $Q^{(i)}$ is bad: that is, $Q^{(i)} > 4d$. Then, $\Pr[A_i] < \frac{\sqrt{2}}{4}$. Hence expected number of bad estimates is at most $\ell \cdot \frac{\sqrt{2}}{4}$.
- For median estimate to be bad, more than half of A_i's have to be bad.
- Let X_i be that random variable that takes value 1 when A_i happens and 0 otherwise.

Error reduction via median trick

Let Z be median of the $\ell = 12 \log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z > 4d] \leq \delta/2$.

- Let A_i be event that estimate $Q^{(i)}$ is <u>bad</u>: that is, $Q^{(i)} > 4d$. Then, $\Pr[A_i] < \frac{\sqrt{2}}{4}$. Hence expected number of bad estimates is at most $\ell \cdot \frac{\sqrt{2}}{4}$.
- For median estimate to be bad, more than half of A_i's have to be bad.
- Let X_i be that random variable that takes value 1 when A_i happens and 0 otherwise.
- Let $X = \sum_{i=1}^{\ell} X_i$.
- Our output is "bad" if and only if X is at least $\ell/2$.

Applying Chernoff bound

- Let $X_1, \ldots, X_k$ be k independent $0/1$-random variables,
- $X = \sum_{i=1}^{k} X_i$, and
- $\mathbf{E}[X] \leq \mu_{\max}$.

Then, for any $0 < \varepsilon < 1$, it holds that:

- $\Pr[X \geq (1+\varepsilon)\mu] \leq e^{\frac{-\varepsilon^2 \mu_{max}}{3}}$

Applying Chernoff bound

- Let $X_1, \ldots, X_k$ be k independent $0/1$-random variables,
- $X = \sum_{i=1}^{k} X_i$, and
- $\mathbf{E}[X] \le \mu_{\max}$.

Then, for any $0 < \varepsilon < 1$, it holds that:

- $\Pr[X \ge (1+\varepsilon)\mu] \le e^{\frac{-\varepsilon^2 \mu_{max}}{3}}$

$$
\begin{aligned}
\Pr[X \ge \ell/2] &\le \Pr[X \ge 2\mu_{\max}] \\
&\le \Pr[X \ge (1+0.99)\mu_{\max}] \\
&\le e^{\frac{-(0.99)^2 \mu_{\max}}{3}} \\
&\le e^{\frac{-(0.99)^2 \sqrt{2}\ell}{12}}
\end{aligned}
$$

Choose $\ell = 12 \cdot (\log \frac{1}{\delta})$. Then, $\Pr[X \ge \ell/2] \le \delta$

Error reduction via median trick

Let Z be median of the $\ell = 12\log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z < d/4] \leq \delta/2$.

Error reduction via median trick

Let Z be median of the $\ell = 12 \log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z < d/4] \leq \delta/2$.

We got $(O(1), \delta)$-estimate

Error reduction via median trick

Let Z be median of the $\ell = 12\log(2/\delta)$ independent estimators.

Lemma

$\Pr[Z < d/4] \leq \delta/2$.

We got $(O(1), \delta)$-estimate

Space complexity: $O(\log(1/\delta)\log^2 n)$.

Summary

- We have seen estimating number of distinct elements
- We used pairwise independent hash family
- Median Trick

Thank You.