Sublinear Algorithms

Lecture 02: Streaming II

Karl Bringmann
May 14, 2020
Recap: Concentration Inequalities

Markov: \[ P[X \geq t] \leq \frac{E[X]}{t} \]

For any \( t > 0 \), assuming \( X \geq 0 \)

Chebyshev: \[ P[|X - E[X]| \geq t] \leq \frac{Var[X]}{t^2} \]

For any \( t > 0 \)

Variance \( Var[X] = E[X^2] - E[X]^2 \)

Chernoff: \[ P[|X - E[X]| \geq t] \leq 2 \exp\left(-\frac{2t^2}{n}\right) \]

For any \( t > 0 \), assuming \( X = X_1 + \cdots + X_n \)

with independent \( X_1, \ldots, X_n \in \{0,1\} \)
Recap: Approximate Counting

monitor a sequence of events, maintain an **approximate counter** of the number of events

**Data structure problem:**

- maintain a number \( n \)
- **update()**: increment \( n \) by 1
- **query()**: output \( \tilde{n} \) with

\[
\mathbb{P}[|\tilde{n} - n| \geq \varepsilon n] \leq \delta
\]

**Solution:** Morris’ Counter (1978)

1) Initialize \( X = 0 \)
2) On update(): Increment \( X \) with probability \( 2^{-X} \)
3) On query(): output \( \tilde{n} = 2^X - 1 \)

**Morris+:** average over \( s = \lceil 2/\varepsilon^2 \rceil \) runs of Morris

**Morris++:** median of \( t = \lceil 8 \log(2/\delta) \rceil \) runs of Morris+

**Space usage:**

\[
O\left(\frac{1}{\varepsilon^2 \log \left(\frac{1}{\delta}\right)} \log \log \left(\frac{n}{\varepsilon \delta}\right)\right) \text{ bits}
\]

with probability at least \( 1 - \delta \)
Recap: Approximate Counting

monitor a sequence of events, maintain an **approximate counter** of the number of events

**Lem:** Morris’ Counter is an **unbiased estimator** of \( n \), that is, \( \mathbb{E}[\tilde{n}] = n \).

**Lem:** We have \( \mathbb{E}[\tilde{n}^2] = \frac{3}{2} n^2 - \frac{1}{2} n \).

\[
\mathbb{P}[|\tilde{n} - n| \geq \varepsilon n] \leq \frac{\text{Var}[\tilde{n}]}{\varepsilon^2 n^2} \leq \frac{1}{2\varepsilon^2}
\]

**Boosting via Chebyshev:**
Morris+ improves variance to \( \frac{\text{Var}[\tilde{n}]}{s} \)

**Boosting via Chernoff:**
Morris++ improves error probability from 1/4 to \( \exp(-t/8) \)

**Solution:** Morris’ Counter (1978)

1) Initialize \( X = 0 \)

2) On update(): Increment \( X \) with **probability** \( 2^{-X} \)

3) On query(): output \( \tilde{n} = 2^X - 1 \)

**Morris+:** average over \( s = \lceil 2/\varepsilon^2 \rceil \) runs of Morris

**Morris++:** median of \( t = \lceil 8 \log(2/\delta) \rceil \) runs of Morris+
Outline

1) Distinct Elements: Idealized Setting
2) Distinct Elements: Theoretical Variant
3) Distinct Elements: Practical Variant
Distinct Elements

determine the number of distinct items among $x_1, \ldots, x_m$

Data structure problem:
maintain set $D$ and its size $t$

update($x$): add $x$ to $D$

query(): output $t$

assume $x_1, \ldots, x_m \in \mathbb{Z}^{\leq n} = \{1, \ldots, n\}$

Solution 1: Store all distinct items
uses $O(t \log n)$ bits of space

Solution 2: Bitvector of length $n$
uses $n$ bits of space

Count the number of distinct items
in a huge database table

Count the number of distinct users
accessing a website (=distinct IP addresses)

exact solution requires $\log\left(\binom{n}{t}\right) \approx t \log\left(\frac{n}{t}\right)$ bits
Approximate Distinct Elements

approximate the number of distinct items among $x_1, \ldots, x_m \in [n]$

Data structure problem:
- maintain set $D$ and its size $t$
- $\text{update}(x)$: add $x$ to $D$
- $\text{query}()$: output $\tilde{t}$ with $\mathbb{P}[|\tilde{t} - t| \geq \epsilon t] \leq \delta$

Let $y_1, \ldots, y_t$ be the distinct items in the stream
Suppose that $y_1, \ldots, y_t$ are random in $[0,1]$
Then we expect $y_i \approx i/t$
So $1/ \min_i y_i \approx t$
Approximate Distinct Elements

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

Data structure problem:

- maintain set \( D \) and its size \( t \)
- \( \text{update}(x) \): add \( x \) to \( D \)
- \( \text{query}() \): output \( \tilde{t} \) with
  \[ \mathbb{P}[|\tilde{t} - t| \geq \varepsilon t] \leq \delta \]

Initially \( X = 1 \)

On update(): \( X = \min \{X, h(x)\} \)

Idealized Setting: FM (Flajolet, Martin 1985)

1) Pick random function \( h: [n] \rightarrow [0,1] \)
2) On update(): Maintain \( X = \min_i h(x_i) \)
3) On query(): Output \( \tilde{t} = \frac{1}{X} - 1 \)

Let \( y_1, \ldots, y_t \) be the distinct items in the stream

Suppose that \( y_1, \ldots, y_t \) are random in \([0,1]\)

Then we expect \( y_i \approx \frac{i}{t} \)

So \( \frac{1}{i} y_i \approx t \)
Analysis of Idealized Setting

**Idealized Assumptions:**

- We can handle real numbers
- We can store a random function $h: [n] \rightarrow [0,1]$
  
  $= n$ many random real numbers

**Idealized Setting:**  
**FM**

1) Pick random function $h: [n] \rightarrow [0,1]$  
2) Maintain $X = \min_i h(x_i)$  
3) Output $\tilde{t} = 1/X - 1$

let $y_1, \ldots, y_t$ be the distinct items among $x_1, \ldots, x_m \in [n]$
Analysis of Idealized Setting

Standard Approach:

Show that FM is an **unbiased estimator** of \( t \), that is, \( \mathbb{E}[\tilde{t}] = t \).

**This is false!**

\[
X \leq h(x_1) \quad \frac{1}{X} \geq \frac{1}{h(x_1)}
\]

\[
\mathbb{E}\left[\frac{1}{X}\right] \geq \mathbb{E}\left[\frac{1}{h(x_1)}\right] = \int_0^1 \frac{1}{x} \, dx = \infty \neq t + 1
\]

---

**Idealized Setting:** FM

1) Pick random function \( h: [n] \rightarrow [0,1] \)
2) Maintain \( X = \min_i h(x_i) \)
3) Output \( \tilde{t} = 1/X - 1 \)

let \( y_1, \ldots, y_t \) be the distinct items among \( x_1, \ldots, x_m \in [n] \)
Analysis of Idealized Setting

A change of perspective:  (assume $t \geq 1$)

Lem:  If $\left| X - \frac{1}{t+1} \right| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1}$ then $\left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t$

Sanity check for $\varepsilon = 0$:

$$\left| X - \frac{1}{t+1} \right| = 0 \iff \left| \frac{1}{X} - 1 - t \right| = 0$$

Let $y_1, \ldots, y_t$ be the distinct items among $x_1, \ldots, x_m \in [n]$

Idealized Setting:  FM

1) Pick random function $h: [n] \to [0,1]$
2) Maintain $X = \min_i h(x_i)$
3) Output $\tilde{t} = 1/X - 1$
Analysis of Idealized Setting

A change of perspective: (assume $t \geq 1$)

**Lem:** If $\left| X - \frac{1}{t+1} \right| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1}$ then $\left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t$

**Proof:**

\[
X - \frac{1}{t+1} \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1}
\]

\[\Rightarrow X \leq \left(1 + \frac{\varepsilon}{3}\right) \cdot \frac{1}{t+1}\]

\[\Rightarrow \frac{1}{X} \geq \frac{1}{1 + \frac{\varepsilon}{3}} (t + 1)\]

\[\Rightarrow \frac{1}{X} \geq \left(1 - \frac{\varepsilon}{3}\right) \cdot (t + 1)\]

using $1 \geq 1 - x^2 = (1 - x)(1 + x)$ for any $x \in \mathbb{R}$

\[\Rightarrow \frac{1}{X} - 1 - t \geq -\frac{\varepsilon}{3} \cdot (t + 1) \geq -\frac{\varepsilon}{3} \cdot 2t \geq -\varepsilon t\]

let $y_1, ..., y_t$ be the distinct items among $x_1, ..., x_m \in [n]$

**Idealized Setting:** FM

1) Pick random function $h : [n] \rightarrow [0, 1]$
2) Maintain $X = \min_i h(x_i)$
3) Output $\tilde{t} = 1/X - 1$
Analysis of Idealized Setting

A change of perspective: (assume \( t \geq 1 \))

**Lem:** If \( |X - \frac{1}{t+1}| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \) then \( \left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t \)

**Proof:**

\[
X - \frac{1}{t+1} \geq -\frac{\varepsilon}{3} \cdot \frac{1}{t+1}
\]

\[
\Rightarrow X \geq \left( 1 - \frac{\varepsilon}{3} \right) \cdot \frac{1}{t+1}
\]

\[
\Rightarrow \frac{1}{X} \leq \frac{1}{1-\frac{\varepsilon}{3}} \cdot (t + 1)
\]

\[
\Rightarrow \frac{1}{X} \leq \left( 1 + \frac{\varepsilon}{2} \right) \cdot (t + 1)
\]

using \( \left( 1 - \frac{x}{3} \right) \left( 1 + \frac{x}{2} \right) = 1 + \frac{x}{6} - \frac{x^2}{6} \geq 1 \) for any \( x \in [0,1] \)

\[
\Rightarrow \frac{1}{X} - 1 - t \leq \frac{\varepsilon}{2} \cdot (t + 1) \leq \frac{\varepsilon}{2} \cdot 2t \leq \varepsilon t
\]

**Idealized Setting:** FM

1) Pick random function \( h: [n] \rightarrow [0,1] \)

2) Maintain \( X = \min_i h(x_i) \)

3) Output \( \tilde{t} = 1/X - 1 \)

Let \( y_1, ..., y_t \) be the distinct items among \( x_1, ..., x_m \in [n] \).
Analysis of Idealized Setting

Lem: If $\left| X - \frac{1}{t+1} \right| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1}$ then $\left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t$

Standard Approach under new perspective:

Lem: FM is an unbiased estimator, that is, $\mathbb{E}[X] = \frac{1}{t+1}$.

Proof:

$\mathbb{E}[X] = \int_0^1 \mathbb{P}[X > z] \, dz$

$= \int_0^1 \mathbb{P}[\text{for all } i: h(x_i) > z] \, dz$

$= \int_0^1 \prod_{i=1}^t \mathbb{P}[h(y_i) > z] \, dz = \int_0^1 (1 - z)^t \, dz = \frac{1}{t + 1}$

Idealized Setting: FM

1) Pick random function $h: [n] \rightarrow [0,1]$
2) Maintain $X = \min_i h(x_i)$
3) Output $\tilde{t} = 1/X - 1$

let $y_1, \ldots, y_t$ be the distinct items among $x_1, \ldots, x_m \in [n]$
Analysis of Idealized Setting

Lem: If \( \left| X - \frac{1}{t+1} \right| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \) then \( \left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t \)

Standard Approach under new perspective:

Lem: FM is an unbiased estimator, that is, \( \mathbb{E}[X] = \frac{1}{t+1} \).

Lem: We have \( \mathbb{E}[X^2] = \frac{2}{(t+1)(t+2)} \leq 2\mathbb{E}[X]^2 \).

Proof: \( \mathbb{E}[X^2] = \int_0^1 \mathbb{P}[X^2 > z] \, dz = \int_0^1 \mathbb{P}[X > \sqrt{z}] \, dz \)

\[(u = 1 - \sqrt{z}) = \int_0^1 (1 - \sqrt{z})^t \, dz = 2 \int_0^1 u^t (1 - u) \, dz = \frac{2}{(t + 1)(t + 2)} \]

let \( y_1, ..., y_t \) be the distinct items among \( x_1, ..., x_m \in [n] \)

Idealized Setting: FM

1) Pick random function \( h : [n] \to [0,1] \)
2) Maintain \( X = \min_i h(x_i) \)
3) Output \( \tilde{t} = 1/X - 1 \)
Analysis of Idealized Setting

**Lem:** If \( \left| X - \frac{1}{t+1} \right| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \) then \( \left| \frac{1}{X} - 1 - t \right| \leq \varepsilon t \)

Standard Approach under new perspective:

**Lem:** FM is an unbiased estimator, that is, \( \mathbb{E}[X] = \frac{1}{t+1} \).

**Lem:** We have \( \mathbb{E}[X^2] = \frac{2}{(t+1)(t+2)} \leq 2\mathbb{E}[X]^2 \).

By Chebyshev:

\[
P \left[ \left| X - \frac{1}{t+1} \right| \geq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \right] \leq \frac{2\mathbb{E}[X]^2 - \mathbb{E}[X]^2}{\varepsilon^2 \mathbb{E}[X]^2} = \frac{9}{\varepsilon^2}
\]

**Idealized Setting:** FM

1) Pick random function \( h: [n] \to [0,1] \)
2) Maintain \( X = \min_i h(x_i) \)
3) Output \( \tilde{t} = 1/X - 1 \)

**Chebyshev:** \( \mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}[X]}{t^2} \)

\( \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \)
Analysis of Idealized Setting

Lem: If \( |X - \frac{1}{t+1}| \leq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \) then \( \frac{1}{X} - 1 - t \leq \varepsilon t \)

\[
\mathbb{P} \left[ \left| X - \frac{1}{t+1} \right| \geq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \right] \leq \frac{2\mathbb{E}[X]^2 - \mathbb{E}[X]^2}{\frac{\varepsilon^2}{9} \cdot \mathbb{E}[X]^2} = \frac{9}{\varepsilon^2}
\]

Boosting via Chebyshev:

FM+ = average over \( \frac{36}{\varepsilon^2} \) runs of FM satisfies

\[
\mathbb{P} \left[ \left| X^+ - \frac{1}{t+1} \right| \geq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \right] \leq \frac{1}{4}
\]

Boosting via Chernoff:

FM++ = median over \( 8 \log(2/\delta) \) runs of FM+ satisfies

\[
\mathbb{P} \left[ \left| X^{++} - \frac{1}{t+1} \right| \geq \frac{\varepsilon}{3} \cdot \frac{1}{t+1} \right] \leq \delta
\]

and thus \( \mathbb{P} \left[ \left| \frac{1}{X^{++}} - 1 - t \right| \geq \varepsilon t \right] \leq \delta \)

Idealized Setting: FM

1) Pick random function \( h: [n] \rightarrow [0,1] \)

2) Maintain \( X = \min_i h(x_i) \)

3) Output \( \tilde{t} = 1/X - 1 \)

Chebyshev: \( \mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}[X]}{t^2} \)

\( \text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \)

let \( y_1, \ldots, y_t \) be the distinct items among \( x_1, \ldots, x_m \in [n] \)
Approximate Distinct Elements

approximate the number of distinct items among $x_1, \ldots, x_m \in [n]$

Data structure problem:
- maintain set $D$ and its size $t$
- $\text{update}(x)$: add $x$ to $D$
- $\text{query}()$: output $\tilde{t}$ with
  \[
  \Pr[|\tilde{t} - t| \geq \varepsilon t] \leq \delta
  \]

Idealized Setting: FM (Flajolet, Martin 1985)
1) Pick random function $h: [n] \to [0,1]
2) Maintain $X = \min_i h(x_i)$
3) Let $X^+$ be average over $\frac{36}{\varepsilon^2}$ independent copies of $X$
4) Let $X^{++}$ be median of $8 \log(2/\delta)$ independent copies of $X^+$
5) On query(): Output $\tilde{t} = 1/X^{++} - 1$

Space:
- $O\left(\frac{1}{\varepsilon^2} \log \left(\frac{1}{\delta}\right)\right)$ real numbers
- $O\left(\frac{1}{\varepsilon^2} \log \left(\frac{1}{\delta}\right)\right)$ random functions

Goal: $O(\log n)$ space
Outline

1) Distinct Elements: Idealized Setting
2) Distinct Elements: Theoretical Variant
3) Distinct Elements: Practical Variant
Hash Function

We assumed access to random function \( h: \mathbb{Z}_n \to [0,1] \)

Cannot handle real numbers!  We only have finitely many bits...

Solution: \( h: \mathbb{Z}_n \to [m] \)

Cannot store random function!  There are \( m^n \) functions \( h: \mathbb{Z}_n \to [m] \)

so storing a random function requires \( \log(m^n) = n \log m \) bits

Solution: pairwise independence
Pairwise Independence

Random variables $X_1, \ldots, X_n$ are **independent** if for any $j_1, \ldots, j_n$ we have

$$\Pr[X_1 = j_1 \text{ and } \ldots \text{ and } X_n = j_n] = \Pr[X_1 = j_1] \cdot \ldots \cdot \Pr[X_n = j_n]$$

Random variables $X_1, \ldots, X_n$ are **pairwise independent** if for any $i \neq i'$ the random variables $X_i$ and $X_{i'}$ are independent.

In other words: For any $i \neq i'$ and any $j, j'$ we have

$$\Pr[X_i = j \text{ and } X_{i'} = j'] = \Pr[X_i = j] \cdot \Pr[X_{i'} = j']$$

**Lem:** For pairwise independent $X_1, \ldots, X_n$ we have

$$\text{Var}[X_1 + \cdots + X_n] = \text{Var}[X_1] + \cdots + \text{Var}[X_n]$$
Pairwise Independence

**Lemma:** For pairwise independent \(X_1, \ldots, X_n\) we have

\[
\text{Var}[X_1 + \cdots + X_n] = \text{Var}[X_1] + \cdots + \text{Var}[X_n]
\]

**Proof:**

\[
\begin{align*}
\text{Var}[X_1 + \cdots + X_n] &= \mathbb{E}[(X_1 + \cdots + X_n)^2] - \mathbb{E}[X_1 + \cdots + X_n]^2 \\
&= \sum_{i,j} \mathbb{E}[X_i \cdot X_j] - \sum_{i,j} \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] \\
&= \sum_i \mathbb{E}[X_i^2] - \sum_i \mathbb{E}[X_i]^2 \\
&= \text{Var}[X_1] + \cdots + \text{Var}[X_n]
\end{align*}
\]

For \(i \neq j\): \(X_i\) and \(X_j\) and independent, so

\[
\mathbb{E}[X_i \cdot X_j] = \mathbb{E}[X_i] \cdot \mathbb{E}[X_j]
\]

Thus all summands with \(i \neq j\) cancel!

What is \(\mathbb{E} \left[ \min_i X_i \right]??\)
Pairwise Independent Hash Function

Let \( m = p \) be a prime with \( m \geq n \)

Let \( \mathcal{H} \) be the set of all functions \( h: [n] \to [m] \) of the form

\[
    h(i) = (a \cdot i + b) \mod p
\]

where \( a, b \in [p] \)

Pick \( h \in \mathcal{H} \) uniformly at random

Each hash value \( h(i) \) is **uniformly distributed** in \([m]\)

The random variables \( h(1), ..., h(n) \) are **pairwise independent**
Pairwise Independent Hash Function

Let \( m = p \) be a prime with \( m \geq n \)

Let \( \mathcal{H} \) be the set of all functions \( h: [n] \rightarrow [m] \) of the form \( h(i) = (a \cdot i + b) \mod p \) where \( a, b \in [p] \)

Pick \( h \in \mathcal{H} \) uniformly at random

Each hash value \( h(i) \) is **uniformly distributed** in \([m]\)

The random variables \( h(1), \ldots, h(n) \) are **pairwise independent**

A function \( h \in \mathcal{H} \) can be represented by the pair \( (a, b) \in [p]^2 \), using \( 2[\log p] \) bits

We can sample a function \( h \in \mathcal{H} \) in time \( O(1) \)

*Small space, efficiently samplable, sufficiently random*
Approximate Distinct Elements

approximate the number of distinct items among $x_1, \ldots, x_m \in [n]$

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime $m$ with $n^3 \leq m \leq n^{O(1)}$

2) Pick pairwise independent hash function $h': [n] \rightarrow [m]$

3) Denote $h(x) := h'(x)/m \in (0, 1]$

4) Maintain a set $L$ containing the $k := [36/\varepsilon^2]$ smallest distinct values among $h(x_1), \ldots, h(x_m)$

5) On query():
   - If $|L| < k$: Output $\tilde{t} = |L|
   - Otherwise: Output $\tilde{t} = k / \max(L)$

**Idealized Setting:** FM (Flajolet,Martin 1985)

1) Pick random function $h: [n] \rightarrow [0,1]$

2) On update(): Maintain $X = \min_i h(x_i)$

3) On query(): Output $\tilde{t} = 1/X - 1$

Initially $L = \emptyset$

On update($x$):
   - If $h(x) \notin L$: $L = L \cup \{h(x)\}$
   - If $|L| > k$: remove largest element from $L$

$$L \approx \frac{\varepsilon^2}{4 \pi^2, \ldots}$$
Approximate Distinct Elements

approximate the number of distinct items among \(x_1, \ldots, x_m \in [n]\)

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime \(m\) with \(n^3 \leq m \leq n^{O(1)}\)
2) Pick pairwise independent hash function \(h': [n] \rightarrow [m]\)
3) Denote \(h(x) := h'(x)/m \in (0,1]\)
4) Maintain a set \(L\) containing the \(k := [36/\varepsilon^2]\) smallest distinct values among \(h(x_1), \ldots, h(x_m)\)
5) On query():
   - If \(|L| < k\): Output \(\hat{\ell} = |L|\)
   - Otherwise: Output \(\hat{\ell} = k / \max(L)\)

**Idealized Setting:** FM (Flajolet,Martin 1985)

1) Pick random function \(h: [n] \rightarrow [0,1]\)
2) On update(): Maintain \(X = \min_i h(x_i)\)
3) On query(): Output \(\tilde{\ell} = 1/X - 1\)

**Space usage:**

- For \(L\): \(O\left(\frac{1}{\varepsilon^2} \log n\right)\)
- For \(h\): \(O(\log n)\)
Approximate Distinct Elements

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime \( m \) with \( n^3 \leq m \leq n^9 \)
2) Pick pairwise independent hash function \( h' : [n] \rightarrow [m] \)
3) Denote \( h(x) := h'(x)/m \) in \((0,1]\)
4) Maintain a set \( L \) of the \( k := [36/\varepsilon^2] \) smallest distinct values among \( h(x_1), \ldots, h(x_m) \)
5) On query():
   - If \( |L| < k \): Output \( \bar{t} = |L| \)
   - Otherwise: Output \( \bar{t} = k/\max(L) \)

**Perfect hash function:**

No hash collisions

\[
\mathbb{P}[\exists x \neq y : h(x) = h(y)] 
\leq \sum_{x \neq y} \sum_{z} \mathbb{P}[h(x) = h(y) = z] 
\leq n^2 m \cdot \frac{1}{m^2} \leq \frac{1}{n}
\]

We condition on: \( h \) is a perfect hash function

If \( |L| < k \) then \( t = |L| \)

If \( |L| \geq k \) then \( t \geq k \)
Approximate Distinct Elements

approximate the number of distinct items among $x_1, \ldots, x_m \in [n]$

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime $m$ with $n^3 \leq m \leq n^6(1)$
2) Pick pairwise independent
   hash function $h': [n] \rightarrow [m]$
3) Denote $h(x) := h'(x)/m \in (0,1]$
4) Maintain a set $L$ of the $k := [36/\varepsilon^2]$ smallest distinct values among $h(x_1), \ldots, h(x_m)$
5) On query():
   - If $|L| < k$: Output $\tilde{\ell} = |L|$
   - Otherwise: Output $\tilde{\ell} = k / \max(L)$

let $y_1, \ldots, y_t$ be the distinct items among $x_1, \ldots, x_m \in [n]$

**Success Probability:** Can assume $t \geq k$

$$Y_i = \begin{cases} 1, & \text{if } h(y_i) < \frac{k}{(1+\varepsilon)t} \\ 0, & \text{otherwise} \end{cases}$$

$$Y = Y_1 + \ldots + Y_t$$

Observe: $\tilde{\ell} > (1 + \varepsilon)t$ can only happen if $Y \geq k$

If $Y < k$, $\max(L) > \frac{k}{(1+\varepsilon)t}$

$$\frac{k}{\max(L)} \leq (1+\varepsilon)t$$
Approximate Distinct Elements

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime \( m \) with \( n^3 \leq m \leq n^9(1) \)
2) Pick pairwise independent hash function \( h' : [n] \rightarrow [m] \)
3) Denote \( h(x) := h'(x)/m \in (0,1] \)
4) Maintain a set \( L \) of the \( k := [36/\varepsilon^2] \) smallest distinct values among \( h(x_1), \ldots, h(x_m) \)
5) On query():
   - If \( |L| < k \): Output \( \tilde{\ell} = |L| \)
   - Otherwise: Output \( \tilde{\ell} = k / \max(L) \)

let \( y_1, \ldots, y_t \) be the distinct items among \( x_1, \ldots, x_m \in [n] \)

**Success Probability:** Can assume \( t \geq k \)

\[
Y_i = \begin{cases} 
1, & \text{if } h(y_i) < \frac{k}{(1+\varepsilon)t} \\
0, & \text{otherwise}
\end{cases}
\]

\[
Y = Y_1 + \ldots + Y_t
\]

Observe: \( \tilde{\ell} > (1 + \varepsilon)t \) can only happen if \( Y \geq k \)

\[
\mathbb{E}[Y_i] = \mathbb{P}[Y_i = 1] \leq \frac{k}{(1+\varepsilon)t} + \frac{1}{m} \leq \frac{k}{(1+\varepsilon/2)t}
\]

\[
\mathbb{E}[Y] \leq \frac{k}{1+\varepsilon/2}
\]

\[
\text{Var}[Y_i] = \mathbb{E}[Y_i^2] - \mathbb{E}[Y_i]^2 < \mathbb{E}[Y_i^2] = \mathbb{E}[Y_i] \leq \frac{k}{(1+\varepsilon/2)t}
\]

\[
\text{Var}[Y] \leq \frac{k}{1+\varepsilon/2}
\]
Approximate Distinct Elements

approximate the number of distinct items among $x_1, \ldots, x_m \in [n]$

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime $m$ with $n^3 \leq m \leq n^0(1)$
2) Pick pairwise independent hash function $h': [n] \rightarrow [m]$
3) Denote $h(x) := h'(x)/m \in (0,1]$
4) Maintain a set $L$ of the $k := \lfloor 36/\varepsilon^2 \rfloor$ smallest distinct values among $h(x_1), \ldots, h(x_m)$
5) On query():
   - If $|L| < k$: Output $\tilde{t} = |L|
   - Otherwise: Output $\tilde{t} = k/\max(L)$

let $y_1, \ldots, y_t$ be the distinct items among $x_1, \ldots, x_m \in [n]$

**Success Probability:**

\[
\mathbb{P}[\tilde{t} > (1 + \varepsilon)t] \leq \mathbb{P}[Y \geq k]
\]
\[
\leq \mathbb{P}
\left|
Y - \mathbb{E}[Y]
\right|
\geq k \left(1 - \frac{1}{1+\varepsilon/2}\right)
\]
\[
\leq \mathbb{P}
\left|
Y - \mathbb{E}[Y]
\right|
\geq \frac{k\varepsilon/2}{1+\varepsilon/2}
\]
\[
\leq \frac{k}{1+\varepsilon/2} \cdot \frac{(1+\varepsilon/2)^2}{(k\varepsilon/2)^2} = \frac{4(1+\varepsilon/2)^2}{\varepsilon^2k} \leq \frac{6}{\varepsilon^2k} \leq \frac{1}{6}
\]

Analogous: $\mathbb{P}[\tilde{t} < (1 - \varepsilon)t] \leq \frac{1}{6}$

**Chebyshev:**

\[
\mathbb{P}[|X - \mathbb{E}[X]| \geq \lambda] \leq \frac{\text{Var}[X]}{\lambda^2}
\]
\[
\mathbb{E}[Y], \text{Var}[Y] \leq \frac{k}{1+\varepsilon/2}
\]
**Approximate Distinct Elements**

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

---

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime \( m \) with \( n^3 \leq m \leq n^{O(1)} \)

2) Pick pairwise independent hash function \( h': [n] \to [m] \)

3) Denote \( h(x) := h'(x)/m \in (0,1] \)

4) Maintain a set \( L \) containing the \( k := \left\lceil \frac{36}{\varepsilon^2} \right\rceil \) smallest distinct values among \( h(x_1), \ldots, h(x_m) \)

5) On query():
   - If \( |L| < k \): Output \( \tilde{t} = |L| \)
   - Otherwise: Output \( \tilde{t} = k / \max(L) \)

---

\[
\mathbb{P}[(1 - \varepsilon)t \leq \tilde{t} \leq (1 - \varepsilon)t] \geq 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}
\]

---

**Boosting via Chernoff:**

TV++ = median of \( O(\log(1/\delta)) \) runs of TV:

\[
\mathbb{P}[(1 - \varepsilon)t \leq \tilde{t}^{++} \leq (1 - \varepsilon)t] \geq 1 - \delta
\]

**Space usage:** \( O\left(\frac{1}{\varepsilon^2} \log \left(\frac{1}{\delta}\right) \log n\right) \)
Outline

1) Distinct Elements: Idealized Setting
2) Distinct Elements: Theoretical Variant
3) Distinct Elements: Practical Variant
Practical Variant
what Google implements

Hyperloglog: (Flajolet et al. 2007)

1) Pick hash function $h: [n] \rightarrow [0,1]$
2) Initialize $M[0], \ldots, M[m-1]$ to $-\infty$
3) On update($x$):
   Split $h(x)$ into $b$ bits and the rest: $h_1(x), h_2(x)$
   Let $\rho$ be the number of leading 0s of $h_2(x)$
   $M[h_1(x)] = \max\{M[h_1(x)], \rho + 1\}$
3) Output $\alpha_m m^2 \left( \sum_j 2^{-M[j]} \right)^{-1}$

where $\alpha_m = \left( m \int_0^\infty \left( \log \left( \frac{2+x}{1+x} \right) \right)^m dx \right)^{-1}$
$\approx 0.72134$

Relative error $\approx \sqrt{\frac{1.04}{m}},$ so $m \approx \frac{1}{\epsilon^2}$
Analysis is very complicated!
Has only been analyzed in idealized setting!

$\rho = \left\lfloor \log\left( \frac{1}{h_2(x)} \right) \right\rfloor$
$2^{M[j]} = \max_{x: h_1(x) = j} 2^{\left\lfloor \log\left( \frac{1}{h_2(x)} \right) \right\rfloor + 1}$
$2^{-M[j]} \approx \min_{x: h_1(x) = j} h_2(x) \approx \chi$
Practical Variant

what Google implements

Hyperloglog: (Flajolet et al. 2007)

1) Pick hash function $h: [n] \rightarrow [0,1]$

1) Pick parameter $m = 2^b$

2) Initialize $M[0], \ldots, M[m-1]$ to $-\infty$

3) On update($x$):

   Split $h(x)$ into $b$ bits and the rest: $h_1(x), h_2(x)$

   Let $\rho$ be the number of leading 0s of $h_2(x)$

   $M[h_1(x)] = \max\{M[h_1(x)], \rho + 1\}$

3) Output $\alpha_m m^2 \left(\sum_j 2^{-M[j]}\right)^{-1}$

in practice:

use 64-bit hash function

$m \approx 128$

Easy to implement, in contrast to some theoretical algorithms

Update time $O(1)$, in contrast to the previously presented algorithms

Space: $\approx m \log \log t$ bits
More Material

- Idealized: [Flajolet, Martin „Probabilistic counting algorithms for data base applications“ 1985]
- Theoretical: [Bar-Yossef, Jayram, Kumar, Sivakumar, Trevisan „Counting distinct elements in a data stream“ 2002]
- Practical: [Flajolet, Fusy, Gandouet, Meunier „Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm“ 2007]
- Theoretically optimal: [Kane, Nelson, Woodruff „An optimal algorithm for the distinct elements problem“ 2010]

- Course Website → Material → A Primer to Randomness
  \[ \text{Constant: } \frac{c^2}{\varepsilon^2} + \frac{\log n}{\varepsilon^2} \]
  \[ \text{Count: } 0(\varepsilon^{-2} + \log n) \]

- Course Website → Material → Link to Summer School on Streaming by Jelani Nelson

- Exercise Sheet 1: Online today/tomorrow, due date is Friday, May 22

See you on Monday!
EXTRA
Approximate Distinct Elements

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1. Pick a prime \( m \) with \( n^3 \leq m \leq n^{0.1} \)
2. Pick pairwise independent hash function \( h' : [n] \to [m] \)
3. Denote \( h(x) := h'(x)/m \in (0,1] \)
4. Maintain a set \( L \) of the \( k := [36/\varepsilon^2] \) smallest distinct values among \( h(x_1), \ldots, h(x_m) \)
5. On query():
   - If \( |L| < k \): Output \( \tilde{\ell} = |L| \)
   - Otherwise: Output \( \tilde{\ell} = k/\max(L) \)

let \( y_1, \ldots, y_t \) be the distinct items among \( x_1, \ldots, x_m \in [n] \)

**Success Probability:** Can assume \( t \geq k \)

\[
Z_i = \begin{cases} 
1, & \text{if } h(y_i) \leq \frac{k}{(1-\varepsilon)t} \\
0, & \text{otherwise}
\end{cases}
\]

\[
Z = Z_1 + \cdots + Z_t
\]

Observe: \( \tilde{\ell} < (1 - \varepsilon)t \) can only happen if \( Z < k \)

\[
\mathbb{E}[Z_i] = \mathbb{P}[Z_i = 1] \geq \frac{k}{(1-\varepsilon)t} - \frac{1}{m} \geq \frac{k}{(1-\varepsilon/2)t}
\]

\[
\mathbb{E}[Z] \geq \frac{k}{1-\varepsilon/2}
\]

\[
\text{Var}[Z_i] = \mathbb{E}[Z_i^2] - \mathbb{E}[Z_i]^2 < \mathbb{E}[Z_i^2]
\]

\[
= \mathbb{E}[Z_i] \leq \frac{k}{(1-\varepsilon/2)t}
\]

\[
\text{Var}[Z] \leq \frac{k}{1-\varepsilon/2}
\]
Approximate Distinct Elements

approximate the number of distinct items among \( x_1, \ldots, x_m \in [n] \)

**Theoretical Variant:** (Bar-Yossef et al. 2002)

1) Pick a prime \( m \) with \( n^3 \leq m \leq n^{0.1} \)
2) Pick pairwise independent hash function \( h' : [n] \rightarrow [m] \)
3) Denote \( h(x) := h'(x)/m \in (0,1] \)
4) Maintain a set \( L \) of the \( k := [36/\varepsilon^2] \) smallest distinct values among \( h(x_1), \ldots, h(x_m) \)
5) On query():
   - If \( |L| < k \): Output \( \tilde{t} = |L| \)
   - Otherwise: Output \( \tilde{t} = k/\max(L) \)

**Success Probability:**

\[
\Pr[\tilde{t} < (1 - \varepsilon)t] \leq \Pr[Z < k]
\leq \Pr[|Z - \mathbb{E}[Z]| \geq k \left(\frac{1}{1-\varepsilon/2} - 1\right)]
\leq \Pr[|Z - \mathbb{E}[Z]| \geq \frac{k \varepsilon/2}{1-\varepsilon/2}]
\leq \frac{k}{1-\varepsilon/2} \cdot \frac{(1-\varepsilon/2)^2}{(k \varepsilon/2)^2} = \frac{4(1-\varepsilon/2)^2}{\varepsilon^2 k} \leq \frac{6}{\varepsilon^2 k} \leq \frac{1}{6}
\]

let \( y_1, \ldots, y_t \) be the distinct items among \( x_1, \ldots, x_m \in [n] \)

**Chebyshev:** \( \Pr[|X - \mathbb{E}[X]| \geq \lambda] \leq \frac{\text{Var}[X]}{\lambda^2} \)

\( \text{Var}[Z] \leq \frac{k}{1-\varepsilon/2} \leq \mathbb{E}[Z] \)