CS 473: Algorithms

Ruta Mehta

University of Illinois, Urbana-Champaign

Spring 2018

Streaming Algorithms

Lecture 12

March 1, 2018

Most slides are courtesy Prof. Chekuri

Streaming Algorithms

A topic that is both very old, and very current!

Dawn of CS..

Data was stored on tapes, and amount of RAM was very small.

- Too much data, too little space.
- Store only summary or sketch of data.

Streaming Algorithms

A topic that is both very old, and very current!

Dawn of CS..

Data was stored on tapes, and amount of RAM was very small.

- Too much data, too little space.
- Store only summary or sketch of data.

Now..

Terabytes of memory, Gigabytes of RAM.

- Data streams: Humongous amount of data (sometimes never ending)!
- Can go over it at most once, and sometimes not even that!
- Store only summary: sub-linear space-time algorithms.

Examples

An internet router sees a stream of packets, and may want to know,

- which connection is using the most packets
- how many different connections
- median of the file sizes transferred since mid-night
- which connections are using more than 0.1% of the bandwidth.

Computing aggregative information about data streams.

Outline

Computation with data streams.

Heavy-hitters

- Majority element (by R. Boyer and J.S. Moore)
- ϵ-heavy hitters – deterministic
- Approximate counting

Counting using hashing – Count-min Sketch (Cormode-Muthukrishnan'05)

- Variant of Bloom filters.

Data Streams

A stream of data elements, $S = a_1, a_2, \ldots$.

Say a_t arrive at time t. Let us assume that a_t's are numbers for this lecture.

Data Streams

A stream of data elements, $S = a_1, a_2, \ldots$.

Say a_t arrive at time t. Let us assume that a_t's are numbers for this lecture.

Denote $a_{[1..t]} = \langle a_1, a_2, \ldots, a_t \rangle$.

Given some function we want to compute it continually, while using limited space.

- at any time t we should be able to query the function value on the stream seen so far, i.e., $a_{[1..t]}$.

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing Sum

$$F(a_{[1..t]}) = \sum_{i=1}^{t} a_i$$

Outputs are: 3, 4, 21, 25, 16, 48, 149, 152, -570, ...

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing Sum

$$F(a_{[1..t]}) = \sum_{i=1}^{t} a_i$$

Outputs are: 3, 4, 21, 25, 16, 48, 149, 152, -570, ...

Keep a counter, and keep adding to it.

After T rounds, the number can be at most $T2^b$. $O(b + \log T)$ space.

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing max

$$F(a_{[1..t]}) = \max_{i=1}^{t} a_i$$

Outputs are: 3, 3, 17, 17, 17, 32, 101, 101, ...

Just need to store b bits.

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing max

$$F(a_{[1..t]}) = \max_{i=1}^{t} a_i$$

Outputs are: 3, 3, 17, 17, 17, 32, 101, 101, ...

Just need to store b bits.

Median?

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing max

$$F(a_{[1..t]}) = \max_{i=1}^{t} a_i$$

Outputs are: 3, 3, 17, 17, 17, 32, 101, 101, ...

Just need to store b bits.

Median? A lot more tricky

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing max

$$F(a_{[1..t]}) = \max_{i=1}^{t} a_i$$

Outputs are: 3, 3, 17, 17, 17, 32, 101, 101, ...

Just need to store b bits.

Median? A lot more tricky

distinct elements?

Examples

$$S = 3, 1, 17, 4, -9, 32, 101, 3, -722, 3, 900, 4, 32, \ldots$$

Computing max

$$F(a_{[1..t]}) = \max_{i=1}^{t} a_i$$

Outputs are: 3, 3, 17, 17, 17, 32, 101, 101, ...

Just need to store b bits.

Median? A lot more tricky

distinct elements? also tricky!

Streaming Algorithms: Framework

$\langle$Initialize summary information$\rangle$

While stream S is not done
  $x \leftarrow$ next element in S
  $\langle$Do something with x and update summary information$\rangle$
  $\langle$Output something if needed$\rangle$

Return $\langle$summary$\rangle$

Streaming Algorithms: Framework

⟨Initialize summary information⟩

While stream S is not done
    $x \leftarrow$ next element in S
    ⟨Do something with x and update summary information⟩
    ⟨Output something if needed⟩

Return ⟨summary⟩

Despite of restrictions, we can compute interesting functions if we can tolerate some error.

Streaming Algorithms: One-sided Error

No false negative

Anything that needs to be considered/counted should be counted.

There may be false positive

We may over count. That is we may consider/count something that shouldn't have been counted.

Part I

Heavy Hitters

Finding the Majority Element

Find the element that occur strictly more than half the time, if any.

Note that at most one such element!

Finding the Majority Element

Find the element that occur strictly more than half the time, if any.

Note that at most one such element!

$$E, D, B, D, D_{\mathbf{5}}, D, B, B, B, B, B_{\mathbf{11}}, E, E, E, E, E_{\mathbf{16}}$$

- At time **5**, it is D.
- At time **11**, it is B
- At time **16**, none!

Treasure hunt

Once upon a time...

Puzzle

Finding a Majority Element

Treasure hunt

Once upon a time... there was a treasure hidden in a cave that different gangs were after. Only one-on-one fight is the unsaid rule (wild west style). Thus, if two members from different gangs face each other, then they shoot each other and both die.

Puzzle

Finding a Majority Element

Treasure hunt

Once upon a time... there was a treasure hidden in a cave that different gangs were after. Only one-on-one fight is the unsaid rule (wild west style). Thus, if two members from different gangs face each other, then they shoot each other and both die.

Which gang will get the treasure?

Suppose more than half the bandits are part of gang ALGO, then?

Puzzle

Finding a Majority Element

Treasure hunt

Once upon a time... there was a treasure hidden in a cave that different gangs were after. Only one-on-one fight is the unsaid rule (wild west style). Thus, if two members from different gangs face each other, then they shoot each other and both die.

Which gang will get the treasure?

Suppose more than half the bandits are part of gang ALGO, then?

Gang ALGO will get the treasure for sure!

Finding the Majority Element

Find the element that accrue strictly more than half the time, if any.

R. Boyer and J. S. Moore Algorithm

Initialize: mem$=\emptyset$ and counter$=0$

Finding the Majority Element

Find the element that accrue strictly more than half the time, if any.

R. Boyer and J. S. Moore Algorithm

Initialize: mem=$\emptyset$ and counter=0

When element a_t arrives
    if (counter == 0)
        set mem=a_t and counter=1

Finding the Majority Element

Find the element that accrue strictly more than half the time, if any.

R. Boyer and J. S. Moore Algorithm

Initialize: mem=$\emptyset$ and counter=0

When element a_t arrives
    if (counter == 0)
        set mem=a_t and counter=1
    else if (a_t == mem) then counter++

Finding the Majority Element

Find the element that accrue strictly more than half the time, if any.

R. Boyer and J. S. Moore Algorithm

Initialize: mem=$\emptyset$ and counter=0

When element a_t arrives
    if (counter == 0)
        set mem=a_t and counter=1
    else if (a_t == mem) then counter++
    else counter−− (discard a_t and a copy of mem)
Return mem.

Finding the Majority Element

Find the element that accrue strictly more than half the time, if any.

R. Boyer and J. S. Moore Algorithm

Initialize: mem=$\emptyset$ and counter=0

When element a_t arrives
    if (counter == 0)
        set mem=a_t and counter=1
    else if (a_t == mem) then counter++
    else counter−− (discard a_t and a copy of mem)
Return mem.

Even if no majority element, something is returned – False positive.

Finding the Majority Element: Example

R. Boyer and J. S. Moore Algorithm

Initialize: mem=$\emptyset$ and counter=0

When element a_t arrives
    if (counter == 0)
      set mem=a_t and counter=1
    else if (a_t == mem) then counter++
    else counter−− (discard a_t and a copy of mem)
Return mem.

$$E, D, B, D, D_5, D, B, B, B, B, B_{11}, E, E, E, E, E_{16}$$

a_t	E	D	B	D	D	D	B	B	B	B	B	...
mem	E	E	B	B	D	D	D	D	B	B	B	...
counter	1	0	1	0	1	2	1	0	1	2	3	...

Finding a Majority Element

Correctness, if majority element

Lemma

If there is a majority element, the algorithm will output it.

Proof.

- Decreasing counter is like throwing away a copy of element in mem.

Finding a Majority Element

Correctness, if majority element

Lemma

If there is a majority element, the algorithm will output it.

Proof.

- Decreasing counter is like throwing away a copy of element in mem.
- We do this every time a_t is different than mem, and there are less than half such a_t.

Finding a Majority Element

Correctness, if majority element

Lemma

If there is a majority element, the algorithm will output it.

Proof.

- Decreasing counter is like throwing away a copy of element in mem.
- We do this every time a_t is different than mem, and there are less than half such a_t.
- Sometimes *mem* may not contain the majority element.

Finding a Majority Element

Correctness, if majority element

Lemma

If there is a majority element, the algorithm will output it.

Proof.

- Decreasing counter is like throwing away a copy of element in mem.
- We do this every time a_t is different than mem, and there are less than half such a_t.
- Sometimes *mem* may not contain the majority element. However, even if we are throwing away the majority element every time, since they are more than half all can't be thrown.

Finding a Majority Element

Correctness, if majority element

Lemma

If there is a majority element, the algorithm will output it.

Proof.

- Decreasing counter is like throwing away a copy of element in mem.
- We do this every time a_t is different than mem, and there are less than half such a_t.
- Sometimes *mem* may not contain the majority element. However, even if we are throwing away the majority element every time, since they are more than half all can't be thrown.

In fact at any time t, mem contains majority element of sub-stream $a_{[1..t]}$, if any. □

Part II

Heavy Hitters

ϵ-Heavy Hitters

Definition

Given a stream $S = a_1, a_2, \ldots$, define count of element e at any time t to be

$$\text{count}_t(e) = |\{i \leq t \mid a_i = e\}|$$

e is called ϵ-heavy hitter at time t if $\text{count}_t(e) > \epsilon t$.

ϵ-Heavy Hitters

Definition

Given a stream $S = a_1, a_2, \ldots$, define count of element e at any time t to be

$$\text{count}_t(e) = |\{i \leq t \mid a_i = e\}|$$

e is called ϵ-heavy hitter at time t if $\text{count}_t(e) > \epsilon t$.

Goal:

Maintain a structure containing all the ϵ-heavy hitters so far.
At any point there are at most $1/\epsilon$ such elements.

ϵ-Heavy Hitters

Definition

Given a stream $S = a_1, a_2, \ldots$, define count of element e at any time t to be

$$\text{count}_t(e) = |\{i \leq t \mid a_i = e\}|$$

e is called ϵ-heavy hitter at time t if $\text{count}_t(e) > \epsilon t$.

Goal:

Maintain a structure containing all the ϵ-heavy hitters so far.
At any point there are at most $1/\epsilon$ such elements.

Crucial Note: false positive are OK, but no false negative

We are NOT allowed to miss any heavy-hitters, but we could store non-heavy-hitters.

ϵ-Heavy Hitters: Example

If $\epsilon = 1/2$ then the majority element!

ϵ-Heavy Hitters: Example

If $\epsilon = 1/2$ then the majority element!

$$E, D, B, D, D_5, D, B, A, B, B, B_{11}, E, E, E, E, E_{16}$$

$1/3$-heavy hitters

- At time 5, it is D.
- At time 11, both B and D.
- At time 15,

ϵ-Heavy Hitters: Example

If $\epsilon = 1/2$ then the majority element!

$$E, D, B, D, D_5, D, B, A, B, B, B_{11}, E, E, E, E, E_{16}$$

$1/3$-heavy hitters

- At time 5, it is D.
- At time 11, both B and D.
- At time 15, none!
- At time 16,

ϵ-Heavy Hitters: Example

If $\epsilon = 1/2$ then the majority element!

$$E, D, B, D, D_5, D, B, A, B, B, B_{11}, E, E, E, E, E_{16}$$

$1/3$-heavy hitters

- At time 5, it is D.
- At time 11, both B and D.
- At time 15, none!
- At time 16, it is E.

ϵ-Heavy Hitters: Example

If $\epsilon = 1/2$ then the majority element!

$$E, D, B, D, D_5, D, B, A, B, B, B_{11}, E, E, E, E, E_{16}$$

$1/3$-heavy hitters

- At time 5, it is D.
- At time 11, both B and D.
- At time 15, none!
- At time 16, it is E.

As time passes, the set of heavy hitters may change completely.

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

When element a_t arrives,
  If ($a_t == T[j]$ for some $j \leq k$), then $C[j]++$.

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

When element a_t arrives,
 If ($a_t == T[j]$ for some $j \leq k$), then $C[j]++$.
 Else if ($C[j] == 0$ for some $j \leq k$), then

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

When element a_t arrives,
- If ($a_t == T[j]$ for some $j \leq k$), then $C[j] + +$.
- Else if ($C[j] == 0$ for some $j \leq k$), then
  - Set $T[j] \leftarrow a_t$ and $C[j] \leftarrow 1$.

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

When element a_t arrives,
- If ($a_t == T[j]$ for some $j \leq k$), then $C[j]++$.
- Else if ($C[j] == 0$ for some $j \leq k$), then
  - Set $T[j] \leftarrow a_t$ and $C[j] \leftarrow 1$.
- Else do $C[j]--$ for all j. (discard a_t and a copy of all $T[j]$)

ϵ-Heavy Hitters: Algorithm

If $\epsilon = 1/2$ then the majority element!

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

Initialize: $C[j] = 0$ and $T[j] = \emptyset$ for all i.

When element a_t arrives,
- If ($a_t == T[j]$ for some $j \leq k$), then $C[j]++$.
- Else if ($C[j] == 0$ for some $j \leq k$), then
  - Set $T[j] \leftarrow a_t$ and $C[j] \leftarrow 1$.
- Else do $C[j]--$ for all j. (discard a_t and a copy of all $T[j]$)

Same as the *Majority algorithm* for $\epsilon = 1/2$.

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} est_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

For each element, count is maintained up to ϵt error!

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] \qquad \text{if } e = T[j] \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $\text{est}_t(e) \leq \text{count}_t(e) \leq \text{est}_t(e) + \epsilon t$

For each element, count is maintained up to ϵt error!

If e is not an ϵ-heavy hitter then $\text{count}_t(e) \leq \epsilon t$, and hence $\text{est}_t(e) = 0$ is correct up to ϵt error.

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

Corollary

For any time t, T *contains all the* ϵ*-heavy hitters in* $a_{[1..t]}$.

Proof.

If e is a heavy hitter at time t then $\text{count}_t(e) > \epsilon t$.

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

Corollary

For any time t, T *contains all the* ϵ*-heavy hitters in* $a_{[1..t]}$.

Proof.

If e is a heavy hitter at time t then $\text{count}_t(e) > \epsilon t$.Using the lemma,

$$\text{est}_t(e) \geq \text{count}_t(e) - \epsilon t$$

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $\text{est}_t(e) \leq \text{count}_t(e) \leq \text{est}_t(e) + \epsilon t$

Corollary

For any time t, T contains all the ϵ-heavy hitters in $a_{[1..t]}$.

Proof.

If e is a heavy hitter at time t then $\text{count}_t(e) > \epsilon t$.Using the lemma,

$$\text{est}_t(e) \geq \text{count}_t(e) - \epsilon t > 0$$

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

Proof.

Counter for e increases only when we see e, $\therefore$ $\text{est}_t(e) \leq \text{count}_t(e)$.

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $\text{est}_t(e) \leq \text{count}_t(e) \leq \text{est}_t(e) + \epsilon t$

Proof.

Counter for e increases only when we see e, $\therefore$ $\text{est}_t(e) \leq \text{count}_t(e)$.
We want $\text{count}_t(e) - \text{est}_t(e) \leq \epsilon t$. It increases by one,

- when we decrease all k counters, and see an element outside T

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] & \text{if } e = T[j] \\ &= 0 & \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $est_t(e) \leq count_t(e) \leq est_t(e) + \epsilon t$

Proof.

Counter for e increases only when we see e, $\therefore$ $\text{est}_t(e) \leq \text{count}_t(e)$.
We want $\text{count}_t(e) - \text{est}_t(e) \leq \epsilon t$. It increases by one,

- when we decrease all k counters, and see an element outside T
- this is like discarding $k + 1$ elements.
- up to time t, we have only t elements to discard

ϵ-Heavy Hitters

Algorithm Analysis

At any time t, our estimates are:

$$\begin{aligned} \text{est}_t(e) &= C[j] && \text{if } e = T[j] \\ &= 0 && \text{otherwise} \end{aligned}$$

Lemma

Estimates satisfy: $\text{est}_t(e) \leq \text{count}_t(e) \leq \text{est}_t(e) + \epsilon t$

Proof.

Counter for e increases only when we see e, $\therefore$ $\text{est}_t(e) \leq \text{count}_t(e)$.
We want $\text{count}_t(e) - \text{est}_t(e) \leq \epsilon t$. It increases by one,

- when we decrease all k counters, and see an element outside T
- this is like discarding $k + 1$ elements.
- up to time t, we have only t elements to discard

So at most $t/(k+1) < t\epsilon$ such increases. □

ϵ-Heavy Hitters: Algorithm

Space usage

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

$\vdots$

Maintains $O(1/\epsilon)$ counters and elements.

ϵ-Heavy Hitters: Algorithm

Space usage

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

$\vdots$

Maintains $O(1/\epsilon)$ counters and elements.
$O(\log t)$ for each counter. $O(\Sigma)$ for each element, where Σ is the description of largest element.

ϵ-Heavy Hitters: Algorithm

Space usage

Set $k = \lceil 1/\epsilon \rceil - 1$. (if $\epsilon = 1/2$ then $k = 1$)

Algorithm

Keep an array $T[1, \ldots, k]$ to hold elements
Keep an array $C[1, \ldots, k]$ to hold their counters

$\vdots$

Maintains $O(1/\epsilon)$ counters and elements.
$O(\log t)$ for each counter. $O(\Sigma)$ for each element, where Σ is the description of largest element.

Total: $O(1/\epsilon(\log t + \Sigma))$.

Recall: maintains counts for all elements up to ϵt error.

Part III

Use of Hash Functions

Maintaining Counts

Problem Statement:

At any time t, estimate the number of times every element appeared so far.

Maintaining Counts

Problem Statement:

At any time t, estimate the number of times every element appeared so far.

If error up to ϵt is OK, then we can use ϵ-heavy hitter algorithm.

Maintaining Counts

Problem Statement:

At any time t, estimate the number of times every element appeared so far.

If error up to ϵt is OK, then we can use ϵ-heavy hitter algorithm.

It takes $O(1/\epsilon(\log t + \Sigma))$ space.

Maintaining Counts

Problem Statement:

At any time t, estimate the number of times every element appeared so far.

If error up to ϵt is OK, then we can use ϵ-heavy hitter algorithm.

It takes $O(1/\epsilon(\log t + \Sigma))$ space.

Can we do better?

Yes – Bloom filter like idea

Recall: Bloom Filter

Storage for inserts and lookups

Sample hash functions $h_1, \ldots, h_d$ independently and uniformly at random from some family $\mathcal{H}$.

Insert(e)
For $i = 1...d$
Set $T_i[h_i(e)] \leftarrow 1$

Lookup(e)
For $i = 1...d$
If ($T_i[h_i(e)] == 0$) then return "No"
Return "Yes"

If e inserted, then Lookup(e) will always return "Yes".

Recall: Bloom Filter

Storage for inserts and lookups

Sample hash functions $h_1, \dots, h_d$ independently and uniformly at random from some family $\mathcal{H}$.

```
Insert(e)
  For i = 1...d
    Set T_i[h_i(e)] ← 1
```

```
Lookup(e)
  For i = 1...d
    If (T_i[h_i(e)] == 0) then return "No"
  Return "Yes"
```

If e inserted, then Lookup(e) will always return "Yes".

e not inserted, but still it can return "Yes" with very low probability.

- Due to some e's being inserted with $h_i(e') = h_i(e)$.
- If $\Pr_{h_i \sim \mathcal{H}}[e$ not inserted and $T_i[h_i(e)] = 1] \leq \alpha$, then combined error probability would be at most α^d.

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

Keep d arrays $C_1, ..., C_d$, each to hold m counters.

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

Keep d arrays $C_1, ..., C_d$, each to hold m counters.

$\mathcal{H}$: **2-universal family** of hash functions mapping U to $\{0, \ldots, m-1\}$. Sample $h_1, \ldots, h_d$ independently and uniformly at random from $\mathcal{H}$.

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

Keep d arrays $C_1, ..., C_d$, each to hold m counters.

$\mathcal{H}$: **2-universal family** of hash functions mapping U to $\{0, \ldots, m-1\}$. Sample $h_1, \ldots, h_d$ independently and uniformly at random from $\mathcal{H}$.

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)] + +
```

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

Keep d arrays $C_1, ..., C_d$, each to hold m counters.

$\mathcal{H}$: **2-universal family** of hash functions mapping U to $\{0, \ldots, m-1\}$. Sample $h_1, \ldots, h_d$ independently and uniformly at random from $\mathcal{H}$.

CMInsert(e)
For $i = 1...d$
Do $C_i[h_i(e)] + +$

CMEstimate(e)
$est \leftarrow \infty$
For $i = 1...d$
$est \leftarrow \min\{est, C_i[h_i(e)]\}$
Return est

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

Keep d arrays $C_1, ..., C_d$, each to hold m counters.

$\mathcal{H}$: **2-universal family** of hash functions mapping U to $\{0, \ldots, m-1\}$. Sample $h_1, \ldots, h_d$ independently and uniformly at random from $\mathcal{H}$.

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)] + +
```

```
CMEstimate(e)
  est ← ∞
  For i = 1...d
    est ← min{est, C_i[h_i(e)]}
  Return est
```

As element a_t arrives at time t, call CMInsert(a_t).

To get count of e at any time t, call CMEstimate(e).

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)] + +
```

```
CMEstimate(e)
  est ← ∞
  For i = 1...d
    est ← min{est, C_i[h_i(e)]}
  Return est
```

At time t, let $\text{est}_t(e) = \text{CMEstimate}(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)]++
```

```
CMEstimate(e)
  est ← ∞
  For i = 1...d
    est ← min{est, C_i[h_i(e)]}
  Return est
```

At time t, let $\text{est}_t(e) = \text{CMEstimate}(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

Observation: $\text{est}_t(e) \geq \text{count}_t(e)$.

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)] + +
```

```
CMEstimate(e)
  est ← ∞
  For i = 1...d
    est ← min{est, C_i[h_i(e)]}
  Return est
```

At time t, let $\text{est}_t(e) = \text{CMEstimate}(e) = \min_{i=1}^{d} C_i[h_i(e)]$.
Observation: $\text{est}_t(e) \geq \text{count}_t(e)$.

Question: How big $(\text{est}_t(e) - \text{count}_t(e))$ can be?

Count Min-Sketch

By G. Cormode and S. M. Muthukrishnan'05

```
CMInsert(e)
  For i = 1...d
    Do C_i[h_i(e)]++
```

```
CMEstimate(e)
  est ← ∞
  For i = 1...d
    est ← min{est, C_i[h_i(e)]}
  Return est
```

At time t, let $\text{est}_t(e) = \text{CMEstimate}(e) = \min_{i=1}^{d} C_i[h_i(e)]$.
Observation: $\text{est}_t(e) \geq \text{count}_t(e)$.

Question: How big $(\text{est}_t(e) - \text{count}_t(e))$ can be?

Recall: Any $e, y \in U$, if $e \neq y$ then $\Pr[h_i(y) = h_i(e)] = \frac{1}{m}\ \forall i$.

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Let $f'_e = \text{est}_t(e)$ and $f_e = \text{count}_t(e)$. We want to bound $(f'_e - f_e)$.

Observations:

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Let $f'_e = \text{est}_t(e)$ and $f_e = \text{count}_t(e)$. We want to bound $(f'_e - f_e)$.

Observations:

Define indicator variable $X_{i,e,y} = [h_i(y) = h_i(e)]$.

$$\mathbf{E}[X_{i,e,y}] = \mathbf{Pr}[h_i(y) = h_i(e)] = 1/m$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Let $f'_e = \mathrm{est}_t(e)$ and $f_e = \mathrm{count}_t(e)$. We want to bound $(f'_e - f_e)$.

Observations:

Define indicator variable $X_{i,e,y} = [h_i(y) = h_i(e)]$.

$$\mathbf{E}[X_{i,e,y}] = \mathbf{Pr}[h_i(y) = h_i(e)] = 1/m$$

Let $X_{i,e} := \sum_{y \neq e} X_{i,e,y} f_y$ be the total over counting at $C_i[h_i(e)]$.

$$C_i[h_i(e)] = X_{i,e} + f_e$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Let $f'_e = \text{est}_t(e)$ and $f_e = \text{count}_t(e)$. We want to bound $(f'_e - f_e)$.

Observations:

Define indicator variable $X_{i,e,y} = [h_i(y) = h_i(e)]$.

$$\mathbf{E}[X_{i,e,y}] = \mathbf{Pr}[h_i(y) = h_i(e)] = 1/m$$

Let $X_{i,e} := \sum_{y \neq e} X_{i,e,y} f_y$ be the total over counting at $C_i[h_i(e)]$.

$$C_i[h_i(e)] = X_{i,e} + f_e$$

and since at most t elements have arrived so far,

$$\mathbf{E}[X_{i,e}] = \sum_{y \neq e} \mathbf{E}[X_{i,e,y}]\, f_y = \frac{1}{m} \sum_{y \neq e} f_y \leq$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Let $f'_e = \text{est}_t(e)$ and $f_e = \text{count}_t(e)$. We want to bound $(f'_e - f_e)$.

Observations:

Define indicator variable $X_{i,e,y} = [h_i(y) = h_i(e)]$.

$$\mathbf{E}[X_{i,e,y}] = \Pr[h_i(y) = h_i(e)] = 1/m$$

Let $X_{i,e} := \sum_{y \neq e} X_{i,e,y} f_y$ be the total over counting at $C_i[h_i(e)]$.

$$C_i[h_i(e)] = X_{i,e} + f_e$$

and since at most t elements have arrived so far,

$$\mathbf{E}[X_{i,e}] = \sum_{y \neq e} \mathbf{E}[X_{i,e,y}]\, f_y = \frac{1}{m} \sum_{y \neq e} f_y \leq \frac{t}{m}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\Pr[C_i[h_i(e)] - f_e \geq \epsilon t] \;=\; \Pr[X_{i,e} \geq \epsilon t] \qquad \text{[definition]}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned} \Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\ &\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \end{aligned}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned} \Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\ &\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\ &\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]} \end{aligned}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned}
\Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\
&\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\
&\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]}
\end{aligned}$$

Recall: $f'_e = \text{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned} \Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\ &\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\ &\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]} \end{aligned}$$

Recall: $f'_e = \mathrm{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

$$\Pr\left[f'_e - f_e \geq \epsilon t\right] =$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned} \Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\ &\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\ &\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]} \end{aligned}$$

Recall: $f'_e = \text{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

$$\Pr\left[f'_e - f_e \geq \epsilon t\right] = \Pr[C_i[h_i(e)] - f_e \geq \epsilon t \text{ for all } i]$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned}
\Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\
&\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\
&\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]}
\end{aligned}$$

Recall: $f'_e = \text{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

$$\begin{aligned}
\Pr\left[f'_e - f_e \geq \epsilon t\right] &= \Pr[C_i[h_i(e)] - f_e \geq \epsilon t \text{ for all } i] \\
&= \Pr[X_{i,e} \geq \epsilon t \text{ for all } i]
\end{aligned}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$
\begin{aligned}
\Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\
&\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\
&\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]}
\end{aligned}
$$

Recall: $f'_e = \text{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

$$
\begin{aligned}
\Pr\left[f'_e - f_e \geq \epsilon t\right] &= \Pr[C_i[h_i(e)] - f_e \geq \epsilon t \text{ for all } i] \\
&= \Pr[X_{i,e} \geq \epsilon t \text{ for all } i] \\
&= \Pi_{i=1}^{d} \Pr[X_{i,e} \geq \epsilon t] && \text{[independence of } h_i\text{'s]}
\end{aligned}
$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

We have $C_i[h_i(e)] = X_{i,e} + f_e$ and $\mathbf{E}[X_{i,e}] \leq \frac{t}{m}$.

Then, for $\epsilon > 0$

$$\begin{aligned}
\Pr[C_i[h_i(e)] - f_e \geq \epsilon t] &= \Pr[X_{i,e} \geq \epsilon t] && \text{[definition]} \\
&\leq \frac{\mathbf{E}[X_{i,e}]}{\epsilon t} && \text{[Markov's inequality]} \\
&\leq \frac{t/m}{\epsilon t} = \frac{1}{m\epsilon} && \text{[derived above]}
\end{aligned}$$

Recall: $f'_e = \mathrm{est}_t(e) = \min_{i=1}^{d} C_i[h_i(e)]$.

$$\begin{aligned}
\Pr\left[f'_e - f_e \geq \epsilon t\right] &= \Pr[C_i[h_i(e)] - f_e \geq \epsilon t \text{ for all } i] \\
&= \Pr[X_{i,e} \geq \epsilon t \text{ for all } i] \\
&= \Pi_{i=1}^{d} \Pr[X_{i,e} \geq \epsilon t] && \text{[independence of } h_i\text{'s]} \\
&\leq \left(\frac{1}{\epsilon m}\right)^{d} && \text{[derived above]}
\end{aligned}$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

$$\Pr[\text{est}_t(e) - \text{count}_t(e) \geq \epsilon t] \leq \left(\frac{1}{\epsilon m}\right)^d$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

$$\Pr[\text{est}_t(e) - \text{count}_t(e) \geq \epsilon t] \leq \left(\frac{1}{\epsilon m}\right)^d \leq \delta$$

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

$$\mathbf{Pr}[\mathrm{est}_t(e) - \mathrm{count}_t(e) \geq \epsilon t] \leq \left(\frac{1}{\epsilon m}\right)^d \leq \delta$$

Set $m = \lceil 2/\epsilon \rceil$ and $d = \lceil \lg 1/\delta \rceil$.

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

$$\mathbf{Pr}[\text{est}_t(e) - \text{count}_t(e) \geq \epsilon t] \leq \left(\frac{1}{\epsilon m}\right)^d \leq \delta$$

Set $m = \lceil 2/\epsilon \rceil$ and $d = \lceil \lg 1/\delta \rceil$.

Space:

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

$$\Pr[\text{est}_t(e) - \text{count}_t(e) \geq \epsilon t] \leq \left(\frac{1}{\epsilon m}\right)^d \leq \delta$$

Set $m = \lceil 2/\epsilon \rceil$ and $d = \lceil \lg 1/\delta \rceil$.

Space: $m * d$ counters each of size $\lg(t) = O(\frac{1}{\epsilon} \lg \frac{1}{\delta} \lg t)$ bits.

Count Min-Sketch: Analysis

By G. Cormode and S. M. Muthukrishnan'05

Lemma

Given $\epsilon, \delta > 0$*, we can estimate* $count_t(e)$*, at any time* t *for any element* e*, up to* ϵt *error with probability at least* $(1 - \delta)$ *using* $O(\frac{1}{\epsilon} \lg \frac{1}{\delta})$ *many counters.*