



UNIVERSITÄT
DES
SAARLANDES



max planck institut
informatik

Sublinear Algorithms

Lecture 01: Introduction & Streaming I



European Research Council
Established by the European Commission

Karl Bringmann

May 07, 2020



Etiquette

- These **slides** will be available on the course website
- The lecture is **recorded** and a video will be made available to all participants
- We start every meeting with everyone's (but mine) **mics and videos off**, to avoid noise and save bandwidth

- Questions:**
- can always be asked in the **chat**, I will keep an eye on the chat window
 - during **breaks** the recording is paused, so it is safe to turn on your video and ask questions
 - during recording, you can also unmute and ask questions, but be aware that you are recorded and the video will be made available to all participants

Organization

Advanced Lecture, 2+1, 5CP

Lecture: Karl Bringmann and Vasileios Nakos
every Thursday 16-18 (on holidays: move to Monday 16-18)
via Zoom + video download

Tutorial: Nick Fischer
every second Monday 16-18
via Zoom

Requirements: basic algorithms lecture, e.g., Grundzüge von Algorithmen und Datenstrukturen

Exam: oral exam
admittance by $\geq 50\%$ of points on 4 exercise sheets

<https://lists.mpi-inf.mpg.de/listinfo/sublinear>

<https://www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/summer20/sublinear-algorithms/>

Course Overview

NP: $O(2^n)$ $O(2^{\sqrt{n}})$ $O(n^n)$

P: $O(n^2)$ $O(n^{100})$
 $O(n \log n)$
 $O(n)$

Space $o(n)$?

#Measurements $o(n)$?

Time $o(n)$?

Course Overview

NP: $O(2^n)$ $O(2^{\sqrt{n}})$ $O(n^n)$

P: $O(n^2)$ $O(n^{100})$
 $O(n \log n)$
 $O(n)$

Space $o(n)$?

#Measurements $o(n)$?

Time $o(n)$?

Streaming Algorithms:

Data stream $x_1, x_2, \dots, x_n$

Make one pass over the stream

Working memory $o(n)/O(\log n)$

$\approx$ low-space data structures

Typical problems:

Compute number of distinct x_i 's

Compute the majority element (if exists)

Compute all numbers that appear $\geq \epsilon n$ times



©Stefan Funke / Wikipedia

Course Overview

NP: $O(2^n)$ $O(2^{\sqrt{n}})$ $O(n^n)$

P: $O(n^2)$ $O(n^{100})$
 $O(n \log n)$
 $O(n)$

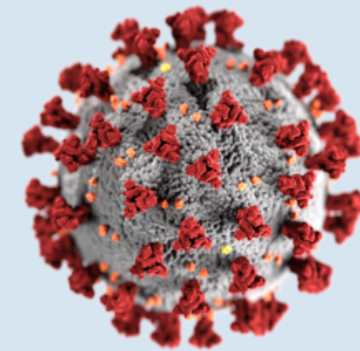
Space $o(n)$?

#Measurements $o(n)$?

Time $o(n)$?

Randomized Trials:

Estimate the infected population by testing random individuals

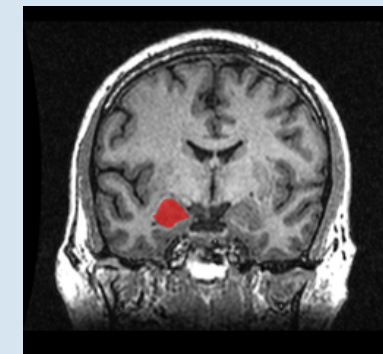


Combinatorial Group Testing:

Mix samples of a group of individuals → test tells us whether at least one individual is positive
Find *all* positive individuals using $o(n)$ group tests

Medical Imaging:

Reconstruct a sparse vector from few Fourier measurements



Course Overview

NP: $O(2^n)$ $O(2^{\sqrt{n}})$ $O(n^n)$

P: $O(n^2)$ $O(n^{100})$
 $O(n \log n)$
 $O(n)$

Space $o(n)$?

#Measurements $o(n)$?

Time $o(n)$?

Property Testing:

Really sublinear time $o(n)$!

“What can we find out about $x_1, x_2, \dots, x_n$
using $o(n)$ random accesses?”

Typical problems:

Is $x_1, x_2, \dots, x_n$ monotone or *far* from monotone?

Is a graph 2-colorable or *far* from 2-colorable?

Course Overview

NP: $O(2^n)$ $O(2^{\sqrt{n}})$ $O(n^n)$

P: $O(n^2)$ $O(n^{100})$
 $O(n \log n)$
 $O(n)$

Space $o(n)$?

#Measurements $o(n)$?

Time $o(n)$?

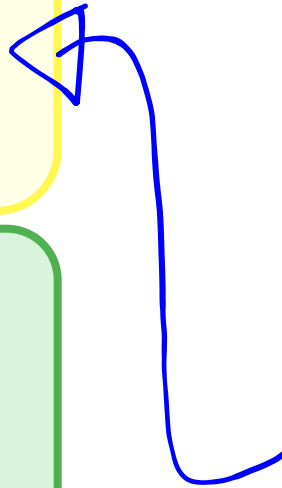
Course Outline:

3x Streaming (Space)

4x Vector Reconstruction (Measurements)

2x Property Testing (Time)

2x Applications



Outline

- 1) Course Overview
- 2) Basic Probability Theory**
- 3) Morris' Counter

Questions?

5min Break?

Basic Probability Theory

Course Website → Material → A Primer to Randomness

**Random
Variable:**

X is a random coin flip

$$\mathbb{P}[X = 0] = \mathbb{P}[X = 1] = \frac{1}{2}$$

$$\mathbb{E}[X] = \frac{1}{2}$$

$X = X_1 + X_2$, where
 X_1, X_2 are random coin flips

$$\mathbb{P}[X = 0] = \mathbb{P}[X = 2] = \frac{1}{4}$$

$$\text{and } \mathbb{P}[X = 1] = \frac{1}{2}$$

$$\mathbb{E}[X] = 1$$

Expectation:

$$\mathbb{E}[X] = \sum_n n \cdot \mathbb{P}[X = n]$$

**Linearity of
Expectation:**

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Event:

$$X = 0$$

$$X \leq 1$$

X is even

Union Bound:

$$\mathbb{P}[A \text{ or } B] \leq \mathbb{P}[A] + \mathbb{P}[B]$$

Concentration Inequalities

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

For any $t > 0$, assuming $X \geq 0$

Concentration Inequalities

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

For any $t > 0$, assuming $X \geq 0$

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

For any $t > 0$

Variance $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Concentration Inequalities

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

For any $t > 0$, assuming $X \geq 0$

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

For any $t > 0$

Variance $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

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

For any $t > 0$, assuming $X = X_1 + \dots + X_n$
with *independent* $X_1, \dots, X_n \in \{0,1\}$

for any values $x_1, \dots, x_n$:

$$\begin{aligned} \mathbb{P}[X_1 = x_1 \text{ and } \dots \text{ and } X_n = x_n] \\ = \mathbb{P}[X_1 = x_1] \cdot \dots \cdot \mathbb{P}[X_n = x_n] \end{aligned}$$

Outline

- 1) Course Overview
- 2) Basic Probability Theory
- 3) Morris' Counter**

Questions?

5min Break?

Counting

Most simple streaming problem

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

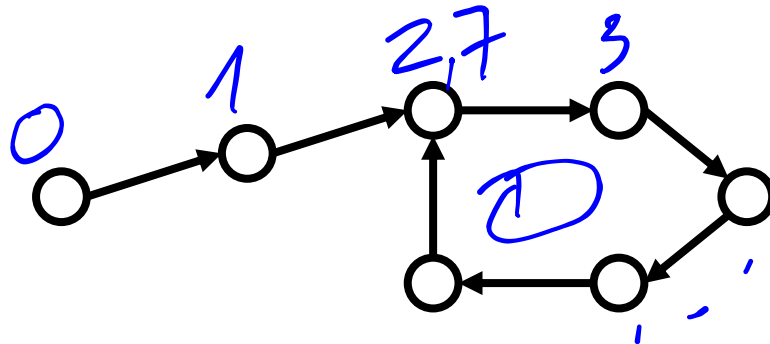
Data structure problem:

maintain a number n

update(): increment n by 1

query(): output n

initially $n = 0$



Solution: Standard Counter

store n using $\lceil \log n \rceil = O(\log n)$ bits

This is optimal: with $< \log n$ bits...

... we must make an error

→ need **approximation**

... we run into infinite loops

→ need **randomization**

Approximate Counting

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

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

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

In other words: $(1 - \varepsilon)n < \tilde{n} < (1 + \varepsilon)n$
with probability at least $1 - \delta$

$\varepsilon, \delta \in (0,1)$ are parameters given to the
algorithm upfront

Solution: Morris' Counter (1978)

1) Initialize $X = 0$

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

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

Intuition: store integer $X \approx \log n$

When to increment?

Increment with probability $\frac{1}{n} \approx 2^{-X}$

Approximate Counting

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

Proof: Consider one update in isolation

Let X, X' be the counter before/after the update

Morris' Counter

1) Initialize $X = 0$

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

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

$$\begin{aligned}\mathbb{E}[2^{X'} | X] &= \frac{1}{2^X} \cdot 2^{X+1} + \left(1 - \frac{1}{2^X}\right) \cdot 2^X \\ &= 2^X + 1\end{aligned}$$

express expectation of $2^{X'}$ in terms of X

$$\mathbb{E}[2^{X_n}] = \mathbb{E}[2^{X_{n-1}}] + 1$$

Thus, inductively after n updates we have: $\mathbb{E}[2^X] = n + 1$

Approximate Counting

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

$$\mathbb{P}[|\tilde{n} - n| \geq \varepsilon n] \leq \frac{\mathbb{E}[\tilde{n}^2] - n^2}{\varepsilon^2 n^2}$$

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

$$\begin{aligned}\mathbb{E}[4^{X'} | X] &= \frac{1}{2^X} \cdot 4^{X+1} + \left(1 - \frac{1}{2^X}\right) \cdot 4^X \\ &= 4^X + 3 \cdot 2^X\end{aligned}$$

$$\mathbb{E}[4^{X_n}] = \mathbb{E}[4^{X_{n-1}}] + 3 \cdot \mathbb{E}[2^{X_{n-1}}] = \mathbb{E}[4^{X_{n-1}}] + 3n = 3 \cdot \binom{n+1}{2}$$

Morris' Counter

1) Initialize $X = 0$

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

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

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

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

$$\mathbb{E}[\tilde{n}^2] = \mathbb{E}[(2^X - 1)^2] = \mathbb{E}[4^X] - 2\mathbb{E}[2^X] + 1$$

Approximate Counting

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

Lem: We have $\mathbb{E}[\tilde{n}^2] \stackrel{<}{=} \frac{3}{2}n^2 - \frac{1}{2}n$.

$$\mathbb{P}[|\tilde{n} - n| \geq \varepsilon n] \leq \frac{\frac{3}{2}n^2 - n^2}{\varepsilon^2 n^2} \leq \frac{1}{2\varepsilon^2}$$

No approximation guarantee for $\varepsilon \leq 0.7$!

Morris' Counter

1) Initialize $X = 0$

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

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

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

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

Questions?

5min Break?

Boosting via Chebyshev

Morris+: *average over s runs of Morris*

- 1) Run s independent copies of Morris' Counter
- 2) Let $\tilde{n}_1, \dots, \tilde{n}_s$ be their estimates
- 3) output $\tilde{n}^+ = \frac{1}{s} \cdot (\tilde{n}_1 + \dots + \tilde{n}_s)$ ||

Lem: For $s \geq 1/(2\varepsilon^2\delta)$ we have

$$\mathbb{P}[|\tilde{n}^+ - n| \geq \varepsilon n] \leq \underline{\delta} \quad ||$$

Proof: Morris+ is an unbiased estimator:

$$\underline{\mathbb{E}[\tilde{n}^+]} = \underline{\frac{1}{s}} \cdot (\mathbb{E}[\tilde{n}_1] + \dots + \mathbb{E}[\tilde{n}_s]) = \underline{n}$$

by linearity of expectation

Morris' Counter computes estimate $\tilde{n}$ s.t.

$$\mathbb{E}[\tilde{n}] = n \text{ (unbiased estimator)}$$

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

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

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

Boosting via Chebyshev

Morris+: *average over s runs of Morris*

- 1) Run s independent copies of Morris' Counter
- 2) Let $\tilde{n}_1, \dots, \tilde{n}_s$ be their estimates
- 3) output $\tilde{n}^+ = \frac{1}{s} \cdot (\tilde{n}_1 + \dots + \tilde{n}_s)$

Lem: For $s \geq 1/(2\varepsilon^2\delta)$ we have
$$\mathbb{P}[|\tilde{n}^+ - n| \geq \varepsilon n] \leq \delta$$

Proof: Morris+ is an unbiased estimator

Variance of Morris+ is $\frac{\text{Var}[\tilde{n}]}{s}$:

$$\text{Var}[\tilde{n}^+] = \frac{1}{s^2} \cdot (\text{Var}[\tilde{n}_1] + \dots + \text{Var}[\tilde{n}_s]) = \frac{\text{Var}[\tilde{n}]}{s}$$

Handwritten note: Var[ñ] with arrows pointing to Var[ñ₁] and Var[ñₛ]

- 1) $\text{Var}[\alpha X] = \alpha^2 \text{Var}[X]$
- 2) $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ for independent X, Y

Morris' Counter computes estimate $\tilde{n}$ s.t.

$$\mathbb{E}[\tilde{n}] = n \text{ (unbiased estimator)}$$

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

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

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

Boosting via Chebyshev

Morris+: *average over s runs of Morris*

- 1) Run s independent copies of Morris' Counter
- 2) Let $\tilde{n}_1, \dots, \tilde{n}_s$ be their estimates
- 3) output $\tilde{n}^+ = \frac{1}{s} \cdot (\tilde{n}_1 + \dots + \tilde{n}_s)$

Lem: For $s \geq \underline{1/(2\varepsilon^2\delta)}$ we have
 $\mathbb{P}[|\tilde{n}^+ - n| \geq \varepsilon n] \leq \delta$

Proof: Morris+ is an unbiased estimator

Variance of Morris+ is $\frac{\text{Var}[\tilde{n}]}{s}$

Morris' Counter computes estimate $\tilde{n}$ s.t.

$$\mathbb{E}[\tilde{n}] = n \text{ (unbiased estimator)}$$

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

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

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

So by Chebyshev:

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

Approximate Counting

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

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 $s = \frac{1}{2\varepsilon^2\delta}$ runs of Morris

once a counter reaches $X = \log\left(\frac{ns}{\delta}\right)$,
it is incremented with probability at most $\frac{\delta}{ns}$, so
by union bound no such counter is ever incremented

Space usage:

$O\left(\frac{1}{\varepsilon^2\delta} \log \log \left(\frac{n}{\varepsilon^2\delta^2}\right)\right)$ bits
with probability at least $1 - \delta$

Boosting via Chernoff

Lem: For $t \geq 8 \log(2/\delta)$ we have

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

Proof:

Each run of Morris+ succeeds with prob. $\geq \frac{3}{4}$

$$Y_i = \begin{cases} 1 & \text{if } i\text{-th run of Morris+ succeeds} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[Y_i] = \mathbb{P}[Y_i = 1] \geq \frac{3}{4}$$

$$Y = Y_1 + \dots + Y_t \quad \mathbb{E}[Y] \geq \frac{3}{4}t$$

$$\mathbb{P}[\text{Morris++ fails}] \leq \mathbb{P}\left[Y \leq \frac{t}{2}\right]$$

$$\leq \mathbb{P}\left[Y \leq \mathbb{E}[Y] - \frac{t}{4}\right] \leq 2 \exp\left(-\frac{2t^2}{16t}\right) \leq \delta$$

Morris+: average over s runs of Morris, then

$$\mathbb{P}[|\tilde{n}^+ - n| \geq \varepsilon n] \leq \frac{1}{2s\varepsilon^2}$$

$$\ell = t/4, n = t$$

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

for $X = X_1 + \dots + X_n$, independent $X_1, \dots, X_n \in \{0,1\}$

Morris++: median over t runs of Morris+

$$\delta = \frac{1}{4}$$

1) Run t copies of Morris+ with $s := \lceil 2/\varepsilon^2 \rceil$

2) output **median** of their estimates $\tilde{n}_1^+, \dots, \tilde{n}_t^+$
(that is, sort and pick the middle value)

Approximate Counting

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

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 $1 - \delta$

Approximate Counting

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

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

We learned about:

- probability basics
- concentration inequalities
- unbiased estimators
- boosting via Chebyshev
- boosting via Chernoff

→ *Primer to Randomness*

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+

More Material

- These slides will be available on the course website
- Video recording will be made available
- Course Website → Material → A Primer to Randomness
- Course Website → Material → Link to Summer School on Streaming by Jelani Nelson
- **Presence Exercise Sheet:** Will be send out in the next couple of days, Tutorial on May 11

See you next week!

EXTRA