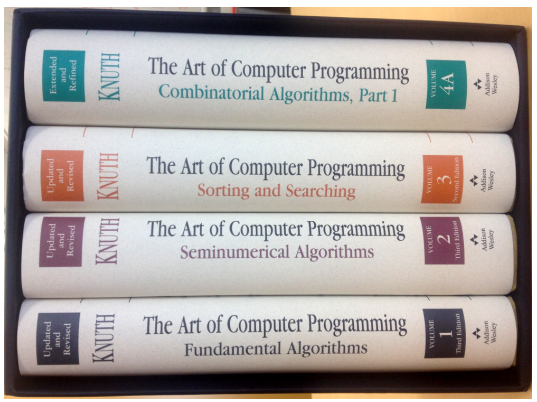
Ewens-like distributions and Analysis of Algorithms

Nicolas Auger, Mathilde Bouvel, Cyril Nicaud, Carine Pivoteau

July 6, 2016

Notion of presortedness

- In practice, data are often **presorted**.
  - No reasons to be uniformly distributed.
  - Few alterations in databases.
- First intuition in [Knuth73] and formalized in [Mannila86].



MEASURES OF PRESORTEDNESS AND OPTIMAL SORTING ALGORITHMS

Extended abstract

Heikki Mannila
Department of Computer Science, University of Helsinki
Tukholmankatu 2, SF-00250 Helsinki 25, Finland

Abstract

The concept of presortedness and its use in sorting are studied. Natural ways to measure presortedness are given and some general properties necessary for a measure are proposed. A concept of a sorting algorithm optimal with respect to a measure of presortedness is defined, and examples of such algorithms are given. An insertion sort is shown to be optimal with respect to three natural measures. The problem of finding an optimal algorithm for an arbitrary measure is studied and partial results are proven.

- In practice :
  - Used in standard libraries
  - Java's developers benchmarks, using spies
  - TimSort

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(4153682 7) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Measures of presortedness

Definition

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences of elements from a set E; $m : E^+ \to \mathbb{Z}^+$ is a **measure of presortedness** iff

1. $m(X) = 0$ if X is sorted.
2. If $n = \ell$ and $x_i < x_j \iff y_i < y_j$, then $m(X) = m(Y)$.
3. If Y is a subsequence of X, then $m(Y) \leq m(X)$.
4. If $X < Y$, then $m(XY) \leq m(X) + m(Y)$.
5. For any element a, $m(aX) \leq |X| + m(X)$.

Two classical measures :

- number of Runs -1, $Runs(4\,15\,368\,27) = 4$
- number of Inversions, $Inv(41536827) = 9$

Adaptiveness of sorting algorithms

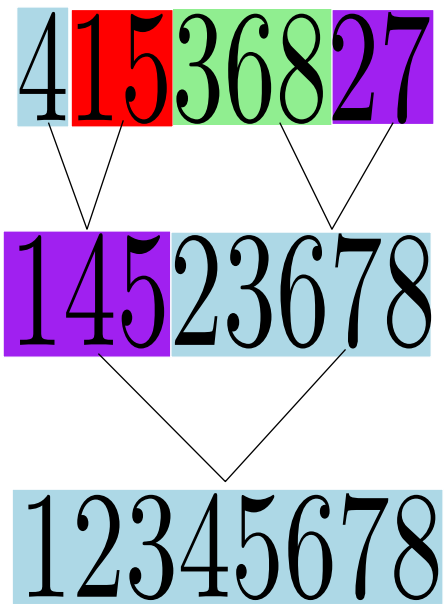
Theorem

Let X be a sequence s.t. $m(X) = k$. Any algorithm uses at least $C(n,k)$ comparisons to sort X, with $C(n,k) \in \Theta(n + \log(\|below_m(n,k)\|)$ and $below_m(n,k) = \{\sigma \in \mathfrak{S}_n : m(\sigma) \leq k\}$.

Definition

A sorting algorithm is **m-optimal** if it reaches this bound.

- Natural Merge Sort [Knuth73]
- $\mathcal{O}(n \log r)$, where r is the number of runs
- Runs-optimal



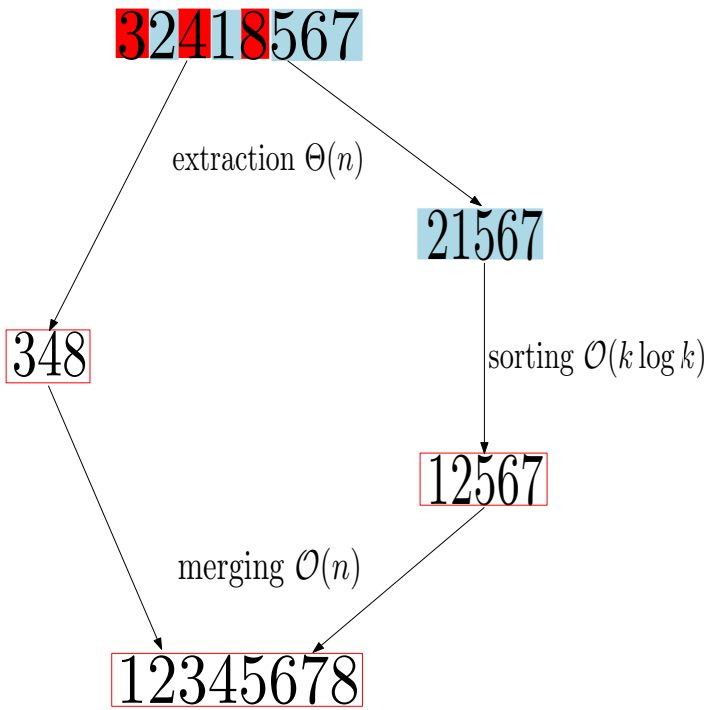
Records as a measure of presortedness

Let $X = (x_1, \ldots, x_n)$ be a sequence; x_i is a **record**(left-to-right maximum) iff $x_j < x_i$ whenever $j < i$.

Lemma

For any sequence X of size n, $m_{rec}(X) = n - \text{record}(X)$ is a measure of presortedness.

Example : For $X = 32418567$, $\text{record}(X) = 3$ and $m_{rec}(X) = 5$.



Complexity $\mathcal{O}(n + k \log k)$

$\|below_{m_{rec}}(n, k)\| \geq k!$

Analysis of algorithms on average

Under the uniform distribution, for most measures m :

- $\|below_m(n, \mathbb{E}[m])\| = \Theta(n!)$.
- $\mathcal{O}(n \log n)$ in average.

Questions

- How to define a probabilistic framework well-suited for presortedness measures ?
- Analysis of algorithms ?

The classical Ewens distribution

Any permutation can be seen as a composition of cycles.
Example : 145263 is composed of 3 cycles : (1), (563) and (42).

We denote cycle(σ) the number of cycles of σ.

Definition (Ewens distribution)

[Ewens72]

- To any $\sigma \in \mathfrak{S}_n$, we associate a weight $\mathrm{w}(\sigma) = \theta^{\mathrm{cycle}(\sigma)}$, where θ is an arbitrary positive real number.
- Total weight : $\sum_{\sigma \in \mathfrak{S}_n} \mathrm{w}(\sigma) = \theta^{(n)}$.
- $\mathbb{P}(\sigma) = \frac{\theta^{\mathrm{cycle}(\sigma)}}{\theta^{(n)}}$.

Notation : $\theta^{(n)} = \theta(\theta+1)\dots(\theta+n-1)$

Generalizing the distribution

Definition (Ewens-like distribution)

- Let χ be any statistic on $\sigma \in \mathfrak{S}_n$.
- To any $\sigma \in \mathfrak{S}_n$, we associate a weight $\mathrm{w}(\sigma) = \theta^{\chi(\sigma)}$.
- Let $\mathrm{W}_n = \sum_{\sigma \in \mathfrak{S}_n} \mathrm{w}(\sigma)$ and $\mathbb{P}(\sigma) = \frac{\mathrm{w}(\sigma)}{\mathrm{W}_n}$.

Generalizing the distribution

Definition (Ewens-like distribution)

- Let χ be any statistic on $\sigma \in \mathfrak{S}_n$.
- To any $\sigma \in \mathfrak{S}_n$, we associate a weight $\mathrm{w}(\sigma) = \theta^{\chi(\sigma)}$.
- Let $\mathrm{W}_n = \sum_{\sigma \in \mathfrak{S}_n} \mathrm{w}(\sigma)$ and $\mathbb{P}(\sigma) = \frac{\mathrm{w}(\sigma)}{\mathrm{W}_n}$.

Analytic combinatorics

Let $F(z,u) = \sum f_{n,k} z^n u^k$, where $f_{n,k} = \|\{\sigma \in \mathfrak{S}_n : \chi(\sigma) = k\}\|$.

$$W_n = n![z^n]F(z,\theta) \qquad \text{and} \qquad \mathbb{E}_n[\chi] = \frac{\theta[z^n]\left.\frac{\mathrm{d}F(z,u)}{\mathrm{d}u}\right|_{u=\theta}}{[z^n]F(z,\theta)}$$

But can be difficult when θ depends on n.

Ewens-like distributions for records

Recall

For any sequence X of size n, $m_{rec}(X) = n - \text{record}(X)$ is a measure of presortedness.

Definition (Ewens-like distribution for records)

- To any $\sigma \in \mathfrak{S}_n$, we associate a weight $\mathrm{w}(\sigma) = \theta^{\text{record}(\sigma)}$.
- Let $\mathrm{W}_n = \sum_{\sigma \in \mathfrak{S}_n} \mathrm{w}(\sigma) = \theta^{(n)}$ and $\mathbb{P}(\sigma) = \frac{\theta^{\text{record}(\sigma)}}{\theta^{(n)}}$.

In the following, we focus on this distribution.

Some probabilities

Results

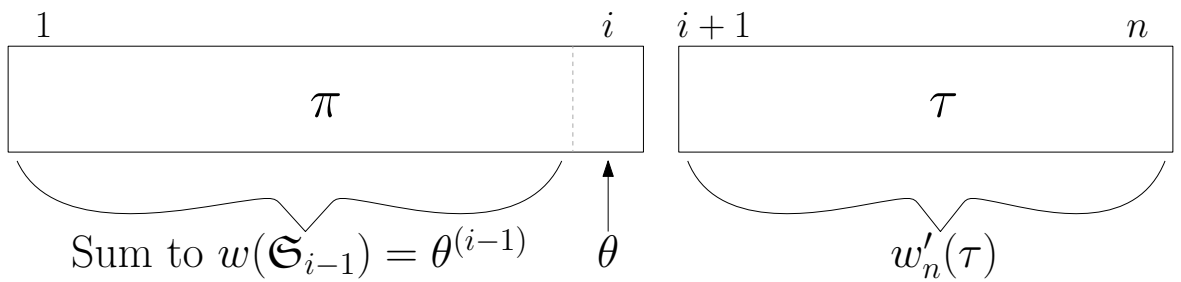
$\mathbb{P}_n(\text{Record at position } i)$	$\frac{\theta}{\theta+i-1}$
$\mathbb{P}_n(\sigma(i-1) > \sigma(i))$	$\frac{(i-1)(2\theta+i-2)}{2(\theta+i-1)(\theta+i-2)}$
$\mathbb{P}_n(\sigma(1) = k)$	$\frac{(n-1)!\theta^{(n-k)}\theta}{(n-k)!\theta^{(n)}}$
$\mathbb{P}_n(inv_j(\sigma) = k)$	$\begin{cases} \frac{\theta}{\theta+j-1} & \text{if } k = 0 \\ \frac{1}{\theta+j-1} & \text{otherwise} \end{cases}$

Some probabilities

Results

$\mathbb{P}_n(\text{Record at position } i)$	$\frac{\theta}{\theta+i-1}$
$\mathbb{P}_n(\sigma(i-1) > \sigma(i))$	$\frac{(i-1)(2\theta+i-2)}{2(\theta+i-1)(\theta+i-2)}$
$\mathbb{P}_n(\sigma(1) = k)$	$\frac{(n-1)!\theta^{(n-k)}\theta}{(n-k)!\theta^{(n)}}$
$\mathbb{P}_n(inv_j(\sigma) = k)$	$\begin{cases} \frac{\theta}{\theta+j-1} & \text{if } k = 0 \\ \frac{1}{\theta+j-1} & \text{otherwise} \end{cases}$

$$\mathbb{P}_n(\text{Record at position } i) = \frac{\theta^{(i-1)}\theta}{\theta^{(i)}} = \frac{\theta}{\theta+i-1}$$



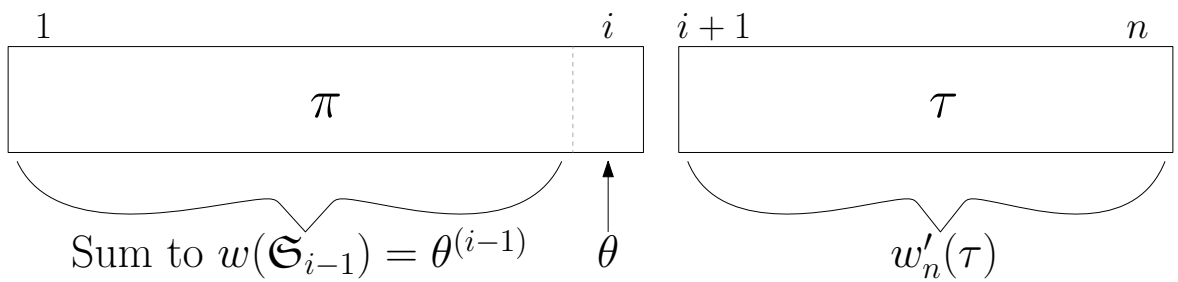
Asymptotic equivalents

Results

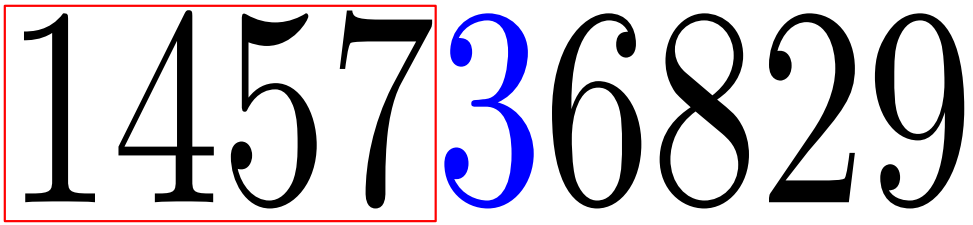
	$\theta = 1$ (uniform)	fixed $\theta > 0$	$\theta := n^\epsilon$, $0 < \epsilon < 1$	$\theta := \lambda n$, $\lambda > 0$	$\theta := n^\delta$ $\delta > 1$
$\mathbb{E}_n$[record]	$\log n$	$\theta \cdot \log n$	$(1-\epsilon) \cdot n^\epsilon \log n$	$\lambda \log(1 + 1/\lambda) \cdot n$	n
$\mathbb{E}_n$[desc]	$n/2$	$n/2$	$n/2$	$n/2(\lambda + 1)$	$n^{2-\delta}/2$
$\mathbb{E}_n[\sigma(1)]$	$n/2$	$n/(\theta + 1)$	$n^{1-\epsilon}$	$(\lambda + 1)/\lambda$	1
$\mathbb{E}_n$[inv]	$n^2/4$	$n^2/4$	$n^2/4$	$n^2/4 \cdot f(\lambda)$	$n^{3-\delta}/6$

With $f(\lambda) = 1 - 2\lambda + 2\lambda^2 \log(1 + 1/\lambda)$.

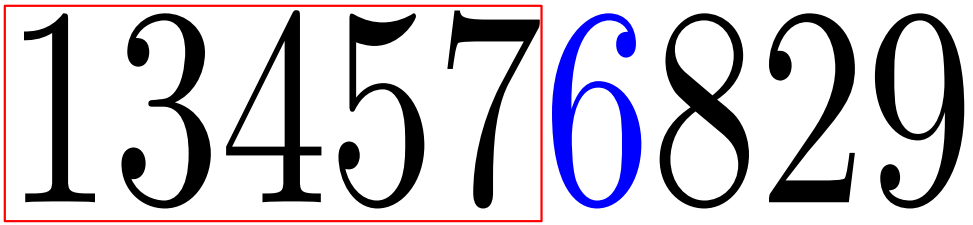
$$\mathbb{P}_n(\text{Record at position } i) = \frac{\theta^{(i-1)}\theta}{\theta^{(i)}} = \frac{\theta}{\theta + i - 1}$$



InsertSort

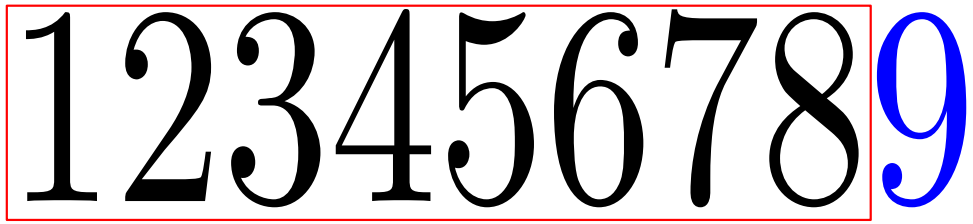


InsertSort



InsertSort

InsertSort


123456789

InsertSort

- Adapts to the number of *inversions*.
- Sorts a sequence X in $\Theta(Inv(X))$ comparisons.

Recall

	$\theta = 1$ (uniform)	fixed $\theta > 0$	$\theta := n^\epsilon$, $0 < \epsilon < 1$	$\theta := \lambda n$, $\lambda > 0$	$\theta := n^\delta$ $\delta > 1$
$\mathbb{E}_n[\text{inv}]$	$n^2/4$	$n^2/4$	$n^2/4$	$n^2/4 \cdot f(\lambda)$	$n^{3-\delta}/6$

With $f(\lambda) = 1 - 2\lambda + 2\lambda^2 \log(1 + 1/\lambda)$.

Unless $\theta \gg n$, InsertSort remains in $\Theta(n^2)$ on average.

Introduction to min/max search

naiveMinMax(T, n)

```
min ← T[1]
max ← T[1]
for i ← 2 to n do
    if T[i] < min do
        min ← T[i]
    if T[i] > max do
        max ← T[i]
return min, max
```

$2n$ comparisons

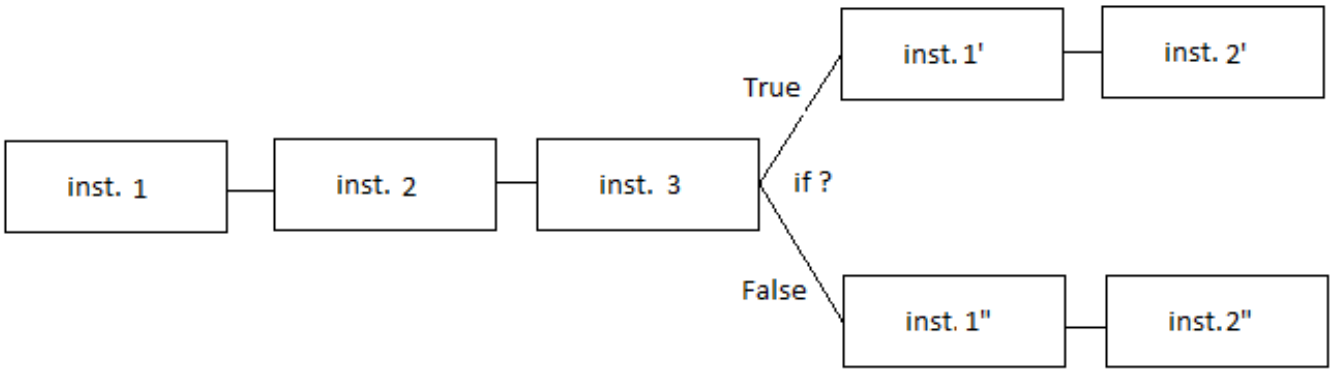
$3/2$-MinMax(T, n)

```
min, max ← T[n], T[n]
for i ← 2 to n by 2 do
    if T[i − 1] < T[i] do
        pMin, pMax ← T[i − 1], T[i]
    else
        pMin, pMax ← T[i], T[i − 1]
    if pMin < min do min ← pMin
    if pMax > max do max ← pMax
return min, max
```

$3n/2$ comparisons

naiveMinMax is faster than $3/2$-MinMax, when the data are uniformly distributed in $[0, 1]$. [Auger,Nicaud,Pivoteau,STACS 2016]

Introduction to modern architecture



Prediction rule : for each if, same as last time

Analysis under the uniform distribution

naiveMinMax(T, n)

```
min ← T[1]
max ← T[1]
for i ← 2 to n do
    if T[i] < min do
        min ← T[i]
    if T[i] > max do
        max ← T[i]
return min, max
```

3/2-MinMax(T, n)

```
min, max ← T[n], T[n]
for i ← 2 to n by 2 do
    if T[i − 1] < T[i] do
        pMin, pMax ← T[i − 1], T[i]
    else
        pMin, pMax ← T[i], T[i − 1]
    if pMin < min do min ← pMin
    if pMax > max do max ← pMax
return min, max
```

misprediction = **alternation** of record and non-record

Results

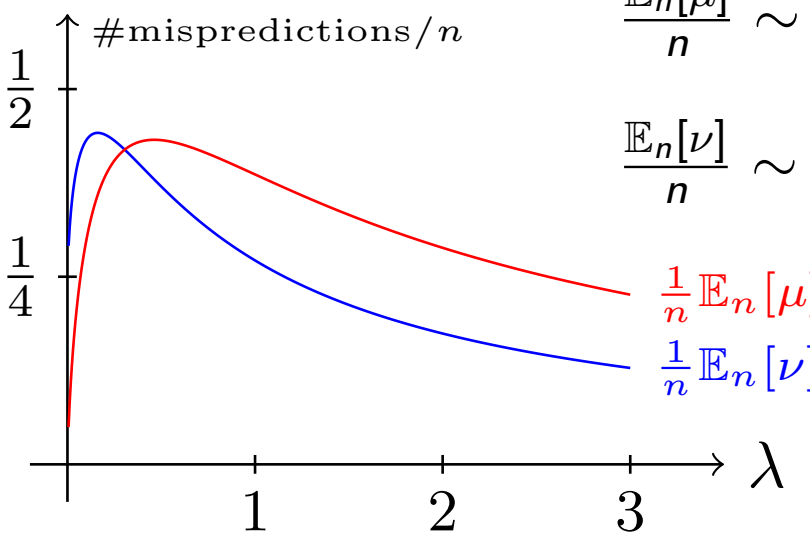
- naiveMinMax generates $\Theta(\log(n))$ mispredictions.
- 3/2-MinMax generates $\Theta(n)$ mispredictions.

What happen if the number of records **increases significantly** ?

Average analysis of the number of mispredictions

When $\theta = \lambda n$ for some real λ and for our prediction rule, we have :

- μ number of mispredictions of naiveMinMax.
- ν number of misprediction of $3/2$-MinMax.



$$\frac{\mathbb{E}_n[\mu]}{n} \sim 2\lambda\left(\log(1+\tfrac{1}{\lambda}) - \frac{1}{(\lambda+1)}\right)$$

$$\frac{\mathbb{E}_n[\nu]}{n} \sim \left(2\lambda\log\left(1+\tfrac{1}{\lambda}\right) - \frac{24\lambda^3+54\lambda^2+32\lambda-3}{12(\lambda+1)^3}\right)$$

$\frac{1}{n}\mathbb{E}_n[\mu]$ naiveMinMax

$\frac{1}{n}\mathbb{E}_n[\nu]$ $3/2$-minMax

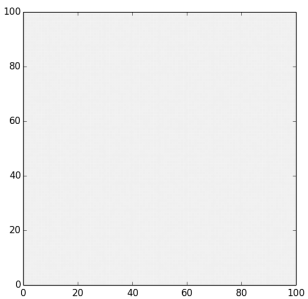
Discussion

Questions

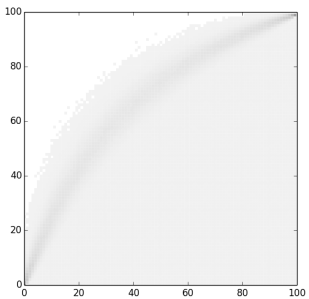
What's next ?

- Ewens-like distribution for other meaningful statistics that take part in (sorting) algorithms.
- For example, the runs for the analysis of TimSort.
- Explain the asymptotic shape of the diagrams below.(done)

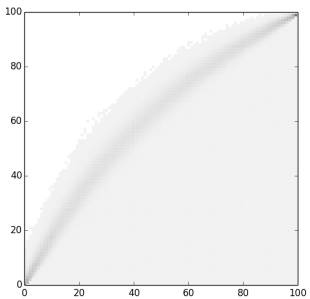
$n = 100$ sample size $= 10000$



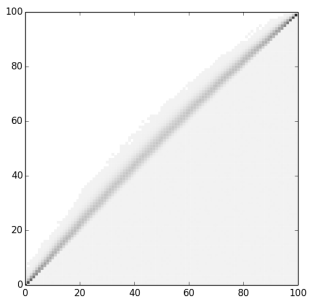
$\theta = 1$



$\theta = 50$



$\theta = 100$



$\theta = 500$