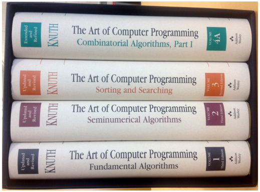
# Ewens-like distributions and Analysis of Algorithms

**Nicolas Auger**, Mathilde Bouvel, Cyril Nicaud, Carine Pivoteau

March 9, 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 python™
  - Oracle's benchmarks, using spies
  - TimSort

# Measures of presortedness

**Definition**

Let $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_\ell)$ two sequences; m 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; m 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; m 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; m 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; m 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; m 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; m 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; m 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; m 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

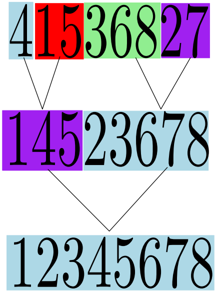
**Theorem**

*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** 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 $m_{rec}(X) = 5$.

**Proof.**

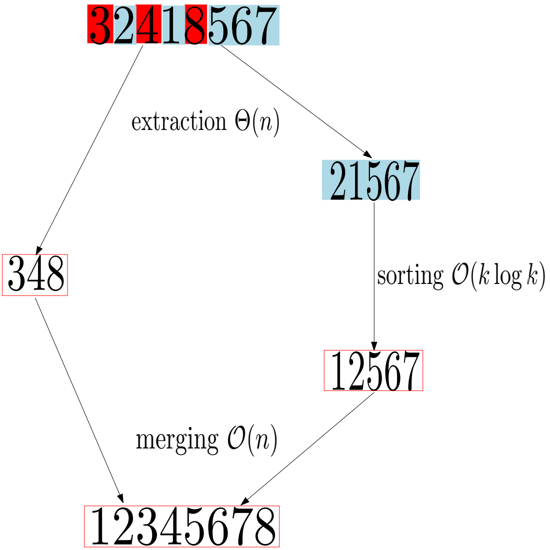
If $Y$ is a subsequence of $X$, then $m_{rec}(Y) \leq m_{rec}(X)$. Two cases :

- Remove a non-record (if we remove 2, $Y = 3418567$, $rec(Y) = 3$ and $m_{rec} = 4$).
- Remove a record (if we remove 8, $Y = 3241567$, $rec(Y) = 5$ and $m_{rec}(Y) = 2$).

The other properties are trivial.

□

# A $m_{rec}$-optimal sorting algorithm




$\|below_{m_{rec}}(n,k)\| \geq k!$ Overall complexity $\mathcal{O}(n + k \log k)$

# 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 $\mathrm{cycle}(\sigma)$ the number of cycles of $\sigma$.

**Definition (Ewens distribution)**

[Ewens72]

- To any $\sigma \in \mathfrak{S}_n$, we associate a weight $\mathrm{w}(\sigma) = \theta^{\mathrm{cycle}(\sigma)}$, where $\theta$ 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 $\chi$ 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 $\chi$ 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 difficult when $\theta$ depends on $n$.

# Ewens-like distributions for records

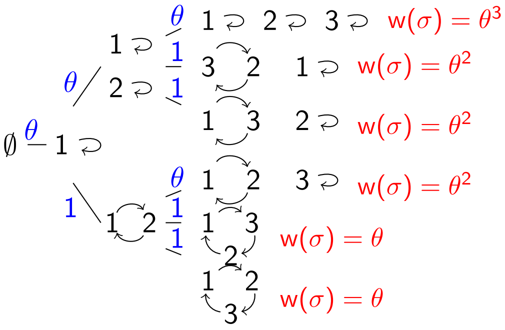
**Recall**

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.




[Ferray2014]

- Generation, in $\mathcal{O}(n)$, following one path in the tree.
- Keep $\sigma$ and $\sigma^{-1}$.
- Choosing a position in a cycle in $\mathcal{O}(1)$.
- Insertion in $\mathcal{O}(1)$.

Sampler for records in $\mathcal{O}(n)$ :

- Fundamental bijection : $\mathbf{145263} \to (\mathbf{1})(\mathbf{6}35)(\mathbf{4}2) \to \mathbf{14}2\mathbf{6}35$.
- Records are already sorted and we read $\sigma^{-1}$ in reverse order.

# Asymptotic equivalents

## Results

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

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

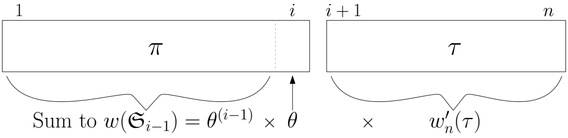
# Asymptotic equivalents

## Results

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

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

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



# InsertSort

$\boxed{1457}368 29$

# InsertSort

$\boxed{13457}\color{blue}{6}\color{black}{829}$

# InsertSort

$$\boxed{1345678}\color{blue}{8}\color{black}{29}$$

$\boxed{13456782}9$

# InsertSort

$\boxed{12345678}9$

# InsertSort

$$\boxed{123456789}$$

# InsertSort

123456789

- 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.

$\textsc{naiveMinMax}(T, n)$

$min \leftarrow T[1]$
$max \leftarrow T[1]$
**for** $i \leftarrow 2$ **to** $n$ **do**
    **if** $T[i] < min$ **do**
        $min \leftarrow T[i]$
    **if** $T[i] > max$ **do**
        $max \leftarrow T[i]$
**return** $min, max$

$2n$ comparisons

$3/2\text{-}\textsc{MinMax}(T, n)$

$min, max \leftarrow T[n], T[n]$
**for** $i \leftarrow 2$ **to** $n$ **by** 2 **do**
    **if** $T[i-1] < T[i]$ **do**
        $pMin, pMax \leftarrow T[i-1], T[i]$
    **else**
        $pMin, pMax \leftarrow T[i], T[i-1]$
    **if** $pMin < min$ **do** $min \leftarrow pMin$
    **if** $pMax > max$ **do** $max \leftarrow pMax$
**return** $min, max$

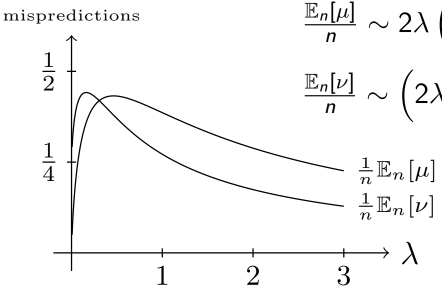
$3n/2$ comparisons

In practice, $\textsc{naiveMinMax}$ is faster than $3/2\text{-}\textsc{MinMax}$, when the data are uniformly distributed in $[0, 1]$.

# Average analysis of the number of mispredictions

When $\theta = \lambda n$ for some real $\lambda$ and for a 1-bit predictor, we have :

- $\mu$ number of mispredictions of naiveMinMax.
- $\nu$ number of misprediction of $3/2$-MinMax.




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

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

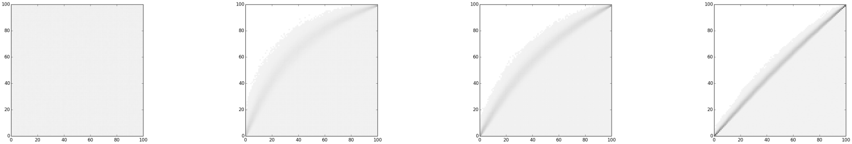
# Discussion

## Questions

What's next ?

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

$n = 100$ sample size $= 10000$



$\theta = 1$ $\theta = 50$ $\theta = 100$ $\theta = 500$