Non-uniform permutations biased according to their records

Mathilde Bouvel
(Loria, CNRS, Univ. Lorraine)

talk based on joint work and work in progress with
Nicolas Auger, Cyril Nicaud and Carine Pivoteau

On-line conference *Analysis of Algorithms* 2021

Non-uniform permutations

Context:

Analysis of algorithms working on arrays of numbers (sorting, . . .)

Average-case analysis of algorithms:

- The uniform distribution on the data set is usually assumed.
- It provides a first answer, but it is not always realistic.
  E.g., sorting algorithms are often used on data which is already "almost sorted". (Ex. of TimSort, wait 30 minutes to know more!)

$\Rightarrow$ Find non-uniform models with good balance between simplicity (so that we can study it) and accuracy (in terms of modeling data)

Some classical models for non-uniform permutations

- Ewens: $\mathbb{P}(\sigma)$ is proportional to $\theta^{\text{number of cycles of } \sigma}$
- Mallows: $\mathbb{P}(\sigma)$ is proportional to $\theta^{\text{number of inversions of } \sigma}$

Our record-biased permutations

Goal: A non-uniform distribution on permutations, which gives higher probabilities to permutations that are "almost sorted".

Record-biased permutations:

- A record is an element larger than all those preceding it.
  Example: **3** **4** 1 2 **6** **8** 7 **9** 5 has 5 records.
- Roughly, a permutation with many records is "almost sorted". More formally, the number of non-records is a measure of presortedness.
- In our model, $\mathbb{P}(\sigma)$ is proportional to $\theta^{\text{number of records of } \sigma}$.
- We focus on the regime where $\theta = \lambda \cdot n$, n being the size of σ.

Remark: Related to the Ewens distribution via Foata's *fundamental bijection*, which sends number of cycles to number of records.
Example: $2\,4\,3\,1\,9\,6\,8\,7\,5 = (3)(4\,1\,2)(6)(8\,7)(9\,5) \to \mathbf{3}\,\mathbf{4}\,1\,2\,\mathbf{6}\,\mathbf{8}\,7\,\mathbf{9}\,5$

Outline of the talk

Goal: Describe properties of the model of record-biased permutations.
Applications to the analysis of algorithms won't be discussed so much.

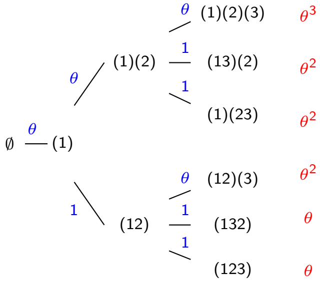
Results obtained:

- Random sampling can be done in linear time.
- Behavior of classical permutation statistics:
  - We obtain their expected values and precise probabilities.
    Applications to analysis of algorithms were presented at AofA 2016:
      - expected running time of InsertionSort,
      - expected number of mispredictions in MinMaxSearch
  - We plan to study their distribution.
- What does a large record-biased permutation typically look like?
  - We describe the (deterministic) permuton limit for our model.

Linear random samplers

Linear-time random samplers

- Ewens-distributed permutations can be sampled in linear time using a variant of the Chinese restaurant process:
  - Insert i from 1 to n.
  - At step i, create a new cycle (i) with probability $\frac{\theta}{\theta+i-1}$, or insert i in an existing cycle, immediately after a previously inserted element, each with probability $\frac{1}{\theta+i-1}$.



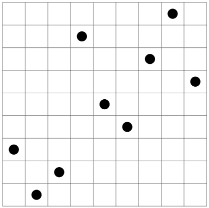
- Using appropriate data structures, we can implement Foata's transform in linear time,
  hence sampling record-biased permutation in linear time.

- We can also do it directly, with appropriate data structures.

Playing with the sampler: a typical diagram arises

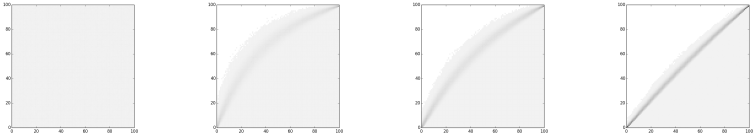
The diagram of a permutation σ of size n is the set of points at coordinates $(i, \sigma(i))$ for $1 \leq i \leq n$.

The normalized diagram of σ is the same picture, rescaled to the unit square.



$\sigma = 3\ 1\ 2\ 8\ 5\ 4\ 7\ 9\ 6$

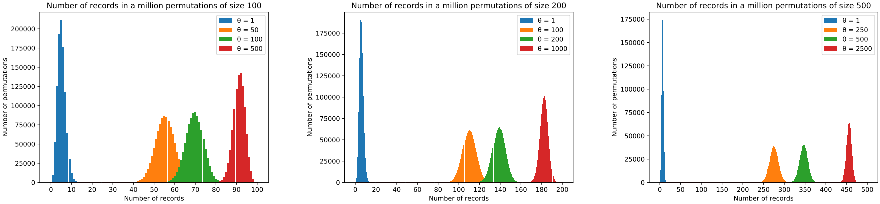
Pictures obtained overlapping 10 000 permutations of size 100 sampled according to the record-biased model with $\theta = 1, 50, 100$ and 500:



We explain it by describing the permuton limit of record-biased permutations.

A record of a permutation σ is given by an index i such that $\sigma(i) > \sigma(j)$ for all $j < i$.

Empirical distribution of the number of records
in record-biased permutations:



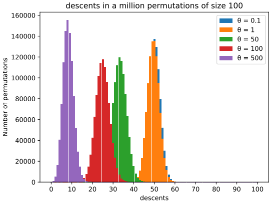
Histograms for 10^6 permutations, of size $n = 100, 200$ and 500,
and for $\theta = 1, \frac{n}{2}, n$ and $5n$.

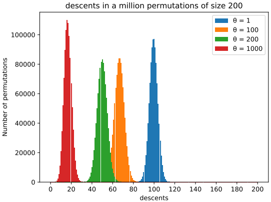
Remark: Corresponds to number of cycles for Ewens distribution, known to be Gaussian for fixed θ.

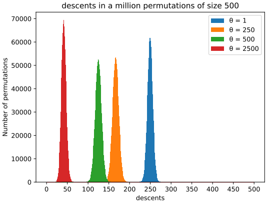
Playing with the sampler: number of descents

A descent of a permutation σ is given by an index i s.t. $\sigma(i-1) > \sigma(i)$.

Empirical distribution of the number of descents
in record-biased permutations:





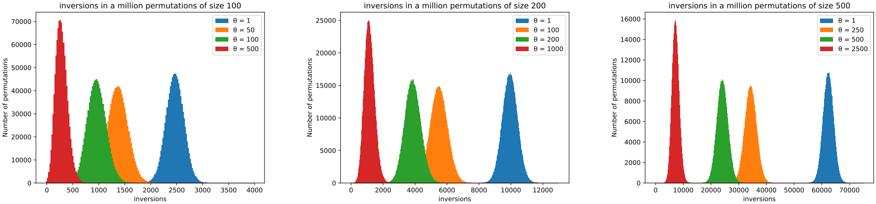


Histograms for 10^6 permutations, of size $n = 100, 200$ and 500,
and for $\theta = 1, \frac{n}{2}, n$ and $5n$.

Remark: Corresponds to number of anti-exceedances (given by i s.t. $\sigma(i) < i$) for Ewens distribution, which can be proved Gaussian for fixed θ.

An inversion of σ is given by a pair i, j s.t. $i < j$ and $\sigma(i) > \sigma(j)$.

Empirical distribution of the number of inversions
in record-biased permutations:

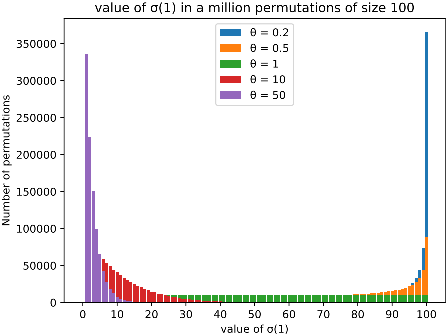


Histograms for 10^6 permutations, of size $n = 100, 200$ and 500,
and for $\theta = 1, \frac{n}{2}, n$ and $5n$.

Remark: No known natural analogue for Ewens distribution

Playing with the sampler: first value

Empirical distribution of the first value $\sigma(1)$ in record-biased permutations:



Histogram for 10^6 permutations, of size $n = 100$, and for $\theta = 0.2, 0.5, 1, 10$ and 50.

Remark: Corresponds to the minimum over all cycles of the maximal value in a cycle for Ewens distribution

Behavior of classical statistics

Example: the number of descents [from AofA 2016]

Proposition: In record-biased permutations of size n, for any $i \in \{2, \dots, n\}$, the probability that there is a descent at position i is $\mathbb{P}\big(\sigma(i-1) > \sigma(i)\big) = \frac{(i-1)(2\theta+i-2)}{2(\theta+i-1)(\theta+i-2)}$.

Corollary: In record-biased permutations of size n, the expected value of the number of descents is $\mathbb{E}[\text{number of descents}] = \frac{n(n-1)}{2(\theta+n-1)}$.

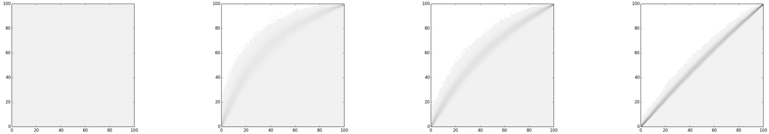
In particular, when $\theta = \lambda n$, $\mathbb{E}[\text{number of descents}] \sim n/2(\lambda+1)$.

Application (using the precise probabilities given by the proposition)**:** Average number of mispredictions in algorithms solving MinMaxSearch.

Question: Can we say more than just the expectation when $\theta = \lambda n$? Can we find the limiting distribution?

More examples: Similar statements for number of records, number of inversions and first value, with applications to the analysis of InsertionSort and MinMaxSearch.

Permuton limit
of record-biased permutations

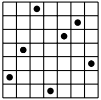


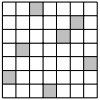
The framework of permutons [Hoppen et al., 2013]

Definition: A permuton μ is a probability measure on the unit square with uniform projections (or marginals):

$$\text{for all } a < b \text{ in } [0,1],\ \mu([a,b]\times[0,1]) = \mu([0,1]\times[a,b]) = b-a.$$

Remark: The normalized diagrams of permutations (denoted σ) are essentially permutons (denoted μ_σ)





Replacing each point $(i/n, \sigma(i)/n)$ by a little square $[(i-1)/n, i/n] \times [(\sigma(i)-1)/n, \sigma(i)/n]$, and distributing the mass 1 uniformly on these little squares

Convergence of a sequence of permutations (σ_n) to a permuton μ:

- inherited from the weak convergence of measures, namely:
- $\sigma_n \to \mu$ when $\sup_{R \text{ rectangle } \subset [0,1]^2} |\mu_{\sigma_n}(R) - \mu(R)| \to 0$ as $n \to +\infty$.
- If each σ_n has size n, taking R of the form $[0, i/n] \times [0, j/n]$ is enough.

Permuton limit of record-biased permutations

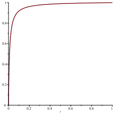
Theorem:
Let σ_n be a random record-biased permutation of size n for $\theta = \lambda n$.
μ_{σ_n} converges in probability to $\mu = \mu_c + \mu_u$ defined below.

Letting $f_\lambda(x) = \frac{x(\lambda+1)}{\lambda+x}$, we define

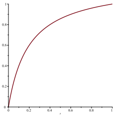
- μ_u is the uniform measure
  of total mass $c_\lambda \int_0^1 f_\lambda$ for $c_\lambda = \frac{1}{\lambda+1}$
  on the area under the curve $y = f_\lambda(x)$;
- μ_c is the measure
  supported by the curve $y = f_\lambda(x)$
  with density $\frac{\lambda}{\lambda+x}$ with respect to Leb_c,
  defined by $Leb_c(x, f_\lambda(x)) = Lebesgue(x)$

Two steps towards this statement:
guessing μ and proving convergence.

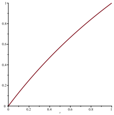
$f_{0.01}$:



$f_{0.2}$:



f_3:



Guessing the limit μ

The pictures suggest to decompose μ as $\mu_u + \mu_c$, with μ_c on a curve, and μ_u uniform under the curve. To determine are:

- the equation $y = f_\lambda(x)$ of the curve,
- how to distribute the mass between μ_c and μ_u.

To find the equation $y = f_\lambda(x)$ of the curve,

- we estimate $\mathbb{P}$(max before position i is j) for $i \approx xn$ and $j \approx yn$;
- we find the relation between x and y which makes this probability not larger than 1, and non-zero once summed over j.

To find the relative measures on the curve and below,

- we compute the measure of the records in σ_n and take the limit in n: this gives the measure $\int_0^1 \frac{\lambda}{\lambda+x} dx$ on the curve;
- we distribute uniformly the mass $c_\lambda \int_0^1 f_\lambda(x) dx$ below the curve, for c_λ s.t. $\int_a^b (\frac{\lambda}{\lambda+x} + c_\lambda f_\lambda(x)) dx = b - a$.

Stay tuned

What was done: [AofA 2016]

- Definition of the record-biased model
- Behavior of statistics (precise probabilities, expectation)
- Applications to the analysis of algorithms

What is new: [AofA 2021]
(but needs to be written down...) and on Arxiv "soon"

- Efficient random samplers
- Permuton limit

What is left to do: hopefully before AofA 2026!

- Is the number of inversions Gaussian?
- Do the Gaussian limiting distributions hold also in the $\theta = \lambda n$ regime?

!! Thank you !!