Coherent random permutations with record statistics

Alexander Gnedin

HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm

2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 157–170

Coherent random permutations with record statistics

Alexander Gnedin

Mathematical Institute, Utrecht University, P.O. Box 80010, 3508 TA Utrecht, The Netherlands, gnedin@math.uu.nl

received 14 Feb 2006, revised 19th January 2008, accepted 19th January 2008.

A two-parameter family of random permutations of $[n]$ is introduced, with distribution conditionally uniform given the counts of upper and lower records. The family interpolates between two versions of Ewens' distribution. A distinguished role of the family is determined by the fact that every sequence of coherent permutations $(\pi_n, n = 1, 2, \ldots)$ with the indicated kind of sufficiency is obtainable by randomisation of the parameters. Generating algorithms and asymptotic properties of the permutations follow from the representation via initial ranks.

Keywords: random permutations, records, sufficiency, Ewens' distribution

1 Introduction

Random permutations with non-uniform distribution appear in a variety of contexts such as combinatorial structures [1], shuffling and sorting algorithms [3, 18], dynamical systems [17] and statistics [11], just to mention a few. Generalisations of the uniform distribution can be designed by assuming some sort of sufficiency, that is requiring the distribution to be uniform conditionally given the value of some statistic of permutation.

One important instance of this kind is Ewens' distribution on the symmetric group $\mathfrak{S}_n$. The distribution assigns probability $\theta^{c-1}/(\theta+1)_{n-1}$ to every permutation $\pi_n \in \mathfrak{S}_n$ with c cycles, where $\theta \geq 0$ is a parameter, see [1, 21]. Ewens' distributions are coherent as n varies, hence can be viewed as a probability on the space $\mathfrak{S}^\infty$, a projective limit of $\mathfrak{S}_n$'s. Moreover, every distribution for $(\pi_n, n = 1, 2, \ldots) \in \mathfrak{S}^\infty$ such that each π_n is uniform given the number of cycles, can be obtained as a mixture over Ewens' family, that is by randomisation of θ, see Gnedin and Pitman [8].

By the virtue of a fundamental bijection $\mathfrak{S}_n \to \mathfrak{S}_n$ the number of cycles is translated into the number of upper records, hence Ewens' distribution may be also viewed as a distribution which assigns probability $\theta^{u-1}/(\theta+1)_{n-1}$ to each permutation $\pi_n \in \mathfrak{S}_n$ with u upper records, a viewpoint due to Kerov [14]. Gnedin and Olshanski [7] explored a similar setting with the number of descents as sufficient statistic; in this case the distinguished role is played by the dicrete-parameter family of a-shuffles (introduced in [3] and appearing in bucket sorting [18]) and their reversals.

This paper was largely motivated by the suggestion in [9, 10] to connect random record models with instances of Ewens' sampling formula. In this direction, we extend Ewens' family by introducing two-parameter distributions $P^{(\theta,\zeta)}$ on $\mathfrak{S}^\infty$ under which every $\pi_n \in \mathfrak{S}_n$ is uniform conditionally given (ℓ, u), for ℓ the number of lower and u the number of upper records of permutation. We show that every probability on $\mathfrak{S}^\infty$ having this kind of sufficiency is obtainable by randomising the parameters θ and ζ. We also show that permutations under $P^{(\theta,\zeta)}$ can be generated by ranking a sequence of real-valued random variables (X_n), whose record sequences follow a two-sided analogue of the GEM distribution for 'stick-breaking' partition of the unit interval.

2 Counting the records

Permutations $\pi_n \in \mathfrak{S}_n$ of $[n] := \{1, \dots, n\}$ will be written in the one-row notation $\pi_n = (\pi_{n1}, \dots, \pi_{nn})$. We call element π_{nj} a lower record of π_n if $\pi_{nj} = \min(\pi_{n1}, \dots, \pi_{nj})$, and we call π_{nj} an upper record if $\pi_{nj} = \max(\pi_{n1}, \dots, \pi_{nj})$. When π_{nj} is a record we say that π_{nj} is a record value and that j is a record time (or a record position). The first entry π_{n1} will be called *center*. We regard the center as *improper* lower and upper record, all other records being *proper*. We denote

$$\mathtt{rec}(\pi_n) = (r_{-\ell}, \dots, r_{-1}, r_0, r_1, \dots, r_u)$$

the two-sided increasing sequence of record values, with distinguished center $r_0 = \pi_{n1}$, proper lower records $r_{-\ell}, \dots, r_{-1}$ and proper upper records $r_1, \dots, r_u$. In this notation ℓ, u count the proper records; for instance, $\mathtt{rec}(3, 2, 7, 6, 1, 4, 8, 5) = (1, 2, \mathbf{3}, 7, 8)$, where the center is boldfaced and $\ell = u = 2$. Clearly, $r_{-\ell} = 1$, $r_u = n$, and the total number of records $\#\mathtt{rec}(\pi_n) = \ell + u + 1$ satisfies $\min(2, n) \le \ell + u + 1 \le n$. The record times of proper lower and upper records will be labelled $t_1, \dots, t_u$ and $t_{-1}, \dots, t_{-\ell}$, respectively, and we denote $t_0 = 1$ the record time associated with the improper record.

Let $\begin{bmatrix} n \\ \ell+1, u+1 \end{bmatrix}$ be the number of permutations $\pi_n \in \mathfrak{S}_n$ with $\ell + 1$ lower and $u + 1$ upper records. This array of combinatorial numbers is symmetric in ℓ and u, and satisfies the recursion

$$\begin{bmatrix} n \\ \ell+1, u+1 \end{bmatrix} = \begin{bmatrix} n-1 \\ \ell, u+1 \end{bmatrix} + \begin{bmatrix} n-1 \\ \ell+1, u \end{bmatrix} + (n-2) \begin{bmatrix} n-1 \\ \ell+1, u+1 \end{bmatrix}. \tag{1}$$

Summing over one of the record counts, say u, yields a signless Stirling number of the first kind

$$\begin{bmatrix} n \\ \ell+1 \end{bmatrix} = \sum_{u=0}^{n-1} \begin{bmatrix} n \\ \ell+1, u+1 \end{bmatrix},$$

equal to the number of permutations with $\ell + 1$ lower records. A more delicate connection to the Stirling numbers appears via the identity

$$\begin{bmatrix} n \\ \ell+1, u+1 \end{bmatrix} = \begin{bmatrix} n-1 \\ \ell+u \end{bmatrix} \binom{\ell+u}{\ell} \tag{2}$$

found in [2, p. 179], where it was derived by manipulation with generating functions.

For our purposes it is important to introduce yet another encoding of permutation into the sequence of *initial ranks*

$$\iota_j := \#\{k : \ k \le j, \, \pi_{nk} \ge \pi_{nj}\}, \quad j \in [n].$$

The correspondence $\pi_n \mapsto (\iota_1, \dots, \iota_n)$ is a well-known bijection between $\mathfrak{S}_n$ and $[1] \times [2] \times \cdots \times [n]$. Note that π_{nj} is a lower record if $\iota_j = 1$, and an upper record if $\iota_j = j$.

In terms of the initial ranks a bijective proof of (2) is easily acquired. To this end, consider the mapping which sends $\pi_n \in \mathfrak{S}_n$ to $\pi'_{n-1} \in \mathfrak{S}_{n-1}$ so that the initial ranks are transformed as $(\iota_1, \dots, \iota_n) \mapsto (\iota'_1, \dots, \iota'_{n-1})$ where $\iota'_{j-1} = \iota_j \mathbf{1}(\iota_j < j)$ for $2 \le j \le n$ (and $\mathbf{1}(\cdots)$ will denote an indicator). Each proper record of π_n is mapped bijectively to a lower record of π'_{n-1}, and the record counts satisfy $\ell(\pi_n) + u(\pi_n) = \ell(\pi'_n) + 1$. It is easily seen that 2^r permutations π_n are mapped to the same π'_{n-1} each time when $\ell(\pi'_{n-1}) + 1 = r$, and of these π_n there are $\binom{r}{\ell}$ permutations with ℓ proper lower records. Because π'_{n-1} with r lower records can be chosen in $\begin{bmatrix} n-1 \\ r \end{bmatrix}$ ways, the identity (2) follows.

When a probability distribution P_n is specified on $\mathfrak{S}_n$, we consider π_n as a random variable. In particular, $P_n^{(1,1)}(\pi_n) \equiv 1/n!$ is the uniform distribution (indices will be explained in the next section). The characteristic feature of the uniform distribution is that the initial ranks are independent, with each ι_j being uniformly distributed on $[j]$. Giving a probabilistic interpretation to (2) we have:

Lemma 1 *Under the uniform distribution $P_n^{(1,1)}$ for π_n, conditionally given the record counts (ℓ, u) and given the positions occupied by $\ell + u$ proper records, all $\binom{\ell+u}{\ell}$ allocations of ℓ lower records within these $\ell + u$ positions are equally likely.*

Recall that *ranking* associates with any sequence of distinct reals $x_1, \dots, x_n$ a sequence of ranks $\pi_{nj} = \#\{i \le n : \ x_i \le x_j\}$, also called the *ranking permutation*. A uniform permutation appears when $x_1, \dots, x_n$ are sampled independently from the uniform distribution on $[0, 1]$ (or from any other nonatomic distribution on reals).

3 The two-parameter family of random permutations

We introduce next a two-parameter deformation of the uniform distribution, for which (ℓ, u) is a sufficient statistic, meaning that given the record counts the distribution of π_n is uniform.

Proposition 2 *For arbitrary positive θ and ζ the formula*

$$P_n^{(\theta,\zeta)}(\pi_n) = \frac{\theta^\ell \zeta^u}{(\theta + \zeta)_{n-1}} \tag{3}$$

defines a distribution on $\mathfrak{S}_n$, which assigns the same probability to every permutation with $\ell + 1$ lower and $u + 1$ upper records.

Proving this amounts to the fact that the probabilities in (3) add to unity, which is equivalent to the formula for the bivariate generating function

$$\sum_{\ell,u} \begin{bmatrix} n \\ \ell + 1, u + 1 \end{bmatrix} \theta^\ell \zeta^u = (\theta + \zeta)_{n-1}, \tag{4}$$

known since at least [4]. Note that for $\zeta = 1$ this specialises as the well-known formula

$$\sum_{\ell=0}^{n-1} \begin{bmatrix} n \\ \ell + 1 \end{bmatrix} \theta^\ell = (\theta + 1)_{n-1}$$

for the generating function of Stirling numbers.

To generate π_n under $P_n^{(\theta,\zeta)}$ one can exploit the representation via initial ranks, with $\iota_1 = 1$ and $\iota_2, \ldots, \iota_n$ sampled independently from distributions

$$\iota_j = \begin{cases} 1 \text{ w.p. } \theta/(\theta+\zeta+j-2), \\ j \text{ w.p. } \zeta/(\theta+\zeta+j-2), \\ r \text{ w.p. } 1/(\theta+\zeta+j-2) \text{ for } r = 2, \ldots, j-1 \end{cases} \tag{5}$$

(w.p.=with probability). Multiplying these out it is seen that (3) is indeed the probability of any sequence $(i_2, \ldots, i_n)$ where $\iota_j = 1$ occurs ℓ times and $\iota_j = j$ occurs u times.

For the uniform distribution, each ι_j should be sampled from the uniform distribution on $[j]$, which is a well-known method to generate uniform permutation. Thus, $P_n^{(\theta,\zeta)}$ deforms $P_n^{(1,1)}$ by tilting the probabilities of extreme values of the initial ranks.

4 Construction for integer parameters

For integer θ, ζ the distribution $P_n^{(\theta,\zeta)}$ can be obtained as a projection of the uniform distribution $P_{n+d}^{(1,1)}$ on $\mathfrak{S}_{n+d}$, where $d = \theta + \zeta - 2$. To ease notation, for the rest of this section the elements of permutation are written with one index.

Fix $(w_1, \ldots, w_{n+d}) \in \mathfrak{S}_{n+d}$. A sequence $(\pi'_j,\ j \in [n])$ (which is a permutation of n integers $\{\theta, \ldots, n+\theta-1\}$) is uniquely defined by the condition that $\{\pi'_1, \ldots, \pi'_j\} \subset \{w_1, \ldots, w_{d+j}\}$ is the subset of integers whose ranks among $\{w_1, \ldots, w_{d+j}\}$ are neither among top $\zeta - 1$ ranks nor among bottom $\theta - 1$ ranks. Here is the inductive definition. Let $s_1, \ldots, s_{n+d}$ be the initial ranks of $w_1, \ldots, w_{n+d}$. At step 1 we define π'_1 to be the element of rank θ among $w_1, \ldots, w_{d+1}$, thus leaving $\zeta - 1$ elements ranked above and $\theta - 1$ ranked below π'_1. At step j the element w_{d+j} is added, if $\theta \le s_{d+j} \le j + \theta - 1$ then $\pi'_j = w_{d+j}$, if $1 \le s_{d+j} \le \theta - 1$ then π'_j is defined to be the element of rank θ among $w_1, \ldots, w_{d+j}$, and if $j+\theta \le s_{d+j} \le j+d$ then π'_j is defined to be the element of rank $j+\theta-1$ among $w_1, \ldots, w_{d+j}$. Understanding the second arrow in $(w_1, \ldots, w_{n+d}) \mapsto (\pi'_1, \ldots, \pi'_n) \mapsto (\pi_1, \ldots, \pi_n)$ as the ranking operation, we have defined a projection $f_n^{(\theta,\zeta)}$ from $\mathfrak{S}_{n+d}$ to $\mathfrak{S}_n$.

Proposition 3 *For positive integers θ, ζ the mapping $f_n^{(\theta,\zeta)}$ sends the uniform distribution on $\mathfrak{S}_{n+d}$ (where $d = \theta + \zeta - 2$) to $P_n^{(\theta,\zeta)}$.*

Proof: In the above, the initial ranks for $(\pi_1, \ldots, \pi_n)$ and $(\pi'_1, \ldots, \pi'_n)$ are the same, and are given for $j = 2, \ldots, n$ by

$$\iota_j = \begin{cases} 1, & \text{if } s_{j+d} \in [1, \theta], \\ s_{j+d} - \theta + 1, & \text{if } s_{j+d} \in [\theta+1, j+\theta-2], \\ j, & \text{if } s_{j+d} \in [j+\theta-1, j+d]. \end{cases}$$

For uniform permutation, s_{j+d} is uniform on $[j+d]$ and these are independent, hence the r_j's are independent with respective probabilities $\theta/(n+d-2)$, $\zeta/(n+d-2)$ for extreme ranks and equal probabilities for other values of ι_j. □

For irrational θ or ζ the distribution $P_n^{(\theta,\zeta)}$ cannot be obtained as a projection of a uniform distribution on some combinatorial object.

5 Coherent permutations

Our view of permutation is biased towards the interpretation as order, rather than mapping. Orders can be obviously restricted from larger sets to smaller. In this direction, we say that permutations π_n and π_m, for $m \le n$, are *coherent* if they determine the same order on $[m]$. A sequence (π_n) of coherent permutations $\pi_n \in \mathfrak{S}_n$ defines a strict order $\lhd$ on the infinite set $\mathbb{N}$: $j \lhd i$ iff $\pi_{nj} < \pi_{ni}$ for all $n \ge \max(j,i)$.

Let $D_{nm} : \mathfrak{S}_n \to \mathfrak{S}_m$ $(n > m)$ be the projection which cuts the last $n-m$ entries of π_n and replaces the first m entries $\pi_{n1}, \dots, \pi_{nm}$ by their ranking permutation. The projection D_{nm} is the same as restricting orders from $[n]$ to $[m]$, hence the coherence means that $D_{nm}(\pi_n) = \pi_m$. The space of all orders on $\mathbb{N}$ has the structure of the projective limit $\mathfrak{S}^\infty := \varprojlim \mathfrak{S}_n$.

Warning. The space $\mathfrak{S}^\infty$ should not be confused with the infinite symmetric group $\mathfrak{S}_\infty$, which is the inductive limit of finite symmetric groups with natural embedding. The elements of $\mathfrak{S}_\infty$ are bijections $\mathbb{N} \to \mathbb{N}$ that displace only finitely many integers.

In terms of the initial ranks, $D_{nm} : (\iota_1, \dots, \iota_n) \mapsto (\iota_1, \dots, \iota_m)$ is just the projection on the first m coordinates. Every infinite sequence (ι_n) determines an order $\lhd$ on $\mathbb{N}$, in which n is ranked ι_nth within the set $[n]$. Therefore $\mathfrak{S}^\infty$ can be identified with the infinite product space $[1] \times [2] \times \dots$ endowed with the discrete product topology (in which $\mathfrak{S}^\infty$ is a metrisable totally disconnected Borel space). When a probability measure is defined on $\mathfrak{S}^\infty$ we view $(\pi_n) \in \mathfrak{S}^\infty$ as a random coherent sequence of permutations, or a random order on $\mathbb{N}$. By the measure extension theorem, distributions P_n on $\mathfrak{S}_n$, defined for every n, determine a unique distribution P on $\mathfrak{S}^\infty$ for a coherent sequence of permutations if and only if the P_n's are compatible with projections.

We denote $P^{(\theta,\zeta)}$ the measure on $\mathfrak{S}^\infty$ under which the initial ranks $\iota_1, \iota_2, \dots$ are independent, with distribution as in Section 3. The distributions $(P_n^{(\theta,\zeta)},\ n = 1, 2, \dots)$ introduced in Proposition 3 are coherent projections of $P^{(\theta,\zeta)}$.

For an order $\lhd$ on $\mathbb{N}$ we shall say that an upper (or lower) record occurs at time n if $\iota_n = n$ (respectively, $\iota_n = 1$). Reversing the order is an automorphism of $\mathfrak{S}^\infty$, which is written as either $\pi_{nj} \mapsto n - \pi_{nj}$ for $j \in [n],\ n \in \mathbb{N}$, or, via the initial ranks, as $\iota_n \mapsto n - \iota_n$ for $n \in \mathbb{N}$. Clearly, reversing the order swaps the types of records, hence maps $P^{(\theta,\zeta)}$ to $P^{(\zeta,\theta)}$.

Remark. Except $D_n := D_{n,n-1}$ there are two other useful n-to-1 projections $D'_n, D''_n : \mathfrak{S}_n \to \mathfrak{S}_{n-1}$, which appear in the context of descent statistics and cycle statistics, respectively [6, 16]. Projection D'_n deletes n in the one-row notation of π_n, and D''_n deletes n in the cycle notation of π_n. The projective limit $\varprojlim(\mathfrak{S}_n, D''_n)$ was introduced in the representation theory of $\mathfrak{S}_\infty$ as the space of *virtual permutations* [16], and D'_n was used in [6]. The isomorphism of three kinds of projective limits is established by means of the commutative diagram

$$\begin{array}{ccccc} \pi_n & \longrightarrow & \pi_n^{-1} & \longrightarrow & (\pi_n^{-1})\widehat{\ } \\ D_n\downarrow & & D'_n\downarrow & & D''_n\downarrow \\ \pi_{n-1} & \longrightarrow & \pi_{n-1}^{-1} & \longrightarrow & (\pi_{n-1}^{-1})\widehat{\ } \end{array}$$

where π_n^{-1} denotes the inverse permutation, and $\widehat{\pi_n}$ denotes the *fundamental bijection* $\mathfrak{S}_n \to \mathfrak{S}_n$ which

translates the one-row notation of permutation into the cycle notation of another permutation by inserting parentheses ')(' before each proper lower record, e.g. $(3,2,7,6,1,4,8,5)\widehat{\ } = (3)(2,7,6)(1,4,8,5)$ (Stanley [23, p. 17] gives a slightly different version of the mapping).

6 Specialisations

Some special values of the parameters θ, ζ and some limits are worth mentioning. We call distribution P on $\mathfrak{S}^\infty$ *degenerate* if $P_n(\pi_n) = 0$ for some n and some $\pi_n \in \mathfrak{S}_n$. All distributions $P^{(\theta,\zeta)}$ for $\theta, \zeta > 0$ are nondegenerate.

The uniform distribution. The measure $P^{(1,1)}$ may be called the uniform distribution on $\mathfrak{S}^\infty$, since every $P_n^{(1,1)}$ is the uniform distribution on $\mathfrak{S}_n$, with $P_n^{(1,1)}(\pi_n) \equiv 1/n!$ for every $\pi_n \in \mathfrak{S}_n$. The corresponding random order $\lhd$ on $\mathbb{N}$ has the characteristic property of *exchangeability*, that is the law of $\lhd$ is invariant under the action of $\mathfrak{S}_\infty$. This order appears by ranking an iid sample (X_n) from the uniform distribution on $[0,1]$ (or some other contunuous distribution on reals). For fixed n there are also other ways to link uniform π_n to a sequence of n random reals [9].

Ewens' distributions $P^{(\theta,1)}$ and $P^{(1,\zeta)}$. Ewens' distribution on $\mathfrak{S}_n$ (also called θ-biased permutation, see [1]) is the one which assigns probability $\theta^{c-1}/(\theta+1)_{n-1}$, to every permutation with c cycles. The partition of n comprised of cycle-sizes of π_n follows then the Ewens sampling formula.

Suppose $\zeta = 1$, so the probabilities (3) become $P_n^{(\theta,1)}(\pi_n) = \theta^\ell/(\theta+1)_{n-1}$ where $\ell+1$ is the number of lower records of π_n. When π_n follows $P_n^{(\theta,1)}$ then also π_n^{-1}, because $\ell(\pi_n) = \ell(\pi_n^{-1})$. To see this, draw permutation in two dimensions as a point scatter $\{(j, \pi_{nj}), j \in [n]\}$. Observe that the records are those points which do not have other points south-west of them. Flip the picture about the diagonal to see that the property is preserved. The inversion combined with the $\widehat{\ }$-mapping in Section 5 transforms the distribution in its conventional 'cycle form'. Therefore we still call $P^{(\theta,1)}$ and $P^{(1,\zeta)}$ Ewens' distributions (this viewpoint was suggested in [14, 15]).

By the same flipping argument, the sequence of lower record times $t_{-\ell}, \ldots, t_{-1}, t_0$ coincides with the decreasing sequence of lower record values of the inverse permutation π_n^{-1}, hence under $P^{(\theta,1)}$ we have further symmetry: $(t_{-\ell}, \ldots, t_{-1}, t_0) \stackrel{d}{=} (r_0, \ldots, r_{-1}, r_{-\ell})$.

Distributions with equal parameters. For $\theta = \zeta$ there is a symmetry between lower and upper records. For distributions $P_n^{(\theta,\theta)}(\pi_n) = \theta^{\ell+u}/(2\theta)_{n+1}$ the minimal sufficient statistic is the total number of records $\ell + u + 1$. Given the value of this statistic, π_n is uniformly distributed.

Bernoulli pyramids $P^{(p,q)\infty}$ $(0 \le p \le 1,\ p+q=1)$. If $\theta, \zeta \to \infty$ but so that $\theta/(\theta+\zeta) \to p$, then under the limiting law the probability of π_n is $p^\ell(1-p)^u$ provided $\ell + u = n-1$, and the probability is zero otherwise. Such π_n has each π_{nj} $(j > 1)$ an upper record with probability p and a lower record with probability $1-p$. Only extreme initial ranks are possible, i.e $i_j \in \{1, j\}$. Such distributions were exploited in optimal stopping [5, 12]. One way to generate such permutation is to split $[n]$ by binomial variable at some integer v, then let $\pi_1 = v$ for the center and then riffle-shuffle $v+1, \ldots, n$ and $v-1, \ldots, 1$ to obtain $\pi_{2n}, \ldots, \pi_{nn}$. In the cases $p = 1$ (respectively, $p = 0$) the distribution concentrates on the permutation $(n, \ldots, 1)$ (respectively, $(1, \ldots, n)$).

Degenerate Ewens' permutations $P^{(\theta,0)},\ P^{(0,\zeta)}$. In the limiting case $\theta \to 0$ (but $\zeta > 0$), the permutation has the form $\pi_n = (1, \pi'_{n-1})$, where π'_{n-1} is a permutation of $\{2, \ldots, n\}$ which upon obvious identification has $P_{n-1}^{(1,\zeta)}$ distribution. In the limiting case $\zeta \to 0$ (but $\theta > 0$), the permutation has the form $\pi_n = (n, \pi'_{n-1})$, where π'_{n-1} is a permutation of $[n-1]$ which has $P_{n-1}^{(\theta,1)}$ distribution.

Permutations with only one proper record $P^{(p,q)_0}$ $(0 \le p \le 1,\ p+q=1)$**.** When both $\theta, \zeta \to 0$ but so that $\theta/(\theta+\zeta) \to p$ for some $p \in [0,1]$, then the limit law of π_n is that of $(\pi_{n1}, \pi_{n2}, \pi'_{n-2})$ where (π_{n1}, π_{n2}) is either $(1,n)$ or $(n,1)$ with probability p and $1-p$, respectively, while π'_{n-2} is a uniform permutation of $\{2,\dots,n-1\}$ independent of (π_{n1}, π_{n2}).

Proposition 4 *The weak closure of the (θ,ζ)-family of distributions on $\mathfrak{S}_\infty$ is comprised of nondegenerate ditributions with $\theta > 0$, $\zeta > 0$, and of the three degenerate types described above.*

Proof: This follows by considering $P_2^{(\theta,\zeta)}$ and $P_3^{(\theta,\zeta)}$. □

7 Characterisation of mixtures by sufficiency

We seek now for a two-parameter generalisation of [8, Theorem 12 (i)], that is we wish to characterise the distributions $P^{(\theta,\zeta)}$ as extreme points of a suitable family of probabilities on $\mathfrak{S}^\infty$ that are conditionally uniform on each $\mathfrak{S}_n$. The following lemma is helpful.

Lemma 5 *Let Q_1 be the law of an independent* 0-1 *sequence $B_1, B_2, \dots$ with B_n* Bernoulli$(1/n)$*. Assume Q is a distribution for $B_1, B_2, \dots$ with the property that, for each n, the conditional law of $(B_1,\dots,B_n)$ given $S_n := B_1 + \dots + B_n$ and given $(B_m,\ m > n)$ under Q is the same as under Q_1. Then Q is a unique mixture of distributions Q_η, $\eta \in [0,\infty]$, under which $B_1, B_2, \dots$ are independent with B_n* Bernoulli$(\eta/(n+\eta-1))$.

Proof: This can be concluded from either [20, p. 269] or [8, Lemma 9]. The key issue is that the convergence $S_n/\log n \to \eta$ holds under Q_η almost surely. □

The first two assertions of the next proposition are equivalent to [8, Theorem 12 (i)] and included here for completeness of exposition.

Proposition 6 *For P a probability on $\mathfrak{S}^\infty$ suppose the distribution of π_n for every $n = 1, 2, \dots$ is uniform conditionally given the value of a statistic* `stat`*. Then the following assertions are true:*

(i) *for* `stat` $= \ell$ *distribution P is a unique mixture of $P^{(\theta,1)}$ $(\theta \in [0,\infty[)$ and $P^{(1,0)_\infty}$,*

(ii) *for* `stat` $= u$ *distribution P is a unique mixture of $P^{(1,\zeta)}$ $(\zeta \in [0,\infty[\,)$ and $P^{(0,1)_\infty}$,*

(iii) *for* `stat` $= \ell + u$ *distribution P is a unique mixture of $P^{(\theta,\theta)}$ $(\theta \in]0,\infty[\,)$, $P^{(\frac12,\frac12)_0}$ and $P^{(\frac12,\frac12)_\infty}$,*

(iv) *for* `stat` $= (\ell, u)$ *distribution P is a unique mixture of nondegenerate distributions $P^{(\theta,\zeta)}$ $(\theta,\zeta \in]0,\infty[\,)$, degenerate distributions $P^{(\theta,0)}$ and $P^{(0,\zeta)}$ $(\theta,\zeta \in]0,\infty[\,)$, and further degenerate distributions $P^{(1,0)_0}$, $P^{(0,1)_0}$ and $P^{(p,q)_\infty}$ $(p \in [0,1])$. The degenerate distributions do not enter provided $P_3 > 0$.*

Proof: We need to show that the described distributions and only they are extreme. Assuming P extreme in the setting of (iv), the tail algebra $\mathcal{F}$ of the process $((\ell(\pi_n), u(\pi_n)),\ n = 1,2,\dots)$ must be trivial. Let $B_n = \mathbf{1}(r_{n+1} \in \{1, n+1\})$ be the indicator of some record at position $n+1$. Under $P^{(1,1)}$ the law of $(B_1, B_2, \dots)$ is Q_2, hence by Lemma 5 and because $\lim S_n/\log n$ is $\mathcal{F}$-measurable the law of (B_n) under P is the same as under Q_η for some η. This says that records occur by a Bernoulli process,

without specifying the types of records. If $\eta = 0$ the situation is clear: there is only one proper record (for $n > 1$) and $P^{(1,0)_0}$, $P^{(0,1)_0}$ are the sole possibilities. Suppose $\eta \neq 0$. A key to recognise how the records are classified in types is the exchangeability. Let I_k be the indicator of the event that the record at $(k+1)$st record time is a lower record. Conditionally given $I_1 + \ldots + I_k = \ell - 1$ all values of the sequence $(I_1, \ldots, I_k)$ have the same probability $1/\binom{k}{\ell-1}$, because by Lemma 1 this is true under $P^{(1,1)}$ by the virtue of a simple stopping times argument. By de Finetti's theorem, there exists a relative frequency of lower records, hence $\ell(\pi_n)/(\ell(\pi_n) + u(\pi_n))$ must converge almost surely. But the limit of this ratio is $\mathcal{F}$-measurable hence constant, say p. Appealing again to Lemma 1 we see that (B_n) and (I_k) are independent, hence the set of positions of lower records is the one obtained by independent thinning with probability p of the occurences of 1's in (B_n). Thus $P = P^{(\theta,\zeta)}$ with $\theta = p\eta,\ \zeta = (1-p)\eta$ (the instance $\eta = \infty$ is included). Part (iii) is shown similarly, with the special feature that $p = 1/2$. □

The result suggests a practical way to generate all possible coherent permutations with a suitable kind of sufficiency. For instance, if we wish to produce a nondegenerate sequence of permutations with every π_n conditionally uniform given (ℓ, u), then we need to first specify a distribution for positive (θ, ζ), to choose the parameters from this distribution and finally to construct permutation from the independent initial ranks whose distributions involve the chosen parameters. It should be noticed, however, that this de-Finetti-type procedure covers all possibilities only under the coherence condition. For each fixed $n > 2$ there exist conditionally uniform distributions on $\mathfrak{S}_n$ which are not mixtures of $P_n^{(\theta,\zeta)}$'s.

Recall that de-Finetti's theorem can be stated as follows: if a 0-1-sequence (B_n) is such that for every n the law of $(B_1, \ldots, B_n)$ is uniform conditionally given the number of 1's then (B_n) is a unique mixture of independent coin-tossing processes. In this spirit, Proposition 6(iv) says that if for (I_n) (with the range of I_n being $[n]$) the law of $(I_1, \ldots, I_n)$ for every n is uniform conditionally given $\ell_n := \{1 < j \leq n : I_j = 1\}$ and $u_n := \{1 < j \leq n : I_j = j\}$ then (I_n) is a unique mixture of independent processes (5) (including the degenerate cases).

To put the characterisation result in the framework of arrays of combinatorial numbers [7, 8], denote $w_n(\ell, u)$ the probability for ℓ lower and u upper proper records in π_n. By the rule of addition of probabilities we have

$$w_n(\ell, u) = w_{n+1}(\ell+1, u) + w_{n+1}(\ell, u+1) + (n-1)w_{n+1}(\ell, u), \quad w_1(0,0) = 1, \tag{6}$$

which is a recursion dual to (1). The set of nonnegative solutions to (6) is a convex compact set, and Proposition 6(iv) describes the set of extreme solutions to (6). Interestingly, the set of extremes is not closed: each distribution $P^{(p,q)_0}$ with $0 < p < 1$ appears as a limit of some nondegenerate $P^{(\theta,\zeta)}$'s, but it is decomposable as a mixture $P^{(p,q)_0} = pP^{(1,0)_0)} + qP^{(0,1)_0}$.

A familiar method of finding the extreme solutions of (6) is based on identifying the Martin boundary. To that end, one explores asymptotic regimes for $\ell' = \ell'(n'), u' = u'(n')$ as $n' \to \infty$, which guarantee for all n, ℓ, u convergence of the ratios

$$\begin{bmatrix} n \\ \ell+1, u+1 \end{bmatrix}_{\ell'+1, u'+1}^{n'} \Big/ \begin{bmatrix} n' \\ \ell'+1, u'+1 \end{bmatrix}, \tag{7}$$

where the numerator is the number of permutations $\pi_{n'}$ of $[n']$ with record counts (ℓ', u'), such that the restriction of $\pi_{n'}$ to $[n]$ has record counts (ℓ, u). Using a monotonicity argument like in [8], the things

can be reversed to show that the convergence $\ell'/\log n'$ and $u'/\log n'$ is necessary and sufficient for the convergence of the ratios (7) for all n, u, ℓ (hence the Martin boundary coincides with the set of extreme solutions).

8 Related distributions and asymptotics

As in the case of the uniform distribution [19], asymptotic properties (as $n \to \infty$) of record counts $\ell = \ell(\pi_n), u = u(\pi_n)$ under $P^{(\theta,\zeta)}$ follow straightforwardly from the representation via independent initial ranks. Thus, both mean and variance of ℓ are asymptotic to $\theta \log n$, and that of u to $\zeta \log n$. Jointly, (ℓ, u) converge in distribution to independent Gaussian variables. The point processes of scaled record times $\{t_k/n : \ k < 0\}$, $\{t_k/n : \ k > 0\}$ converge to independent Poisson processes with intensities $\theta \mathrm{d}t/t$, $\zeta \mathrm{d}t/t$ (for $t \in [0, 1]$), respectively.

The behaviour of each π_{nj} under $P^{(\theta,\zeta)}$ as n varies is that of a process with exchangeable 0-1 increments, known as Pólya's urn model. That is to say, each sequence $(\pi_{nj}, \ n \geq j)$ is a nondecreasing inhomogeneous Markov chain on integers, which starts at some random initial rank $\pi_{jj} = \iota_j$ at time j, and at time n either jumps from some rank $\pi_{nj} = v$ to $v+1$ with probability $(v-1+\theta)/(n-2+\theta+\zeta)$, or otherwise remains at v.

The law of $\mathtt{rec}(\pi_n)$ can be expressed in terms of Pólya-Eggenberger distributions

$$\mathrm{PE}_n^{(\theta,\zeta)}(r) := \binom{n-1}{r-1} \frac{(\theta)_{n-1}(\zeta)_{r-1}}{(\theta+\zeta)_{n-1}} \qquad r \in [n].$$

The distribution of the center $r_0 = \pi_{n1}$ is $\mathrm{PE}_n^{(\theta,\zeta)}$. Conditionally given r_0, the lower and upper record sequences are independent. The sequence of lower records $r_{-1}, \ldots, r_{-\ell}$ is a homogeneous decreasing Markov chain on integers which starts at r_0 and terminates at 1, each time descending from the generic r to $r-d$ with probability $\mathrm{PE}_r^{(\theta,1)}(d)$. In a similar way, the sequence of upper records $r_1, \ldots, r_u$ is a homogeneous increasing Markov chain on integers which starts at r_0 and terminates at n, each time ascending from some r to $r+d$ with probability $\mathrm{PE}_{n-r+1}^{(\zeta,1)}(d)$.

Asymptotics of the record values follow from the well known properties of Pólya urns. Recall that beta(a, b) distribution with parameters $a > 0, \ b > 0$ is the distribution on $[0, 1]$ with density $x^{a-1}(1-x)^{b-1}/\mathrm{B}(a, b)$, where $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

Proposition 7 *As $n \to \infty$, under $P^{(\theta,\zeta)}$ the scaled record values of π_n converge,*

$$\frac{r_k}{n} \to \rho_k \ \text{ a.s. } \ (k \in \mathbb{Z}).$$

The distribution of ρ_0 is beta(θ, ζ). *Given ρ_0 the sequences $(\rho_k, k < 0)$ and $(\rho_k, k > 0)$ are independent and representable as*

$$\rho_k = \rho_0 T_k T_{k+1} \cdots T_{-1} \quad (k < 0), \qquad \rho_k = 1 - (1-\rho_0) Z_1 Z_2 \cdots Z_k \quad (k > 0),$$

where T_k's are beta$(\theta, 1)$, *Z_k's are* beta$(\zeta, 1)$ *and the variables ρ_0, T_k $(k < 0)$ and Z_k $(k > 0)$ are all independent.*

Let $\mathcal{S}$ be the space of two-sided nondecreasing sequences $(x_k,\ k \in \mathbb{Z})$, $x_k \in [0,1]$, endowed with the product topology of $[0,1]^{\mathbb{Z}}$. Padding $\mathtt{rec}(\pi_n)$ by infinitely many 1's on the left and infinitely many n's on the right, and scaling by n makes $n^{-1}\mathtt{rec}(\pi_n)$ a random element of $\mathcal{S}$

$$n^{-1}\mathtt{rec}(\pi_n) = (\ldots, 1/n, 1/n, r_{-\ell}/n \ldots, r_{-1}/n, r_0/n, r_1/n, \ldots, r_u/n, 1, 1, \ldots).$$

Proposition 7 is a strong law of large numbers which says that $n^{-1}\mathtt{rec}(\pi_n)$ converge in $\mathcal{S}$ almost surely to a limit (ρ_k).

Recall that GEM(θ) distribution is the law of the sequence of gaps obtained by breaking $[0,1]$ at atoms of the Poisson point process with intensity $\theta \mathrm{d}x/x$ ($x \in [0,1]$) [1, 21]. The decreasing sequence of atoms has the same distribution as the sequence of 'stick-breaking' products $D_1, D_1D_2, \ldots$, with the D_j's being iid beta($\theta, 1$). The two-sided sequence $(\rho_k,\ k \in \mathbb{Z})$ is obtained in a similar way, by splitting $[0,1]$ at ρ_0, and further partitioning the intervals $[0,\rho_0]$ and $[\rho_0,1]$ by two independent beta stick-breakings with parameters θ and ζ. By analogy, the sequence of gaps $\rho_{k+1} - \rho_k,\ k \in \mathbb{Z}$, may be regarded as a two-sided version of the GEM distribution.

Generalising the classical case of sampling from iid uniforms [22, Proposition 4.11.2], the distribution of the bivariate point process of upper record values and their durations follows from the spraying property of Poisson processes. (The latter also applies for lower records, of course.) Thus, given ρ_0 the point processes $\{(\rho_k, t_{k+1} - t_k),\ k \geq 0\}$ and $\{(\rho_k, t_{k-1} - t_k),\ k \leq 0\}$, are independent Poisson, with intensity measures $\zeta x^{j-1}\mathrm{d}x$ on $[\rho_0, 1] \times \mathbb{N}$ and $\theta(1-x)^{j-1}\mathrm{d}x$ on $[0,\rho_0] \times \mathbb{N}$, respectively. In particular, by the projection property of Poisson processes, given ρ_0 the conditional distribution of the number of pairs of neighbouring lower records $\#\{k \leq 0 : t_{k-1} - t_k = 1\}$ is Poisson($\theta\rho_0$) (an equivalent result is shown in [13, Corollary 3.1] by a computation of moments).

9 Generating permutations from continuous variates

Under $P^{(\theta,\zeta)}$ not only the scaled record values converge (see Proposition 7), but also scaled permutations $(\pi_{nj}/n,\ j \in \mathbb{N})$ converge almost surely to some random sequence $(X_j) \in [0,1]^\infty$. In the case of uniform distribution $P^{(1,1)}$, the sequence (X_j) is just iid uniform$[0,1]$, and (π_n) can be generated by ranking (X_j), as we already mentioned. Under any $P^{(\theta,\zeta)}$, (X_j) can be produced by a kind of shuffling of the sequences of record values $(\rho_k,\ k \geq 0)$, $(\rho_k,\ k < 0)$ and another independent sequence of uniform variables. Here and henceforth, under shuffling of a few sequences we understand a sequence which is comprised of terms of all these sequences arranged in such a way that each of the sequences enters in its original order.

Construction 8 Let (W_n) be iid uniform$[0,1]$, independent of (ρ_k). We define a new sequence (X_n) where some W_n's are used, and some are replaced by ρ_k's which will appear as upper and lower record values. Start with $X_1 = \rho_1$. Suppose before step $n+1$ the values $\rho_{-\ell}, \ldots, \rho_u$ have been included into $X_1, \ldots, X_n$; then $\rho_u = \max(X_1, \ldots, X_n)$ and $\rho_{-\ell} = \min(X_1, \ldots, X_n)$. At step $n+1$ we let $X_{n+1} = \rho_{u+1}$ if $\pi_{n+1} > \rho_u$, or $X_{n+1} = \rho_{-\ell-1}$ if $\pi_{n+1} < \rho_{-\ell}$, or $X_{n+1} = \pi_{n+1}$ otherwise. Define a coherent sequence of permutations (π_n) by ranking (X_n).

It is obvious that, given (ρ_k), the sequence (X_n) resulting from the construction has the same law as iid uniform$[0,1]$ sequence conditioned on its two-sided sequence of record values (see [10] for the one-sided case of upper records). This works for any θ, ζ because conditionally given (ρ_k) the distribution of (π_n) under any $P^{(\theta,\zeta)}$ is the same as under the uniform distribution $P^{(1,1)}$.

For every fixed n a similar procedure yields uniform permutation π_n conditioned on $\mathtt{rec}(\pi_n)$. Start with setting $\pi_{n1} = r_0$. At each step $j > 1$ we will have $\pi_{n1}, \dots, \pi_{n,j-1}$ already determined, with some maximum $\max(\pi_{n1}, \dots, \pi_{n,j-1}) = r_{u'}$ and some minimum $\min(\pi_{n1}, \dots, \pi_{n,j-1}) = r_{-\ell'}$. At step $j \in \{2, \dots, n\}$ a value v is chosen uniformly at random from $[n] \setminus \{\pi_{n1}, \dots, \pi_{n,j-1}\}$. If $v < r_{-\ell'}$ let $\pi_{nj} = r_{-\ell'-1}$, if $v > r_{u'}$ let $\pi_{nj} = r_{u'+1}$, and if $r_{-\ell'} < v < r_{u'}$ let $\pi_{nj} = v$. The sampled value v is replaced each time v breaks the last upper or lower record. In n steps the increasing sequences $(r_{-\ell}, \dots, r_{-1})$, $(r_1, \dots, r_u)$ are shuffled with other elements of $[n]$. It is intiutively clear and not hard to show that, as n becomes large, $n^{-1}\mathtt{rec}(\pi_n) = n^{-1}(\dots, 1, r_{-\ell}, \dots r_{-1}, r_0, r_1, \dots, r_u, n, \dots)$ will converge in $\mathcal{S}$ to (ρ_k). This is just because sampling from large finite sets will have nearly the same effect as independent uniform choices from $[0, 1]$.

Apparently, from the viewpoint of statistical theory of extremes the sequence (X_n) is rather exotic, as it is chosen just to simulate desired behaviour of records. This differs general $P^{(\theta,\zeta)}$ from the uniform distribution $P^{(1,1)}$, when 'injecting' some extrinsic (ρ_k) is not at all necessary since the uniform sample (W_n) supplies automatically appropriate record values, so $(X_n) \stackrel{d}{=} (W_n)$. Still, in the case of integer parameters there is a simpler way to produce appropriate (X_n) from a sequence of uniforms, as parallels the construction of permutations in Proposition 3.

Construction for integer values of the parameters. The idea is to assume some 'prehistorical' sample of uniforms. Suppose $\theta \geq 1, \zeta \geq 1$ are integers. For $d = \theta + \zeta - 2$ let $V_1, \dots, V_d, W_1, W_2, \dots$ be iid uniform$[0, 1]$. At step 1 choose X_1 as the value of rank θ among $V_1, \dots, V_d, W_1$. At each step n we will have $\max(X_1, \dots, X_n)$ equal to the $(n - \theta + 1)$th order statistic in $V_1, \dots, V_d, W_1, \dots, W_n$, and $\min(X_1, \dots, X_n)$ equal to the θth order statistic in $X_1, \dots, X_d, W_1, \dots, W_n$. Now, if $W_{n+1} > \max(X_1, \dots, X_n)$ we set X_{n+1} equal to the $(n+\theta-1)$th order statistic in $V_1, \dots, V_d, W_1, \dots, W_n, W_{n+1}$, if $W_{n+1} < \min(X_1, \dots, X_n)$ we set X_{n+1} equal to the θth order statistic in the same sequence, and otherwise let $X_{n+1} = W_{n+1}$. This works, since there are always θ spacings below $\min(X_1, \dots, X_n)$ and ζ spacings above $\max(X_1, \dots, X_n)$, thus the resulting ranking is as in the proof of Proposition 3.

The described process shows that, for integer $\theta \geq 1, \zeta \geq 1$, Proposition 7 is a consequence of properties of the uniform order statistics. For all other values of θ, ζ the result can be interpolated from the integer case, because the law of each π_n is a rational function of the parameters of beta laws for T_k, Z_k.

References

[1] R. Arratia, A.D. Barbour and S. Tavaré, *Logarithmic combinatorial structures: a probabilistic approach*, European Math. Soc. Publ. House, Zürich, 2003.

[2] F.N. David and D.E. Barton, *Combinatorial chance*, Griffin & Co, London, 1962.

[3] P. Diaconis, M. McGrath and J. Pitman, Riffle shuffles, cycles and descents, *Combinatorica*, 15 (1995) 11-29.

[4] F.G. Foster and A. Stuart, Distribution-free tests in time-series based on the breaking of records, *J. R. Stat. Soc. Ser. B* 16 (1954) 1-22.

[5] A. Gnedin and U. Krengel, A stochastic game of optimal stopping and order selection, *Ann. Appl. Prob.* 5 (1995) 310-321.

[6] A. Gnedin and G. Olshanski, Coherent random permutations and the boundary problem for the graph of zigzag diagrams, *Int. Math. Res. Notes* Article ID 51968 (2006) 39 pp.

[7] A. Gnedin and G. Olshanski, The boundary of the Eulerian number triangle, *Moscow Math. J.* 6 (2006) 461-475.

[8] A. Gnedin and J. Pitman, Exchangeable Gibbs partitions and Stirling triangles, *Zapiski POMI (St. Petersburg Dept. Steklov Math. Inst.)* 325 (2005) 82-105, also *J. Math. Sci.* 138 (2006) 5674-5685.

[9] C.M. Goldie, Records, permutations and greatest convex minorants, *Math. Proc. Camb. Phil. Soc.* 106 (1989) 169-177.

[10] C.M. Goldie and J. Bunge, Record sequences and their applications, in *Handbook of Statistics vol. 19,* (*Stochastic Processes: Theory and Methods*, Shanbhag, D.N. and Rao, C.R. eds), pp. 277-308, North Holland, Amsterdam, 1999.

[11] J. Hajek, Z. Sidak and P.K. Sen, *Theory of rank tests*, Academic Press, 1998.

[12] T.P. Hill and D.P. Kennedy, Sharp inequalities for optimal stopping with rewards based on ranks, *Ann. Appl. Prob.* 2 (1992) 503-517.

[13] L. Holst, On the number of consecutive successes in Bernoulli trials, (2006) preprint.

[14] S.V. Kerov, Subordinators and the permutation actions with quasi-invariant measure, *J. Math. Sci.* 87 (1997) 4094-4117.

[15] S.V. Kerov and N.V. Tsilevich, A random subdivision process generates virtual permutations with Ewens distribution, *J. Math. Sci.* 87 (1997) 4082-4093.

[16] S.V. Kerov, G. Olshanski and A.M. Vershik, Harmonic analysis on the infinite symmetric group, *Comptes Rend. Acad. Sci. Paris*, 316 (1993) 773-778.

[17] S. Lalley, Riffle shuffles and their associated dynamical systems, *J. Theor. Prob.* **12** (1999) 903-932.

[18] H. Mahmoud, P. Flajolet, P. Jacquet and M. Régnier, Analytic variations on bucket selection and sorting. *Acta. Inform.* 36 (2000) 735-760.

[19] V.B. Nevzorov, *Records*, Transl. Math. Monographs, Providence, AMS, 2001.

[20] J. Pitman, An extension of de Finetti's theorem, *Adv. Appl. Prob.* 10 (1978) 268-270.

[21] J. Pitman, *Combinatorial stochastic processes*, Lecture Notes Math. vol. 1875, Springer, NY, 2006.

[22] S. Resnick, *Adventures in stochastic processes*, Birkhäuser, Boston, 1992.

[23] R. Stanley, *Enumerative Combinatorics* vol. 1, Wadsworth & Brooks/Cole, 1986.