



Constrained exchangeable partitions

Alexander Gnedin

► To cite this version:

Alexander Gnedin. Constrained exchangeable partitions. Fourth Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities, 2006, Nancy, France. pp.391-398, 10.46298/dmtcs.3479 . hal-01184681

HAL Id: hal-01184681

<https://inria.hal.science/hal-01184681>

Submitted on 17 Aug 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Constrained exchangeable partitions

Alexander Gnedin^{1†}

¹ *Mathematical Institute, Utrecht University, P.O. Box 80010, 3508 TA Utrecht, The Netherlands*

For a class of random partitions of an infinite set a de Finetti-type representation is derived, and in one special case a central limit theorem for the number of blocks is shown.

Keywords: exchangeability, paintbox, stick-breaking

1 Introduction

Under a *partition* of the set \mathbb{N} we shall mean a sequence (b_1, b_2, \dots) of subsets of \mathbb{N} such that (i) the sets b_j are disjoint, (ii) $\cup_j b_j = \mathbb{N}$, (iii) if $b_k = \emptyset$ then also $b_{k+1} = \emptyset$ and (iv) if $b_{k+1} \neq \emptyset$ then $\min b_k < \min b_{k+1}$. Condition (iv) says that the sequence of minimal elements of the blocks is increasing. One can think of partition as a mapping which sends a generic element $j \in \mathbb{N}$ to one of the infinitely many blocks, in such a way that conditions (iii) and (iv) are fulfilled.

A random partition $\Pi = (B_k)$ of \mathbb{N} (so, with random blocks B_k) is a random variable with values in the set of partitions of \mathbb{N} . This concept can be made precise by means of a projective limit construction and the measure extension theorem. To this end, one identifies Π with consistent partitions $\Pi_n := \Pi|_{[n]}$ ($n = 1, 2, \dots$) of finite sets $[n] := \{1, \dots, n\}$. Note that the restriction Π_n , which is obtained by removing all elements not in $[n]$, still has the blocks in the order of increase of their least elements.

There is a well developed theory of exchangeable partitions [1; 13; 17]. Recall that $\Pi = (B_j)$ is *exchangeable* if the law of Π is invariant under all bijections $\sigma : \mathbb{N} \rightarrow \mathbb{N}$. Partitions with weaker symmetry properties have also been studied. Pitman [16] introduced *partially exchangeable* random partitions of \mathbb{N} with the property that the law of Π is invariant under all bijections $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ that preserve the order of blocks, meaning that the sequence of the least elements of the sets $\sigma(B_1), \sigma(B_2), \dots$ is also increasing. Pitman [16] derived a de Finetti-type representation for partially exchangeable partitions and established a criterion for their exchangeability. Kerov [14] studied a closely related structure of virtual permutations of \mathbb{N} , which may be seen as partially exchangeable partitions with some total ordering of elements within each of the blocks. Kallenberg [13] characterised *spreadable* partitions whose law is invariant under increasing injections $\sigma : \mathbb{N} \rightarrow \mathbb{N}$.

In this note we consider constrained random partitions of \mathbb{N} which satisfy the condition that, for a fixed integer sequence $\rho = (\rho_1, \rho_2, \dots)$ with $\rho_k \geq 1$, each block B_k contains ρ_k least elements of $\cup_{j \geq k} B_j$, for every k with $B_k \neq \emptyset$. It is easy to check that this condition holds if and only if the sequence comprised of ρ_1 least elements of B_1 , followed by ρ_2 least elements of B_2 and so on, is itself an increasing sequence. We shall focus on the constrained partitions with the following symmetry property.

Definition 1 For a given sequence ρ , we call Π *constrained exchangeable* if Π is a constrained partition with respect to ρ and the law of Π is invariant under all bijections $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ that preserve this property.

Since the law of Π is uniquely determined by the laws of finite restrictions Π_n , the constrained exchangeability of Π amounts to the analogous property of Π_n 's for each $n = 1, 2, \dots$. To gain a feeling of the property, the reader is suggested to check that for $\rho = (1, 2, 1, \dots)$ the partition Π_8 assumes the values $(\{1, 3, 5\}, \{2, 4, 6\}, \{7, 8\})$ and $(\{1, 2, 3\}, \{4, 5, 8\}, \{6, 7\})$ with the same probability.

Every partition of \mathbb{N} is constrained with respect to $\rho = (1, 1, \dots)$, and every constrained exchangeable partition with this ρ is partially exchangeable in the sense of Pitman [16]. In principle, any constrained exchangeable partition may be reduced to some Pitman's partially exchangeable partition by isolating $\rho_k - 1$ least elements of B_k in $\rho_k - 1$ singleton blocks, for each $\rho_k > 1$, but this viewpoint will not be adopted here.

For general $\rho \neq (1, 1, \dots)$ the constrained exchangeable partitions which are also exchangeable are rather uninteresting, since they cannot have infinitely many blocks:

[†] gnedin@math.uu.nl

Proposition 2 *Let Π be a constrained partition with respect to some ρ which has $\rho_k > 1$ for some k . If Π is exchangeable then Π has at most k nonempty blocks.*

Proof: Suppose $B_k \neq \emptyset$, then by Kingman's representation of exchangeable partitions [17] the set $\cup_{j \geq k} B_j$ contains infinitely many elements. For the same reason $\#B_k \geq 2$ implies that B_k is an infinite set, and that partition Π' obtained by restricting Π to $\cup_{j \geq k} B_j$ and re-labelling the elements of $\cup_{j \geq k} B_j$ by \mathbb{N} in increasing order is an exchangeable partition of \mathbb{N} . But then with probability one Π' is the trivial single-block partition, because elements 1 and 2 are always in the same block. \square

In many contexts where random partitions appear, exchangeability is an obvious kind of symmetry. Constrained exchangeability may appear when some initial elements of the blocks play a special role of 'establishing' the block. To illustrate, consider the following situation. Suppose there is a sequence of independent random points sampled from some distribution on \mathbb{R}^d . Define D_1 as the convex hull of the first ρ_1 points, D_2 as the convex hull of the first ρ_2 points not in D_1 , D_3 as the convex hull of the first ρ_3 points not in $D_1 \cup D_2$, etc. Divide \mathbb{R}^d in disjoint nonempty subsets $G_1 = D_1, G_2 = D_2 \setminus D_1, G_3 = D_3 \setminus (D_1 \cup D_2), \dots$. A constrained exchangeable partition Π of \mathbb{N} is defined then by assigning to block B_k the indices of ρ_k initial points that determine D_k and the indices of all further points that hit G_k .

Of course, there is nothing special in the convex hulls construction, and any other way of 'spanning' a spatial domain D_k on ρ_k sample points and then 'peeling' the space in G_k 's will also result in a constrained exchangeable partition. An example of this kind related to multidimensional records will be given.

In what follows we extend Pitman's [16] sequential realisation of partitions via frequencies of blocks, to cover arbitrary constrained exchangeable partitions. Generalising a result on exchangeable partitions [6] we shall also derive a central limit theorem for the number of blocks of finite partitions $\Pi_n = \Pi|_{[n]}$ in one important case of partitions induced by a 'stick-breaking' scheme.

2 Constrained sampling

We fix throughout a sequence of positive integers ρ . Recall that a *composition* is a finite sequence of positive integers called parts, e.g. $(3, 1, 2)$ is a composition of $6 = 3 + 1 + 2$ with three parts. We say that a composition $\lambda = (\lambda_1, \dots, \lambda_\ell)$ is a *constrained composition of n* if $\lambda_j \geq \rho_j$ for $j = 1, \dots, \ell - 1$ and $|\lambda| := \sum \lambda_j = n$.

For each λ a constrained composition of n , the following random algorithm, which may be called *constrained sampling*, yields another constrained composition μ of $n - 1$. Imagine a row of boxes labeled $1, \dots, \ell$ and occupied by $\lambda_1, \dots, \lambda_\ell$ white balls. Let $\Lambda_j := \lambda_j + \dots + \lambda_\ell$, $j \leq \ell$. At the first step, ρ_1 balls in box 1 are re-painted black and then a white ball is drawn uniformly at random from all $\Lambda_1 - \rho_1$ white balls. If the ball drawn was in box 1, the ball is deleted and the new composition is $\mu = (\lambda_1 - 1, \lambda_2, \dots, \lambda_\ell)$, and if the ball drawn was in some other box, it is returned to the box and the process continues, so that at the second step ρ_2 balls in box 2 are re-painted black and a white ball is drawn uniformly at random from boxes $2, \dots, \ell$. If the second ball drawn was in box 2, the ball is deleted and the new composition is $\mu = (\lambda_1, \lambda_2 - 1, \lambda_3, \dots, \lambda_\ell)$, and so on. If the procedure does not terminate in $\ell - 1$ steps, then a ball is deleted from the last box and the new composition is $\mu = (\lambda_1, \dots, \lambda_{\ell-1}, \lambda_\ell - 1)$. By this description, for $j < \ell$ the transition probability from λ to $\mu = (\dots, \lambda_j - 1, \dots)$ is

$$\frac{\Lambda_2}{(\Lambda_1 - \rho_1)} \cdots \frac{\Lambda_j}{(\Lambda_{j-1} - \rho_{j-1})} \frac{(\lambda_j - \rho_j)}{(\Lambda_j - \rho_j)},$$

while the transition probability from λ to $\mu = (\dots, \lambda_\ell - 1)$ is

$$\frac{\Lambda_2}{(\Lambda_1 - \rho_1)} \cdots \frac{\Lambda_\ell}{(\Lambda_{\ell-1} - \rho_{\ell-1})}.$$

A random sequence $\mathcal{C} = (\mathcal{C}_n)$ of constrained compositions of integers $n = 1, 2, \dots$ is called *consistent* if \mathcal{C}_{n-1} has the same law as the composition derived from \mathcal{C}_n by the above constrained sampling procedure, for each $n > 1$. Every consistent sequence (\mathcal{C}_n) is an inverse Markov chain with some co-transition probabilities depending only on ρ . By analogy with [5] a consistent sequence \mathcal{C} will be called a *constrained composition structure*.

If the constraints are determined by $\rho = (1, 1, \dots)$, the constrained sampling amounts to a co-transition rule related to the partially exchangeable partitions in [16]. The unconstrained sampling (corresponding to $\rho = (0, 0, \dots)$) leads to composition structures studied in [5; 9; 10].

3 Basic representation

For Π a constrained partition of \mathbb{N} with blocks B_1, B_2, \dots we define, for each $n = 1, 2, \dots$, a composition \mathcal{C}_n of n as the finite sequence of positive values in $\#(B_1 \cap [n]), \#(B_2 \cap [n]), \dots$. We call this composition the *shape* of Π_n and write $\mathcal{C}_n = \text{shape}(\Pi_n)$.

The number of constrained partitions of $[n]$ with shape λ is equal to

$$d(\lambda) := \prod_{j=1}^{\ell-1} \binom{\Lambda_j - \rho_j}{\lambda_j - \rho_j}. \quad (1)$$

Similarly, the number of partitions of $[n]$ with shape $\lambda = (\lambda_1, \dots, \lambda_\ell)$ and whose restriction on $[n']$ (for $n' < n$) has shape $\mu = (\mu_1, \dots, \mu_k)$ is equal to

$$d(\lambda, \mu) := \left[\prod_{j=1}^{\ell-1} \binom{M_j - \Lambda_j}{\mu_j - \lambda_j} \right] \binom{M_\ell - \Lambda_\ell - (\rho_\ell - \lambda_\ell)_+}{\mu_\ell - \rho_\ell \vee \lambda_\ell} \left[\prod_{j=\ell+1}^{k-1} \binom{M_j - \rho_j}{\mu_j - \rho_j} \right], \quad (2)$$

where $M_j = \mu_j + \dots + \mu_k$, $j \leq k$.

Introduce a function of compositions

$$p(\lambda) := \mathbb{P}(\text{shape}(\Pi_n) = \lambda).$$

It is easy to check that the consistency of Π_n 's with respect to restrictions implies that the \mathcal{C}_n 's are consistent in the sense of constrained sampling, therefore appealing to Kolmogorov's measure extension theorem we have:

Proposition 3 *The formula*

$$\mathbb{P}(\Pi_n = \cdot) = p(\text{shape}(\cdot))/d(\text{shape}(\cdot))$$

establishes a canonical homeomorphism between the distributions of constrained exchangeable partitions of \mathbb{N} and constrained composition structures. Conditionally given $\mathcal{C}_n = \text{shape}(\Pi_n) = \lambda$ the distribution of Π_n is uniform on the set of constrained partitions of $[n]$ with shape λ .

The following basic construction modifies the one exploited in [14; 16]. Let (P_1, P_2, \dots) be an arbitrary sequence of random variables satisfying $P_k \geq 0$ and $\sum_k P_k \leq 1$. A constrained exchangeable partition Π directed by (P_k) is defined as follows. Conditionally given (P_k) the partition is obtained by successive extension of Π_n to Π_{n+1} , for each $n = 1, 2, \dots$, according to the rules: given Π_n with $\text{shape}(\Pi_n) = (\lambda_1, \dots, \lambda_\ell)$, the element $n+1$

- (i) joins the block B_j , $j < \ell$, with probability P_j ,
- (ii) if $\lambda_\ell < \rho_\ell$ joins the block B_ℓ with probability $1 - \sum_{j=1}^{\ell-1} P_j$,
- (iii) and if $\lambda_\ell \geq \rho_\ell$ joins the block B_ℓ with probability P_ℓ or starts the new block $B_{\ell+1}$ with probability $1 - \sum_{j=1}^{\ell} P_j$.

Explicitly, for the function p of compositions we have the formula

$$p(\lambda) = d(\lambda) \mathbb{E} \left\{ \left[\prod_{j=1}^{\ell-1} \left(1 - \sum_{i=1}^{j-1} P_i \right)^{\rho_j} P_j^{\lambda_j - \rho_j} \right] \left(1 - \sum_{i=1}^{\ell-1} P_i \right)^{\rho_\ell \wedge \lambda_\ell} P_\ell^{(\lambda_\ell - \rho_\ell)_+} \right\}. \quad (3)$$

The next de Finetti-type result states that the construction covers all possible constrained exchangeable partitions. The proof is only sketched, since it follows the same lines as in [14; 16].

Proposition 4 *For Π a constrained exchangeable partition, the normalised shapes $\text{shape}(\Pi_n)/n$ (considered as sequences padded by infinitely many zeroes) converge in the product topology with probability one to some random limit (P_1, P_2, \dots) satisfying $P_j \geq 0$ and $\sum_j P_j \leq 1$. Conditionally given (P_k) the partition Π is recovered according to the above rules (i)-(iii).*

Proof: The key point is to show the existence of frequencies. This can be concluded from de Finetti's theorem for 0–1 exchangeable sequences by noting that the indicators $1(m \text{ belongs to block } B_k)$ for $m > n$ are conditionally exchangeable given that the block B_k has at least ρ_k representatives in $[n]$. Alternatively, one can use, as in [14], more direct Martin boundary arguments which exploit the explicit formulas (1) and (2) to show that the pointwise limit of ratios $d(\cdot, \mu)/d(\mu)$, as $m = |\mu| \rightarrow \infty$, exists if and only if μ_j/m converge for every j . \square

4 The formation sequence

Nacu [15] established that the law of a partially exchangeable partition is uniquely determined by the law of the increasing sequence of the least elements of blocks. We show that a similar result holds for every constrained exchangeable partition Π with general ρ .

We define the *formation sequence* to be the sequence obtained by selecting the ρ_k th least element of the block B_k of Π , for $k = 1, 2, \dots$. For a composition λ let $q(\lambda)$ be the probability that the formation sequence starts with elements $\lambda_1, \lambda_1 + \lambda_2, \dots, \lambda_1 + \dots + \lambda_\ell$. Let (P_j) be the frequencies as in Section 3, and introduce the variables

$$H_k = 1 - \sum_{j=1}^k P_j, \text{ so } P_k = H_{k-1} - H_k,$$

where we set $H_0 = 1$. Then

$$q(\lambda_1, \dots, \lambda_\ell) = \mathbb{E} \left[\prod_{j=1}^{\ell-1} \binom{\lambda_{j+1} - 1}{\rho_{j+1} - 1} H_j^{\rho_{j+1}} (1 - H_j)^{\lambda_{j+1} - \rho_{j+1}} \right]. \quad (4)$$

Comparing this with (3) written in the same variables (where $H_0 = 1$) we obtain for constrained compositions

$$p(\lambda) = d(\lambda) \mathbb{E} \left\{ \left[\prod_{j=1}^{\ell-1} H_{j-1}^{\rho_j} (H_{j-1} - H_j)^{\lambda_j - \rho_j} \right] H_{\ell-1}^{\rho_\ell \wedge \lambda_\ell} (H_{\ell-1} - H_\ell)^{(\lambda_\ell - \rho_\ell)_+} \right\}, \quad (5)$$

which leads to the following conclusion:

Proposition 5 *There is an invertible linear transition from p to q . Hence each of these two functions on compositions uniquely determines the law of Π .*

Proof: The substantial part of the claim is showing that we can compute p from q . To that end, start by observing that p is uniquely determined by the values on compositions of the type $(\lambda_1, \dots, \lambda_{\ell-1}, \rho_\ell)$. To see that this follows from the consistency for various n , argue by induction in $m = 0, \dots, \rho_\ell$ for compositions $(\dots, \rho_\ell - m)$. Now, for such compositions whose last part meets the constraint exactly, (5) and (4) involve the same factors of the type $H_j^{\rho_j}$, hence p can be reduced to q by expanding each factor $(H_{j-1} - H_j)^k = ((1 - H_j) - (1 - H_{j-1}))^k$ using the binomial formula. \square

5 The paintbox

Paintbox representations based on the uniform sampling from $[0, 1]$ are often used to model exchangeable structures and their relatives [5; 8; 9; 17]. We shall design a version that is appropriate for constrained exchangeable partitions.

Let $1 = H_0 \geq H_1 \geq H_2 \geq \dots \geq 0$ be an arbitrary nonincreasing random sequence. Let (U_n) be a sequence of independent $[0, 1]$ -uniform random points, independent of (H_k) . We define a new sequence $(U_n) \mid_\rho (H_k)$ with some of U_n 's replaced by H_k 's, as follows. Replace U_1, \dots, U_{ρ_1} by H_1 . Then replace the first ρ_2 entries which belong to $U_{\rho_1}, U_{\rho_1+1}, \dots$ and hit $[0, H_1[$ by H_2 . Inductively, when H_1, \dots, H_k get used, respectively, ρ_1, \dots, ρ_k times, keep on screening uniforms until replacing the first ρ_{k+1} points hitting $[0, H_k[$ by H_{k+1} . Eventually all H_k 's will enter the resulting sequence.

The construction has an interpretation in terms of the classical theory of records (see [7; 14] for a special case).

Proposition 6 *Conditionally given (H_k) , the sequence $(U_n) \mid_\rho (H_k)$ has the same distribution as (U_n) conditioned on the event that the sequence of lower records in (U_n) is (H_k) , with the record value H_k repeated ρ_k times.*

In this framework, we define a partition Π by assigning to block B_k all integers which label the entries of $(U_n) \mid_\rho (H_k)$ falling in $[H_k, H_{k-1}[$. Given (H_k) , the chance for U_n to hit $[H_j, H_{j-1}[$ is $P_j = H_{j-1} - H_j$, therefore the construction is equivalent to that defined by the rules (i)-(iii) above.

6 Stick-breaking partitions

Explicit evaluation of the function p is possible when the frequencies involve a kind of independence. To this end, it is convenient to introduce yet another set of variables (W_k) (sometimes called *residual fractions*) which satisfy $H_k = W_1 \cdots W_k$, $W_k \in [0, 1]$. In these variables (5) becomes

$$p(\lambda) = d(\lambda) \mathbb{E} \left\{ \left[\prod_{j=1}^{\ell-1} W_j^{\Lambda_j - \lambda_j} (1 - W_j)^{\lambda_j - \rho_j} \right] (1 - W_\ell)^{(\lambda_\ell - \rho_\ell)_+} \right\} \quad (6)$$

where $\Lambda_j = \lambda_j + \dots + \lambda_\ell$. As in [16], if the W_k 's are independent, (6) assumes the product form

$$p(\lambda) = \prod_{k=1}^{\ell} q_k(\Lambda_k : \lambda_k) \quad (7)$$

with the *decrement matrices*

$$q_k(n : m) = \binom{n - \rho_k}{m - \rho_k} \mathbb{E} [(1 - W_k)^{m - \rho_k} W_k^{n - m}] , \quad 1 \leq m \leq n, \quad (8)$$

and the convention $\binom{-i}{-j} = 1 (i = j)$ for negative arguments of the binomial coefficients. In fact, (7) forces representation (8) (this fact is implicit in [14; 16] in the case of partially exchangeable partitions):

Proposition 7 *A constrained composition structure (\mathcal{C}_n) satisfies (7) with some decrement matrices q_k , $k = 1, 2, \dots$, if and only if there exist independent $[0, 1]$ -valued random variables (W_k) such that (8) holds.*

Proof: We argue the ‘only if’ part. For $n < \rho_1$ we have $q_1(n : n) = 1$ by definition. For $n \geq \rho_1$ the constrained sampling consistency yields

$$q_1(n : m) = \frac{m + 1 - \rho_1}{n + 1 - \rho_1} q_1(n + 1 : m + 1) + \frac{n + 1 - m}{n + 1 - \rho_1} q_1(n + 1 : m)$$

which is the familiar Pascal-triangle recursion in the variables $n - \rho_1$, $m - \rho_1$, therefore the integral representation (8) follows as a known consequence of the Hausdorff moments problem. The case $k > 1$ is completely analogous. The independence of the W_k 's is obvious from (7). \square

We note in passing that the product formula (7) with a single decrement matrix leads, in a related setting of regenerative composition structures, to a nonlinear recursion and a very different conclusion [9]. See [11] for product formulas of another kind in the exchangeable case.

Suppose now that W_k 's are independent and have beta(a_k, b_k) distributions, whose density is

$$(1 - s)^{a_k - 1} s^{b_k - 1} / B(a_k, b_k).$$

The rows of the decrement matrices are then Pólya-Eggenberger distributions

$$q_k(n : m) = \binom{n - \rho_k}{m - \rho_k} \frac{(a_k)_{m - \rho_k} (b_k)_{n - m}}{(a_k + b_k)_{n - \rho_k}}, \quad m = 1, \dots, n.$$

For instance, taking positive integer a_k, b_k and $\rho_k = a_k + b_k - 1$, a partition Π is constructed as follows: replace U_1, \dots, U_{ρ_1} by the value H_1 equal to the b_1 th minimal order statistic of these points, then replace the first ρ_2 uniforms that hit $[0, H_1[$ by the value H_2 equal to the b_2 th minimal order statistic of these hits, etc, thus defining a partition via $(U_n) |_{\rho}(H_k)$. A distinguished class of structures of this kind is the Ewens-Pitman two-parameter family of exchangeable partitions [12; 16; 17] with $\rho = (1, 1, \dots)$, $a_k = \theta + k\alpha$ and $b_k = 1 - \alpha$ (for suitable α and θ); the product formula simplifies in this case due to a major telescoping of factors.

7 Counting the blocks

Let K_n be the number of blocks of Π_n , which in the $(U_j) |_{\rho}(H_k)$ -representation coincides with the number of intervals $]H_1, H_0],]H_2, H_1], \dots$ discovered by the n first terms of the sequence. Conditionally given (H_k) , K_n is the number of certain independent geometric summands which sum to no more than n . In

particular, the difference between the k th and the $(k+1)$ st entries of the formation sequence follows the negative binomial distribution with parameters ρ_k, H_k .

We shall proceed by assuming a ‘stick-breaking’ scheme $H_k = W_1 \cdots W_k$, $k = 1, 2, \dots$, with independent identically distributed W_k ’s. We assume further that the logarithmic moments $\mu = \mathbb{E}[-\log W_1]$, $\sigma^2 = \text{Var}[-\log W_1]$ are both finite. The idea is to derive a CLT for K_n from the standard CLT for renewal processes [4]. Similar technique was used in [6; 7], but in the new situation we need to also limit the growth of ρ_k as $k \rightarrow \infty$.

We will show that K_n is asymptotic to $J_n := \max\{k : H_k > 1/n\}$. The last quantity is indeed asymptotically Gaussian with the mean $(\log n)/\mu$ and the variance $(\log n)/(\sigma^2 \mu^{-3})$ because J_n is just the number of renewal epochs within $[0, \log n]$ of the renewal process with steps $-\log W_k$. In loose terms, we will exploit a ‘cut-off phenomenon’: typically, only a few points out of n uniforms fall below $1/n$, while for $k < J_n$ essentially all intervals get hit, with exponentially growing occupancy numbers when scanned backwards in k from $k = J_n$.

Proposition 8 *Suppose Π is directed by $H_k = W_1 \cdots W_k$, where for $k = 1, 2, \dots$ the W_k ’s are i.i.d. with finite logarithmic moments $\mu = \mathbb{E}[-\log W_1]$, $\sigma^2 = \text{Var}[-\log W_1]$. If*

$$\log \left[\sum_{j=1}^k \rho_j \right] = o(k), \text{ as } k \rightarrow \infty, \quad (9)$$

then the strong law of large numbers holds, i.e. $K_n \sim \mu^{-1} \log n$ a.s.. Moreover, the random variable $(K_n - \mathbb{E}[K_n])/\sqrt{\text{Var}[K_n]}$ converges in law to the standard Gaussian distribution, whereas the moments satisfy

$$\mathbb{E}[K_n] \sim \frac{\log n}{\mu}, \quad \text{Var}[K_n] \sim \frac{\log n}{\sigma^2 \mu^{-3}}. \quad (10)$$

Proof: By the construction of $(U_j) \mid \rho(H_k)$, we have a dichotomy: $U_n \in]H_k, H_{k-1}]$ implies that either U_n will enter the transformed sequence or will get replaced by some $H_i \geq H_k$. Let $U_{1n} < \dots < U_{nn}$ be the order statistics of U_1, \dots, U_n . It follows that

- (i) if $U_{jn} > H_k$ then $K_n \leq j + k$,
- (ii) if $U_{mn} < H_k$ for $m = \sum_{i=1}^k \rho_i$ then $K_n \geq k$.

Define ξ_n by $U_{\xi_n, n} < 1/n < U_{\xi_n+1, n}$ and recall that J_n was defined by $H_{J_n+1} \leq 1/n < H_{J_n}$, thus ξ_n is the number of uniforms to the left of $1/n$, and J_n follows the CLT. Clearly, J_n and ξ_n are independent and ξ_n is binomial($n, 1/n$). By (i), we have $K_n \leq J_n + \xi_n$ where ξ_n is approximately Poisson(1), which yields the desired upper bound.

The lower bound is more delicate. Introduce

$$\psi_n = c \sum_{j=1}^{\lfloor \log n \rfloor} \rho_j$$

where c should be selected sufficiently large. Then by the assumption (9) $\log \psi_n = o(\log n)$, which is enough to assure that the number, say L_n , of the H_k ’s larger than ψ_n/n is still asymptotic to J_n . Because L_n is close to Gaussian with moments as in (10), an easy large deviation estimate implies that the inequality $L_n < (c/2) \log n$ holds with probability at least $1 - n^{-2}$. On the other hand, the number of uniforms smaller ϕ_n/n is also close to Gaussian with both central moments about ϕ_n , hence, in view of $\phi_n > c \log n$, a similar estimate shows that this number is at least $(c/2) \log n$ with probability at least $1 - n^{-2}$. By application of (ii) with $k = L_n$ we see that the lower bound $K_n > L_n$ holds up to an event of probability $O(n^{-2})$. This completes the proof of the CLT. Finally, since both J_n and L_n are asymptotic to $\mu^{-1} \log n$ almost surely, the Borel-Cantelli lemma implies that the same is valid for K_n . \square

8 A continuous time process

The sequential construction of Π from the frequencies (P_k) can be embedded in continuous time by letting the elements $1, 2, \dots$ arrive at epochs of a rate-1 Poisson process on \mathbb{R}_+ . Let R_t be the total frequency of

the blocks which are not represented by the elements arrived before t , then $R = (R_t)$ is a nonincreasing pure-jump process with piecewise-constant paths and $R_0 = 1$.

Suppose as in Section 7 that W_k 's are independent and identically distributed. The process R is then easy to describe: if after the $(k-1)$ st jump the process R is in state s then the time in this state has distribution $\text{gamma}(\rho_k, s)$, and thereafter the state is changed to sW_k . The sojourns in consecutive states $1, W_1, W_1W_2, \dots$ are independent. The instance $\rho = (1, 1, \dots)$ corresponds to a known self-similar Markov process which appears as a 'tagged particle' process in random fragmentation models [1]. For general ρ the process is no longer Markovian, as one needs to also include the time spent in the current state to summarise the history.

A minor adjustment of Proposition 8 to the continuous-time setting allows to conclude that under the same assumptions the number of jumps during the time $[0, T]$ is approximately Gaussian, as $T \rightarrow \infty$. In fact, the process R is well defined for arbitrary positive values ρ_k ($k = 1, 2, \dots$), in which case an analogous CLT is readily acquired by interpolation from the case of integer ρ_k 's.

9 Example: the chain records

Next is an example of Pitman's partially exchangeable partitions, so the constraint is $\rho = (1, 1, \dots)$. Consider a Borel space \mathcal{Z} endowed with a distribution μ and some measurable strict partial order \prec . For a sample V_1, V_2, \dots from (\mathcal{Z}, μ) , we say that a *chain record* occurs at index j if either $j = 1$, or $j > 1$ and V_j is \prec -smaller than the last chain record in the sequence V_1, \dots, V_{j-1} . The instance of \mathbb{R}^d with the natural coordinate-wise partial order was discussed in [7].

Let R_k , $k = 1, 2, \dots$, be the sample values when the chain records occur; the sequence (R_k) is a 'greedy' falling chain of the partially ordered sample (V_j) . Introducing the lower sets $L_v := \{u \in \mathcal{Z} : u \prec v\}$, we define $D_k := L_{R_k}$, $G_k := D_k \setminus D_{k-1}$ (where $D_0 := \emptyset$), and we define a constrained exchangeable partition $\Pi = (B_k)$ as in Section 1. The frequencies of B_k 's are $P_k = \mu(L_{G_k})$, and we have $H_k = \mu(L_{R_k})$, as is easily seen.

To guarantee a 'stick-breaking' form of (H_k) , as in Section 6, we need to assume a self-similarity property of the sampling space. We may call $(\mathcal{Z}, \mu, \prec)$ *regenerative* if (i) $\mu(L_v) > 0$ for μ -almost all points $v \in \mathcal{Z}$, and (ii) the lower section L_v with conditional measure $\mu(\cdot)/\mu(L_v)$ is isomorphic, as a partially ordered probability space, to the whole space $(\mathcal{Z}, \mu, \prec)$. Since all L_v 's are in this sense the same, the H_k 's undergo stick-breaking with i.i.d. residual fractions whose distribution is the same as that of L_{V_1} . Under the hypothesis of Proposition 8, the number of chain records among the first n sample points is approximately Gaussian, since this number coincides with the number of blocks of Π_n . A class of regenerative spaces is comprised of the Bollobás-Brightwell box-spaces [3], which have all intervals $\{u : v \prec u \prec w\}$ for $v \prec w$ isomorphic to the whole space (and not only lower sections).

Further examples of regenerative spaces appear, in a disguise, in the context of multidimensional data structures like quad-trees or simplex-trees [2]. More generally, constrained exchangeability appears in connection with data structures which allow multiple key storage at a node of the search tree.

References

- [1] J. Bertoin, *Random fragmentation and coagulation processes*, Cambridge Univ. Press., 2006.
- [2] L. Devroye, Universal limit laws for depths in random trees, *SIAM J. Comp.* 28 (1999) 409-432.
- [3] B. Bollobás and G. Brightwell, Box spaces and random partial orders, *Trans. Amer. Math. Soc.* 124 (1991) 59-72.
- [4] W. Feller, *An introduction to probability theory and its applications*, vol. 2, Wiley, NY, 1971.
- [5] A. V. Gnedin, The representation of composition structures, *Ann. Probab.* 25 (1997) 1437-1450.
- [6] A.V. Gnedin, The Bernoulli sieve, *Bernoulli* 10 (2004) 79-96.
- [7] A.V. Gnedin, Counting the chain records: the product case, 2005 arXiv: math.PR/0510042.
- [8] A.V. Gnedin and G.I. Olshanski, Coherent random permutations with descent statistic and the boundary problem for the graph of zigzag diagrams, *Int. Math. Res. Notes* (2006) Article ID 51968.
- [9] A. V. Gnedin and J. Pitman, Regenerative composition structures, *Ann. Probab.* 33 (2005) 445-479.

- [10] A. V. Gnedin and J. Pitman, Regenerative partition structures, *Elec. J. Comb.* 11 (2005) paper R12.
- [11] A. V. Gnedin and J. Pitman, Gibbs exchangeable partitions and Stirling triangles, *Zapiski Nauchnykh Seminarov POMI* 325 (2005) 82-103.
- [12] A. V. Gnedin and J. Pitman, Self-similar and Markov composition structures, *Zapiski Nauchnykh Seminarov POMI* 326 (2005) 59-84.
- [13] O. Kallenberg, *Probabilistic symmetries and invariance principles*, Springer, NY, 2005.
- [14] S. Kerov, Subordinators and permutations action with quasi-invariant measure, *Zapiski Nauchnykh Seminarov POMI* 223 (1995) 181-218 (translated in *J. Math. Sci.* (New York)).
- [15] S. Nacu, Increments of random partitions, *Combinatorics, Probability and Computing* (2006) arXiv:math.PR/0310091
- [16] J. Pitman, Exchangeable and partially exchangeable random partitions. *Prob. Th. Rel. Fields* 102 (1995) 145-158.
- [17] J. Pitman, *Combinatorial stochastic processes* (Lect. Notes for 2002 St. Flour Course), Springer L. Notes Math. 2006.