Combinatorics, Probability and Computing (2006) 15, 589–595. © 2006 Cambridge University Press
doi:10.1017/S0963548305007455 Printed in the United Kingdom

Increments of Random Partitions

ŞERBAN NACU[†]

Department of Statistics, University of California, Berkeley, CA 94720, USA
(e-mail: serban@stat.berkeley.edu
http://stat-www.berkeley.edu/~serban)

Received 10 October 2003; revised 19 November 2003

For any partition of $\{1,2,\ldots,n\}$ we define its *increments* $X_i, 1 \leqslant i \leqslant n$ by $X_i = 1$ if i is the smallest element in the partition block that contains it, $X_i = 0$ otherwise. We prove that for partially exchangeable random partitions (where the probability of a partition depends only on its block sizes in order of appearance), the law of the increments uniquely determines the law of the partition. One consequence is that the Chinese Restaurant Process CRP(θ) (the partition with distribution given by the Ewens sampling formula with parameter θ) is the only exchangeable random partition with independent increments.

1. Introduction

Random partitions have been studied extensively during the past thirty years, and have found various applications in population biology, Bayesian statistics, combinatorics, and statistical physics; see [2] for an in-depth survey. Exchangeable random partitions were introduced by Kingman, motivated by applications in genetics. Partially exchangeable random partitions were introduced by Pitman in [1]. We recall their definition.

Definition. Consider a random partition $\Pi_n = \{A_1, A_2, \ldots, A_k\}$ of $[n] = \{1,2,\ldots,n\}$, where the blocks A_i are listed in order of appearance (*i.e.*, in increasing order of their smallest element). Π_n is called *partially exchangeable* if its probability only depends on the ordered block sizes

$$P(\Pi_n) = p(|A_1|, \ldots, |A_k|) \tag{1.1}$$

for some function p taking values on the set of *compositions* of n (ordered sets of positive integers that sum up to n). Π_n is called *exchangeable* if p is symmetric, so its probability depends on the sizes of the blocks but not their order.

[†] Research supported in part by NSF Grant DMS-0071448.

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

A natural way to look at a partition is to construct it one element at a time. We start with a single block $\{1\}$, then the next element 2 either joins the existing block or starts a new one, and so on until n. For $i = 1, \ldots, n$ let $X_i = 1$ if i starts a new block, $X_i = 0$ otherwise. (Alternatively, $X_i = 1$ if i is the smallest element in its block, $X_i = 0$ otherwise.) We always define $X_1 = 1$.

Definition. Let Π_n be a partition of $\{1, 2, \ldots, n\}$. Let $X_i = 1$ if i is the smallest element in its block, 0 otherwise. We call the sequence $(X_1, \ldots, X_n)$ the *increments* of Π_n.

Clearly the partition determines the increments, but not *vice versa*. The increments do not even determine the block sizes: the partitions $\{\{1, 3, 4\}, \{2\}\}$ and $\{\{1, 3\}, \{2, 4\}\}$ have the same increments $X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0$.

For a random partition, the law of the partition induces a law for its increments on the probability space $\{0, 1\}^n$. We prove that if the partition is partially exchangeable, the law of the increments *does* determine the law of the partition.

Theorem 1.1. *If Π_n is a partially exchangeable partition of $[n]$, the distribution of its increments $(X_1, \ldots, X_n)$ uniquely determines the distribution of Π_n.*

Hence we obtain a correspondence between the set of partially exchangeable laws for partitions, and the set of laws for binary sequences of zeros and ones. This correspondence is as close to a bijection as we could possibly hope for, in the following sense. There are exactly 2^{n-1} compositions of n, so according to (1.1), the law of the partition is determined by the 2^{n-1} values of p, which are arbitrary except for the constraint that P is a probability measure (so $p \geqslant 0$ and a weighted sum of its values is equal to 1). There are 2^{n-1} possible sequences of increments (since $X_1 = 1$ always), so their law is also determined by 2^{n-1} numbers; the constraint on them turns out to be a system of linear inequalities (plus the obvious constraint that they sum up to 1).

As an application of Theorem 1.1, we answer a question raised by Jim Pitman.

Definition. Let $\theta > 0$. The *Chinese Restaurant Process* with parameter θ is the exchangeable random partition of $[n]$ with distribution given by

$$p(n_1, \ldots, n_k) = \theta^k \prod_{i=1}^{k} (n_i - 1)! / \prod_{i=0}^{n-1} (\theta + i). \tag{1.2}$$

We denote it by CRP(θ).

Equation (1.2) is equivalent to the well-known Ewens sampling formula, and the distribution of the block sizes of CRP(θ) is also referred to as the Ewens partition structure. It has also been referred to as the Blackwell–MacQueen distribution. See [3] for a survey. The name 'Chinese Restaurant Process' was introduced by Lester Dubins and Jim Pitman in the early 1980s.

CRP(θ) has several equivalent descriptions; for example, for $\theta = 1$ it is the partition induced by the cycles of an uniform random permutation. See [2] for details. The

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

description we are interested in is in terms of its increments X_i. It follows from (1.2) that those are independent and satisfy

$$P(X_i = 1) = \theta/(i - 1 + \theta). \tag{1.3}$$

Hence CRP(θ) admits a simple construction one element at a time: i starts a new block with probability $\theta/(i-1+\theta)$ (and joins an existing block with probability proportional to the size of the block). It is easy to prove that this constructs an exchangeable partition. Jim Pitman asked whether this is the only exchangeable random partition with the property that X_i are independent. We prove that the answer is 'yes'.

Theorem 1.2. *The Chinese Restaurant Process* CRP(θ) *is the only exchangeable random partition with independent increments.*

Hence in this sense, the Chinese Restaurant Process is the simplest exchangeable random partition.

Several other random partitions admit simple representations in terms of their increments X_i. If the increments are not independent, the next simplest case is to assume some kind of Markov structure. For example, we can require that the partial sums $S_i = X_1 + \cdots + X_i$ form a Markov chain. This is the same as requiring that the probability that $i+1$ starts a new block depends only on i and on the number of already existing blocks S_i. One process that satisfies this is Pitman's two-parameter generalization of CRP(θ), described in [1, 2]. In this case we have

$$P(X_{i+1} = 1 | S_i = k) = (k\alpha + \theta)/(i + \theta),$$

where the parameters satisfy $0 \leqslant \alpha < 1, \theta > -\alpha$. For $\alpha = 0$ we obtain CRP(θ). It is an open question to describe all random partitions for which S_i is a Markov chain.

2. Proof of Theorem 1.1

Let $\Pi_n = \{A_1, A_2, \ldots, A_k\}$ be a partition of $[n]$ with k blocks. We list the blocks A_i in order of appearance, so $1 \in A_1$, the smallest element not in A_1 is in A_2, and so on. Let $B(\Pi_n) = (|A_1|, \ldots, |A_k|)$ the sizes of its blocks in order of appearance.

Let $(X_1, \ldots, X_n)$ be the increments of the partition. Since there are k blocks, there are exactly k ones and $n-k$ zeros among the increments. We can encode such a binary sequence by the distance between consecutive 1s. For example, the sequence $1, 0, 0, 1, 1, 0, 1, 0$ will be encoded $3, 1, 2, 2$, as the distance between the first and the second 1 is three, the distance between the second and the third 1 is one, and so on. Formally, if a_i is the smallest element in A_i, then we know $X_{a_i} = 1$ for all $1 \leqslant i \leqslant k$, so we encode the increments as the k-tuple $D(\Pi_n) = (a_2 - a_1, a_3 - a_2, \ldots, a_k - a_{k-1}, n + 1 - a_k)$. There is a bijection between such k-tuples and binary sequences with $X_1 = 1$, so from now on we will identify the sequence of increments with its encoding, and work directly with k-tuples.

Let $S_{n,k}$ be the set of k-tuples whose elements add up to n. We are interested in the relationship between $B(\Pi_n)$ and $D(\Pi_n)$; both are in $S_{n,k}$. We define a partial order relation

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

on $S_{n,k}$ as follows: $(y_1,\dots,y_n) \geqslant (z_1,\dots,z_n)$ if and only if $y_1 \geqslant z_1, y_1+y_2 \geqslant z_1+z_2,\dots,y_1+y_2+\cdots+y_n \geqslant z_1+z_2+\cdots+z_n$. Then we have the following result.

Lemma 2.1. *For any partition* Π_n, $B(\Pi_n) \geqslant D(\Pi_n)$.

Proof. Let $B(\Pi_n)=(b_1,\dots,b_k)$, $D(\Pi_n)=(d_1,\dots,d_k)$. Fix m with $1 \leqslant m \leqslant k$. Clearly $\sum_{i=1}^m b_i = |\bigcup_{i=1}^m A_i|$ and $\sum_{i=1}^m d_i = a_{m+1}-a_1 = a_{m+1}-1$, where a_{m+1} is the smallest element in A_{m+1}. Since the blocks A_i are listed in order of appearance, a_{m+1} is the smallest element outside $\bigcup_{i=1}^m A_i$, so it is at most $|\bigcup_{i=1}^m A_i|+1$; equality occurs if and only if $\bigcup_{i=1}^m A_i = \{1,2,\dots,a_{m+1}-1\}$. □

Now consider a partially exchangeable law for Π_n, defined as in (1.1) by a function p. This induces a law for the increments of Π_n, and by using the encoding of binary sequences into k-tuples discussed above, this induces a law on k-tuples:

$$q(d_1,\dots,d_k) = P(D(\Pi_n)=(d_1,\dots,d_k)). \tag{2.1}$$

Summing up over all partitions with the same block structure, we obtain

$$q(d_1,\dots,d_k) = \sum_{(b_1,\dots,b_k)\in S_{n,k}} p(b_1,\dots,b_k)r(d_1,\dots,d_k;b_1,\dots,b_k), \tag{2.2}$$

where $r(a_1,\dots,a_k;b_1,\dots,b_k)$ denotes the number of partitions Π_n with blocks $B(\Pi_n)=(b_1,\dots,b_k)$ and increments encoded as $D(\Pi_n)=(d_1,\dots,d_k)$.

This gives a system of linear equations in the qs and the ps; we will show it can be solved to compute p in terms of q. Consider the dictionary order on $S_{n,k}$: $(y_1,\dots,y_k) \geqslant_d (z_1,\dots,z_k)$ if and only if $y_1 > z_1$ or $y_1=z_1,\dots,y_m=z_m$ and $y_{m+1} > z_{m+1}$ for some m or $y_i = z_i$ for all i. It is easy to see that if $(y_1,\dots,y_k) \geqslant (z_1,\dots,z_k)$ in the partial order previously defined, then $(y_1,\dots,y_k) \geqslant_d (z_1,\dots,z_k)$ in the dictionary order. Hence from Lemma 2.1 we obtain the following result.

Lemma 2.2. *For* $\mathbf{y},\mathbf{z} \in S_{n,k}$, $r(\mathbf{y};\mathbf{z}) = 0$ *unless* $\mathbf{z} \geqslant \mathbf{y}$, *and* $r(\mathbf{y};\mathbf{z}) = 1$ *if* $\mathbf{z}=\mathbf{y}$.

Now $S_{n,k}$ is totally ordered under the dictionary order so we can arrange its elements in decreasing order. But then the lemma says that the matrix of the system of linear equations is triangular and its diagonal elements are all 1, so it is trivially invertible and p can be computed in terms of q. Explicitly, if the elements of $S_{n,k}$ are $\mathbf{y_1} >_d \mathbf{y_2} >_d \mathbf{y_3} > \cdots$, then $p(\mathbf{y_1}) = q(\mathbf{y_1})$ and for $i \geqslant 2$, $p(\mathbf{y_i}) = q(\mathbf{y_i}) - \sum_{j=1}^{i-1} p(\mathbf{y_j})r(\mathbf{y_i};\mathbf{y_j})$. This completes the proof of Theorem 1.1. □

The following result gives an explicit formula and a generating function for the coefficients $r(\cdot;\cdot)$.

Proposition 2.3. *For* $(d_1,\dots,d_k) \in S_{n,k}$ *and* $(b_1,\dots,b_k) \in S_{n,k}$, *let* $\mathcal{M}$ *be the set of* $k \times k$ *square matrices* $M=(m_{ij})$ *with the following properties:*

(i) M *is upper triangular, so* $m_{ij}=0$ *if* $i > j$,

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

(ii) *all other entries are nonnegative integers, so* $m_{ij} \geqslant 0$ *if* $i \leqslant j$,
(iii) *for all* i, *the* i*-th row sums up to* $b_i - 1$,
(iv) *for all* i, *the* i*-th column sums up to* $d_i - 1$.

Then

$$r(d_1, \ldots, d_k; b_1, \ldots, b_k) = \sum_{M \in \mathcal{M}} \prod_i (d_i - 1)! / \prod_{i,j} (m_{ij})! \tag{2.3}$$

Hence the following generating function identity holds:

$$x_1^{d_1 - 1}(x_1 + x_2)^{d_2 - 1} \cdots (x_1 + x_2 + \cdots + x_k)^{d_k - 1} = \sum_{\mathbf{b} \in S_{n,k}} r(d_1, \ldots, d_k; b_1, \ldots, b_k) x_1^{b_1 - 1} x_2^{b_2 - 1} \cdots x_k^{b_k - 1} \tag{2.4}$$

Proof. We need to count the number of partitions with block structure **b** and increment structure **d**. Let A_i be the blocks of such a partition, in order of appearance. Knowing **d** is equivalent to knowing the smallest element of each A_i, call them a_i. Let $D_i = \{a_i, a_i + 1, \ldots, a_{i+1} - 1\}$ and let $n_{ij} = |A_i \bigcap D_j|$. Of the elements in D_j, a_j must belong to A_j; all others may be assigned to any A_i with $i \leqslant j$, and the number of ways in which this can be done is $(d_j - 1)!/(n_{1j}! n_{2j}! \cdots n_{j-1\,j}! (n_{jj} - 1)!)$. If we let $m_{ij} = n_{ij}$ if $i \neq j$, $m_{ii} = n_{ii} - 1$, then we obtain the desired formula. The generating function identity follows easily. □

3. Partitions with independent increments

We now prove Theorem 1.2.

Proof. Let Π_n be an exchangeable random partition of $[n]$ such that its increments $X_1, \ldots, X_n$ are independent. As above, let p be the function that describes its distribution as in (1.1), and let q be the joint distribution of the increments. Consider sequences of increments with $k = n - 1$; that is, there is only one zero among $X_1, \ldots, X_n$. If the zero is at the beginning ($X_2 = 0$) then the partition must be $\Pi_n = \{\{1, 2\}, \{3\}, \ldots, \{n\}\}$. Hence

$$q(1, 0, 1, \ldots, 1) = p(2, 1, \ldots, 1).$$

If the zero is at the end ($X_n = 0$) then n could be in any of the $n - 1$ pre-existing blocks, so there are $n - 1$ choices for Π_n. Since Π_n is exchangeable, all these choices are equally likely. Hence

$$q(1, 1, 1, \ldots, 1, 0) = (n - 1)p(2, 1, \ldots, 1).$$

But if we let $u_n = P(X_n = 1)$, by independence we also have

$$q(1, 0, 1, \ldots, 1) = u_1(1 - u_2)u_3 \cdots u_n$$

and

$$q(1, 1, 1, \ldots, 1, 0) = u_1 u_2 \cdots u_{n-1}(1 - u_n).$$

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

If $u_2 \neq 0, 1$ then it follows easily by induction that $u_n \neq 0$ for all n, and hence

$$(n-1)(1-u_2)/u_2 = (1-u_n)/u_n,$$

so if we let $\theta = u_2/(1-u_2)$ then $u_n = \theta/(n-1+\theta)$. Hence the increments of the process have the same law as the increments of CRP(θ), so the process must be CRP(θ).

It remains to consider the cases when u_2 is 0 or 1. If $u_2 = 1$ then by induction $u_n = 1$, so all blocks are singletons (this is CRP(θ) in the limit case $\theta = \infty$). If $u_2 = 0$ then all $u_n = 0$ (Π_n cannot contain the singleton $\{1\}$ so by exchangeability it cannot contain the singleton $\{n\}$ either) so there is only one block (CRP(θ) for $\theta = 0$). □

4. Another binary representation

Theorem 1.1 allows us to obtain a partially exchangeable partition from any law on random binary sequences satisfying certain constraints. While we have proved the theorem for partitions of the finite set $[n]$, the result is easily extended to infinite partitions. We thus obtain a correspondence between the set of distributions of partially exchangeable partitions of $\mathbb{N}$, and the set of distributions of infinite binary sequences. By Theorem 1.1, the correspondence is one-to-one; it is not onto.

There is another way to associate infinite binary sequences to partitions, which is discussed in detail in [4, Chapter 4]. Given a binary sequence, the problem is to construct an exchangeable random partition Π, so that the distribution of the (unordered) block sizes of its restriction Π_n to $[n]$ is the same as the distribution of the 'gaps' (distances) between consecutive 1s in the binary sequence. More precisely, we have the following definition.

Definition. Let $Y_1, Y_2, \ldots$ be a random infinite sequence with $Y_1 = 1$ and $Y_n \in \{0,1\}\ \forall n$, and let Π be an exchangeable random partition of $\mathbb{N}$. Let $1 = n_1 < n_2 < \cdots$ be the locations of the 1s in the sequence Y_n. For any fixed n, let $n_k = \max\{i : i \leqslant n, X_i = 1\}$. Then

$$n = (n_2 - n_1) + \cdots + (n_k - n_{k+1}) + (n + 1 - n_k)$$

is a partition of n into k integers. If this has the same distribution as the partition of n induced by restricting Π to $[n]$, then we say that Π has a *gap representation* by Y.

The two binary representations are related, but not identical. In particular, while any random partition has a representation by increments, it is not known under what conditions a random partition admits a gap representation. Also, note that the gap representation is interesting only for infinite partitions and sequences; for fixed n, any law for the sequence easily translates into a law for an exchangeable partition, as we are free to specify probabilities for all possible block sizes. The problem is whether these laws are compatible as n varies.

In [4], gap representations are constructed for various partitions, including CRP(θ), for which the gap representation has $P(Y_n = 1) = \theta/(n-1+\theta)$ and the Y_n are independent. It is also proved that the only exchangeable random partition which admits a gap

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press

representation via a sequence Y of independent binary random variables is CRP(θ). The similarity to Theorem 1.2 may seem surprising, but it is explained by the following result.

Proposition 4.1. *Suppose* Π *has a gap representation by* Y, *and let* X *be the increments of* Π. *Then*

$$X_1 + \cdots + X_n \stackrel{d}{=} Y_1 + \cdots + Y_n, \quad \forall n \geqslant 1. \tag{4.1}$$

Proof. Both sides of the identity are equal in distribution to the number of blocks in Π_n. □

If X and Y are each sequences of independent variables, then (4.1) implies they have the same distribution (in fact, it is enough to assume that the sequences of partial sums $X_1 + \cdots + X_n$ and $Y_1 + \cdots + Y_n$ are (possibly non-homogenous) Markov chains).

Acknowledgements

Many thanks to Jim Pitman for suggesting the problem and for useful discussions. Thanks also to Noam Berger.

References

[1] Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probab. Theory Rel. Fields* **102** 145–158.
[2] Pitman, J. (2002) *Combinatorial Stochastic Processes*. Lecture Notes for the St. Flour School.
[3] Tavaré, S. and Ewens, W. J. (1998) The Ewens Sampling Formula. In *Encyclopedia of Statistical Sciences*, Update Vol. 2 (S. Kotz, C. B. Read, D. L. Banks, eds), Wiley, New York, pp. 230–234.
[4] Young, J. (1995) Partition-valued stochastic processes with applications. PhD Thesis, UC Berkeley.

https://doi.org/10.1017/S0963548305007455 Published online by Cambridge University Press