Combinatorics, Probability and Computing (2001) 10, 213–217. Printed in the United Kingdom
© 2001 Cambridge University Press

A Characterization of GEM Distributions

ALEXANDER GNEDIN[1] and SERGEI KEROV[2]†

[1] School of Mathematics and Statistics, University of Newcastle,
Newcastle upon Tyne NE1 7RU, England
(e-mail: Alexander.Gnedin@ncl.ac.uk)

[2] Steklov Mathematical Institute at St. Petersburg (POMI),
Fontanka 27, St. Petersburg, 191011, Russia
(e-mail: kerov@pdmi.ras.ru)

Received 29 November 1999; revised 21 June 2000

Let (X_n) be a residual allocation model with i.i.d. residual fractions U_n. For W a random variable with values in $[0,1]$ and independent of (X_n), we define another sequence (Y_n) by setting

$$(Y_1, Y_2, Y_3, \ldots) = \begin{cases} (W, X_1 - W, X_2, X_3, \ldots), & \text{if } W < X_1, \\ (X_1 + X_2, X_3, X_4, \ldots), & \text{if } W \geqslant X_1. \end{cases}$$

Under minor regularity assumptions we show that (X_n) and (Y_n) have the same probability law if and only if this law is a GEM distribution. In this case, the distribution of W and the U_ns is Beta$(1,\theta)$ for some $\theta > 0$.

1. The GEM distribution named after Griffiths, Engen and McCloskey is a probability law for the sequence (X_n) arising as a *residual allocation model* (RAM)

$$X_1 = U_1, \quad X_n = (1-U_1)(1-U_2)\ldots(1-U_{n-1})\,U_n, \ \ n = 2,3,\ldots \tag{1}$$

where the *residual fractions* $U_1, U_2, \ldots$ are i.i.d. with Beta$(1,\theta)$ density

$$\theta(1-x)^{\theta-1}, \qquad 0 < x < 1,$$

for some $0 < \theta < \infty$. The GEM distribution appears in connection with the famous Poisson–Dirichlet distribution and the Ewens sampling formula, and plays a key role

† Partially supported by grant RFBR-NNIO 96-01-00096.

https://doi.org/10.1017/S0963548301004692 Published online by Cambridge University Press

in various sampling and combinatorial problems (see Chapter 41 of [1] for history and background).

A characterization result by McCloskey says that the GEM distribution is the only RAM which has i.i.d. residual fractions and is invariant under a *size-biased permutation*. That is to say, if (X_n) is regarded as a random discrete distribution over $\{1,2,\ldots\}$ and $N_1, N_2, \ldots$ are different integers, as they appear for the first time in the course of sequential sampling from this distribution, then the probability law of (X_{N_n}) coincides with that of (X_n). See Pitman [5] for a proof of the characterization and its extension to arbitrary independent fractions and Gnedin [2] for general properties of the size-biased permutations.

In this note, we introduce a transformation which we call *split-and-merge*, and characterize the GEM sequences as the only RAMs which are invariant under this transformation. In sharp contrast to the size-biased permutation, which affects only the arrangement of components in the sequence (X_n) and not their sizes, the split-and-merge transformation alters the sizes by either splitting X_1 in two components or merging X_1 with X_2. The motivation for this development comes from our study of asymptotic combinatorial properties of the Young–Fibonacci graph, where a discrete version of the split-and-merge transformation governs the evolution of Fibonacci words, and the GEM $(1/2)$ distribution appears as a limit of the Plancherel measure [3].

2. Let (X_n) be a random positive sequence with probability zero for ties, such that $\sum_n X_n = 1$. For a random variable W with values in $[0,1]$ and independent of (X_n), we define the *split-and-merge transformation* $(Y_n) = T(W,(X_n))$ by setting

$$(Y_n) = \begin{cases} (W, X_1 - W, X_2, X_3, \ldots), & \text{if } W < X_1, \\ (X_1 + X_2, X_3, X_4, \ldots), & \text{if } W \geqslant X_1. \end{cases}$$

It is very intuitive to think of T in terms of the interval partitions. Consider the partition of $[0,1]$ into sub-intervals by random *cuts* $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$. Pick a random point W from a given probability distribution. If W falls in the first partition interval then a new cut W is added; otherwise the cut X_1 is erased and the first two intervals are merged.

We call (X_n) *invariant* under the split-and-merge transformation directed by W if (Y_n) has the same probability distribution as (X_n). The transformation $T(W,\cdot)$ can be regarded as a stochastic kernel defined on the infinite-dimensional simplex $\Delta = \{(x_i) : \sum x_i \leqslant 1, x_i \geqslant 0\}$ and a finite-invariant measure on Δ corresponds to some invariant sequence (X_n).

We will assume from here on that the directing variable W has a strictly increasing distribution function on $[0,1]$. Our main result says that GEM sequences are the only RAM sequences with i.i.d. continuously distributed residual fractions that are invariant with respect to some $T(W,\cdot)$. More precisely, we will prove the following theorem.

Theorem 1. *Suppose that* (X_n) *is a residual allocation model* (1) *where* $U_1, U_2, \ldots$ *are i.i.d. random variables with a continuous distribution function. Then the sequence* (X_n) *is invariant*

https://doi.org/10.1017/S0963548301004692 Published online by Cambridge University Press

under the split-and-merge transformation directed by W if and only if W and the U_ns have the same Beta $(1, \theta)$ *distribution with some $\theta > 0$.*

3. We will find it convenient to deal with the distribution functions

$$f(t) = P(1 - U_n < t) \quad \text{and} \quad g(t) = P(1 - W < t)$$

of the complementary variables $1 - U_n$ and $1 - W$. The first step in proving the theorem is to show that the invariance can be translated as an integral equation on f and g.

Lemma 2. *A RAM sequence with i.i.d. residual fractions is invariant if and only if f and g satisfy the equation*

$$\int_a^1 (f(u) - f(bu))\, dg(u) + (1 - f(b)) \int_a^1 (1 - f(au^{-1}))\, g(u)\, df(u) = (1 - f(a))\,(1 - f(b)) \quad (2)$$

for $a, b \in [0, 1]$.

Proof. The residual fractions for the transformed sequence $(Y_n) = T(W, (X_n))$ are

$$V_1 = Y_1; \qquad V_n = \frac{Y_n - Y_{n-1}}{1 - Y_{n-1}}, \qquad n = 2, 3, \ldots.$$

One can easily express V_n in terms of the variables U_n as

$$(V_1, V_2, V_3, V_4, \ldots) = \begin{cases} (W, (U_1 - W)/(1 - W), U_2, U_3, \ldots), & \text{if } W < U_1, \\ U_1 + (1 - U_1)\, U_2, U_3, U_4, U_5, \ldots), & \text{if } W \geqslant U_1. \end{cases}$$

The invariance holds if and only if the sequence (V_n) has the same distribution as (U_n). Clearly, the variables $V_3, V_4, \ldots$ are independent and have the same distribution as $U_3, U_4, \ldots$; moreover, they are also independent of V_1 and V_2. Therefore, we only need to examine whether the joint distribution of V_1, V_2 coincides with that of U_1, U_2.

To determine the joint distribution of V_1, V_2 write the probability

$$P(V_1 < 1 - a, V_2 < 1 - b) \quad (3)$$

as the sum of two terms. The first term is

$$\begin{aligned} P(V_1 < 1 - a, \quad V_2 < 1 - b, \quad W < U_1) \\ &= P(1 - W > a, \quad 1 - W > 1 - U_1 > (1 - W)b) \\ &= \int_a^1 \big(f(u) - f(bu)\big)\, dg(u), \end{aligned}$$

and the second term is

$$\begin{aligned} P(V_1 < 1 - a, \quad V_2 < 1 - b, \quad W \geqslant U_1) \\ &= P((1 - U_1)(1 - U_2) > a, \quad 1 - U_3 > b, \quad 1 - W < 1 - U_1) \\ &= \big(1 - f(b)\big)\, P(1 - U_2 > a/(1 - U_1), \quad 1 - W \leqslant 1 - U_1) \\ &= (1 - f(b)) \int_a^1 (1 - f(au^{-1}))\, g(u)\, df(u). \end{aligned}$$

https://doi.org/10.1017/S0963548301004692 Published online by Cambridge University Press

We see that (3) is equal to the left-hand side of (2). Because $P(U_1 < 1-a, U_2 < 1-b) = (1-f(a))(1-f(b))$, the statement follows. □

Now, we need to verify that a sequence (X_n) with GEM(θ) distribution is invariant under $T(W, \cdot)$, provided W has the Beta$(1,\theta)$ probability law. To this end, we just plug

$$f(u) = g(u) = u^\theta, \qquad u \in [0,1]$$

into (2). Since for such f we have $f(u) - f(bu) = u^\theta(1-b^\theta) = f(u)(1-f(b))$, the integral equation takes the form

$$\int_a^1 f(u)\, dg(u) + \int_a^1 g(u)\, df(u) = 1 - f(a) + \int_a^1 f(au^{-1})\, g(u)\, df(u).$$

Integrating by parts, we write this as

$$\int_a^1 f(au^{-1})\, g(u)\, df(u) = f(a)\,\big(1-g(a)\big). \tag{4}$$

Because $f(au^{-1})g(u) = f(a)$, both sides are equal and the invariance follows.

4. The uniqueness claim follows from the next lemma.

Lemma 3. *Suppose that the distribution functions $f(t)$ and $g(t)$ satisfy equation* (2). *Then for some $\theta > 0$ the variables U_n and W have one and the same* Beta$(1,\theta)$ *distribution.*

Proof. Write (2) as

$$\int_a^1 (f(u) - f(bu))dg(u) = (1-f(b))\,(1-f(a) - \int_a^1 (1-f(au^{-1}))\, g(u)\, df(u))$$

and observe that the right-hand side is of the form $(1-f(b))h(a)$. Setting $b = 0$ yields

$$h(a) = \int_a^1 f(u)\, dg(u),$$

and the equation reads as

$$\int_a^1 f(bu)\, dg(u) = f(b) \int_a^1 f(u)\, dg(u).$$

Varying a we obtain

$$f(bu) = f(b)\, f(u) \tag{5}$$

for g-almost all $u \in [0,1]$. Since g was assumed to be strictly monotone, any set of the full g-measure is dense in $[0,1]$. It now follows from the monotonicity of f that (5) holds for arbitrary b and u. The only monotone solution of this equation is $f(u) = u^\theta$ with some $\theta > 0$.

Let us now check that $g(u) = f(u)$ for all $u \in [0,1]$. We can start with equation (4), which in view of (5) takes the form

$$f(a) \int_a^1 g(u)\, \frac{df(u)}{f(u)} = f(a)\,\big(1-g(a)\big).$$

https://doi.org/10.1017/S0963548301004692 Published online by Cambridge University Press

Dividing by $f(a)$ and differentiating, we arrive at

$$g'(a) + \frac{\theta g(a)}{a} = 0.$$

It is clear that $g(a) = a^\theta$ is the only solution with the initial condition $g(0) = 0$. □

Theorem 1 follows from Lemmas 2 and 3 and the invariance.

5. An invariant sequence exists for arbitrary split-and-merge transformation, provided W has a continuous distribution function G. To see this, consider the generalized inverse $G^{\leftarrow}$ which transforms the standard uniform distribution into G. If (X_n) has the GEM(1) distribution, then the sequence $(G^{\leftarrow}(X_n))$ is invariant, as it follows readily from the invariance of (X_n) under the split-and-merge transformation directed by uniform W.

From this viewpoint, each GEM(θ) sequence is a $G^{\leftarrow}$-image of the GEM(1)-sequence, although in general $G^{\leftarrow}$ does not keep the RAM structure. It is therefore natural to ask about the uniqueness of the $T(\cdot, W)$-invariant sequence in the broad sense, that is, without restricting to the class of residual allocation models. In [4] we have proved that, for W uniformly distributed, the uniqueness is indeed the case: for each $c \in (0, 1]$ there is a unique nonnegative invariant sequence (X_n) satisfying $\sum_n X_n \equiv c$. For $c = 1$ the unique invariant sequence is GEM(1), and for $c < 1$ it is just a truncated version of the GEM(1) sequence. By a transformation technique the uniqueness result is easily extended to arbitrary continuous distributions G.

Acknowledgement

The work of S. Kerov was partially supported by grant RFBR–NNIO and the University of Göttingen. Helpful comments by one referee are gratefully acknowledged.

References

[1] Johnson, N., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*, Wiley, New York.
[2] Gnedin, A. (1998) On convergence and extensions of size-biased permutations. *J. Appl. Probab.* **35** 642–650.
[3] Gnedin, A. and Kerov, S. (2000) The Plancherel measure of the Young–Fibonacci graph. *Math. Proc. Camb. Phil. Soc.* **129** 433–446.
[4] Gnedin, A. and Kerov, S. (2000) Fibonacci solitaire. Preprint.
[5] Pitman, J. (1996) Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Probab.* **28** 525–539.

https://doi.org/10.1017/S0963548301004692 Published online by Cambridge University Press