A Characterization of GEM Distributions

Alexander Gnedin and Sergei Kerov *
University of Göttingen and *POMI, St. Petersburg*

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

$$(Y_1, Y_2, Y_3, Y_4, \ldots) = \begin{cases} (W, X_1 - W, X_2, X_3, \ldots) & \text{if } W < X_1, \\ (X_1 + X_2, X_3, X_4, X_5, \ldots) & \text{if } W \geq X_1. \end{cases}$$

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_n's is beta $(1, \theta)$ for some $\theta > 0$.

Keywords: GEM distribution, residual allocation model, Poisson-Dirichlet distribution, size-biased permutation, merge-and-split transformation.

AMS 1991 Subject Classification: Primary 62E10.

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 is closely connected with the famous Poisson-Dirichlet distribution and the Ewens sampling formula. It plays a key role in various sampling and combinatorial problems (see Chapter 41 of [1] for the 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 *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

*Partially supported by the Grant RFBR–NNIO 96-01-00096, and RFBR 96-15-96060

(X_n). See Pitman [4] for a proof of the characterization and its extension to independent arbitrarily distributed residual fractions and Gnedin [2] for general properties of size-biased permutations.

In this note we introduce a transformation which we call *split-and-merge transformation*, and we characterize the GEM sequences as the only RAM which are invariant under this transformation. In sharp contrast with the size-biased permutation which affects only the arrangement of components in the sequence (X_n) but not their sizes, the split-and-merge transformation affects the sizes by either splitting X_1 in two components or merging X_1 with X_2. The impetus for this development comes from our study of asymptotic combinatorics of the Young-Fibonacci modular lattice, where a discrete version of the split-and-merge transformation governes 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 sequence with probability zero for ties and such that $0 < X_n < 1$ and $\sum_n X_n = 1$. For a random variable W in $[0,1]$ independent of (X_n) we define the *split-and-merge transformation* $(Y_n) = T(W,(X_n))$ *directed by* W by setting

$$(Y_1, Y_2, Y_3, Y_4, \ldots) = \begin{cases} (W, X_1 - W, X_2, X_3, \ldots), & \text{if } W < X_1, \\ (X_1 + X_2, X_3, X_4, X_5, \ldots), & \text{if } W \geq X_1. \end{cases}$$

An intuitive way to think of T is in terms of the interval partitions. Consider the partitioning of $[0,1]$ by random cuts $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$. Pick a random point W from some 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). One can regard $T(W, \cdot)$ as a Markov transition operator on an infinite dimensional simplex. In these terms, each stationary distribution for T corresponds to some invariant sequence (X_n).

3. We will assume in the sequel that the directing variable W has a strictly increasing distribution function on $[0,1]$. Here is our main result.

Theorem 1 *Assume 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 under the split-and-merge transformation directed by W if and only if W and the U_n's have the same beta$(1,\theta)$ ditribution for some $\theta > 0$.*

The proof of the theorem will be given in Sections 4-6.

Corollary 2 *GEM sequences are the only RAM sequences with i.i.d. continuously distributed residual fractions which are invariant under some merge-and-split process.*

4. 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 $1-W$, where U_n and W are as in the condition of Theorem 1. The first step to the proof is to show that the invariance can be translated as an integral equation on f and g.

Proposition 3 *A RAM sequence with i.i.d. residual fractions is invariant if and only if f and g satisfy the equation*

$$\begin{aligned} \int_a^1 (f(u)-f(bu))\, dg(u) & \quad + \\ (1-f(b)) \int_a^1 \left(1-f\left(\frac{a}{u}\right)\right) g(u)\, df(u) & = (1-f(a))\,(1-f(b)) \end{aligned} \tag{2}$$

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

Proof. The residual fractions for the trasformed 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

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

The invariance holds iff 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 independent of V_1 and V_2. Therefore, we only need to say when 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 we write the probability

$$P(V_1<1-a, V_2<1-b) \tag{3}$$

as the sum of two terms,

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

and

$$\begin{aligned} P(V_1<1-a, \quad V_2<1-b, \quad W\ge U_1) & = \\ P((1-U_1)(1-U_2)>a, \quad 1-U_3>b, \quad 1-W<1-U_1) & = \\ (1-f(b))\, P(1-U_2>a/(1-U_1), \quad 1-W\le 1-U_1) & = \\ (1-f(b)) \int_a^1 \left(1-f\left(\frac{a}{u}\right)\right) g(u)\, df(u). & \end{aligned}$$

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

5. To verify that a sequence (X_n) with GEM(θ) distribution is invariant under $T(W,\cdot)$ with beta $(1,\theta)$ distribution for W, we just need to plug

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

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

$$\int_a^1 f(u)\, dg(u) + \int_a^1 g(u)\, df(u) = 1 - f(a) + \int_a^1 f\left(\frac{a}{u}\right) g(u)\, df(u),$$

or, integrating by parts,

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

Since $f(a/u)g(u) = f(a)$, both sides are equal and we are done.

6. The uniqueness claim follows from the next proposition.

Proposition 4 *Suppose that the distribution functions $f(t) = P(1-U_n < t)$ and $g(t) = P(1-W < t)$ satisfy equation (2). Then the variables U_n, W have equal beta$(1,\theta)$ distributions for some $\theta > 0$.*

Proof. Write (2) as

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

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 can be written 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 $[0,1]$, thus it 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 the equation (4) which takes the form

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

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

$$a\, g'(a) + \theta g(a) = 0.$$

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

7. An invariant sequence exists for arbitrary split-and-merge transformation, provided W has a continuous distribution function G. Indeed, let ϕ be a monotone function which transforms the uniform distribution into G. For (X_n) the GEM (1)-sequence, the sequence $(\phi(X_n))$ is invariant, as it follows readily from the invariance of (X_n) under the split-and-merge transformation directed by uniformly distributed W.

From this viewpoint, each GEM(θ)-sequence is a ϕ-image of the GEM (1)-sequence. In general, however, ϕ does not keep the RAM structure, thus it is natural to ask about the uniqueness of the invariant sequence, without restricting to the residual allocation models. We have strong indications that the invariant sequence is always unique for arbitrary continuous G.

8. As a final remark we mention a related invariance property of the Poisson-Dirichlet (PD) distribution.

Let (X_n) be a decreasing random sequence with $\sum_n X_n = 1$ and $(\widetilde{X}_n)$ be a sized-biased permutation of (X_n). Consider the sequence (Y_n) obtained by arranging components of $T(W, (\widetilde{X}_n))$ in decreasing order. Given (X_n), the component X_i splits into parts W and $X_i - W$ with probability $X_i\, G(X_i)$, while X_i merges with X_j $(j \neq i)$ with probability

$$X_i\, X_j\, \frac{1 - G(X_i)}{1 - X_i} + X_i\, X_j\, \frac{1 - G(X_j)}{1 - X_j}.$$

Known relations between GEM and PD distributions taken together with the invariance of GEM imply that the sequences (Y_n) and (X_n) have the same distributions provided (X_n) is governed by PD(θ) and W by beta$(1, \theta)$. In the particular case $\theta = 1$ this invariance was recently observed by N. Tsilevich (see [5], Example (4.2)) in connection with the action of a transposition on the space of virtual permutations.

References

[1] Johnson, N., Kotz, S. and Balakrishnan, N. (1997) *Multivariate Discrete Distributions*, Wiley, NY.

[2] Gnedin, A. (1998). On convergence and extensions of size-biased permutations, *J. Appl. Prob.* **35**.

[3] Gnedin, A. and Kerov, S. (1998). The Plancherel Measure of the Young-Fibonacci Graph, Preprint.

[4] Pitman, J. (1996). Random discrete distributions invariant under size-biased permutation *Adv. Appl. Prob.* **28**, 525-539.

[5] Tsilevich, N. (1998). Stationary measures on the space of virtual permutations for an action of the infinite symmetric group, Preprint St. Petersburg Dept. Steklov Math. Inst., PDMI-13/1998