Sequential construction of samples from residual allocation models

Yuri Yakubovich
St. Petersburg State University, Russia
(ongoing joint work with Jim Pitman, Berkeley, California)

Polynomial Computer Algebra '2019
18 April, 2019

Residual allocation models

Consider a sequence of random variables $P_\bullet = (P_1, P_2, \dots)$ such that $P_j > 0$ and $\sum_j P_j = 1$ a.s.
Given $P_\bullet$ think of it as a discrete probability distribution and consider a sample $(X_1, X_2, \dots)$ of i.i.d.$(P_\bullet)$ random variables, that is $\mathbb{P}[X_i = j | P_\bullet] = P_j$. If the distribution of $P_\bullet$ is not degenerate (it is indeed random) then the unconditional sample is not independent, but is exchangeable, that is for any n the distribution of $(X_1, \dots, X_n)$ is invariant under permutation of coordinates. By De Finetti's theorem, any exchangeable distribution on positive integers can be obtained this way.

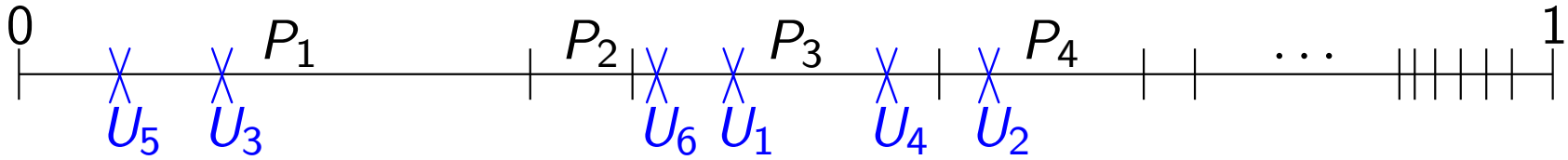
Any probability distribution $P_\bullet$ can be represented in a stick-breaking form:

$$P_1 = H_1, \quad P_2 = (1 - H_1)H_2, \quad P_3 = (1 - H_1)(1 - H_2)H_3, \ \dots$$

for some sequence $H_\bullet$ known as discrete hazard rates. If H_j are random and independent this model of a random discrete distribution is known as *residual allocation model* (RAM).

Kingman's paintbox

A convenient way to obtain samples from random discrete distribution $P_\bullet$ is by the so-called Kingman's paintbox construction. Given $P_\bullet$ consider the partition of the interval $[0,1]$ on subintervals of lengths $P_1, P_2, \ldots$. Consider also an independent sample $U_1, U_2, \ldots$ of the uniform$([0,1])$ random variables.

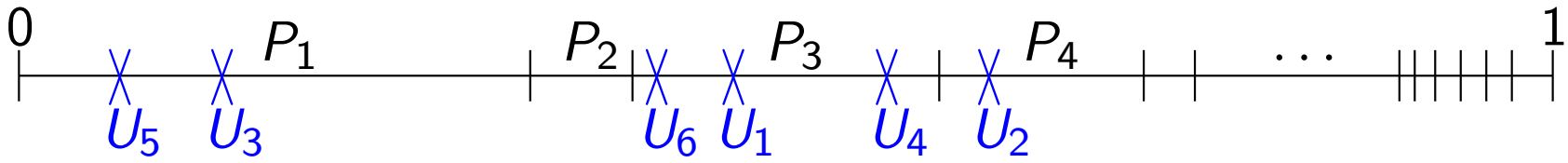


Define $X_i = j$ if U_i falls into jth interval. Then $\mathbb{P}[X_i = j | P_\bullet] = P_j$.

So in the picture above $n = 6$, and $(X_1, \ldots, X_6) = (3, 4, 1, 3, 1, 3)$.

Alternative description of samples

The sample can be also described in various different ways:



The simplest is by specifying counts of samples in each interval:

$$N_{n:j} = \#\{i \in \{1, \ldots, n\} : X_i = j\}, \qquad j \in \mathbb{N}.$$

Here $N_{6:1} = 2$, $N_{6:3} = 3$, $N_{6:4} = 1$ and all other counts are zero.
Given $N_{n:j} = n_j$, each particular sample with these counts come with equal probability $\binom{n}{n_1, n_2, n_3, \ldots}^{-1}$.
Another possibility is to consider gaps $G_{n:i}$ between the order statistics of the sample, $X_{n:1} \leq \cdots \leq X_{n:n}$, defined as

$$G_{n:1} = X_{n:1} - 1; \qquad G_{n:i} = X_{n:i} - X_{n:i-1}, \quad i = 2, \ldots, n.$$

Here $G_{n:\bullet} = (0, 0, 2, 0, 0, 1)$.

Consecutive construction of the sample

It turns out that the sample from a RAM can be constructed step-by-step in the following way.
Introduce binomial moments of hazards:

$$\mu_j(n,m) = \mathbb{E}H_j^n(1-H_j)^m. \tag{1}$$

Suppose we already know the sample $X_1,\dots,X_n$ of size $n \geq 0$, which has counts $N_{n:j} = n_j$. (If $n = 0$ then all counts are zero; in general $n = \sum_{j\geq 1} n_j$). Consider tail sums $s_k = \sum_{j\geq k} n_j$.

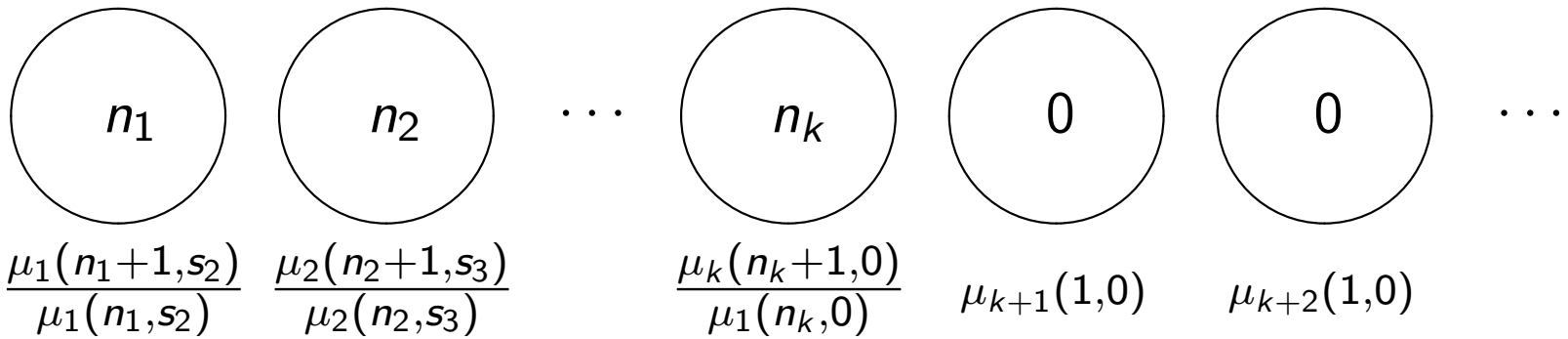
Thm. In this settings for any $j \in \mathbb{N}$

$$\mathbb{P}[X_{n+1} = j | X_1,\dots,X_n, X_{n+1} \geq j] = \frac{\mu_j(n_j+1, s_{j+1})}{\mu_j(n_j, s_{j+1})}. \tag{$*$}$$

Hence the next sample X_{n+1} can be obtained as follows: for each $j = 1,2,\dots$ test whether $X_{n+1} = j$ with probability $(*)$, if not, proceed to the next j.

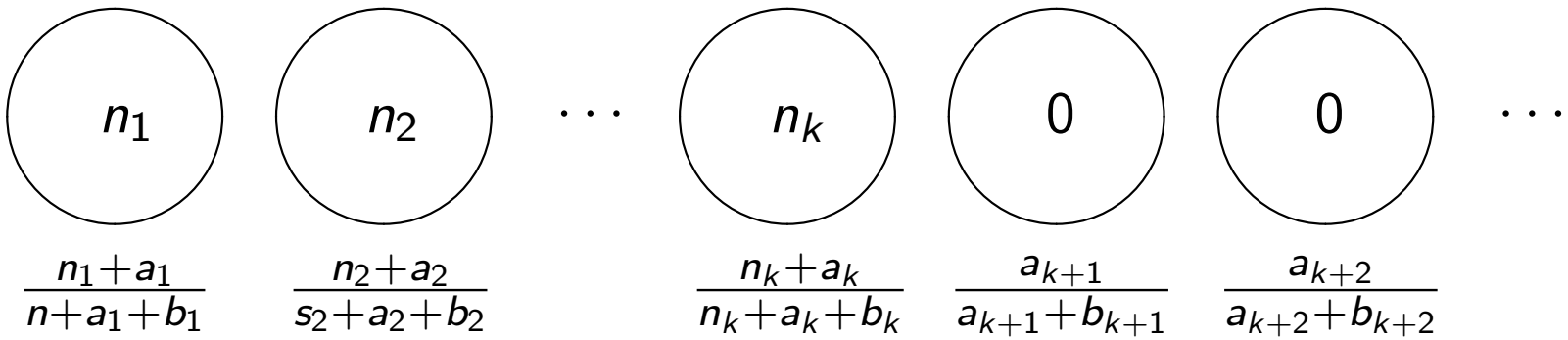
Ordered Chinese restaurant metaphor

A customer enters the restaurant with an infinite row of tables, goes along them and chooses whether to sit at the current table with the probability depending on the number of customers seated at the current table and on all further tables (and on the number of the table):



Ordered Chinese restaurant metaphor

A customer enters the restaurant with an infinite row of tables, goes along them and chooses whether to sit at the current table with the probability depending on the number of customers seated at the current table and on all further tables (and on the number of the table):



In the case when H_j has beta(a_j, b_j) distribution with density $B(a_j, b_j)^{-1} h^{a_j-1}(1-h)^{b_j-1}$ for $0 < h < 1$, the probabilities to choose a table become just simple quotients. This is the case, for the celebrated GEM(α, θ) model ($0 \leq \alpha < 1$, $\theta > -\alpha$) when H_j has beta$(1-\alpha, \theta + j\alpha)$ distribution.

Some occurrences of GEM distributions

Random permutations
Scaled cycle lengths of the uniform permutation of $[n]$ (in order of least elements) converges in distribution to GEM$(0, 1)$ as $n \to \infty$.

Some occurrences of GEM distributions

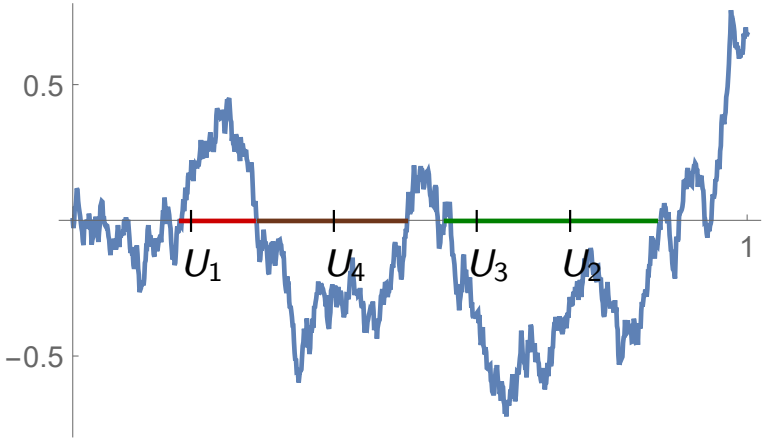
Random permutations
Scaled cycle lengths of the uniform permutation of $[n]$ (in order of least elements) converges in distribution to GEM$(0, 1)$ as $n \to \infty$.

Non-uniform permutations
If one weights the permutation by $\theta^{\text{number of cycles}}$ it gives GEM$(0, \theta)$ in the limit.
(Ewens's sampling formula)

Some occurrences of GEM distributions

Random permutations
Scaled cycle lengths of the uniform permutation of $[n]$ (in order of least elements) converges in distribution to GEM$(0, 1)$ as $n \to \infty$.

Brownian excursions



The size-biased reordering of excursion lengths of a standard Brownian motion has GEM$(1/2, 0)$ distribution.

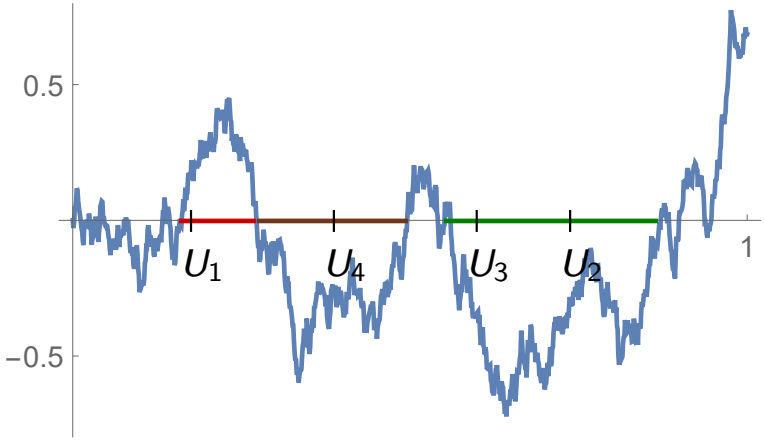
Non-uniform permutations
If one weights the permutation by $\theta^{\text{number of cycles}}$ it gives GEM$(0, \theta)$ in the limit.
(Ewens's sampling formula)

Some occurrences of GEM distributions

Random permutations
Scaled cycle lengths of the uniform permutation of $[n]$ (in order of least elements) converges in distribution to GEM$(0, 1)$ as $n \to \infty$.

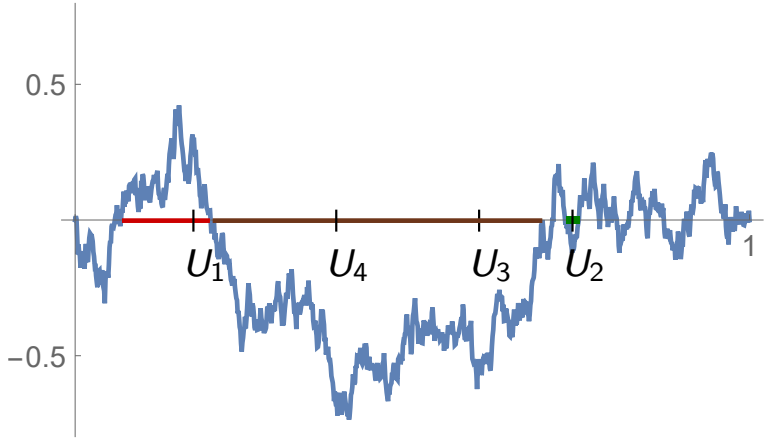
Non-uniform permutations
If one weights the permutation by $\theta^{\text{number of cycles}}$ it gives GEM$(0, \theta)$ in the limit.
(Ewens's sampling formula)

Brownian excursions



The size-biased reordering of excursion lengths of a standard Brownian motion has GEM$(1/2, 0)$ distribution.

Brownian bridge excursions



The size-biased reordering of excursion lengths of a standard Brownian bridge has GEM$(1/2, 1/2)$ distribution.

The descending reordering GEM(α, θ) random frequencies is known as the Poisson–Dirichlet(α, θ) distribution.

Probability in terms of gaps

Things become even easier when all hazards are identically distributed: $H_j \stackrel{d}{=} H$ for all j. Denote their common binomial moments by $\mu(n,m) = \mathbb{E}[H^n(1-H)^m]$.
Recall the description of the sample order statistics in term of gaps:

$$G_{n:1} = X_{n:1} - 1; \qquad G_{n:i} = X_{n:i} - X_{n:i-1}, \quad i = 2, \ldots, n.$$

Thm. For a RAM with i.i.d. factors there is a recursive formula for the probability $p_n(g_1, \ldots, g_n)$ to have a specific gap sequence $(g_1, \ldots, g_n)$:

$$p_n(0, \ldots, 0) = \mu(n, 0), \qquad (n = 1, 2, \ldots)$$

and for $n = 1, 2, \ldots$ and $(g_1, \ldots, g_n)$ with $\sum_i g_i > 0$

$$p_n(g_1, \ldots, g_n) = \binom{n}{\ell} \mu(\ell, n-\ell)\, p_{n-\ell}(g_\ell - 1, \ldots, g_n). \qquad (2)$$

where $\ell := \min\{i : g_i > 0\}$.

Gaps in GEM$(0, \theta)$ model

In GEM(α, θ) with $H_j \sim$ beta$(1-\alpha, \theta + j\alpha)$ factors are i.i.d. only when $\alpha = 0$.

Cor. For sampling from GEM$(0, \theta)$, the gaps probability function is

$$p_n(g_1, \ldots, g_n) = \frac{(1)_n}{(1+\theta)_n} \prod_{i=1}^{n} \left(\frac{\theta}{n+1-i+\theta} \right)^{g_i}. \qquad (3)$$

Hence gaps are independent in GEM$(0, \theta)$ model and $G_{n:i}$ has the geometric$(\frac{n+1-i}{n+1-i+\theta})$ distribution.

It gives an alternative way to generate a sample from GEM$(0, \theta)$ model: generate n independent geometric random variables with indicated parameters, and reconstruct the order statistics of $X_1, \ldots, X_n$ as their partial sums.

Thank you!