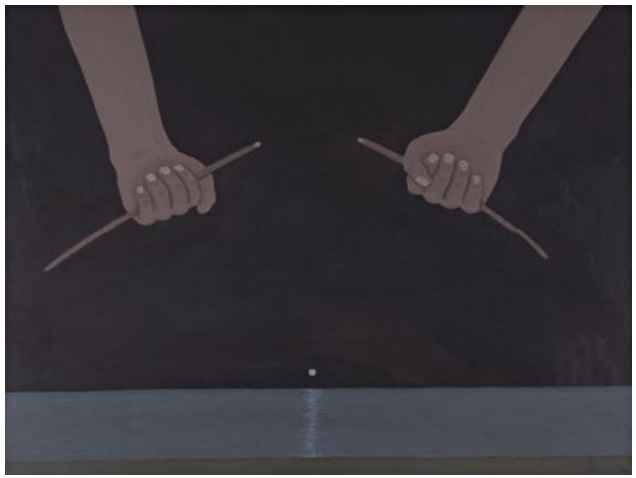
Limit laws for stick-breaking partitions

Alexander Gnedin

(joint work with Alex Iksanov and Uwe Rösler)



Stick-breaking

W the generic random factor with values in $(0,1)$

$W_1, W_2, \ldots$ independent copies of W

The multiplicative renewal process with points

$$W_1 \ldots W_j, \quad j = 1, 2, \ldots$$

breaks the unit stick in fragments of positive sizes

$$P_j = W_1 \ldots W_{j-1}(1 - W_j), \quad j = 1, 2, \ldots$$

The fragments are labelled $1, 2, \ldots$ in the left-to-right order on $(0,1)$.

Since $\sum_j P_j = 1, \ P_j > 0$, we can view (P_j) as a random discrete probability distribution.

A regenerative occupancy model

Consider fragment j as a box with frequency P_j.

n balls are thrown in boxes $1, 2, \dots$, in such a way that conditionally given frequencies (P_j) each ball falls in box j with probability P_j, independently of other balls.

The allocation of balls-in-boxes, read in the left-to-right order on $(0, 1)$, is a *weak composition* of integer n, sometimes written like infinite vector

$$(3, 2, 1, 2, 0, 1, 0, 0, 0, \dots)$$

or as a stars-and-bars word

$$***|**|*|**||*||||\dots$$

Without account of the order of boxes the allocation is a *partition* of n, written as $(3, 2, 2, 1, 1)$ or $1^2 2^2 3$. These 'stick-breaking' partitions constitute the simplest class of *regenerative partition structures* studied recently by Barbour, G, Pitman, Yor.

Quantities of interest in the analysis of random fractals, fragmentation, split and coalescent trees, and applications like genetic diversity, data structures, queues, species sampling:

- K_n the number of boxes occupied by at least one ball
- $K_{n,r}$ the number of boxes occupied by exactly r balls, $r = 1, 2, \ldots$

We shall connect them to also

- K_n^* the largest index of occupied box
- $K_{n,0}$ the number empty boxes with index less than K_n^*

These are related via

$$K_n = \sum_{r \geq 1} K_{n,r}, \quad n = \sum_r r\, K_{n,r}, \quad K_{n,0} = K_n^* - K_n.$$

The beta case

In the classical stick-breaking model it is assumed that

$$W \stackrel{d}{=} \mathrm{beta}(\theta, 1), \qquad \theta > 0$$

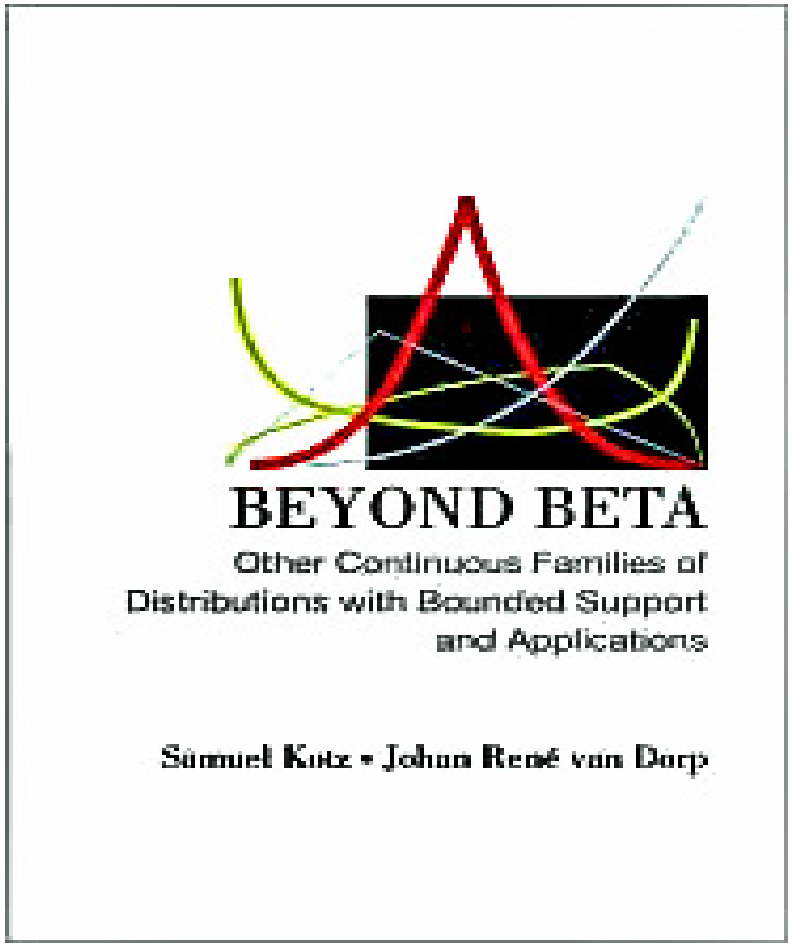
(uniform distribution when $\theta = 1$). Basic facts in this case:

- the multiplicative renewal process $\{W_1 \dots W_j\}$ is a scale-invariant Poisson process with rate θ/x, $x \in (0, 1)$,
- the partition/composition has distribution widely known as Ewens' sampling formula,
- $\mathbb{E}K_n \sim \theta \log n, \quad \mathrm{Var}K_n \sim \theta \log n,$
- $(K_n - \mathbb{E}K_n)/\sqrt{\mathrm{Var}K_n} \stackrel{d}{\to} \mathcal{N}(0, 1),$
- $(K_{n,1}, K_{n,2}, \dots) \to (\mathrm{Poisson}(\theta/j),\ j \geq 1)$ with independent components.

See the ABT-book (Arratia-Barbour-Tavaré)

thus we are interested in what happens

thus we are interested in what happens



...in fact, even beyond beta$(\theta, 1)$.

For general W, the logarithmic moments

$$\mu = \mathbb{E}|\log W|, \qquad \sigma^2 = \mathrm{Var}(\log W), \quad \nu = \mathbb{E}|\log(1-W)|.$$

may be finite or infinite.

We shall assume throughout $\nu < \infty$. This controls the number of empty boxes in the view of

$$\lim_{n\to\infty} \mathbb{E}K_{n,0} = \frac{\nu}{\mu}.$$

A cutoff phenomenon

Under the assumption $\nu < \infty$

- all boxes up to the last filled, K_n^*th, are nonempty, with a small number of exceptions $K_{n,0}$,
- K_n and K_n^* have the same limit law (if any), with the same scaling/centering.

In view of the representation

$$K_n^* = \#\{j : W_1 \dots W_j > \min(U_1, \dots, U_n)\}$$

with $U_1, \dots, U_n$ i.i.d. uniform$[0,1]$,

- K_n^* is approximable by the number of epochs on $[0, \log n]$ of the additive renewal process with step distribution $|\log W|$.

This leads to a complete classification of possible limit laws for K_n and conditions of convergence.

Normal limit ($\mu < \infty, \sigma^2 < \infty$)

Theorem

If $\mu < \infty, \sigma^2 < \infty$ then

$$\frac{K_n - a_n}{b_n} \xrightarrow{d} \mathcal{N}(0,1)$$

with the standard scaling/centering by the moments, that is

$$a_n = \frac{\log n}{\mu}, \quad b_n = (\mu^{-3}\sigma^2 \log n)^{1/2}$$

which is a situation analogous to the standard beta$(\theta, 1)$ case.

Example: for the general two-parameter beta

$$W \stackrel{d}{=} \mathrm{beta}(\theta, \zeta)$$

with density $x^{\theta-1}(1-x)^{\zeta-1}/\mathrm{Beta}(\theta,\zeta)$, the logarithmic moments are finite,
$\mu = \Psi(\theta+\zeta) - \Psi(\theta), \quad \nu = \Psi(\theta+\zeta) - \Psi(\zeta), \quad \sigma^2 = \Psi'(\theta) - \Psi'(\theta+\zeta)$
(where $\Psi = \Gamma'/\Gamma$).

So the limit law of K_n is normal.
For the number of empty boxes we have a limit

$$K_{n,0} \stackrel{d}{\to} K_{\infty,0},$$

with the generating function computable in the 'Ewens' case' $\zeta = 1$ as

$$\mathbb{E}s^{K_{\infty,0}} = \frac{\Gamma(1+\theta)\Gamma(1+\theta-\theta s)}{\Gamma(1+2\theta-\theta s)}, \qquad s \in [0,1].$$

Normal limit ($\mu < \infty,\ \sigma^2 = \infty$)

Theorem

If $\mu < \infty, \sigma^2 = \infty$ *and for some* L *slowly varying at* ∞

$$\int_x^1 |\log y|^2 \mathbb{P}(W \in dy) \sim L(|\log x|), \qquad x \to 0,$$

then

$$\frac{K_n - a_n}{b_n} \xrightarrow{d} \mathcal{N}(0,1)$$

with

$$a_n = \frac{\log n}{\mu}, \quad b_n = \mu^{-3/2} c_{\lfloor \log n \rfloor}$$

where c_n *is any sequence with*

$$\lim_{n\to\infty} nL(c_n)/c_n^2 = 1.$$

then, it turns out that

then, it turns out that

not all limits by stick-breaking are normal...

Other limits only exist under the assumption of regular variation

$$\mathbb{P}(W \le x) \sim |\log x|^{-\alpha} L(|\log x|), \qquad x \to 0$$

with some $0 \le \alpha < 2$.

Stable limit $(1 < \alpha < 2)$

Theorem

Suppose

$$\mathbb{P}(W \le x) \sim |\log x|^{-\alpha} L(|\log x|), \qquad x \to 0$$

holds with slowly varying L and $1 < \alpha < 2$. Then $\frac{K_n - a_n}{b_n} \xrightarrow{d} \mathcal{S}_\alpha$, where $\mathcal{S}_\alpha$ is asymmetric α-stable, with characteristic function

$$t \mapsto \exp\{-|t|^\alpha \Gamma(1-\alpha)(\cos(\pi\alpha/2) + i \sin(\pi\alpha/2)\operatorname{sgn}(t))\},$$

$$a_n = \frac{\log n}{\mu}, \quad b_n = \mu^{-(\alpha+1)/\alpha} c_{\lfloor \log n \rfloor},$$

where c_n is any sequence satisfying

$$\lim_{n\to\infty} nL(c_n)/c_n^\alpha = 1.$$

Stable limit ($\alpha = 1, \mu < \infty$)

Theorem

Suppose

$$\mathbb{P}(W \le x) \sim |\log x|^{-1} L(|\log x|), \qquad x \to 0$$

holds with slowly varying L. Then $\frac{K_n - a_n}{b_n} \xrightarrow{d} \mathcal{S}_1$, *where* $\mathcal{S}_1$ *is asymmetric* 1*-stable, with characteristic function*

$$t \mapsto \exp\{-|t|(\pi/2 - i \log|t| \operatorname{sgn}(t))\}.$$

In the case $\mu < \infty$ *one should take*

$$a_n = \frac{\log n}{\mu}, \qquad b_n = \frac{c_{\lfloor \log n \rfloor}}{\mu^2},$$

where c_n *be any sequence satisfying*

$$\lim_{n \to \infty} nL(c_n)/c_n = 1.$$

Stable limit $(\alpha = 1, \mu = \infty)$

Theorem

If $\alpha = 1$ as above and $\mu = \infty$ then for $c(z)$ satisfying $zL(c(z))/c(z) \to 1$ (as $z \to \infty$) define

$$d(z) := z \int_{\exp(-c(z))}^{1} \mathbb{P}(W \le y) y^{-1} dy.$$

For ψ asymptotic inverse to d (i.e. $\psi(d(z)) \sim \psi(d(z)) \sim z, \quad z \to \infty$), the 1-stable limit of $(K_n - a_n)/b_n$ holds with

$$a_n = \psi(\log n), \quad b_n = \psi(\log n) c(\psi(\log n))/\log n.$$

Example: in the case

$$\mathbb{P}(W \le x) = \frac{1}{1 - \log x}, \quad x \in (0, 1)$$

we have $\alpha = 1$, $\mu = \infty$, and

$$\frac{(\log\log n)^2}{\log n} K_n - \log\log n - \log\log\log n \quad \xrightarrow{d} \quad \mathcal{S}_1.$$

Mittag-Leffler limit ($0 \le \alpha < 1$)

Theorem

Suppose

$$\mathbb{P}(W \le x) \sim |\log x|^{-\alpha} L(|\log x|), \qquad x \to 0$$

holds with slowly varying L and $0 \le \alpha < 1$. Then $\frac{K_n}{b_n} \xrightarrow{d} \mathcal{M}_\alpha$, where $\mathcal{M}_\alpha$ is the Mittag-Leffler distribution , with moments

$$\frac{n!}{\Gamma^n(1-\alpha)\Gamma(1+n\alpha)}, \qquad n = 1, 2, \ldots$$

and

$$b_n = \frac{(\log n)^\alpha}{L(\log n)}.$$

In particular, for $\alpha = 0$ the limit distribution is exponential.

Modified stick-breaking

Suppose $\mu = \mathbb{E}|\log W| < \infty$.

Take W_0 with density $\mathbb{P}(W \le x)/(\mu x)$ on $(0,1)$, $W_j \stackrel{d}{=} W$ for $j > 0$, and all W's independent. The point process with points

$$W_0 W_1 \dots W_{j-1}, \qquad j = 1, 2, \dots$$

extends to a scale-invariant ordinary point process $\mathcal{P}$ on $\mathbb{R}_+$.

An occupancy model is defined with gaps in $\mathcal{P}$ being 'boxes', and points of independent unit Poisson process $\mathcal{U}$ being 'balls'.

Let $Z_n^{(1)}, Z_n^{(2)}, \ldots$, where $Z_n^{(1)} > 0$, be the occupancy numbers, i.e. parts of the composition, labelled in the reversed order, and let $Z^{(1)}, Z^{(2)}, \ldots$ be the occupancy numbers in the $\mathcal{P}$-$\mathcal{U}$ point process model.

Theorem

If $\mu < \infty$, as $n \to \infty$

$$(Z_n^{(1)}, Z_n^{(2)}, \ldots) \xrightarrow{d} (Z^{(1)}, Z^{(2)}, \ldots),$$

where the rhs has distribution given by

$$\mathbb{P}(Z^{(1)} = n_1, \ldots, Z^{(k)} = n_k) = \frac{1}{\mu(n_1 + \cdots + n_\ell)} \times p(n_1 + \cdots + n_\ell : n_\ell) p(n_1 + \cdots + n_{\ell-1} : n_{\ell-1}) \ldots p(n_1 : n_1)$$

for $n_1 > 0, n_2, \ldots, n_\ell \geq 0$; $\ell > 0$ and

$$p(n : m) := \binom{n}{m} \mathbb{E}[W^{n-m}(1-W)^m], \quad 0 \leq m \leq n.$$

The small-counts asymptotics

Theorem

If $\mu < \infty$, then as $n \to \infty$,

$$(K_{n,0}, K_{n,1}, \ldots)$$

converge in distribution to the analogous counts in the $\mathcal{P}$-$\mathcal{U}$-model.

The size of the meander gap between the rightmost $\mathcal{P}$-point on $(0,1)$ and 1 has the same distribution as the size-biased pick from the frequencies in the modified stick-breaking model (Pitman/Yor '96, G/Pitman '05); from this

$$\lim_{n\to\infty} \mathbb{E}[K_{n,r}] = \frac{1}{\mu r}, \qquad r \geq 1.$$

beyond the beta

beyond the beta

we know too little about the limit distributions of $K_{n,r}$'s...

Related results to compare

- in the 2-parameter ($0 \le \alpha < 1, \theta > 0$) Pitman-Yor model with independent but not i.i.d. factors $W_j \stackrel{d}{=} \mathrm{beta}(\theta + j\alpha, 1-\alpha)$ the limits of $K_n, K_{n,r}$'s exist on the n^α-scale,
- for *nonrandom* frequencies of boxes (P_j),
  - K_n converge in distribution iff $\mathrm{Var} K_n \to \infty$, in which case the limit is normal;
  - typically $(K_{n,r}, r \ge 1)$ do not have a joint limit unless (P_j) satisfy a condition of regular variation, in which case the joint limit is a particular multivariate normal (G/Barbour '09).

Application: counting collisions in some Λ coalescents

In the Λ-coalescent with finite parameter measure Λ on $(0,1)$, when there are b particles each m-tuple $(2 \leq m \leq b)$ is merging at rate

$$\int_0^1 x^{m-2}(1-x)^{b-m}\Lambda(dx)$$

(Pitman '99, Sagitov '99).

Let C_n be the total number of collisions, and τ_n the absorption time, for the coalescent starting with n particles, and terminating with a single particle. Suppose the measure can be scaled so that

$$\int_0^1 x^{-2}\Lambda(dx) = 1.$$

It can be shown that

- 'Almost all' collisions directly involve some of the original n particles, whose number decays like in the stick-breaking model,
- $(C_n - a_n)/b_n$ has the same limit law (if any) as $(K_n - a_n)/b_n$, where K_n is the number of occupied boxes, by stick-breaking with the generic factor $W \stackrel{d}{=} x^{-2}\Lambda(dx)$.
- A similar result is true for τ_n, with a different scaling in the finite-moments case.

Example: for $x^{-2}\Lambda(dx) = \mathrm{beta}(\theta, \zeta)$ with $\theta > 0, \zeta > 0$

$$\frac{C_n - (\log n)/\mu}{(\sigma^2 \mu^{-3} \log n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

for $\mu = \Psi(\theta + \zeta) - \Psi(\zeta)$, $\sigma^2 = \Psi'(\theta) - \Psi'(\theta + \zeta)$.
Similarly, for the absorption time

$$\frac{\tau_n - (\log n)/\mu}{\{\mu^{-3}(\mu^2 + \sigma^2) \log n\}^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

Thank you!

Thank you!