Ann. Inst. Statist. Math.
Vol. 54, No. 3, 607–620 (2002)

LIMIT PROCESSES WITH INDEPENDENT INCREMENTS FOR THE EWENS SAMPLING FORMULA

GUTTI JOGESH BABU[1]* AND EUGENIJUS MANSTAVIČIUS[2]**

[1]*Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, PA 16802-2111, U.S.A.*
[2]*Department of Mathematics, Vilnius University, Naugarduko str. 24, LT2006 Vilnius, Lithuania*

(Received June 16, 2000; revised February 16, 2001)

Abstract. The *Ewens sampling formula* in population genetics can be viewed as a probability measure on the group of permutations of a finite set of integers. Functional limit theory for processes defined through partial sums of dependent variables with respect to the Ewens sampling formula is developed. Techniques from probabilistic number theory are used to establish necessary and sufficient conditions for weak convergence of the associated dependent process to a process with independent increments. Not many results on the necessity part are known in the literature.

Key words and phrases: Random partitions, cycle, population genetics, permutations, law of large numbers, probabilistic number theory, Skorohod topology, slowly varying function, functional limit theorem.

1. Introduction

Let $\boldsymbol{S}_n$ denote the symmetric group of permutations on $\{1,\ldots,n\}$. Erdös and Turán (1965) established the weak convergence result,

$$\frac{1}{n!}\#\left(\sigma\in\boldsymbol{S}_n : \log Ord(\sigma)-\frac{1}{2}\log^2 n \le \frac{y}{\sqrt{3}}\log^{3/2} n\right)\to\Phi(y), \tag{1.1}$$

as $n\to\infty$, where Φ denotes the standard normal distribution function, and $Ord(\sigma)$ denotes the order of a permutation σ. Since each $\sigma\in\boldsymbol{S}_n$ can be uniquely represented (up to the order) by the product $\sigma=\varkappa_1\cdots\varkappa_w$ of independent cycles $\varkappa_i$ (Feller (1968), X.6), where $w=w(\sigma)$ denotes the total number of cycles, $\boldsymbol{S}_n$ can be partitioned into equivalence classes of conjugate elements. As the elements in a conjugate class will all have the same number k_j of cycles of length j for all $1\le j\le n$, the conjugate class containing an element $\sigma\in\boldsymbol{S}_n$ can be identified by the vector $\bar{k}:=(k_1,\ldots,k_n)$, where $0\le k_j=k_j(\sigma)$ denotes the number of cycles of σ of length j. Thus the space $\bar{\boldsymbol{S}}_n$ of conjugate classes can be taken as the set of vectors $\bar{k}=(k_1,\ldots,k_n)$, of non-negative integers representing partitions of n,

$$1k_1+\cdots+nk_n=n. \tag{1.2}$$

*Research supported in part by NSA grant MDA904-97-1-0023, NSF grants DMS-9626189, DMS-0101360, and by National Research Council's 1997-99 Twinning fellowship.

**Research supported in part by Lithuanian Science and Studies Fund and by National Research Council's 1997-99 Twinning fellowship.

Since $Ord(\sigma)$ and $w(\sigma)$ depend only on the conjugate class containing σ, these can be treated as functions on $\tilde{\boldsymbol{S}}_n$. Note that the function $\sum_{k_j(\sigma)\geq 1} \log j$ is closely related to $\log Ord$ considered by Erdös and Turán (1965) in (1.1).

A general family of measures $\nu_{n,\theta}$, $\theta > 0$, on $\tilde{\boldsymbol{S}}_n$ was described by Ewens (1972) in connection with models in population genetics. The measure $\nu_{n,\theta}$ on $\tilde{\boldsymbol{S}}_n$, known as the Ewens sampling formula is given by

$$\nu_{n,\theta}(k_1,\ldots,k_n) := \frac{n!}{\theta_{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{k_j} \frac{1}{k_j!}, \tag{1.3}$$

where $\theta > 0$ and $\theta_{(n)} = \theta(\theta+1)\cdots(\theta+n-1)$. Clearly, $\nu_{n,\theta}$ for $\theta = 1$ is the measure on $\tilde{\boldsymbol{S}}_n$ induced by the uniform measure on $\boldsymbol{S}_n$ considered in (1.1). If the probability $\theta^{w(\sigma)}/\theta_{(n)}$ is assigned to $\sigma \in \boldsymbol{S}_n$, then the measure of the class $\bar{k}$ is given by the Ewens sampling formula (1.3). It is well known that the asymptotic distribution of $k_j(\cdot)$ for a fixed $j \geq 1$ is Poisson with parameter θ/j. The relation (1.2) makes $k_j(\cdot)$, $1 \leq j \leq n$ a dependent sequence. The dependency is rather strong for $\varepsilon n \leq j \leq n$. The details about the Ewens sampling formula and its relation to population genetics can be found in the monographs by Ewens (1979) and Kingman (1980).

The first functional limit theorem in the case $\theta = 1$ for the function $w(\sigma)$, was established by DeLaurentis and Pittel (1985). The case of general θ for the function $w(\sigma)$ was examined by Hansen (1990) and Donnelly *et al.* (1991). A short proof of Hansen's theorem is given in Section 2.C of Arratia and Tavaré (1992). Convergence of more general partial sum processes to the Brownian motion was investigated by Babu and Manstavičius (1999). It was shown that an analog of the Lindeberg condition is necessary and sufficient for the weak convergence of the processes. However, by constructing an example it is shown that the Lindeberg condition is not necessary for the one-dimensional central limit theorem.

In this paper, we consider arbitrary limit processes with independent increments. The main difficulties arise in proving necessity of the conditions. We base our analysis on an idea that originated in probabilistic number theory (Timofeev and Usmanov (1984)). For the sufficiency part, as in Arratia and Tavaré (1992), Babu and Manstavičius (1999), and as in the earlier papers on probabilistic number theory of Kubilius (1964), Babu (1973), Philipp (1973), Manstavičius (1984, 1985), we exploit the idea of approximating truncated sums of dependent and the corresponding independent random variables. For a recent account of results on probabilistic number theory see Tenenbaum (1995).

2. Results

Let $h_j(k)$ be a real double sequence, $k \geq 0$, $j \geq 1$ such that $h_j(0) = 0$ for each j. Set for brevity $a(j) = h_j(1)$, and $u^* = (1 \wedge |u|)\operatorname{sgn} u$, where $a \wedge b := \min\{a, b\}$. Throughout this paper the limits are taken as $n \to \infty$ and we assume that the normalizing factors $\beta(n) > 0$ satisfy $\beta(n) \to \infty$. The sequence $\{\beta(n)\}$ need not be monotone. Define

$$B(u,n) = \sum_{j\leq u} \left(\frac{a(j)}{\beta(n)}\right)^{*^2} \frac{1}{j}, \qquad A(u,n) = \theta \sum_{j\leq u} \left(\frac{a(j)}{\beta(n)}\right)^{*} \frac{1}{j}$$

and

$$y(t) := y_n(t) = \max\{l \leq n : B(l,n) \leq tB(n,n)\}, \qquad t \in [0,1]. \tag{2.1}$$

We shall consider the weak convergence (denoted by $\Rightarrow$) of the process

$$H_n := H_n(\sigma, t) = \frac{1}{\beta(n)} \sum_{j \le y(t)} h_j(k_j(\sigma)) - A(y(t), n), \qquad t \in [0,1]$$

under the measure $\nu_{n,\theta}$, in the space $\boldsymbol{D}[0,1]$ endowed with the Skorohod topology (Billingsley (1968)). Let $\mathcal{D}$ denote the Borel σ-field generated by the Skorohod topology. The corresponding process X_n with independent increments is defined by

$$X_n := X_n(t) = \sum_{j \le y(t)} X_{nj} - A(y(t), n), \qquad t \in [0,1],$$

where $X_{nj} = a(j)\xi_j/\beta(n)$, and ξ_j are independent Poisson random variables with $\boldsymbol{E}\xi_j = \frac{\theta}{j}$.

In general, the limiting behavior of the dependent process H_n is different from the corresponding X_n. Since the summands $h_j(k_j(\sigma))$ with indices $j \le \varepsilon n$ are nearly independent (Lemma 2 below gives a quantitative version of this statement), the main influencing factors are the dependent summands $h_j(k_j(\sigma))$ for $\varepsilon n \le j \le n$. Intuitively, if the normalizing sequence $\beta(n)$ satisfies $\beta(\varepsilon n) \sim \beta(n)$, where $0 < \varepsilon < 1$ is fixed, then the effect of the normalized sum over $[\varepsilon n, n]$ of the random variables on the limit behavior of H_n and X_n should be negligible. Conversely, if $h_j(k_j(\sigma))$ for $\varepsilon n \le j \le n$ has influence on the limiting behavior of H_n, then the limiting process had to have dependent increments, at least, in the neighborhood of the point $t = 1$. Thus in the case of limit processes with independent increments the condition above on the normalizing sequence $\beta(n)$ should be necessary. This is presented below.

THEOREM 1. *In order that $H_n \Rightarrow X$, where X is a process with independent increments such that the distribution of $X(1)$ is non-degenerate, it is necessary and sufficient that the following two conditions are satisfied:*

(I) *$\beta(n)$ is slowly varying in the sense of Karamata;*

(II) *the sequence of functions*

$$\Psi_n(u) := \sum_{\substack{j \le n \\ a(j) < u\beta(n)}} \left(\frac{a(j)}{\beta(n)}\right)^{*^2} \frac{1}{j}$$

converges weakly to some non-decreasing function $\Psi(u)$, $0 = \Psi(-\infty) < \Psi(+\infty) < \infty$, so that $\Psi_n(\pm\infty) \to \Psi(\pm\infty)$.

Moreover, if Conditions (I) *and* (II) *hold, then the limiting process X satisfies*

$$\boldsymbol{E}e^{i\lambda X(t)} = \exp\left\{\int_{\boldsymbol{R}} (e^{i\lambda u} - 1 - i\lambda u^*) u^{*^{-2}} dM_t(u)\right\}, \qquad \lambda \in \boldsymbol{R}, \tag{2.2}$$

where

$$M_t(u) = \int_{-\infty}^{u/k(t)} (vk(t))^{*^2} v^{*^{-2}} d\Psi(v) \qquad \text{and} \qquad k(t) = \lim \beta(y_n(t))/\beta(n).$$

The equation $M_t(+\infty) = t\Psi(+\infty)$ can be used to compute $k(t)$.

Remark. If Condition (I) holds, then $H_n \Rightarrow X$ if and only if $X_n \Rightarrow X$. One of the cases, when Condition (I) is given by Condition (II) implicitly is treated in Babu and Manstavičius (1999). Condition (II) is essentially needed to establish weak convergence of X_n.

For properties of slowly varying functions, see Bingham *et al.* (1989). The sufficiency part of Theorem 1 is contained in the following result of independent interest.

THEOREM 2. *Suppose that* $B(n,n) - B(un,n) = o(1)$ *for each fixed* $0 < u < 1$. *Then Condition* (II) *of Theorem* 1, *is equivalent to* $H_n \Rightarrow X$ *as well as* $X_n \Rightarrow X$, *where the distribution of* $X(1)$ *is non-degenerate. Moreover, if Condition* (II) *holds, then the limiting process* X *is a process with independent increments satisfying* (2.2).

In the case of limiting processes with independent increments, the normalizing constants $\beta(n)$ are necessarily slowly varying. Here one expects that on the average $a(j)$ are small. Apparently the converse also holds. This idea due to Timofeev and Usmanov (1986) is exploited to strengthen the necessity part of Theorem 1.

THEOREM 3. *Suppose that*

$$\sum_{j\le n}\left(\frac{a(j)}{n^{\varepsilon}}\right)^{*^2}\frac{1}{j} = o(1) \tag{2.3}$$

for each positive $\varepsilon > 0$. *In order that* $H_n \Rightarrow X$, *where distribution of* $X(1)$ *is non-degenerate, it is necessary and sufficient that Conditions* (I) *and* (II) *of Theorem* 1 *are satisfied.*

It is likely that the counter-example constructed in Babu and Manstavičius (1999) or the functions with $a(j) = j^c$ with $c \neq 1$, $c > 0$ (Manstavičius (1996)) may converge to processes with dependent increments. We shall return to the topic of limiting processes with dependent increments elsewhere. As mentioned earlier, the values of functions on σ with long cycles, will have a dominating effect. In this sense, the Ewens formula has interesting connection to the Poisson-Dirichlet and Griffiths-Engen-McCloskey distributions (see Hirth (1997*a*,*b*,*c*)).

An Example to illustrate Theorem 1. If $\beta(n) = \log n$ and $a(j) = \{j\sqrt{2}\}^{-1}$, then Conditions (I) and (II) are satisfied, with $\Psi(u) = 0$, $= u$ or $= 2 - (1/u)$ according as $u \le 0$, $0 \le u \le 1$ or $u \ge 1$. To establish this we clearly have $\Psi_n(u) = \Psi(u) = 0$ for $u \le 0$. It is well known (Drmota and Tichy (1997), Corollary 1.65) that

$$\Delta_x(v) := \frac{1}{x}\sum_{j\le x}\mathbf{1}(\{j\sqrt{2}\} < v) - v = O((\log x)/x), \tag{2.4}$$

uniformly in $v \in [0,1]$. Hence summing by parts we obtain, for $0 < u \le 1$ and $\log n \ge 1/u$, that

$$\begin{aligned} F_n(u) :&= \sum_{j\le n}\frac{1}{j}\mathbf{1}(a(j) > u\log n) \\ &= \frac{1}{u\log n}(\log n + 1) + \Delta_n\left(\frac{1}{u\log n}\right) + \int_1^n \frac{1}{x}\Delta_x\left(\frac{1}{u\log n}\right)dx. \end{aligned}$$

As a result, $uF_n(u)$ is uniformly bounded in $0 < u \leq 1$. Further, $uF_n(u) \to 1$ for each fixed $0 < u \leq 1$. Here we have exploited the relation that $\Delta_x(v) \to 0$ as $v \to 0$ for each $x \geq 1$, and that

$$\sup_{0 \leq v \leq 1} \int_1^\infty \frac{1}{x} |\Delta_x(v)| dx < \infty.$$

Thus for $0 < u \leq 1$,

$$\Psi_n(u) = -\int_0^u s^2 dF_n(s) = -u^2 F_n(u) + 2\int_0^u sF_n(s)ds \to u,$$

and for $u > 1$,

$$\Psi_n(u) = \Psi_n(1) + F_n(1) - F_n(u) \to 2 - (1/u).$$

Also note that $\Psi_n(+\infty) \to 2 = \Psi(+\infty)$. Hence Condition (II) holds in this case.

To find $k(t)$ for our example, we use $M_t(\infty) = t\Psi(\infty)$, to obtain

$$\int_{-\infty}^{\infty} (uk(t))^{*^2} u^{*^{-2}} d\Psi(u)$$
$$= k^2(t) \int_0^1 du + k^2(t) \int_1^{1/k(t)} u^2 \frac{du}{u^2} + \int_{1/k(t)}^{\infty} \frac{du}{u^2} = 2k(t) = 2t.$$

So $k(t) = t$ and $\log y_n(t)/\log n \to t$. A simple algebra leads to

$$M_t(u) = \begin{cases} 0, & \text{if } u < 0 \\ tu, & \text{if } 0 \leq u < 1 \\ 2t - t/u, & \text{if } u \geq 1 \end{cases}$$

showing that our example models a homogeneous Cauchy process.

3. Auxiliary results

It will be shown later that the main results can easily be reduced to considering processes with $h_j(k) = kh_j(1) = ka(j)$, for all $j \geq 1$ and $k \geq 0$. Let

$$\hat{H}_n := \hat{H}_n(\sigma, t) := \frac{1}{\beta(n)} \sum_{j \leq y(t)} a(j)k_j(\sigma) - A(y(t), n). \tag{3.1}$$

For $0 \leq r \leq n$, let $\hat{H}_n^r := \hat{H}_n^r(\sigma, t)$, $H_n^r := H_n^r(\sigma, t)$ and $X_n^r := X_n^r(t)$ be the processes obtained from $\hat{H}_n$, H_n and X_n respectively by substituting $y(t) \wedge r$ for $y(t)$. For any two non-negative functions f and g, it is often convenient to use the notation $f(n) \ll g(n)$ instead of $f(n) = O(g(n))$. The basic techniques and the main steps used in proving the theorems are presented as several lemmas. In proving the main results, the condition $B(n,n) - B(un,n) = o(1)$ reduces the problem of weak convergence of X_n to that of truncated sums X_n^r. Then we use Lemma 2 to show that the effect of truncation is negligible. Lemma 7 helps in establishing that $\beta(n)$ is a slowly varying function.

LEMMA 1. *For any $\gamma > 0$, and $0 \leq r \leq n$,*

$$P_{n,r}(\gamma) := P\left(\sup_t |X_n(t) - X_n^r(t)| \geq \gamma\right) \ll B(n,n) - B(r,n) + o(1). \tag{3.2}$$

PROOF. Note that as $\beta(n) \to \infty$

$$\left|\sum_{j\le u} \boldsymbol{E}X^*_{nj} - A(u,n)\right| \le \theta \sum_{j=1}^{\infty} \left|\frac{a(j)}{\beta(n)}\right|^* \frac{|e^{-\theta/j}-1|}{j} + \sum_{j=1}^{\infty}\sum_{k=2}^{\infty} \left|\frac{a(j)k}{\beta(n)}\right|^* \left(\frac{\theta}{j}\right)^k \frac{e^{-\theta/j}}{k!} = o(1) \tag{3.3}$$

holds uniformly in $u \ge 1$, which yields

$$\max_{0\le r\le n} \max_{r\le k\le n} \left|\sum_{r<j\le k} \boldsymbol{E}X^*_{nj} - (A(k,n) - A(r,n))\right| = o(1). \tag{3.4}$$

In addition, we also have

$$\sum_{r<j\le n} P(X_{nj} \ne X^*_{nj}) \le \sum_{\substack{j\le n \\ |a(j)|\ge\beta(n)}} \frac{\theta}{j} + \sum_{j=1}^{\infty}\sum_{k=2}^{\infty} \left|\frac{a(j)k}{\beta(n)}\right|^* \left(\frac{\theta}{j}\right)^k \frac{e^{-\theta/j}}{k!}. \tag{3.5}$$

Thus, for all $\gamma > 0$, we have by (3.4) and (3.5), that

$$\begin{aligned} P_{n,r}(\gamma) &\le P\left(\bigcup_{r<j\le n} (X_{nj} \ne X^*_{nj})\right) + P\left(\max_{r<k\le n}\left|\sum_{r<j\le k} (X^*_{nj} - \boldsymbol{E}X^*_{nj})\right| \ge \gamma/2\right) + o(1) \\ &\ll B(n,n) - B(r,n) + o(1). \end{aligned}$$

This completes the proof.

LEMMA 2. (Arratia *et al.* (1992), Theorem 3) *For $0 < \varepsilon < 1$ and $2 \le r \le \varepsilon n$,*

$$\sup_{D\in\mathcal{D}} |\nu_{n,\theta}(\hat{H}^r_n \in D) - P(X^r_n \in D)| \le \varepsilon\theta(\theta + (1-\varepsilon)^{-1}).$$

Recall that $\mathcal{D}$ denotes the Borel σ-field on $\boldsymbol{D}[0,1]$.

Let for brevity, $\mathcal{L}(\mathcal{I})$ be the linear space of real functions g on $I \subset \boldsymbol{R}$ with finite supremum norm.

LEMMA 3. (Babu and Manstavičius (1999)) *Let*

$$h(\sigma,t) = h_1(k_1(\sigma),t) + \cdots + h_n(k_n(\sigma),t)$$

where $h_j(k,t)$, $t \in I \subset \boldsymbol{R}$, be a set of real array of functions on I such that $h_j(0,t) = 0$ and $h_j(k,\cdot) \in \mathcal{L}(\mathcal{I})$, for $k \ge 0$, $j \le n$, $t \in I$. Denote $\Xi_n(t) = h_1(\xi_1,t) + \cdots + h_n(\xi_n,t)$. Then for any $g \in \mathcal{L}(\mathcal{I})$ and $x \ge 0$,

$$\nu_{n,\theta}\left(\sup_{t\in I} |h(\sigma,t) - g(t)| \ge x\right) \le C(\theta)\left(P^{\theta\wedge 1}\left(\sup_{t\in I} |\Xi_n(t) - g(t)| \ge x/3\right) + n^{-\theta}\right).$$

Here $C(\theta)$ is a positive constant depending only on θ and $P^{\theta\wedge 1}(D) = (P(D))^{\theta\wedge 1}$.

At present we could not get rid of the exponent $\theta \wedge 1$. However, this inequality is sufficient for our purpose.

In the necessity part of Theorem 1 we will use an estimate of the mean values of multiplicative functions defined on permutations having only long cycles.

LEMMA 4. (Babu and Manstavičius (1999)) *For $b(j) \in \boldsymbol{C}$, $1 \le j \le n$, and $\sigma \in \boldsymbol{S}_n$, let*

$$f(\sigma) = \prod_{j=1}^{n} b(j)^{k_j(\sigma)}.$$

If $b(j) = 1$ for all but $j \in J \subset (n/2, n]$, then

$$M_n(f) := \frac{1}{\theta_{(n)}} \sum_{\sigma \in \boldsymbol{S}_n} \theta^{w(\sigma)} f(\sigma) = 1 + \sum_{j \in J} \frac{b(j) - 1}{j} d_{jn},$$

where

$$d_{jn} = \theta \frac{n! \theta_{(n-j)}}{\theta_{(n)}(n-j)!}.$$

LEMMA 5. *The measures $\nu_{n,\theta} \cdot H_n^{-1}$ and $\nu_{n,\theta} \cdot \hat{H}_n^{-1}$ can only converge simultaneously and to the same limit.*

PROOF. Since $\beta(n) \to \infty$ we have by Lemma 3, for any $\gamma > 0$ and $K > 2$, that

$$\begin{aligned}
\nu(\gamma; n, \theta) : &= \nu_{n,\theta}\left(\sup_t |H_n(\sigma, t) - \hat{H}_n(\sigma, t)| > \gamma\right) \\
&\ll P^{\theta\wedge 1}(\exists j \le K : \xi_j \ge K) + P^{\theta\wedge 1}(\exists j \ge K : \xi_j \ge 2) \\
&\quad + P^{\theta\wedge 1}\left(\sum_{j \le K} (|h_j(\xi_j)| + |a(j)|\xi_j) \ge \gamma\beta(n)/3, \xi_i \le K \forall i \le K\right) + o(1) \\
&\ll \left(\sum_{j \le K} \sum_{k \ge K} e^{-\theta/j} \frac{\theta^k}{j^k k!}\right)^{\theta\wedge 1} + \left(\sum_{j \ge K} \sum_{k \ge 2} e^{-\theta/j} \frac{\theta^k}{j^k k!}\right)^{\theta\wedge 1} + o_K(1) \\
&\ll K^{-\theta\wedge 1} + o_K(1).
\end{aligned}$$

Hence $\nu(\gamma; n, \theta) = o(1)$ for arbitrary $\gamma > 0$. This establishes Lemma 5.

For the infinitesimal independent array of random variables $X_{nj}, 1 \le j \le n$, we recall the following result.

LEMMA 6. *Let $0 \le x(n) \le z(n) \le n$ be any sequences of real numbers. For some sequence $\{c_n\}$ of real numbers, the weak law of large numbers*

$$P\left(\left|\sum_{x(n) < j \le z(n)} \frac{a(j)}{\beta(n)} \xi_j - c_n\right| \ge \gamma\right) = o(1) \tag{3.6}$$

for each $\gamma > 0$, *holds if and only if*

$$B(z(n), n) - B(x(n), n) = o(1). \tag{3.7}$$

If (3.6) *holds, then it holds with*

$$c_n = \sum_{x(n)<j\le z(n)} \left(\frac{a(j)}{\beta(n)}\right)^* \frac{\theta}{j}.$$

Further, if (3.7) *holds then for each* $\gamma > 0$,

$$\nu_{n,\theta}\left(\left|\sum_{x(n)<j\le z(n)} \frac{a(j)}{\beta(n)} k_j(\sigma) - \sum_{x(n)<j\le z(n)} \left(\frac{a(j)}{\beta(n)}\right)^* \frac{\theta}{j}\right| \ge \gamma\right) = o(1). \tag{3.8}$$

PROOF. The first part follows from Theorem 4 of Chapter 9 of Petrov (1975). Lemma 3 and (3.6) yield (3.8).

Without loss of generality by Lemma 5, we assume from now on that

$$h_j(k_j(\sigma)) = a(j)k_j(\sigma).$$

LEMMA 7. *If* $H_n(1)$ *converges weakly to a non-degenerate random variable, then*

$$\liminf_{n\to\infty} B(n, n) > 0. \tag{3.9}$$

If the sequence $\{\nu_{n,\theta} \cdot H_n^{-1}\}$ *is tight in* $\boldsymbol{D}[0, 1]$, *then*

$$B(n, n) \ll 1 \tag{3.10}$$

and for any $r(n) \to \infty$ *with* $r(n) = o(n)$,

$$\beta(r(n)) \ll \beta(n). \tag{3.11}$$

Further, if $H_n(1)$ *converges weakly, and if* (3.9), (3.10) *and*

$$B(n, n) - B(nu, n) \to 0, \quad \text{for some } \frac{3}{4} < u < 1 \tag{3.12}$$

hold, then

$$\beta(un) \ll \beta(n). \tag{3.13}$$

PROOF. The inequality (3.9) follows from (3.8). To prove (3.10), now suppose it is false. Then for a subsequence $n := n_k \to \infty$, $B(n, n) \to \infty$. Set $s_n := B(n, n)^{-1/2}$ and $r = r(n) := y(s_n)$. If $r > \varepsilon n$ for some $0 < \varepsilon < 1$, then clearly we have

$$B(\varepsilon n, n) \le B(r, n) = B(y(s_n), n) \le s_n B(n, n) \le B(n, n)^{1/2}$$

and

$$B(n,n) - B(\varepsilon n, n) \le 1 + \log(1/\varepsilon) \ll 1.$$

These inequalities together imply $B(n,n) - B(n,n)^{1/2} \ll 1$, which violates the assumption $B(n,n) \to \infty$. Thus, if $B(n,n) \to \infty$ for some subsequence $n := n_k \to \infty$, then $r(n) = o(n)$ and $s_n \to 0$. Consequently, Lemma 2 and the tightness of the family of measures $\nu_{n,\theta} \cdot H_n^{-1}$ (see Billingsley (1968)) imply

$$P\left(\left|\sum_{j\le r} X_{nj} - A(r,n)\right| \ge \gamma\right) = \nu_{n,\theta}\left(\left|\sum_{j\le r} a(j)k_j(\sigma)/\beta(n) - A(r,n)\right| \ge \gamma\right) + o(1)$$
$$= o(1)$$

for any $\gamma > 0$. Hence by Lemma 6, $B(r,n) = o(1)$. But as $\beta(n) \to \infty$, we have

$$B(r,n) = B(r+1,n) + o(1) \ge B(n,n)^{1/2} + o(1),$$

which contradicts the earlier supposition that $\limsup B(n,n) = \infty$. This establishes (3.10).

To prove (3.11), let on the contrary $\eta_n := \beta(r)/\beta(n) \to \infty$, for some $r := r(n) \to \infty$, and $r = o(n)$ on some subsequence $n := n' \to \infty$. Now tightness of the family $\{\nu_{n,\theta} \cdot H_n^{-1}\}$ implies stochastic boundedness in $\boldsymbol{D}[0,1]$, hence $\nu_{n,\theta}(|H_n(\sigma,t_n)| \ge \gamma\eta_n) = o(1)$ for an arbitrary $\gamma > 0$ and any sequence $t_n \in [0,1]$. If $t_n = B(r,n)/B(n,n)$, then $y(t_n) = r$. By Lemma 2 we obtain that

$$P(|\eta_n^{-1} X_n(t_n)| \ge \gamma) = \nu_{n,\theta}(|H_n(t_n)| \ge \gamma\eta_n) + o(1) = o(1). \tag{3.14}$$

As $\eta_n^{-1} X_n(t_n) = \sum_{j\le r}(a_j/\beta(r))\xi_j - c_r'$, for some sequence of real numbers $\{c_r'\}$, we have by (3.14) and Lemma 6,

$$B(r,r) = \sum_{j\le r}\left(\frac{a(j)}{\beta(r)}\right)^{*^2}\frac{1}{j} = o(1).$$

This contradicts (3.9) establishing (3.11).

To prove (3.13), we substitute nu for n in (3.12) several times to arrive at

$$\sum_{u^K n \le j \le n}\left(\frac{a(j)}{\beta(n,K)}\right)^{*^2}\frac{1}{j} = o(1) \tag{3.15}$$

for any fixed $K \ge 1$ with $\beta(n,K) := \max_{0\le l\le K}\beta(u^l n)$. From (3.15) we can get a strictly increasing sequence of integers $n_m > 2^m$ such that for all $n \ge n_m$,

$$\sum_{u^m n \le j \le n}\left(\frac{a(j)}{\beta(n,m)}\right)^{*^2}\frac{1}{j} < \frac{1}{m}. \tag{3.16}$$

Let $K = K(n) := m$ for $n_m \le n < n_{m+1}$ and $\tilde{\beta}(n) = \beta(n,K)$. As $nu^K \ge u^m n_m \ge (2u)^m$, clearly $u^K n \to \infty$ and (3.15) holds with $K = K(n) \to \infty$. Note that $\tilde{\beta}(n) =: \beta(u^{l_0} n)$ for some $1 \le l_0 = l(n,K) \le K$ and $r = r(n) := u^K n = o(n)$. If $\tilde{\beta}(n)/\beta(n) \to \infty$ for a subsequence $n := n' \to \infty$, then $(\beta(n)/\tilde{\beta}(n))H_n(\cdot,1) \Rightarrow 0$ and by (3.15),

$$\frac{\beta(n)}{\tilde{\beta}(n)} H_n^r(\cdot,1) - A_n \Rightarrow 0, \tag{3.17}$$

for some sequence of real numbers A_n. By (3.15), (3.17), and Lemmas 1 and 2, it follows that $\tilde{\beta}(n)^{-1}\sum_{j=1}^{n} a(j)\xi_j - c_n \Rightarrow 0$ for some sequence of real numbers c_n. Lemma 6 now yields that

$$o(1) = \sum_{j\le n}\left(\frac{a(j)}{\tilde{\beta}(n)}\right)^{*^2}\frac{1}{j} \ge B(u^{l_0}n, u^{l_0}n),$$

which violates (3.9). This proves boundedness of $\tilde{\beta}(n)/\beta(n)$, which in turn implies (3.13). This completes the proof of Lemma 7.

4. Proofs of the results

Recall that as mentioned earlier, we assume without loss of generality that $h_j(k_j(\sigma)) = a(j)k_j(\sigma)$. We first establish Theorem 2.

PROOF OF THEOREM 2. Let n_k be a strictly increasing sequence tending to infinity and satisfying,

$$B(n,n) - B(nk^{-1},n) < k^{-1},$$

for all $n \ge n_k$, $k \ge 1$. By taking $r = r(n) = nk^{-1}$ for $n_k \le n < n_{k+1}$, we get that $r(n) = o(n)$ and $B(n,n) - B(r,n) = o(1)$. By Lemma 1, the processes X_n and X_n^r can converge only simultaneously and to the same limit. By Lemma 3 we also have

$$\nu_{n,\theta}\left(\sup_{0\le t\le 1}|H_n(\sigma,t) - H_n^r(\sigma,t)| \ge \gamma\right) \ll P_{n,r}^{\theta\wedge 1}(\gamma) + o(1) = o(1).$$

Hence the processes H_n and H_n^r can converge only simultaneously and to the same limit. Thus by Lemma 2, we have $H_n \Rightarrow X$ if and only if $X_n \Rightarrow X$. The conditions for the weak convergence of $X_n \Rightarrow X$ may be established from the general results on weak convergence to additive process (see for example, Sato (1999)). For the form of time function considered here, the derivation is not trivial and it requires some additional work. Such details are given in Manstavičius (1985). Thus, the assertion of Theorem 2 now follows from Theorem 1 of Manstavičius (1985) on the equivalence of Condition (II) and $X_n \Rightarrow X$.

PROOF OF THEOREM 1. *Sufficiency.* Let $0 < u \le 1$ be fixed. By Condition (I), we have for $\beta(un) \le 2\beta(n)$ all large n. Hence $(|v|\beta(un)/\beta(n))^* \le 2|v^*|$. The representation

$$B(un,n) = \int_{\boldsymbol{R}}\left(v\frac{\beta(un)}{\beta(n)}\right)^{*^2} v^{*^{-2}}\,d_v \sum_{\substack{j\le un\\ a(j)<v\beta(un)}}\left(\frac{a(j)}{\beta(un)}\right)^{*^2}\frac{1}{j} \tag{4.1}$$

$$= \int_{\boldsymbol{R}}\left(v\frac{\beta(un)}{\beta(n)}\right)^{*^2} v^{*^{-2}}\,d_v\Psi_{un}(v),$$

the dominated convergence theorem, and Condition (II) imply that

$$B(un,n) \to \int_{\boldsymbol{R}} d\Psi(v) = \Psi(+\infty)$$

for each $0 < u \leq 1$. Hence an application of Theorem 2 establishes sufficiency part of Theorem 1. Note that from (2.2) we can infer that $X(1)$ has a non-degenerate distribution.

To prove *necessity part*, recall that we now have $H_n \Rightarrow X$. Hence there exists a countable set $T \subset [0,1)$ such that $H_n(\cdot,t) - H_n(\cdot,s) \Rightarrow X(t) - X(s)$ for all $t, s \notin T$. By Lemma 7,

$$0 < c_0 < B(n,n) < c_1 < \infty, \tag{4.2}$$

for some constants c_0, c_1 and for all sufficiently large n. We shall now prove that

$$\tau_n := \sup\{u \leq 1 : y(u) \leq (2/3)n\} \to 1. \tag{4.3}$$

We then use (4.3) to establish

$$B(n,n) - B(\varepsilon n, n) = o(1) \quad \text{for each} \quad 0 < \varepsilon < 1. \tag{4.4}$$

If on the contrary $\tau_n \to t' < 1$ for some subsequence $n := n' \to \infty$, then we can choose points s, $t \notin T$ such that $t' < s < t < 1$. Since $(2/3)n \leq y(s) \leq y(t) \leq n$ for sufficiently large n, we use Lemma 4 for each of the characteristic functions of the increments of H_n in the intervals $[s,1]$, $[s,t]$, $[t,1]$. As the limit process has independent increments, we arrive at the equality

$$\begin{aligned} &1 + \sum_{y(s)<j\leq n} \frac{1}{j}(e^{i\lambda a(j)/\beta(n)} - 1)d_{jn} + o(1) \\ &\quad = \left(1 + \sum_{y(s)<j\leq y(t)} \frac{1}{j}(e^{i\lambda a(j)/\beta(n)} - 1)d_{jn}\right)\left(1 + \sum_{y(t)<j\leq n} \frac{1}{j}(e^{i\lambda a(j)/\beta(n)} - 1)d_{jn}\right) \end{aligned}$$

uniformly in $|\lambda| \leq K$ for any $K > 0$. Hence uniformly in $|\lambda| \leq K$,

$$\left(\sum_{y(s)<j\leq y(t)} \frac{1}{j}(e^{i\lambda a(j)/\beta(n)} - 1)d_{jn}\right)\left(\sum_{y(t)<j\leq n} \frac{1}{j}(e^{i\lambda a(j)/\beta(n)} - 1)d_{jn}\right) = o(1).$$

Since these sums are uniformly bounded, there exists a further subsequence such that the real part of one of these sums tend to zero on the λ set A of Lebesgue measure at least K. The set A is symmetric with respect to the origin and by the inequality $1 - \cos(u+v) \ll (1-\cos u) + (1 - \cos v)$ we see that $A \pm A \subset A$. Hence $A = \boldsymbol{R}$ by Steinhaus's lemma (see Bingham *et al.* (1989), Theorem 1.1.1). If, for simplicity, this was the first sum, since $d_{jn} \geq c(\theta) > 0$ for $(2/3)n \leq j \leq n$, we have

$$\sum_{y(s)<j\leq y(t)} \frac{1}{j}(1 - \cos(\lambda a(j)/\beta(n))) = o(1).$$

We now consider two parts of the sum; one over j satisfying $|a(j)| < \beta(n)$ and the other over $|a(j)| \geq \beta(n)$. Using the inequality $1 - \cos u \geq cu^2$, $c > 0$, for $|u| \leq 1$ in the first sum case and integrating over $\lambda \in [0,2]$ in the second case, we obtain

$$(t-s)B(n,n) = \sum_{y(s)<j\leq y(t)} \left(\frac{a(j)}{\beta(n)}\right)^{*^2} \frac{1}{j} + o(1) = o(1),$$

contradicting (4.2). So $\tau_n \to 1$. Hence we have, for all large n and $(3/4) \le u \le 1$, that

$$1 + o(1) \le \tau_n \le \frac{B(y(\tau_n)+1, n)}{B(n,n)} \le \frac{B((2/3)n+1, n)}{B(n,n)} \le \frac{B(un,n)}{B(n,n)} \le 1.$$

Since by (4.2), $B(n,n)$ is bounded, it follows that for each $(3/4) \le u \le 1$,

$$B(n,n) - B(un,n) = o(1). \tag{4.5}$$

Consequently, (3.13) holds by Lemma 7. Now (3.13) and a repeated use of

$$\begin{aligned} B(n,n) - B(u^2n,n) &= B(un,n) - B(u^2n,n) + B(n,n) - B(un,n) \\ &= B(un,n) - B(u^2n,n) + o(1) \\ &\ll \big(B(un,un) - B(u^2n,un)\big) + o(1) = o(1) \end{aligned}$$

yield (4.4). By Theorem 2, this leads to Condition (II). Another application of Theorem 2 yields that $X_n \Rightarrow X$. Hence $X_{un}(1) \Rightarrow X(1)$ and by (4.4), $X_n^{un}(1) \Rightarrow X(1)$ for each $0 < u < 1$. As $X(1)$ has a non-degenerate distribution, it follows that $\beta(un)/\beta(n) \to 1$ for all $0 < u < 1$, which implies Condition (I). This completes the proof of Theorem 1.

PROOF OF THEOREM 3. It suffices to deal with the necessity part only. The idea is rather simple. From (2.3) and $H_n \Rightarrow X$ we need to construct a subsequence $n := n' \to \infty$ such that $X_n \Rightarrow X$. Since in this case, X being the limit process of a partial sum process of independent summands, it has independent increments. Theorem 1 then yields Theorem 3. Nevertheless the details are a bit involved. We divide the proof into several steps.

Since $X(1)$ has non-degenerate distribution, we first note that (3.9) holds by Lemma 7. Now suppose for some $0 < \delta < 1$,

$$\beta(n) \ge 2\beta(\delta n) \quad \text{for all} \quad n \ge n_0 = n_0(\delta). \tag{4.6}$$

We use (4.6) repeatedly for $\delta^i n$, $i = 1, 2, \dots, t$, where $\delta^{t-1} n \ge n_0 > \delta^t n$, to get $\beta(n) \ge 2^t \beta(\delta^t n) = 2^t c_1$, where $c_1 = c_1(\delta) > 0$. Therefore

$$\beta(n) \ge c_1 \exp\{t \log 2\} \ge c_1 \exp\left\{\frac{\log 2}{\log \delta}(\log n_0 - \log n)\right\} =: c_2 n^\rho \tag{4.7}$$

where $c_2 > 0$ and $\rho = -(\log 2)/\log \delta > 0$ provided $n \ge n_0$. By (2.3) we get that

$$B(n,n) \ll \sum_{j \le n} \left(\frac{a(j)}{n^\rho}\right)^{*^2} \frac{1}{j} = o(1),$$

which contradicts (3.9). Hence (4.6) is false for all $0 < \delta < 1$. Thus by taking $\delta = 1/k$, $k = 2, 3, \dots$, we get an increasing sequence n_k such that $n_k > k^2$ and $\beta(n_k) < 2\beta(n_k/k)$. If we put $r := r(n_k) = n_k/k$, then as the subsequence $n := n_k \to \infty$, we have

$$\beta(n) < 2\beta(r), \quad r = o(n) \quad \text{and} \quad r \to \infty. \tag{4.8}$$

We now partition the interval $[r, n]$ by the points $r_i := r(\log(n/r))^i$, $i = 0, 1, \dots, m-1$, so that $r_m > n$. Thus $m > (\log(n/r))/\log\log(n/r) \to \infty$ as $n := n_k \to \infty$. Since $B(n,n)$ is bounded, there exists at least one index $0 \le s \le m-1$ such that

$$B(r_{s+1}, n) - B(r_s, n) = o(1). \tag{4.9}$$

By (3.11), (4.8) and as $r = o(r_{s+1})$, we have

$$\left(\frac{a(j)}{\beta(r_{s+1})}\right)^{*^2} = \left(\frac{a(j)}{\beta(n)} \cdot \frac{\beta(n)}{\beta(r)} \cdot \frac{\beta(r)}{\beta(r_{s+1})}\right)^{*^2} \ll \left(\frac{a(j)}{\beta(n)}\right)^{*^2}.$$

Hence from (4.9),

$$B(r_{s+1}, r_{s+1}) - B(r_s, r_{s+1}) \ll B(r_{s+1}, n) - B(r_s, n) = o(1). \tag{4.10}$$

Since $r_s = o(r_{s+1})$, we have by (3.8) and (4.10) that if $H_n \Rightarrow X$, then $H^{r_s}_{r_{s+1}} \Rightarrow X$. Lemmas 1 and 2 along with another application of (4.10) now imply that $X_{r_{s+1}} \Rightarrow X$. Since X is a limit process of some partial sum processes of independent summands, it must have independent increments. Theorem 3 now follows from Theorem 1.

Acknowledgements

We thank the two referees for constructive comments and suggestions which improved the paper.

References

Arratia, R. and Tavaré, S. (1992). Limit theorems for combinatorial structures via discrete process approximations, *Random Structures Algorithms*, **3**, 321–345.

Arratia, R., Barbour, A. D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula, *Ann. Appl. Probab.*, **2**, 519–535.

Babu, G. J. (1973). A note on the invariance principle for additive functions, *Sankhyā Ser. A*, **35**, 307–310.

Babu, G. J. and Manstavičius, E. (1999). Brownian motion for random permutations, *Sankhyā Ser. A*, **61**, 312–327.

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1989). *Regular variation*, 2nd ed., Cambridge University Press, Cambridge, Massachusetts.

DeLaurentis, J. M. and Pittel, B. G. (1985). Random permutations and the Brownian motion, *Pacific J. Math.*, **119**, 287–301.

Donnelly, P., Kurtz, T. G. and Tavaré, S. (1991). On the functional central limit theorem for the Ewens Sampling Formula, *Ann. Appl. Probab.*, **1**, 539–545.

Drmota, M. and Tichy, R. F. (1997). *Sequences, Discrepancies and Applications*, Lecture Notes in Mathematics, **1651**, Springer, Berlin.

Erdös, P. and Turán, P. (1965). On some problems of a statistical group theory I, *Z. Wahrsch. Verw. Gebiete*, **4**, 175–186.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology*, **3**, 87–112.

Ewens, W. J. (1979). *Mathematical Population Genetics*, Springer, Berlin.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Volume 1, 2nd ed., Wiley, New York.

Hansen, J. C. (1990). A functional central limit theorem for the Ewens Sampling Formula, *J. Appl. Probab.*, **27**, 28–43.

Hirth, U. M. (1997*a*). A Poisson approximation for the Dirichlet distribution, the Ewens sampling formula and the Griffith-Engen-McCloskey law by the Stein-Chen coupling method, *Bernoulli*, **3**, 225–232.

Hirth, U. M. (1997*b*). Probabilistic number theory, the GEM/Poisson-Dirichlet distribution and the arc-sine law, *Combin. Probab. Comput.*, **6**, 57–77.

Hirth, U. M. (1997*c*). From GEM back to Dirichlet via Hoppe's urn, *Combin. Probab. Comput.*, **6**, 185–195.

Kingman, J. F. C. (1980). *Mathematics of Genetic Diversity*, SIAM, Philadelphia, Pennsylvania.

Kubilius, J. (1964). *Probabilistic Methods in the Theory of Numbers*, Translations of Mathematical Monographs, **11**, AMS, Providence, Rhode Island.

Manstavičius, E. (1984). Arithmetic simulation of stochastic processes, *Lithuanian Math. J.*, **24**, 276–285.

Manstavičius, E. (1985). Additive functions and stochastic processes, *Lithuanian Math. J.*, **25**, 52–61.

Manstavičius, E. (1996), Additive and multiplicative functions on random permutations, *Lithuanian Math. J.*, **36**(4), 400–408.

Petrov, V. V. (1975). *Sums of Independent Random Variables*, Springer, New York.

Philipp, W. (1973). Arithmetic functions and Brownian motion, *Proc. Sympos. Pure Math.*, **24**, 233–246.

Sato, K.-I. (1999). *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge.

Tenenbaum, G. (1995). *Introduction to Analytic and Probabilistic Number Theory*, Cambridge University Press, Cambridge.

Timofeev, N. M. and Usmanov, H. H. (1984). Arithmetic modelling of random processes with independent increments, *Dokl. Akad. Nauk Respub. Tadzhikistan*, **27**(10), 556–559 (Russian).

Timofeev, N. M. and Usmanov, H. H. (1986). On a class of arithmetic models of random processes, *Dokl. Akad. Nauk Respub. Tadzhikistan*, **29**(6), 330–334 (Russian).