ON THE ASYMPTOTIC DISTRIBUTION OF LARGE PRIME FACTORS

PETER DONNELLY AND GEOFFREY GRIMMETT

ABSTRACT

A random integer N, drawn uniformly from the set $\{1, 2, \ldots, n\}$, has a prime factorization of the form $N = \alpha_1 \alpha_2 \ldots \alpha_M$ where $\alpha_1 \geqslant \alpha_2 \geqslant \ldots \geqslant \alpha_M$. We establish the asymptotic distribution, as $n \to \infty$, of the vector $\mathbf{A}(n) = (\log \alpha_i / \log N : i \geqslant 1)$ in a transparent manner. By randomly re-ordering the components of $\mathbf{A}(n)$, in a size-biased manner, we obtain a new vector $\mathbf{B}(n)$ whose asymptotic distribution is the GEM distribution with parameter 1; this is a distribution on the infinite-dimensional simplex of vectors $(x_1, x_2, \ldots)$ having non-negative components with unit sum. Using a standard continuity argument, this entails the weak convergence of $\mathbf{A}(n)$ to the corresponding Poisson–Dirichlet distribution on this simplex; this result was obtained by Billingsley [3].

1. *Introduction*

Let N be a positive integer, and let α be the largest prime factor of N. The typical behaviour of α, when N is large, is such that $\log \alpha / \log N$ has a certain asymptotic distribution over the interval $(0, 1)$; that is to say, α behaves roughly in the manner of N^W where W is a random variable whose distribution is the first component of the Poisson–Dirichlet distribution with parameter 1. It was apparently Dickman [9] who was the first to study such a question (the distribution function of W is sometimes named after him), and this matter has received some attention since then. Ramaswami [19] established the existence, and some properties, of the limiting distribution of $\log \alpha / \log N$ (in an appropriate limit); de Bruijn [6, 7] provided more precise asymptotic information. This distribution arises also in studying least kth power non-residues (see Norton [18] for a review).

In all this work, the limit in question is the usual one, amounting in probabilistic terms to the following. Let n be a positive integer, and let $N(n)$ be a random integer chosen uniformly from the set $\{1, 2, \ldots, n\}$. We are concerned with the asymptotic distributions of certain functions of $N = N(n)$, in the limit as $n \to \infty$.

Work on the *largest* prime divisor of N was extended by Billingsley [3] to the sequence $(\alpha_k : k \geqslant 1)$ of prime divisors of N, written in non-increasing order. He explored the asymptotics of the joint distribution of the vector $(\alpha_1, \alpha_2, \ldots, \alpha_r)$. More precisely, he studied the vector $\mathbf{A}(n) = (\log \alpha_k / \log N : k \geqslant 1)$, and calculated its asymptotic finite-dimensional distributions. His result has been rediscovered in different forms by Knuth and Trabb Pardo [17] and Vershik [21].

Our purpose in this paper is to present an elementary and especially transparent proof of Billingsley's theorem, via a different route to that used by him. Our proof is

Received 13 November 1991.

1991 *Mathematics Subject Classification* 11N25.

The first author was supported in part by SERC Advanced Fellowship B/AF/1255. The second author was supported in part by Cornell University and by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

based on the following observation. The (Poisson–Dirichlet) distribution that arises as the limit for normalized prime divisors is somewhat complicated, and therefore presents some difficulty in its study (see Griffiths [13] and Donnelly and Joyce [10] for accounts of this distribution and its properties). On the other hand, the Poisson–Dirichlet distribution may be represented as an invertible function of another distribution, called the GEM distribution. In contrast to the Poisson–Dirichlet distribution, the GEM distribution has a relatively simple structure in terms of a sequence of independent uniform random variables. Using completely elementary means, we shall prove that a certain (random) function of $\mathbf{A}(n)$ converges to the GEM distribution, and we shall deduce by a general argument that $\mathbf{A}(n)$ has the Poisson–Dirichlet distribution in the limit.

We remark, as have others, that the Poisson–Dirichlet and GEM distributions occur in various contexts. They occur in the study of the normalized cycle lengths of a random permutation, the Poisson–Dirichlet distribution as the limit of ordered cycle lengths, and the GEM distribution as the limit of the cycle lengths when the cycles are labelled in increasing order of their smallest members (for random permutations, this labelling is equivalent to a random size-biased ordering similar to that which we shall encounter in the next section); see [20, 10].

In greater generality, the Poisson–Dirichlet and GEM distributions form one-parameter families indexed by a parameter θ. As in this paper, the limiting distributions for random permutations are such that $\theta = 1$. Other values of θ arise in other settings. One encounters the case $\theta = \frac{1}{2}$ in studying the sizes of components of a random mapping; see, for example, [1]. The case of general θ arises in the study of neutral models in population genetics, the value of θ being related to the rate of mutation; see [16, 11].

We give a formal description of the Poisson–Dirichlet and GEM distributions in the next section, where we state our result and some of its consequences. This is followed by a brief account, in a few lines, of the calculation required for the proof. Section 3 contains a formal proof.

2. *Statement of result*

We write $\mathbb{N} = \{1, 2, \ldots\}$ and denote by

$$C = \prod_{i \in \mathbb{N}} [0, 1]$$

the $\mathbb{N}$-dimensional unit cube. We endow C with the product Euclidean topology and the Borel σ-field. Write the vector $\mathbf{x} \in C$ in the form $\mathbf{x} = (x_1, x_2, \ldots)$, and denote by

$$\Delta = \left\{ \mathbf{x} \in C : \sum_{i=1}^{\infty} x_i = 1 \right\}, \quad T = \{ \mathbf{x} \in \Delta : x_1 \geqslant x_2 \geqslant \ldots \},$$

the simplex of vectors in C with unit sum, and the set of vectors in Δ with non-increasing components. We give to Δ and T the topologies which they inherit as subsets of the topological space C.

We now introduce the relevant probability measures on these spaces. Let $U_1, U_2, \ldots$ be a sequence of independent random variables each having the uniform

distribution on [0, 1]. We denote by λ the probability measure on C associated with the random element $(U_1, U_2, \ldots)$; clearly λ is Lebesgue measure on C. Next, we define

$$X_1 = U_1, \quad X_2 = (1-U_1)\,U_2, \quad X_3 = (1-U_1)(1-U_2)\,U_3, \ldots,$$

noting that $\mathbf{X} = (X_1, X_2, \ldots)$ belongs almost surely to Δ. We write γ for the probability measure on Δ induced by $\mathbf{X}$: thus

$$\gamma(A) = P(\mathbf{X} \in A) \quad \text{for Borel subsets } A \text{ of } \Delta.$$

The measure γ is called the GEM distribution with parameter 1. Finally, we write $X_{(1)}, X_{(2)}, \ldots$ for the order statistics (in non-increasing order) of the X_i, that is to say, we have that $X_{(1)} \geqslant X_{(2)} \geqslant \ldots$ and the $X_{(j)}$ are a rearrangement of the X_i. We denote by π the probability measure on T associated with the vector $\mathbf{X}_{(\,)} = (X_{(1)}, X_{(2)}, \ldots)$: thus

$$\pi(A) = P(\mathbf{X}_{(\,)} \in A) \quad \text{for Borel subsets } A \text{ of } T,$$

and π is the Poisson–Dirichlet distribution (having parameter 1).

If, instead of the uniform distribution, we had taken as common density function of the U_i the function $f(u) = \theta(1-u)^{\theta-1}, 0 \leqslant u \leqslant 1$, then the consequent measures γ and π would be the GEM and Poisson–Dirichlet distributions with parameter θ.

We turn now to the question of prime factorization. Let n be a positive integer; later we shall take the limit as $n \to \infty$. Let $N(n)$ be a random integer chosen uniformly from the set $\{1, 2, \ldots, n\}$. The integer $N(n)$ has a prime factorization in the form

$$N(n) = \prod_p p^{A(p,\,n)}$$

where $A(p, n)$ is the multiplicity of the prime p. All summations and products over p are deemed to be over the set Π of prime numbers. If $N(n) = 1$, we write $A(p, n) = 0$ for all p. Let

$$M(n) = \sum_p A(p, n)$$

be the number of prime divisors of $N(n)$, counted according to their multiplicities. Writing $\alpha_1, \alpha_2, \ldots, \alpha_{M(n)}$ for the prime divisors of $N(n)$, listed in non-increasing order, we place the α_j in a random order in the following manner. The first term is chosen at random from the sequence of α_j in a size-biased way; more specifically, the prime α_j is chosen with probability proportional to $\log \alpha_j$. Having chosen the first term, the second is chosen similarly from the remaining divisors, and so on. Taking into account the multiplicities of each prime factor of $N(n)$, this amounts to the following. The first term, written $D_1(n)$, has distribution

$$P(D_1(n) = p \mid N(n)) = \frac{A(p, n)\log p}{\log N(n)}, \quad p \in \Pi.$$

Conditional on $N(n)$ and $D_1(n)$, the second term $D_2(n)$ has mass function

$$P(D_2(n) = p \mid N(n), D_1(n)) = \frac{A_2(p, n)\log p}{\log\{N(n)/D_1(n)\}}, \quad p \in \Pi,$$

where

$$A_2(p, n) = \begin{cases} A(p, n) & \text{if } D_1(n) \neq p \\ A(p, n) - 1 & \text{if } D_1(n) = p. \end{cases}$$

More generally, the kth term $D_k(n)$ has conditional mass function

$$P(D_k(n) = p \mid N(n), D_1(n), \ldots, D_{k-1}(n)) = \frac{A_k(p,n)\log p}{\log R_k(n)}, \quad p \in \Pi, \tag{2.1}$$

where $A_k(p,n) = \max\{0, A(p,n) - |\{i < k: D_i(n) = p\}|\}$, and

$$R_k(n) = \frac{N(n)}{D_1(n)\,D_2(n)\ldots D_{k-1}(n)} = \exp\left(\sum_p A_k(p,n)\log p\right).$$

Note that $R_1(n) = N(n)$. We obtain in this way a sequence $D_1(n), D_2(n), \ldots, D_{M(n)}(n)$ of prime divisors of $N(n)$, and we shall study the asymptotic distribution of this vector in the limit as $n \to \infty$. Rather than working directly with this sequence, it is more convenient to work with the random vector $\mathbf{B}(n) = (B_1(n), B_2(n), \ldots)$ defined as follows: we set $B_1(n) = \log D_1(n)/\log N(n)$, and more generally

$$B_i(n) = \begin{cases} \dfrac{\log D_i(n)}{\log R_i(n)} & \text{if } 1 \leqslant i \leqslant M(n), \\ 0 & \text{if } i > M(n). \end{cases}$$

Note that $0 \leqslant B_i(n) \leqslant 1$ for all i, and therefore $\mathbf{B}(n) \in C$.

THEOREM 1. *The vector* $\mathbf{B}(n)$ *converges weakly to Lebesgue measure* λ *on* C, *in the limit as* $n \to \infty$.

We abuse terminology here and later by speaking of the weak convergence of a random vector rather than of its distribution.

We give a formal proof of the theorem in the next section. This proof is rather simple, and is based upon the elementary calculation which follows. See [2] for a general discussion of weak convergence of probability measures.

In order to prove the theorem, it suffices to show that, for each k, the vector $(B_1(n), B_2(n), \ldots, B_k(n))$ converges weakly to the product of k uniform measures. Consider the case when $k = 1$, and let $0 < a < 1$. The following rough calculation may be made valid. By neglecting the question of the *multiplicities* of the prime divisors of $N(n)$, we have that

$$\begin{aligned} P(B_1(n) \leqslant a) &= P(D_1(n) \leqslant N(n)^a) \\ &= \sum_{m=1}^{n} \sum_{\substack{p \leqslant m^a \\ p \mid m}} P(D_1(n) = p, N(n) = m) \\ &\simeq \sum_{p \leqslant n^a} \sum_{\substack{m:\, p \mid m, \\ p^{1/a} \leqslant m \leqslant n}} \frac{1}{n} \cdot \frac{\log p}{\log m} \\ &\simeq \sum_{p \leqslant n^a} \left(\frac{\log p}{n}\right)\left(\frac{n}{p \log n}\right) \\ &\sim \frac{\log(n^a)}{\log n} = a. \end{aligned}$$

The theorem will be proved by an elaboration of this basic calculation, in which we deal with the question of multiplicities, extend the calculation to general k, and tighten up the asymptotic analysis a little.

We make two remarks about the above calculation. First, its success hinges upon the elementary fact that

$$\sum_{p \leqslant K} \frac{\log p}{p} = \log K + O(1) \quad \text{as } K \to \infty;$$

see [14, Theorem 425]. Secondly, the calculation succeeds for the function $B_1(n)$ of the size-biased random variable $D_1(n)$ precisely because of the conjunction of logarithms. Amongst the prime divisors p of $N(n)$, the variable $D_1(n)$ takes the value p with a probability proportional to $\log p$, while the density of primes in the region of an integer p is approximately $1/\log p$.

The following corollaries are immediate consequences of Theorem 1; the second is implicit in the paper of Billingsley [3], and appears without proof in Vershik [21].

COROLLARY 2. *In the notation above, let*

$$C_i(n) = \begin{cases} \dfrac{\log D_i(n)}{\log N(n)} & \text{if } 1 \leqslant i \leqslant M(n), \\ 0 & \text{otherwise.} \end{cases}$$

The vector $\mathbf{C}(n) = (C_1(n), C_2(n), \ldots)$ *converges weakly to* γ, *the GEM distribution with parameter* 1, *as* $n \to \infty$.

Proof. For each k, the vector $(C_1(n), C_2(n), \ldots, C_k(n))$ is a continuous function of $(B_1(n), B_2(n), \ldots, B_k(n))$; therefore the limiting distribution of the first vector is given by the corresponding function of that of the second. The claim now follows from Theorem 1, using the fact that the GEM distribution is specified by its finite-dimensional distributions.

COROLLARY 3. *In the notation above, let*

$$A_i(n) = \begin{cases} \dfrac{\log \alpha_i}{\log N(n)} & \text{if } 1 \leqslant i \leqslant M(n), \\ 0 & \text{otherwise.} \end{cases}$$

The vector $\mathbf{A}(n) = (A_1(n), A_2(n), \ldots)$ *converges weakly to* π, *the Poisson–Dirichlet distribution with parameter* 1, *as* $n \to \infty$.

Proof. The function which rearranges the components of a vector $\mathbf{x} \in \Delta$ into non-increasing order is a continuous mapping from Δ into T (see [10]). The claim follows from Corollary 2.

We close this section with several remarks.

(a) From Corollary 3 and known properties of the Poisson–Dirichlet distribution, various results follow. For example, as was stated by Knuth and Trabb Pardo [17],

$$P(\alpha_k \leqslant N(n)^a) \longrightarrow F_k(a), \quad 0 \leqslant a \leqslant 1,$$

where F_k is the marginal distribution of the kth component of the Poisson–Dirichlet distribution with parameter 1. Knuth and Trabb Pardo have derived certain properties of the F_k. Related properties have been derived independently and in other

contexts; see, for example, Shepp and Lloyd [20] and earlier papers for the problem of random permutations, and Watterson [22] for general Poisson–Dirichlet distributions. Related work appears in Diaconis [8].

(b) It is a consequence of the bounded convergence theorem that all (joint) moments of the components of the vector $\mathbf{A}(n)$ converge to the corresponding moments of the Poisson–Dirichlet distribution.

It is a useful property of the Poisson–Dirichlet distribution (with parameter θ) that

$$E(|\{j: X_{(j)} \in A\}|) = \int_A \phi(y)\,dy$$

for any measurable subset A of $[0,1]$, where $\phi(y) = \theta y^{-1}(1-y)^{\theta-1}$ is called the frequency spectrum (see [12]). In our case $\theta = 1$, and consequently

$$E\left(\sum_{j=1}^{\infty} g\left(\frac{\log \alpha_j}{\log N(n)}\right)\right) \longrightarrow \int_0^1 \frac{g(y)}{y}\,dy \quad \text{as } n \longrightarrow \infty$$

for all appropriate functions g (see, for example, [15]). This, together with its multivariate generalizations, provides a convenient method for calculating the asymptotic expectations of certain symmetric functions of the vector $\mathbf{A}(n)$.

(c) There is another characterization of the Poisson–Dirichlet distribution which may prove to be useful. Let $Y_1, Y_2, \ldots$ be the points of a non-homogeneous Poisson process on $(0,\infty)$ with mean measure density $\theta e^{-y}/y$, these points being written in decreasing order. It is easily seen that the Y_i are a.s. bounded, and that $Z = \sum_{i=1}^{\infty} Y_i < \infty$ a.s. (actually, Z has a gamma distribution). It may be shown that the vector $(Y_1, Y_2, \ldots)/Z$ has the Poisson–Dirichlet distribution with parameter θ, and is independent of Z. The Poisson–Dirichlet distribution is not easy to work with, and this representation can be of value; see [13].

(d) Since the GEM distribution may be specified in terms of a family of independent, identically distributed random variables, it is usually easier to work with this distribution rather than with the Poisson–Dirichlet distribution. For example, a law of large numbers and a central limit theorem for these distributions is easily derived from the fact that the GEM distribution may be expressed in terms of an independent, identically distributed sequence. As a further example, Theorem 6 of [21] is an immediate consequence of an exact asymptotic result for the GEM distribution. One obtains, in the above notation, that

$$P(D_1(n)\,D_2(n)\ldots D_k(n) \geqslant n^b) \longrightarrow \frac{1}{k!}\Gamma(k+1, -\log(1-b))$$

as $n \to \infty$, where Γ is the incomplete gamma function and all logarithms are natural (use Theorem 1, together with the fact that $-\log U$ is exponentially distributed if U is uniform on $(0,1)$). This implies Vershik's Theorem 6 (part 1), since

$$\alpha_1\,\alpha_2\ldots\alpha_k \geqslant D_1(n)\,D_2(n)\ldots D_k(n).$$

(e) In principle, the error terms in the proof of Theorem 1 may be estimated, and thus one would arrive at an estimate for the rate of convergence.

(f) Wunderlich and Selfridge [23] argued heuristically that the distribution of the second largest prime factor α_2 of $N(n)$ should behave in the manner of the largest factor of a number drawn at random from $\{1, 2, \ldots, N(n)/\alpha_1\}$. As indicated by the calculations of Knuth and Trabb Pardo [17, equations (9.5) and (9.6)], this is not

correct. However, Theorem 1 demonstrates that this heuristic argument *is* correct when applied to the size-biased random permutation $D_1(n), D_2(n), \ldots$ of the sequence $\alpha_1, \alpha_2, \ldots$ of prime factors.

(g) The results above provide information about the *large* prime factors of a 'typical' integer, but provide no information concerning prime factors having smaller order. The distribution of prime factors of an integer N, these primes being of order $\exp[(\log N)^a]$ where $0 < a < 1$, is described by the Erdős–Kac central limit theorem and the subsequent invariance principle of Billingsley; see [**4**, **5**] and the references therein.

3. *Proof of Theorem* 1

Let $k \geqslant 1$, and write $\mathbf{B}_k(n)$ for the vector $(B_1(n), B_2(n), \ldots, B_k(n))$. It suffices to prove for general k that

$$P(\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}) \longrightarrow \prod_{i=1}^{k} (b_i - a_i) \quad \text{for all } \mathbf{a}, \mathbf{b} \in (0,1)^k \text{ such that } \mathbf{a} < \mathbf{b}; \tag{3.1}$$

we write $\mathbf{v} < \mathbf{w}$ (respectively $\mathbf{v} \leqslant \mathbf{w}$) if $v_i < w_i$ (respectively $v_i \leqslant w_i$) for all i. We may restrict ourselves to vectors $\mathbf{a}, \mathbf{b}$ satisfying $\mathbf{0} < \mathbf{a} < \mathbf{b} < \mathbf{1}$, in the light of the fact that Lebesgue measure on the cube $[0,1]^k$ has its entire mass on the interior of the cube. We shall prove that

$$\liminf_{n \to \infty} P(\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}) \geqslant \prod_{i=1}^{k} (b_i - a_i) \quad \text{for } \mathbf{a} < \mathbf{b}, \tag{3.2}$$

and we claim that this implies (3.1). Suppose on the contrary that (3.2) holds, but that there exist $\mathbf{c}, \mathbf{d} \in (0,1)^k$ such that $\mathbf{c} < \mathbf{d}$ and

$$\limsup_{n \to \infty} P(\mathbf{c} < \mathbf{B}_k(n) \leqslant \mathbf{d}) > \prod_{i=1}^{k} (d_i - c_i). \tag{3.3}$$

We may partition the k-dimensional cube $[0,1]^k$ as the finite disjoint union $E_1 \cup E_2 \cup \ldots \cup E_K$ in such a way that $E_1 = \prod_{i=1}^{k} (c_i, d_i]$ and each E_j is the product of sub-intervals of $[0,1]$. Now

$$1 = P(\mathbf{0} \leqslant \mathbf{B}_k(n) \leqslant \mathbf{1}) = \sum_{j=1}^{K} P(\mathbf{B}_k(n) \in E_j).$$

Denote the interior of E_j by $I_j = \prod_{i=1}^{k} (x_i, y_i)$, find ε satisfying $0 < \varepsilon < y_i - x_i$ for all i, and note that

$$P(\mathbf{B}_k(n) \in E_j) \geqslant P\left(\mathbf{B}_k(n) \in \prod_{i=1}^{k} (x_i, y_i - \varepsilon]\right);$$

therefore, by (3.2),

$$\liminf_{n \to \infty} P(\mathbf{B}_k(n) \in E_j) \geqslant \prod_{i=1}^{k} (y_i - \varepsilon - x_i) \longrightarrow \lambda_k(E_j) \quad \text{as } \varepsilon \downarrow 0,$$

where λ_k is k-dimensional Lebesgue measure. Using (3.3), we find that there exists a sequence of values of n along which

$$1 = \sum_{j=1}^{K} \lim P(\mathbf{B}_k(n) \in E_j) > \sum_{j=1}^{K} \lambda_k(E_j) = 1,$$

a contradiction.

It remains to prove (3.2). Let $\mathbf{a}, \mathbf{b} \in (0,1)^k$ satisfy $\mathbf{a} < \mathbf{b}$. We have from the definition of the $B_i(n)$ that $\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}$ if and only if

$$R_i(n)^{a_i} < D_i(n) \leqslant R_i(n)^{b_i} \quad \text{for } 1 \leqslant i \leqslant k,$$

where $R_i(n) = N(n)/\{D_1(n) D_2(n) \dots D_{i-1}(n)\}$. Therefore

$$P(\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}) = \sum_{\mathbf{p}, m} P(N(n) = m, D_i(n) = p_i \text{ for } 1 \leqslant i \leqslant k) \tag{3.4}$$

where the sum is over all vectors $\mathbf{p} = (p_1, p_2, \dots, p_k) \in \Pi^k$ and all positive integers m satisfying $1 \leqslant m \leqslant n$ and

$$\left(\frac{m}{p_1 p_2 \cdots p_{i-1}}\right)^{a_i} < p_i \leqslant \left(\frac{m}{p_1 p_2 \cdots p_{i-1}}\right)^{b_i} \quad \text{for } 1 \leqslant i \leqslant k. \tag{3.5}$$

Let $0 < \varepsilon < 1$, and note that

$$P(N(n) \geqslant \varepsilon n) \geqslant 1 - \varepsilon. \tag{3.6}$$

Consequently, by restricting ourselves to values of m exceeding εn, we lose at most ε of probability. It follows from (3.4) to (3.6) that

$$P(\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}) \geqslant \sum_{\mathbf{p}, m} P(N(n) = m, D_i(n) = p_i \text{ for } 1 \leqslant i \leqslant k) \tag{3.7}$$

where the summation is over all vectors $\mathbf{p}$ and integers m such that $\varepsilon n \leqslant m \leqslant n$ and

$$n_i^{a_i} < p_i \leqslant (\varepsilon n_i)^{b_i} \quad \text{for } 1 \leqslant i \leqslant k, \tag{3.8}$$

where $n_i = n/(p_1 p_2 \dots p_{i-1})$. For such $\mathbf{p}$ and m,

$$P(N(n) = m, D_i(n) = p_i \text{ for } 1 \leqslant i \leqslant k) \geqslant \frac{1}{n} \prod_{i=1}^{k} \frac{\log p_i}{\log\{m/(p_1 p_2 \cdots p_{i-1})\}} \tag{3.9}$$

if $p_1 p_2 \dots p_k \mid m$, and this probability is 0 otherwise; the inequality arises from the fact that the relevant multiplicities $A_j(p_i, n)$ in (2.1) are at least 1. The inequality in (3.9) remains valid when m is replaced by n on the right-hand side. Substitute the ensuing inequality into (3.7), and sum over multiples m of $p_1 p_2 \dots p_k$, to obtain that the summation in (3.7) is at least

$$\left(1 - \varepsilon - \frac{1}{n_{k+1}}\right) \sum_{\mathbf{p}} \prod_{i=1}^{k} \frac{\log p_i}{p_i \log n_i} \tag{3.10}$$

where the summation is over all sequences $\mathbf{p}$ satisfying (3.8).

We shall require lower bounds for the n_i. We have from (3.8) that

$$\frac{n_i}{n_{i+1}} = p_i \leqslant (\varepsilon n_i)^{b_i} \leqslant n_i^{b_i},$$

and hence

$$n_{i+1} \geqslant n_i^{1-b_i} \geqslant n^{\nu} \quad \text{for } 0 \leqslant i \leqslant k, \tag{3.11}$$

where $\nu = \prod_{i=1}^{k} (1 - b_i) > 0$.

It is a standard result [**14**, Theorem 425] that

$$\sum_{p \leqslant K} \frac{\log p}{p} = \log K + O(1) \quad \text{as } K \longrightarrow \infty. \tag{3.12}$$

By (3.12), we may find M such that

$$\sum_{m^{a_i}<p\leqslant(em)^{b_i}} \frac{\log p}{p\log m} \geqslant b_i - a_i - \varepsilon \quad \text{for } 1 \leqslant i \leqslant k$$

whenever $m \geqslant M$. Hence, using (3.11),

$$S_i(n_i) = \sum_{n_i^{a_i}<p\leqslant(en_i)^{b_i}} \frac{\log p}{p\log n_i}$$

satisfies

$$S_i(n_i) \geqslant b_i - a_i - \varepsilon \quad \text{if } n \geqslant M^{1/\nu}.$$

Returning to (3.10) and summing over $p_k, p_{k-1}, \ldots, p_1$ in order, we deduce that

$$P(\mathbf{a} < \mathbf{B}_k(n) \leqslant \mathbf{b}) \geqslant (1-\varepsilon-n^{-\nu})\prod_{i=1}^{k}(b_i - a_i - \varepsilon) \quad \text{if } n \geqslant M^{1/\nu}.$$

Take the limits as $n \to \infty$ and $\varepsilon \downarrow 0$ to obtain (3.2), as required.

Acknowledgements. This work was commenced while the authors were attending a London Mathematical Society Symposium on 'Probabilistic Methods in Combinatorics' at the University of Durham, supported by the SERC. We thank Harry Kesten for his useful remarks at this symposium.

References

1. D. ALDOUS, 'Exchangeability and related topics', *Ecole d'Eté de Probabilités de Saint-Flour XIII*, Lecture Notes in Mathematics 1117 (Springer, Berlin, 1985) 1–198.
2. P. BILLINGSLEY, *Convergence of probability measures* (John Wiley, New York, 1968).
3. P. BILLINGSLEY, 'On the distribution of large prime factors', *Period. Math. Hungar.* 2 (1972) 283–289.
4. P. BILLINGSLEY, 'Prime numbers and Brownian motion', *Amer. Math. Monthly* 80 (1973) 1099–1115.
5. P. BILLINGSLEY, 'The probability theory of additive arithmetic functions', *Ann. Probability* 2 (1974) 749–791.
6. N. G. DE BRUIJN, 'On the number of positive integers $\leqslant x$ and free of prime factors $> y$'. *Nederl. Akad. Wetensch. Proc. Ser.* A 54 (1951) 50–60; *Indag. Math.* 13.
7. N. G. DE BRUIJN, 'On a function occurring in the theory of primes', *J. Indian Math. Soc.* 15 (1951) 25–32.
8. P. DIACONIS, 'Average running time of the fast Fourier transform', *J. Algorithms* 1 (1980) 187–208.
9. K. DICKMAN, 'On the frequency of numbers containing prime factors of a certain relative magnitude', *Ark. Mat., Astronomi och Fysik* 22 (1930) 1–14.
10. P. DONNELLY and P. JOYCE, 'Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex', *Stochastic Processes Appl.* 31 (1989) 89–103.
11. P. DONNELLY and P. JOYCE, 'Consistent ordered sampling distributions: characterization and convergence', *Advances in Appl. Probability* 23 (1991) 229–258.
12. W. J. EWENS, 'The sampling theory of selectively neutral alleles', *Theoret. Population Biology* 3 (1972) 87–112.
13. R. C. GRIFFITHS, 'On the distribution of points in a Poisson–Dirichlet process', *J. Appl. Probability* 25 (1988) 336–345.
14. G. H. HARDY and E. M. WRIGHT, *An introduction to the theory of numbers* (Clarendon Press, Oxford, 1979).
15. P. JOYCE and S. TAVARÉ, 'A convergence theorem for symmetric functionals of random partitions', *J. Appl. Probability* 29 (1992) 280–290.
16. J. F. C. KINGMAN, 'The population structure associated with the Ewens sampling formula', *Theoret. Population Biology* 11 (1977) 274–283.
17. D. E. KNUTH and L. TRABB PARDO, 'Analysis of a simple factorization algorithm', *J. Theoret. Comput. Sci.* 3 (1976) 321–348.
18. K. K. NORTON, 'Numbers with small prime factors, and the least kth power non-residue', *Mem. Amer. Math. Soc.* 106 (1971) 9–27.
19. V. RAMASWAMI, 'The number of positive integers $\leqslant x$ and free of prime divisors $> x^c$, and a problem of S. S. Pillai', *Duke Math. J.* 16 (1949) 99–109.

20. L. A. SHEPP and S. P. LLOYD, 'Ordered cycle lengths in a random permutation', *Trans. Amer. Math. Soc.* 121 (1966) 340–357.

21. A. M. VERSHIK, 'The asymptotic distribution of factorizations of natural numbers into prime divisors', *Soviet Math. Dokl.* 34 (1987) 57–61.

22. G. A. WATTERSON, 'The sampling theory of selectively neutral alleles', *Advances in Appl. Probability* 6 (1974) 463–488.

23. M. L. WUNDERLICH and J. L. SELFRIDGE, 'A design for a number theory package with an optimized trial division routine', *Comm. ACM* 17 (1974) 272–276.

School of Mathematical Sciences
Queen Mary and Westfield College
Mile End Road
London E1 4NS

Statistical Laboratory
16 Mill Lane
Cambridge CB2 1SB