The Annals of Applied Probability
1998, Vol. 8, No. 4, 975–994

LARGE DEVIATION PRINCIPLES FOR SOME RANDOM COMBINATORIAL STRUCTURES IN POPULATION GENETICS AND BROWNIAN MOTION[1]

By Shui Feng and Fred M. Hoppe

McMaster University

Large deviation principles are established for some random combinatorial structures including the Ewens sampling formula and the Pitman sampling formula. A path-level large deviation principle is established for the former on the cadlag space $\mathbf{D}([0,1], R)$ equipped with the uniform convergence topology, and the rate function is the same as for a Poisson process justifying the Poisson process approximation for the Ewens sampling formula at the large deviation level. A large deviation principle for the total number of parts in a partition is obtained for the Pitman formula; here the rate function depends only on one of the two parameters which display the different roles of the two parameters at different scales. In addition to these large deviation results, we also provide an embedding scheme which gives the Pitman sampling formula. A product of this embedding is an intuitive alternate proof of a result of Pitman on the limiting total number of parts.

1. Introduction. The purposes of this paper are to establish large deviation principles for some random combinatorial structures and to provide a biological context in which they occur.

A partition π of a positive integer n is an unordered representation of n as a sum of positive integers $n = n_1 + n_2 + \cdots + n_k$. Depending on the context, there are three common equivalent ways to describe π:

1. as an unordered set of "occupancy numbers" $\{n_1, \ldots, n_k\}$;
2. as a decreasing sequence $n_{(1)} \geq n_{(2)} \geq \cdots \geq n_{(k)}$;
3. as a multiplicity vector or "allelic partition" (cf. [12]) $m = (m_1, \ldots, m_n)$, where $m_j = m_j(\pi) = \sharp\{i\colon n_{(i)} = j\}$ is the number of times integer j appears in $\{n_1, \ldots, n_k\}$. Evidently, $\sum_{i=1}^n m_i = k$ and $\sum_{i=1}^n i m_i = n$.

A random partition of n is a random variable Π_n with values in the finite set of all partitions of n. One of the random partitions Π_n^θ considered in this paper is the Ewens sampling formula introduced in [6]:

$$P_n^\theta(m_1, m_2, \ldots, m_n) = P\big[\Pi_n^\theta = (m_1, \ldots, m_n)\big] = \frac{n!}{[\theta]^n}\Pi_{i=1}^n \frac{\theta^{m_i}}{i^{m_i}(m_i!)}, \tag{1.1}$$

where $\theta > 0$ is a parameter and $[\theta]^n = \theta(\theta+1)\cdots(\theta+n-1)$ is the ascending factorial. It describes the partition when a sample of size n is taken from a

Received November 1996; revised August 1997.

[1] Supported by the Natural Sciences and Engineering Research Council of Canada.

AMS 1991 *subject classifications*. Primary 60F10; secondary 05A17, 92D10.

Key words and phrases. Ewens sampling formula, large deviation principle, population genetics, Pitman sampling formula, random partitions.

selectively neutral haploid population which has evolved toward equilibrium in a sense made precise by a number of models in which (1.1) emerges. This remarkable distribution also arises elsewhere, for instance, in the context of Bayesian statistics [1], random permutations [18] and in a Pólya-like urn model [10].

The number of parts K_n^θ of Π_n^θ is another random variable. In genetics K_n^θ is the number of alleles in a sample of size n, while in the context of random permutations K_n^θ represents the number of cycles in the cycle representation of a permutation chosen uniformly from all $n!$ permutations (in which case $\theta = 1$).

There have been many studies on the behavior of K_n^θ as $n \to \infty$. Goncharov [8] obtained the central limit theorem

$$\frac{K_n^1 - \log n}{\sqrt{\log n}} \Rightarrow_{\mathscr{D}} N(0,1). \tag{1.2}$$

A far-reaching generalization of (1.2) was obtained by Hansen [9]. If we write $K_n^\theta(t) = \sum_{i=1}^{[n^t]} m_i(t)$ [so $K_n^\theta(t)$ is, say, the number of alleles having less than or equal to n^t copies] and let

$$Y_n^\theta(t) = \frac{K_n^\theta(t) - \theta t \log n}{\sqrt{\theta \log n}}, \qquad 0 \le t \le 1,$$

then Hansen proved that $Y_n^\theta(\cdot)$ converges weakly to Wiener measure on the Skorohod space $D[0,1]$, which is a functional central limit theorem for the process $K_n^\theta(\cdot)$. Previously, DeLaurentis and Pittel [3] obtained this result for $\theta = 1$ and, subsequently, Donnelly, Kurtz and Tavaré [5] provided a nice alternate proof based on a Poisson embedding using a model of Karlin and McGregor [11].

With these preliminaries in hand we can state one of the principal results in this paper, a path-level large deviation principle for the Ewens sampling formula.

THEOREM 1.1. *Let* $\mathbf{D}[0,1]$ *be the space* $D[0,1]$ *equipped with the uniform convergence topology and let* ν_n^θ *be the law of* $K_n^\theta(t)/\log n$ *under* P_n^θ *in* (1.1). *Define*

$$\begin{aligned} S_\theta(f) &= \int_0^1 I(\dot f(t))\,dt \\ &= \begin{cases} \displaystyle\int_0^1 \dot f(t) \log\left(\frac{\dot f(t)}{\theta}\right) dt + \theta - f(1), & \textit{if } f(0) = 0 \textit{ and } f \textit{ is} \\ & \qquad \textit{absolutely continuous}, \\ \infty, & \textit{otherwise}. \end{cases} \end{aligned}$$

Then the sequence $\{\nu_n^\theta\}_{n\ge1}$ *satisfies a large deviation principle on space* $\mathbf{D}[0,1]$ *with rate function* $S_\theta(\cdot)$ *and speed* $\log n$; *that is, for any Borel set* $A \subset \mathbf{D}[0,1]$,

$$-\inf_{f\in A^\circ} S_\theta(f) \le \liminf_{n\to\infty} \frac{1}{\log n} \nu_n^\theta(A) \le \limsup_{n\to\infty} \frac{1}{\log n} \nu_n^\theta(A) \le -\inf_{f\in \bar A} S_\theta(f),$$

where $\bar A$ *is the closure of* A, A° *the interior of* A.

REMARK. The nontrivial aspect of this result is to establish the large deviation principle in the uniform convergence topology. Actually our proof shows that the process $K_n^{\theta}(t)/\log n$ is exponentially equivalent to the Poisson process, which justifies the Poisson approximation in [2] at the large deviation level.

Pitman [15, 17] described a two-parameter version of (1.1) with parameters $0 \le \alpha < 1$, $\theta > -\alpha$, defined by

$$\begin{aligned} P_n^{\alpha,\theta}(m_1, m_2, \dots, m_n) &= P\big[\Pi_n^{\alpha,\theta} = (m_1, m_2, \dots, m_n)\big] \\ &= \frac{n!}{[\theta]^n} \Pi_{l=0}^{k-1}(\theta + l\alpha) \Pi_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{m_i}}{(i!)^{m_i}(m_i!)}, \end{aligned} \tag{1.3}$$

where $k = \sum_{i=1}^n m_i$.

The random partition $\Pi_n^{\alpha,\theta}$ arose in the study of stable processes with index α, and, in particular, the case $\alpha = \frac{1}{2}$ is related to the zeros of Brownian motion and Brownian bridge. When $\alpha = 0$, $\Pi_n^{0,\theta} = \Pi_n^{\theta}$.

Let $K_n^{\alpha,\theta}$ denote the number of parts of $\Pi_n^{\alpha,\theta}$. In the Ewens case ($\alpha = 0$, $\theta > 0$),

$$\lim_{n\to\infty} \frac{K_n^{\theta}}{\log n} = \theta \quad \text{a.s.},$$

but, in contrast, for $\alpha > 0$ and general θ (see [16]),

$$\lim_{n\to\infty} \frac{K_n^{\alpha,\theta}}{n^{\alpha}} = S_{\alpha,\theta} \quad \text{a.s.}, \tag{1.4}$$

where $S_{\alpha,\theta}$ is related to the Mittag–Leffler distribution.

We can now state the second principal result in this paper, the following marginal large deviation principle for Pitman's sampling formula.

THEOREM 1.2. *For $\alpha \in (0,1)$, $\theta + \alpha > 0$, the sequence $\{\nu_n^{\alpha,\theta}\}_{n\ge 1}$ satisfies a large deviation principle with rate function $I^{\alpha}(\cdot)$ and speed n, where $\nu_n^{\alpha,\theta}$ is the law of $K_n^{\alpha,\theta}/n$ under $P_n^{\alpha,\theta}$*:

$$I^{\alpha}(x) = \sup_{\lambda}\{\lambda x - \Lambda_{\alpha}(\lambda)\} \tag{1.5}$$

and

$$\Lambda_{\alpha}(\lambda) = \begin{cases} -\log[1 - (1 - e^{-\lambda})^{1/\alpha}], & \textit{if } \lambda > 0, \\ 0, & \textit{otherwise}. \end{cases} \tag{1.6}$$

REMARKS. Since Λ_{α} is increasing in α for fixed λ, we get $\Lambda_{\alpha}(\lambda) \le \lambda$ which implies that $I^{\alpha}(x) \ge \sup_{\lambda>0}\{\lambda(x-1)\}$. Hence, for $x > 1$, $I^{\alpha}(x) = \infty$. By letting $\lambda \to -\infty$, we can also get that $I^{\alpha}(x) = \infty$ for $x < 0$. Note that, for $\alpha > 0$, I^{α} does not depend on θ. An intuitive explanation for this will follow from the embedding scheme in Section 2. When $\alpha = \frac{1}{2}$, the rate function has the following explicit expression:

$$I^{1/2}(x) = \begin{cases} (2-x)\log\dfrac{2}{2-x} + (1-x)\log(1-x), & \text{if } x \in [0,1], \\ \infty, & \text{otherwise.} \end{cases}$$

The case $\alpha = 0$ needs separate handling and is given in the remark at the end of Section 3.

In addition to the preceding two large deviation results, we provide a biological basis for Pitman's sampling formula as a particular example of a general class of models introduced in [11], in the same fashion as [19] provides an embedding interpretation of the urn model [10] leading to the Ewens sampling formula. We have obtained other results in this direction, building on our model, which will appear elsewhere [7]. This embedding is described in Section 2. It leads to a simple intuitive proof of the existence of $\lim_{n\to\infty} K_n^{\alpha,\theta}/n^{\alpha}$. Theorem 1.2, the large deviation principle for $K_n^{\alpha,\theta}/n^{\alpha}$, is proved in Section 3. Finally, in Section 4, we prove Theorem 1.1, the path-level large deviation principle for the Ewens sampling formula. It is not entirely clear to us what is the correct formulation of the path-level large deviation principle for the Pitman case and this requires additional research.

2. An embedding scheme for the two-parameter formula.

We consider a population comprising a of various number of different types (mutants, alleles, in a biological context, say) which is evolving continuously in time. Following [11], there is an input process $I(t)$ describing how new mutants enter the population and a stochastic structure $x(t)$ with $x(0) = 1$ and the convention $x(t) = 0$ if $t < 0$ prescribing the growth pattern of each mutant population.

Mutants arrive at the times $0 \le T_1 < T_2 < \cdots$ and initiate lines according to independent versions of $x(t)$. Thus let $\{x_i(t)\}$ be independent copies of $x(t)$ with $x_i(t)$ being initiated by the ith mutant. Then $x_i(t - T_i)$ will be the size at time t of the ith mutant line. The process $N(t)$ represents the total population size at time t:

$$N(t) = \sum_{i=1}^{I(t)} x_i(t - T_i).$$

For $N(t) \ge 1$ the mutant lines induce a random partition $\Pi(t)$ of the integer $N(t)$. Let

$$m_i(t) = \sharp\{j\colon T_j \le t \text{ and } x_j(t - T_j) = i\}$$

so that

$$\Pi(t) = (m_1(t), \ldots, m_{N(t)}(t))$$

is the corresponding random allelic partition.

For the purposes of the present paper, the following specific model suffices but we have more extensive results for the general formulation which will be described elsewhere [7]:

1. $I(t)$ is a pure birth process with $I(0) = 0$ and infinitesimal birth rate
$$r_k = \lim_{h\to 0} \frac{1}{h} P\{I(t+h) - I(t) = 1 | I(t) = k\},$$
where
$$r_0 = \beta_0, \qquad r_k = \alpha(k-1) + \beta \quad \text{for } k \geq 1,$$
with $\beta_0 > 0$, $\beta > 0$ and $\alpha \geq 0$ arbitrary parameters. (Only β, α are relevant in the sequel, β_0 being merely a delay parameter until the population size reaches 1.)
2. The process $x(t)$ is also a pure birth process but starting at $x(0) = 1$ and with infinitesimal birth rate $\ell_n = n - \alpha$ for $n \geq 1$.
3. The cumulative process $N(t)$ is then again a pure birth starting at $N(0) = 0$ with infinitesimal rates determined by the "competing Poissons" $\{I(t), x_1(t - T_1), \dots, x_{I(t)}(t - T_{I(t)})\}$. Let
$$\rho_n = \lim_{h\to 0} \frac{1}{h} P\{N(t+h) - N(t) = 1 | N(t) = n\}.$$
Obviously, $\rho_0 = \beta_0$ and, for small $h > 0$,
$$\begin{aligned} &P\{N(t+h) - N(t) = 1 | N(t) = n\} \\ &\quad = E\biggl[\biggl(\alpha(I(t) - 1) + \beta + \sum_{i=1}^{I(t)} (x_i(t - T_i) - \alpha)\biggr) h + o(h) \bigg| N(t) = n\biggr] \\ &\quad = (n + \beta - \alpha)h + o(h). \end{aligned}$$
Thus, for $n \geq 1$, $\rho_n = n + \beta - \alpha$.

Define
$$\tau_n = \inf\{t \geq 0 \colon N(t) = n\} \quad \text{for } n \geq 1.$$
Then the random partition
$$\Pi_n = \Pi(\tau_n) = (m_1(\tau_n), \dots, m_n(\tau_n))$$
will be a random partition of n. Based on our construction, we have
$$\begin{aligned} &P\bigl\{\Pi_{n+1} = (m_1 + 1, \dots, m_n, 0) \big| \Pi_n = (m_1, \dots, m_n)\bigr\} \\ &\qquad = \frac{\alpha(\sum_{i=1}^n m_i) + \beta - \alpha}{n + \beta - \alpha} \end{aligned} \tag{2.1}$$

if $m_i \geq 1$, $1 \leq i < n$,

$$P\{\Pi_{n+1} = (m_1, \ldots, m_i - 1, m_{i+1} + 1, \ldots, m_n, 0) \big| \Pi_n = (m_1, \ldots, m_n)\} \\ = \frac{m_i(i-\alpha)}{n+\beta-\alpha}, \tag{2.2}$$

$$P\{\Pi_{n+1} = (m_1, \ldots, m_{n-1}, 0, 1) \big| \Pi_n = (m_1, \ldots, m_n)\} = \frac{n-\alpha}{n+\beta-\alpha} \tag{2.3}$$

if $m_n = 1$, where the change of state in (2.1) corresponds to the introduction of a new mutant, while in (2.2) and (2.3) one of the existing mutant populations is augmented by 1.

If we reparametrize by introducing $\theta = \beta - \alpha$, then these conditional probabilities become (15) in [15]. So we have established the following theorem.

THEOREM 2.1. *The distribution of the random partition $\Pi_n(\tau_n)$ is given by the Pitman formula*

$$P[\Pi(\tau_n) = (m_1, \ldots, m_n)] = \frac{n!}{[\theta]^n} \Pi_{l=0}^{k-1}(\theta + l\alpha) \Pi_{j=1}^{n} \frac{([1-\alpha]^{j-1})^{m_j}}{(j!)^{m_j}(m_j!)},$$

where $k = \sum_{i=1}^{n} m_i$.

Each of the processes $I(t)$, $x(t)$ and $N(t)$ has structure analogous to a linear birth process with immigration, generically denoted by $Y(t)$. This is a time-homogeneous Markov chain with infinitesimal parameter

$$\lambda_n = \lim_{h \to 0} \frac{1}{h} P[Y(t+h) - Y(t) = 1 | Y(t) = n] = \lambda n + c, \tag{2.4}$$

where $\lambda > 0$, $c > 0$. Typically, $Y(0) = 0$. We record the known result

$$P[Y(t) = n | Y(0) = 0] = \binom{\frac{c}{\lambda} + n - 1}{n} e^{-ct}(1 - e^{-\lambda t})^n, \qquad n = 0, 1, \ldots. \tag{2.5}$$

By applying a change of scale in Theorem 5 of [19], we deduce that if $Y(0) = 0$, then

$$\lim_{t \to \infty} e^{-\lambda t} Y(t) = W_{c/\lambda} \quad \text{a.s.}, \tag{2.6}$$

where W_d has the gamma density

$$f(x) = \frac{e^{-x} x^{d-1}}{\Gamma(d)}, \qquad x > 0.$$

Sometimes it is convenient to start with $Y(0) = 1$. With the same infinitesimal parameters λ_n, we now have in place of (2.5) the marginal distribution

$$\begin{aligned} &P[Y(t) = n | Y(0) = 1] \\ &\quad = \binom{\frac{c}{\lambda} + n - 1}{n-1} e^{-(\lambda+c)t}(1 - e^{-\lambda t})^{n-1}, \qquad n = 1, 2, \dots, \end{aligned} \tag{2.7}$$

and in place of (2.6)

$$\lim_{t\to\infty} e^{-\lambda t} Y(t) = W_{1+c/\lambda} \quad \text{a.s.} \tag{2.8}$$

We make explicit here the following interesting observation that the parameter range now allows negative values for c, namely, $c > -\lambda$, and we can no longer have the nice interpretation of $Y(t)$ as representing binary splitting at rate λ with immigration at rate c. Instead it is $Y(t) - 1$ which undergoes binary splitting with rate λ and immigration with rate $\lambda + c$.

Applying (2.5)–(2.8) to $I(t)$, $x(t)$ and $N(t)$, we get that:

(a) For $I(t)$, one has

$$I(0) = 0, \qquad \lambda_0 = \beta_0, \qquad \lambda_n = \alpha n + \beta - \alpha, \qquad n \ge 1.$$

In view of the delay, let $I^*(t) =_{\mathscr{D}} I(t + T_1)$. Then $I^*(t)$ is a linear birth process with immigration starting with $I^*(0) = 1$ and with infinitesimal rates $\lambda_n^* = \alpha n + \beta - \alpha$. Thus

$$\lim_{t\to\infty} e^{-\alpha t} I(t) = \lim_{t\to\infty} e^{-\alpha T_1} e^{-\alpha(t-T_1)} I^*(t - T_1) = e^{-\alpha T_1} W_\beta \quad \text{a.s.}, \tag{2.9}$$

where T_1 and W_β are independent and T_1 has an exponential distribution with mean β_0^{-1}.

(b) For $x(t)$, one has

$$\begin{aligned} x(0) = 1, \qquad \lambda_n = n - \alpha, \qquad n \ge 1, \\ \lim_{t\to\infty} e^{-t} x(t) = W_{1-\alpha} \quad \text{a.s.} \end{aligned} \tag{2.10}$$

(c) For $N(t)$, one has

$$\begin{aligned} N(0) = 0, \qquad \lambda_0 = \beta_0, \qquad \lambda_n = n + \beta - \alpha, \qquad n \ge 1, \\ \lim_{t\to\infty} e^{-t} N(t) = e^{-T_1} W_{1+\beta-\alpha} \quad \text{a.s.} \end{aligned} \tag{2.11}$$

With these results in hand we are ready to prove the following theorem.

THEOREM 2.2. *For* $0 < \alpha < 1$, $\theta > -\alpha$,

$$\lim_{n\to\infty} \frac{K_n^{\alpha,\theta}}{n^\alpha} = S_{\alpha,\theta} \quad \textit{a.s.}, \tag{2.12}$$

where $P[0 < S_{\alpha,\theta} < \infty] = 1$.

REMARKS. This theorem is due to Pitman [16] who obtained his results by moment calculations and martingale convergence. The following simple proof helps explain the factor n^α.

PROOF. By applying (2.9) and (2.11) we have

$$\lim_{t\to\infty}\frac{I(t)}{(N(t))^\alpha}=\lim_{t\to\infty}\frac{e^{-\alpha t}I(t)}{[e^{-t}N(t)]^\alpha}=\frac{W_{\theta+\alpha}}{(W_{1+\theta})^\alpha}\quad \text{a.s.,} \tag{2.13}$$

which gives (2.12). □

3. Large deviation principle for the total number of parts.

PROOF OF THEOREM 1.2. The main part in the proof of Theorem 1.2 is the following lemma. Let $\Lambda_\alpha(\lambda)$ be defined as in (1.6). Then Theorem 1.2 follows from Lemma 3.1, the facts that $\{\lambda;\ \Lambda_\alpha(\lambda)<\infty\}=R$ and $\Lambda_\alpha(\lambda)$ is differentiable and Theorem 2.3.6 (Gärtner–Ellis theorem) in [4]. □

Next we will prove Lemma 3.1.

LEMMA 3.1.

$$\lim_{n\to\infty}\frac{1}{n}\log E\big[\exp(\lambda K_n^{\alpha,\theta})\big]=\Lambda_\alpha(\lambda). \tag{3.1}$$

Before proving the lemma we first state some facts used in the proof:

$$\binom{a+i}{i}<\binom{b+i}{i}\quad \text{for } 0<a<b, \tag{3.2}$$

$$\left(\frac{1}{1-x}\right)^n=\sum_{i=0}^\infty x^i\binom{i+n-1}{n-1}, \tag{3.3}$$

$$\binom{i+n}{n-1}<n\binom{i+n}{n}, \tag{3.4}$$

for $x\in(0,1)$ and $\theta=0$, we have

$$E\left[\left(\frac{1}{1-x}\right)^{K_n^{\alpha,0}}\right]=\sum_{i=0}^\infty x^i\binom{\alpha i+n-1}{n-1}. \tag{3.5}$$

In the proof of (3.5) we make use of (3.3) and the following equality obtained in [17]:

$$E[[K_n^{\alpha,0}]^i]=\frac{\Gamma(i)[\alpha i]^n}{\alpha\Gamma(n)}. \tag{3.6}$$

PROOF OF LEMMA 3.1. First assume $\theta = 0$. Then, for $\lambda \le 0$, we have by Jensen's inequality

$$\log E\bigl[\exp(\lambda K_n^{\alpha,0})\bigr] \ge \lambda E[K_n^{\alpha,0}].$$

By (3.6), $E[K_n^{\alpha,0}]$ grows like n^α and therefore

$$\lim_{n\to\infty} \frac{1}{n} \log E\bigl[\exp(\lambda K_n^{\alpha,0})\bigr] = 0 = \Lambda_\alpha(\lambda).$$

Next, when $\lambda > 0$, let $x = 1-e^{-\lambda} < 1$. We first consider the case of $0 < \alpha < 1$ with α rational. We will show that

$$E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{\alpha,0}}\biggr] = C_{\alpha,x,n}\biggl(\frac{1}{1-x^{1/\alpha}}\biggr)^n \tag{3.7}$$

for some constant $C_{\alpha,x,n}$ which grows, for fixed α, x, algebraically, at most, in n. Thus

$$\frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{\alpha,0}}\biggr] = \frac{1}{n}\log C_{\alpha,x,n} + \log\biggl(\frac{1}{1-x^{1/\alpha}}\biggr),$$

and (3.1), for rational α, follows in the limit as $n \to \infty$.

Once we have established (3.7) for rational α, then for irrational α we can approximate by rationals above and below, say $r/m < \alpha < s/l$. By the monotonicity of $E[(1/(1-x))^{K_n^{\alpha,0}}]$, by (3.2) and (3.5),

$$\begin{aligned}\frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{r/m,0}}\biggr] &< \frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{\alpha,0}}\biggr]\\ &< \frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{s/l,0}}\biggr]\end{aligned} \tag{3.8}$$

and letting $n \to \infty$,

$$\begin{aligned}\log\biggl(\frac{1}{1-x^{m/r}}\biggr) &\le \liminf_{n\to\infty}\frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{\alpha,0}}\biggr]\\ &\le \limsup_{n\to\infty}\frac{1}{n}\log E\biggl[\biggl(\frac{1}{1-x}\biggr)^{K_n^{\alpha,0}}\biggr]\\ &\le \log\biggl(\frac{1}{1-x^{l/s}}\biggr).\end{aligned} \tag{3.9}$$

We now invoke the continuity of the function $\log(1/(1-x^{1/t}))$ in t, thereby deducing (3.1) for all $\alpha \in (0,1)$ from its validity for rational α.

It remains to prove (3.7) for rational α. Let $\alpha = r/m$. We break the series in (3.5) into rm components, each of which has the same growth rate. These components are chosen so that the combinatorial factor $\binom{\alpha i+n-1}{n-1}$ may be replaced with $\binom{i+n-1}{n-1}$ and the resulting series is summable in closed form by (3.3).

Introduce the notation

$$\begin{aligned} H_n\left(x;\frac{r}{m}\right) &= \sum_{k=0}^{\infty} x^k \binom{kr/m+n-1}{n-1} \\ &= \sum_{j=0}^{m-1}\sum_{i=0}^{\infty} x^{mi+j}\binom{((mi+j)/m)r+n-1}{n-1} \qquad (3.10) \\ &= \sum_{j=0}^{m-1} x^j \sum_{i=0}^{\infty} x^{mi}\binom{(j/m)r+ri+n-1}{n-1}. \end{aligned}$$

We first evaluate the $j=0$ term in (3.10), denoting it by

$$A_n(x;r,m) = \sum_{i=0}^{\infty} x^{mi}\binom{ri+n-1}{n-1}. \qquad (3.11)$$

The right-hand side of (3.11) comprises every rth term in the larger series

$$\left(\frac{1}{1-x^{m/r}}\right)^n = \sum_{k=0}^{\infty}(x^{m/r})^k\binom{k+n-1}{n-1} \qquad (3.12)$$

and we will argue that (3.11) and (3.12) are of the same order in n.

Write

$$\begin{aligned} \left(\frac{1}{1-x^{m/r}}\right)^n &= \sum_{l=0}^{r-1}\sum_{i=0}^{\infty}(x^{m/r})^{ri+l}\binom{ri+l+n-1}{n-1} \qquad (3.13) \\ &= \sum_{l=0}^{r-1} x^{ml/r}\sum_{i=0}^{\infty}(x^{m/r})^{ri}\binom{ri+l+n-1}{n-1}. \end{aligned}$$

From (3.2), for $0 \le l \le r-1$,

$$\binom{ri+n-1}{n-1} < \binom{ri+l+n-1}{n-1} < \binom{ri+r+n-1}{n-1}.$$

We substitute these inequalities into (3.13), obtaining

$$\begin{aligned} \sum_{l=0}^{r-1} x^{ml/r}\sum_{i=0}^{\infty} x^{mi}\binom{ri+n-1}{n-1} &< \sum_{l=0}^{r-1} x^{ml/r}\sum_{i=0}^{\infty}(x^{m/r})^{ri}\binom{ri+l+n-1}{n-1} \\ &< \sum_{l=0}^{r-1} x^{ml/r}\sum_{i=0}^{\infty}(x^{m/r})^{ri}\binom{ri+r+n-1}{n-1}. \end{aligned}$$

Hence

$$\begin{aligned}
A_n(x;r,m)\sum_{l=0}^{r-1} x^{ml/r} &< \left(\frac{1}{1-x^{m/r}}\right)^n \\
&< \sum_{l=0}^{r-1} x^{ml/r} \sum_{i=0}^{\infty} (x^{m/r})^{ri} \binom{r(i+1)+n-1}{n-1} \\
&= \sum_{l=0}^{r-1} x^{ml/r} x^{-m} \sum_{i=0}^{\infty} (x^{m/r})^{r(i+1)} \binom{r(i+1)+n-1}{n-1} \\
&= \sum_{l=0}^{r-1} (x^{m/r})^{l-r} \sum_{j=1}^{\infty} (x^{m/r})^{rj} \binom{rj+n-1}{n-1} \\
&\le \sum_{l=0}^{r-1} (x^{m/r})^{l-r} \sum_{j=0}^{\infty} (x^{m/r})^{rj} \binom{rj+n-1}{n-1} \\
&= A_n(x;r,m) \sum_{l=0}^{r-1} (x^{m/r})^{l-r}.
\end{aligned} \tag{3.14}$$

This takes care of the $j = 0$ term in (3.10). For $1 \le j < m-1$, we begin with

$$\binom{ri+n-1}{n-1} < \binom{ri+\frac{rj}{m}+n-1}{n-1} < \binom{ri+r+n-1}{n-1},$$

and repeatedly apply (3.4) to deduce

$$\binom{ri+r+n-1}{n-1} < [n]_r \binom{ri+r+n-1}{n+r-1},$$

where $[n]_r = n(n-1)\cdots(n-r+1)$ is the descending factorial. Thus

$$\begin{aligned}
\sum_{i=0}^{\infty} x^{mi} \binom{ri+n-1}{n-1} &< \sum_{i=0}^{\infty} x^{mi} \binom{ri+rj/m+n-1}{n-1} \\
&< [n]_r \sum_{i=0}^{\infty} x^{mi} \binom{ri+r+n-1}{n+r-1},
\end{aligned}$$

which can be expressed, using our notation, as

$$A_n(x;r,m) < \sum_{i=0}^{\infty} x^{mi} \binom{ri+rj/m+n-1}{n-1} < [n]_r A_{n+r-1}(x;r,m).$$

Taking the sum on j, we have

$$A_n(x;r,m)\sum_{j=0}^{m-1} x^j < H_n\left(x;\frac{r}{m}\right) < A_{n+r-1}(x;r,m)[n]_r \sum_{j=0}^{m-1} x^j. \tag{3.15}$$

Combining (3.14) and (3.15), we arrive at the bounds

$$\begin{aligned}\frac{\sum_{j=0}^{m-1} x^j}{\sum_{l=0}^{r-1}(x^{m/r})^{l-r}}\left(\frac{1}{1-x^{m/r}}\right)^n &< H_n\left(x; \frac{r}{m}\right)\\ &< \frac{[n]_r \sum_{j=0}^{m-1} x^j}{\sum_{l=0}^{r-1}(x^{m/r})^l}\left(\frac{1}{1-x^{m/r}}\right)^n.\end{aligned} \tag{3.16}$$

This proves our assertion for α rational, stated at the outset of the proof, that

$$E\left[\left(\frac{1}{1-x}\right)^{K_n^{r/m,0}}\right] \equiv H_n\left(x; \frac{r}{m}\right) = C_{r/m,x,n}\left(\frac{1}{1-x^{m/r}}\right)^n.$$

Obviously,

$$\frac{1}{n}\log\left[E\left(\frac{1}{1-x}\right)^{K_n^{r/m,0}}\right] \to -\log(1-x^{m/r})$$

as $n \to \infty$ and our previous continuity argument establishes (3.1).

Finally, we consider the case of $\theta \neq 0$. Let

$$m = \inf\{n \geq 1\colon \theta \leq n\alpha\}.$$

Then by direct calculation we have

$$c_1(\alpha,\theta)\frac{1}{k-1}\frac{\Pi_{l=1}^{k-1} l\alpha}{(n-1)!} \leq \frac{\Pi_{l=1}^{k-1} l\alpha}{(\theta)_n} \leq c_2(\alpha,\theta)\Pi_{l=k}^{k+m-1}[l]\frac{\Pi_{l=1}^{k-1} l\alpha}{(n-1)!}, \tag{3.17}$$

where

$$c_1(\alpha,\theta) = \frac{(\theta+\alpha)(n-1)!}{\alpha(\theta)_n},$$

$$c_2(\alpha,\theta) = \frac{(n-1)!}{(\theta)_n}\frac{1}{\Pi_{l=1}^m l}.$$

For any $\varepsilon > 0$ there exists a $k_0 > 0$ such that for all $k > k_0$ we have

$$(k-1)^{-1} \geq e^{-\varepsilon k}, \qquad \Pi_{l=k}^{k+m-1}[l] \leq e^{\varepsilon k}.$$

This, combined with the special form of $c_1(\alpha,\theta)$ and $c_2(\alpha,\theta)$, implies that the case of $\theta \neq 0$ is the same as the case of $\theta = 0$. Thus the lemma follows. □

REMARK 1. Here are simple alternate proofs for the cases of $\theta = 0$, $\alpha = \frac{1}{2}$ and $\theta = \alpha = \frac{1}{2}$. Pitman [16] shows that $K_n^{1/2,0}$ can also be derived from the zeros of Brownian motion in the case of $\theta = 0$, $\alpha = \frac{1}{2}$, and from the zeros of Brownian bridge in the case of $\theta = \alpha = \frac{1}{2}$. For Brownian motion

$$P\{K_n^{1/2,0} = k\} = \binom{2n-k-1}{n-1}2^{k+1-2n}, \tag{3.18}$$

and for Brownian bridge

$$P\{K_n^{1/2,\,1/2} = k\} = \frac{k(n-1)!}{(3/2)_{n-1}}\binom{2n-k-1}{n-1}2^{k+1-2n}. \tag{3.19}$$

Thus, for $\lambda > 0$, we have

$$\begin{aligned}
&E\big[\exp(\lambda K_n^{1/2,\,0})\big]\\
&\quad= \sum_{k=1}^{n} e^{\lambda k}\binom{2n-k-1}{n-1}2^{k+1-2n}\\
&\quad= \sum_{k=1}^{n} e^{\lambda n}\binom{2n-k-1}{n-1}\left(\frac{1}{2e^{\lambda}}\right)^{n-k}\left(\frac{1}{2}\right)^{n-1}\\
&\quad= \frac{e^{\lambda n}}{(1-1/2e^{\lambda})^{n-1}}2^{-(n-1)}\sum_{k=1}^{n}\binom{2n-k-1}{n-1}\left(\frac{1}{2e^{\lambda}}\right)^{n-k}\left(1-\frac{1}{2e^{\lambda}}\right)^{n-1}\\
&\quad= \frac{e^{\lambda n}}{(1-1/2e^{\lambda})^{n-1}}2^{-(n-1)}\sum_{k=1}^{n}\binom{2n-k-1}{n-1}\left(\frac{1}{2e^{\lambda}}\right)^{n-k}\left(1-\frac{1}{2e^{\lambda}}\right)^{n-1}\\
&\quad= \frac{e^{\lambda n}}{(1-1/2e^{\lambda})^{n-1}}2^{-(n-1)}\sum_{j=0}^{n-1}\binom{j+n-1}{n-1}\left(\frac{1}{2e^{\lambda}}\right)^{j}\left(1-\frac{1}{2e^{\lambda}}\right)^{n-1},
\end{aligned} \tag{3.20}$$

where in the last equality we made the substitution of $j = n-k$.

Consider a sequence of independent trials, each of which results in a success with probability $p = 1-1/(2e^{\lambda})$, or a failure with probability $q = 1/(2e^{\lambda})$. The fact that $\lambda > 0$ implies that $p > \frac{1}{2}$. Let A be the event that the nth success occurs before the $2n$th trial; B be the event that a failure follows the nth success. Then we have $\lim_{n\to\infty} P\{A\} = 1$, $P\{B|A\} = q$ and

$$P\{A\cap B\} = \sum_{j=0}^{n-1}\binom{j+n-1}{n-1}q^{j}p^{n-1}pq \to q \quad \text{as } n\to\infty. \tag{3.21}$$

This combined with (3.20) implies that

$$\frac{1}{n}\log E\big[\exp(\lambda K_n^{1/2,\,0})\big] \to \log\frac{1/2e^{\lambda}}{1-1/2e^{\lambda}} = \Lambda_\alpha(\lambda),$$

which is exactly (3.1) and the case of Brownian bridge follows because of the same α.

REMARK 2. For the Ewens sampling formula ($\alpha = 0$, $\theta > 0$), the sequence $\{\mu_n^{\theta}\}_{n\geq 1}$ satisfies a large deviation principle with speed $\log n$ and rate function

$$I_\theta(x) = \begin{cases} x\log\dfrac{x}{\theta} - x + \theta, & \text{for } x > 0,\\ \theta, & \text{for } x = 0,\\ +\infty, & \text{for } x < 0,\end{cases}$$

where μ_n^θ is the law of $K_n^\theta/\log n$ under P_n^θ in (1.1). The proof is straightforward by using the representation of K_n^θ as a sum of independent components (see [18] and [16]) and the generating function formula for the Stirling numbers of the first kind.

4. Path-level large deviation principle.

4.1. *Poisson embedding.* The Poisson embedding used in [5] is a special case of the embedding scheme developed in Section 2 with $I(t)$ being a homogeneous Poisson process (corresponding to $\alpha = 0$) on R_+ with parameter θ and $x(t)$ being a pure birth process with rate 1. In this particular case, one has the following:

THEOREM 4.1 ([5]). *Let*

$$\bar{K}_n(t) = \sum_{k=1}^{I(\tau_n)} I_{\{x_k(\tau_n - T_k)\le n^t\}}.$$

Then, for any $0 < \delta < \frac{1}{2}$ *and* $c > 1$,

$$\bar{K}_n(t) \le I(\tau_n) - I(\tau_n - t\log n - (\log n)^\delta) + R_1(n),$$

$$\bar{K}_n(t) \ge I(\tau_n) - I(\tau_n - t\log n + (\log n)^\delta) - R_2(n),$$

$$R_1(n) \le R_1^c(n), \qquad R_2(n) \le R_2^c(n),$$

where

$$R_1(n) = \sum_{k=1}^{I(\tau_n)} I_{\{x_k(\tau_n - T_k) < \exp[\tau_n - T_k - (\log n)^\delta]\}}, \tag{4.1}$$

$$R_2(n) = \sum_{k=1}^{I(\tau_n)} I_{\{x_k(\tau_n - T_k) > \exp[\tau_n - T_k + (\log n)^\delta]\}}, \tag{4.2}$$

$$R_1^c(n) = \sum_{k=1}^{I(c\log n)} I_{\{\inf_{t\in[0,1]}\exp[-t]x_k(t) < \exp[-(\log n)^\delta]\}} + R_3^c(n), \tag{4.3}$$

$$R_2^c(n) = \sum_{k=1}^{I(c\log n)} I_{\{\sup_{t\in[0,1]}\exp[-t]x_k(t) > \exp[(\log n)^\delta]\}} + R_3^c(n), \tag{4.4}$$

$$R_3^c(n) = I(\tau_n) I_{\{\tau_n > c\log n\}}. \tag{4.5}$$

Here $I(t) = 0$ for $t < 0$.

From the construction in Section 2, we see that $K_n(t)$ introduced in Section 1 equals $\bar{K}_n(t)$ in law.

4.2. *Large deviation for the Poisson process.* Let $\mathbf{D}[0,1]$ be the space of all real-valued functions on $[0,1]$ that are right continuous and have left limit on $[0,1)$ and are left continuous at $t=1$, equipped with the uniform convergence topology and metric

$$||f-g|| = \sup_{t\in[0,1]}\{|f(t)-g(t)|\} \quad \text{for any } f, g \in \mathbf{D}[0,1].$$

Let $I(t)$ be a Poisson process with parameter θ and let $\{a_n\}_{n\geq 1}$ be a sequence of positive numbers that goes to ∞ as $n\to\infty$. Let $Z_n(t) = I(a_n t)/a_n$ and let Q_n be the law of $Z_n(t)$ on space $\mathbf{D}[0,1]$. Then one has the following:

THEOREM 4.2 ([14]). *The sequence $\{Q_n\}_{n\geq 1}$ satisfies a large deviation principle on space $\mathbf{D}[0,1]$ with rate function $S(\cdot)$ and speed a_n.*

REMARK. This theorem in its present form is proved in [14]. A weaker version when $\mathbf{D}[0,1]$ is equipped with the weak topology is obtained in [13].

Next we will prove a generalized version of this theorem. Let $\{b_n\}_{n\geq 1}$ be a sequence of positive numbers such that $\lim_{n\to\infty} b_n/a_n = 0$. Let $\bar{Z}_n(t) = I(a_n t + b_n)/a_n$ and $\bar{Q}_n$ be the law of $\bar{Z}_n(t)$ on space $\mathbf{D}[0,1]$. We have the following:

THEOREM 4.3. *The sequence $\{\bar{Q}_n\}_{n\geq 1}$ satisfies a large deviation principle on space $\mathbf{D}[0,1]$ with rate function $S(\cdot)$ and speed a_n.*

PROOF. Without loss of generality, we may assume that $\inf_{n\geq 1} b_n > 0$. In general, we may add 1 to the sequence and use the same argument as below. The main idea of the proof is to show that Q_n and $\bar{Q}_n$ are exponentially equivalent and then to apply Theorem 4.2.13 in [4]. Let $\mathscr{P}_n$ be the joint distribution of $Z_n(t)$ and $\bar{Z}_n(t)$. Then for any $\delta > 0$ we have by using the Markov property and Doob's inequality

$$\begin{aligned}
&\mathscr{P}_n\{||\bar{Z}_n(\cdot) - Z_n(\cdot)|| > \delta\}\\
&\quad = P\Big\{\sup_{t\in[0,1]}[I(a_n t + b_n) - I(a_n t)] > a_n\delta\Big\}\\
&\quad \leq P\Big\{\sup_{s,t\in[0,a_n+b_n],\, t-s\in[0,b_n]}(I(t)-I(s)) > a_n\delta\Big\}\\
&\quad \leq \sum_{l=0}^{[a_n/b_n]} P\Big\{\sup_{t\in[lb_n,(l+2)b_n\wedge(a_n+b_n)]}(I(t)-I(lb_n)) > a_n\delta\Big\}\\
&\quad \leq \left(1+\left[\frac{a_n}{b_n}\right]\right)\sup_{k\geq 0} P^k\Big\{\sup_{t\in[0,2b_n]}(I(t)-I(0)) > a_n\delta\Big\}\\
&\quad = \left(1+\left[\frac{a_n}{b_n}\right]\right)P\Big\{\sup_{t\in[0,2b_n]} I(t) > a_n\delta\Big\}
\end{aligned}$$

$$
\begin{aligned}
&= \left(1 + \left[\frac{a_n}{b_n}\right]\right) P\Big\{ \sup_{t\in[0,\,2b_n]} \exp[\lambda(I(t) - \theta t)] > \exp[\lambda(a_n\delta - 2\theta b_n)]\Big\} \\
&\le \left(1 + \left[\frac{a_n}{b_n}\right]\right) \exp[\lambda(-a_n\delta + 2\theta b_n)] E[\exp[\lambda(I(2b_n) - 2\theta b_n)]] \\
&= \left(1 + \left[\frac{a_n}{b_n}\right]\right) \exp[-\lambda a_n \delta] \exp[2\theta b_n(e^{\lambda} - 1)],
\end{aligned}
$$

where $[a_n/b_n]$ is the integer part of a_n/b_n and P^k represents the law of the Poisson process that starts at k. This implies that

$$
\limsup_{n\to\infty} a_n^{-1} \log \mathscr{P}_n\{\|\bar{Z}_n(\cdot) - Z_n(\cdot)\| > \delta\} \le -\lambda\delta. \tag{4.6}
$$

Letting $\lambda \to \infty$, we get the exponential equivalence of Q_n and $\bar{Q}_n$. By applying Theorem 4.2.13 in [4], we get the result. □

4.3. *The main result.* In this section we will prove Theorem 1.1. To prove this theorem, we need the following two lemmas.

LEMMA 4.4. *For any fixed* $\rho > 0$ *we have*

$$
\lim_{c\to\infty} \limsup_{n\to\infty} \frac{1}{\log n} \log P\{R_1^c(n) > \rho \log n\} = -\infty, \tag{4.7}
$$

$$
\lim_{c\to\infty} \limsup_{n\to\infty} \frac{1}{\log n} \log P\{R_2^c(n) > \rho \log n\} = -\infty, \tag{4.8}
$$

$$
\lim_{c\to\infty} \limsup_{n\to\infty} \frac{1}{\log n} \log P\{R_3^c(n) > \rho \log n\} = -\infty. \tag{4.9}
$$

PROOF. First let us assume that (4.9) is true. By direct calculation we have, for any $c > 1$,

$$
\begin{aligned}
&P\Bigg\{ \sum_{k=1}^{I(c\log n)} I_{\{\inf_{t\in[0,\,1]} \exp[-t]x_k(t) < \exp[-(\log n)^{\delta}]\}} > \rho \log n \Bigg\} \\
&\quad \le \exp[-\lambda\rho\log n] E\Bigg\{\exp\Bigg[\lambda \sum_{k=1}^{I(c\log n)} I_{\{\inf_{t\in[0,\,1]} \exp[-t]x_k(t) < \exp[-(\log n)^{\delta}]\}}\Bigg]\Bigg\} \\
&\quad = \exp[-\lambda\rho\log n] \sum_{i=1}^{\infty} \frac{(\theta c \log n)^i}{i!} \exp[-\theta c \log n] \\
&\qquad\qquad \times \prod_{k=1}^{i} \exp\left[\lambda I_{\{\inf_{t\in[0,\,1]} \exp[-t]x_k(t) < \exp[-(\log n)^{\delta}]\}}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\le \exp[-\lambda\rho\log n]\sum_{i=1}^{\infty}\frac{(\theta c\log n)^i}{i!}\exp[-\theta c\log n]\\
&\qquad\qquad \times\prod_{k=1}^{i}\bigl[1+\exp(\lambda)P\bigl\{I_{\{\inf_{t\in[0,1]}\exp[-t]x_k(t)<\exp[-(\log n)^\delta]\}}\bigr\}\bigr]\\
&\le \exp[-\lambda\rho\log n]\sum_{i=1}^{\infty}\frac{(\theta c\log n)^i}{i!}\exp[-\theta c\log n]\\
&\qquad\qquad \times\left[1+2\exp(\lambda)\exp\left[-\frac{(\log n)^\delta}{2}\right]\right]^i\\
&=\exp[-\lambda\rho\log n]\exp\left[2\theta c\log n\ \exp(\lambda)\exp\left[-\frac{(\log n)^\delta}{2}\right]\right].
\end{aligned}
$$

In the last inequality we used the martingale inequality (3.9) in [5]. Letting $n\to\infty$, and then $\lambda\to\infty$, we get

$$
\limsup_{n\to\infty}\frac{1}{\log n}\log P\left\{\sum_{k=1}^{I(c\log n)} I_{\{\inf_{t\in[0,1]}\exp[-t]x_k(t)<\exp[-(\log n)^\delta]\}}>\rho\log n\right\} = -\infty. \tag{4.10}
$$

This, combined with (4.3) and (4.9), implies (4.7).

Next we prove (4.8). By Doob's inequality,

$$
\begin{aligned}
&P\left\{\sum_{k=1}^{I(c\log n)} I_{\{\sup_{t\in[0,1]}\exp[-t]x_k(t)>\exp[(\log n)^\delta]\}}>\rho\log n\right\}\\
&\quad\le \exp[-\lambda\rho\log n]\sum_{i=1}^{\infty}\frac{(\theta c\log n)^i}{i!}\exp[-\theta c\log n]\\
&\qquad\qquad \times\prod_{k=1}^{i}\exp\bigl[\lambda I_{\{\sup_{t\in[0,1]}\exp[-t]x_k(t)>\exp[(\log n)^\delta]\}}\bigr]\\
&\quad\le \exp[-\lambda\rho\log n]\sum_{i=1}^{\infty}\frac{(\theta c\log n)^i}{i!}\exp[-\theta c\log n]\\
&\qquad\qquad \times\prod_{k=1}^{i}\bigl[1+\exp(\lambda)P\bigl\{I_{\{\sup_{t\in[0,1]}\exp[-t]x_k(t)>\exp[(\log n)^\delta]\}}\bigr\}\bigr]\\
&\quad\le \exp[-\lambda\rho\log n]\\
&\qquad\times\sum_{i=1}^{\infty}\frac{(\theta c\log n)^i}{i!}\exp[-\theta c\log n][1+\exp(\lambda)\exp[-(\log n)^\delta]]^i\\
&\quad=\exp[-\lambda\rho\log n]\exp[\theta c\log n\exp(\lambda)\exp[-(\log n)^\delta]].
\end{aligned}
$$

Letting $n \to \infty$, and then $\lambda \to \infty$, we get

$$\limsup_{n\to\infty} \frac{1}{\log n} \log P\left\{ \sum_{k=1}^{I(c\log n)} I_{\{\sup_{t\in[0,1]} \exp[-t] x_k(t) > \exp[(\log n)^\delta]\}} > \rho \log n \right\} = -\infty. \tag{4.11}$$

This, combined with (4.4) and (4.9), implies (4.8).

Finally, we turn to the proof of (4.9). Noting that τ_n can be represented as the sum of n independent exponential random variables with parameters $\theta, \theta+1, \dots, \theta+n-1$, one has $\lim_{n\to\infty} \tau_n/\log n = 1$ almost surely as $n \to \infty$ and, for $\lambda < \theta$,

$$\lim_{n\to\infty} \frac{1}{\log n} \log E\left[e^{\lambda \tau_n}\right] = \lim_{n\to\infty} \frac{1}{\log n} \log \frac{\Gamma(\theta+n)}{\Gamma(\theta-\lambda+n)} = \lambda. \tag{4.12}$$

By using Theorem 2.3.6 (Gärtner–Ellis theorem) in [4], we find that $\tau_n/\log n$ satisfies a large deviation upper bound with rate function

$$\bar{I}(x) = \begin{cases} \theta(x-1), & \text{if } x \geq 1, \\ \infty, & \text{otherwise.} \end{cases}$$

Thus we get

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{\log n} \log P\{R_3^c(n) > \rho \log n\} &\leq \limsup_{n\to\infty} \frac{1}{\log n} \log P\{\tau_n \geq c \log n\} \\ &\leq -\inf_{x\geq c} \bar{I}(x) = -\theta(c-1), \end{aligned}$$

which implies (4.9). □

LEMMA 4.5. *Let*

$$f_n^1(t) = \frac{1}{\log n}[I(\tau_n) - I(\tau_n - t\log n - (\log n)^\delta)]$$

and

$$f_n^2(t) = \frac{1}{\log n}[I(\tau_n) - I(\tau_n - t\log n + (\log n)^\delta)].$$

Then f_n^1 and f_n^2 are exponentially equivalent.

PROOF. By using an argument similar to that used in the proof of (4.6), we have, for any $\rho > 0$,

$$\begin{aligned} P\{\|f_n^1 - f_n^2\| \geq \rho\} = P\Big\{ \sup_{t\in[0,1]} \big[& I(\tau_n - t\log n + (\log n)^\delta) \\ & - I(\tau_n - t\log n - (\log n)^\delta)\big] \geq \rho \log n \Big\} \end{aligned}$$

$$\begin{aligned}&\leq P\Big\{\sup_{s,\,t\in[0,\,c\log n+(\log n)^{\delta}],\,t-s\in[0,2(\log n)^{\delta}]}[I(t)-I(s)]\geq\rho\log n\Big\}\\&\quad+P\{\tau_n>c\log n\}\\&\leq\tfrac{1}{2}(1+c(\log n)^{1-\delta})\exp[-\lambda\rho\log n]\exp[4\theta(\log n)^{\delta}(e^{\lambda}-1)]\\&\quad+P\{\tau_n>c\log n\}.\end{aligned}$$

Since $0<\delta<\frac{1}{2}$ and $c>1$ and λ are arbitrary, by letting $n\to\infty$, and then $\lambda\to\infty$, finally $c\to\infty$, we get

$$\limsup_{n\to\infty}\frac{1}{\log n}P\{||f_n^1-f_n^2||\geq\rho\}=-\infty.$$

Thus we proved that f_n^1 and f_n^2 are exponentially equivalent. □

PROOF OF THEOREM 1.1. Since $K_n^{\theta}(t)$ and $\bar{K}_n(t)$ have the same law, it suffices to verify the result for the sequence $\{\bar{K}_n(t)/\log n\}_{n\geq 1}$. Lemma 4.4 implies that $R_1(n)/\log n$ and $R_2(n)/\log n$ are superexponentially small. This combined with Theorem 4.1 and Lemma 4.5 implies that $f_n^1(t)$ $\bar{K}_n(t)/\log n$ and $f_n^2(t)$ are exponentially equivalent. Note that $f_n^1(t)$ and $I((\log n)t+(\log n)^{\delta})/\log n$ have the same law. An application of Theorem 4.3 gives the result. □

REMARK. A reader of this paper has suggested an alternate approach to prove Theorem 1.1 by adapting the projective limit technique and an induction based on the approximate independence of $K_n^{\theta}(s)$ and $K_n^{\theta}(t)-K_n^{\theta}(s)$ for $s,t\in[0,1]$. This would lead to a large deviation principle in the pointwise convergence topology. To strengthen the result to the uniform convergence topology, one would then need to check the exponential tightness in this stronger topology. Our approach is more direct.

Acknowledgments. The authors would like to thank Jim Pitman for providing the factorial moment calculation (3.6), and a referee for helpful suggestions.

REFERENCES

[1] ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

[2] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2** 519–535.

[3] DELAURENTIS, J. M. and PITTEL, B. G. (1985). Random permutations and Brownian motion. *Pacific J. Math.* **119** 287–301.

[4] DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations and Applications*. Jones and Bartlett, Boston.

[5] 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.

[6] EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* **3** 87–112.

[7] FENG, S. and HOPPE, F. M. (1996). Models for partition structures. Unpublished manuscript.

[8] GONCHAROV, V. L. (1944). Some facts from combinatorics. *Izv. Akad. Nauk SSSR Ser. Mat.* **8** 3–48.

[9] HANSEN, J. C. (1990). A functional central limit theorem for the Ewens sampling formula. *J. Appl. Probab.* **27** 28–43.

[10] HOPPE, F. M. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.* **20** 91–94.

[11] KARLIN, S. and MCGREGOR, J. (1967). The number of mutant forms maintained in a population. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 415–438. Univ. California Press, Berkeley.

[12] KINGMAN, J. F. C. (1978). Random partitions in population genetics. *Proc. Roy. Soc. London Ser. A* **361** 1–20.

[13] LYNCH, J. and SETHURAMAN, J. (1987). Large deviations for processes with independent increments. *Ann. Probab.* **15** 610–627.

[14] MOGULSKII, A. A. (1993). Large deviations for processes with independent increments. *Ann. Probab.* **21** 202–215.

[15] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158.

[16] PITMAN, J. (1996). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli* **3** 79-96.

[17] PITMAN, J. (1997). Notes on the two parameter generalization of the Ewens random partition structure. Unpublished manuscript.

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

[19] TAVARÉ, S. (1987). The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.* **25** 161–168.

DEPARTMENT OF MATHEMATICS AND STATISTICS
MCMASTER UNIVERSITY
HAMILTON, ONTARIO
CANADA L8S 4K1
E-MAIL: shuifeng@mcmail.cis.mcmaster.ca
hoppe@mcmail.cis.mcmaster.ca