Available online at www.sciencedirect.com
ScienceDirect
Stochastic Processes and their Applications 119 (2009) 2082–2094
stochastic processes and their applications
www.elsevier.com/locate/spa

Poisson–Dirichlet distribution with small mutation rate

Shui Feng*

Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1

Received 13 March 2008; received in revised form 6 October 2008; accepted 3 November 2008
Available online 13 November 2008

Abstract

A large deviation principle is established for the Poisson–Dirichlet distribution when the mutation rate θ converges to zero. The rate function is identified explicitly, and takes on finite values only on states that have finite number of alleles. This result is then applied to the study of the asymptotic behavior of the homozygosity, and the Poisson–Dirichlet distribution with selection. The latter shows that several alleles can coexist when selection intensity goes to infinity in a particular way as θ approaches zero.

MSC: primary 60F10; secondary 92D10

Keywords: Poisson–Dirichlet distribution; Dirichlet process; Homozygosity; Large deviations; Selection

1. Introduction

For $\theta > 0$, let $V_1(\theta) \geq V_2(\theta) \geq \cdots$ be the points of a nonhomogeneous Poisson process with mean measure density

$$\theta v^{-1} e^{-v},\ v > 0.$$

Set

$$V(\theta) = \sum_{i=1}^{\infty} V_i(\theta),$$

and

$$\mathbf{P}(\theta) = (P_1(\theta), P_2(\theta), \ldots) = \left(\frac{V_1(\theta)}{V(\theta)}, \frac{V_2(\theta)}{V(\theta)}, \ldots\right). \tag{1.1}$$

* Tel.: +1 905 525 9140; fax: +1 905 522 0935.
E-mail address: shuifeng@mcmaster.ca.

0304-4149/$ - see front matter
doi:10.1016/j.spa.2008.11.002

Then $\mathbf{P}(\theta)$ and $V(\theta)$ are independent, and $V(\theta)$ is a $Gamma(\theta, 1)$-distributed random variable. The law of $\mathbf{P}(\theta)$ is called the Poisson–Dirichlet distribution with parameter θ, and is denoted by $PD(\theta)$. Clearly $PD(\theta)$ is a probability on the space

$$\nabla = \left\{ \mathbf{p} = (p_1, p_2, \ldots) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i = 1 \right\},$$

equipped with the subspace topology of $[0, 1]^{\infty}$. Let

$$\bar{\nabla} = \left\{ \mathbf{p} = (p_1, p_2, \ldots) : p_1 \geq p_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} p_i \leq 1 \right\}$$

be the closure of ∇ in $[0, 1]^{\infty}$, equipped with the corresponding subspace topology. Then $PD(\theta)$ can be extended naturally to $\bar{\nabla}$.

The labeled version of the Poisson–Dirichlet distribution, the Dirichlet process, was introduced in [1] and is defined as the law of

$$\Xi_{\theta,\nu} = \sum_{k=1}^{\infty} P_k(\theta)\delta_{\xi_k}, \tag{1.2}$$

where ξ_k, $k = 1, \ldots$ is a sequence of i.i.d. random variables, independent of $\mathbf{P}(\theta)$, with a common distribution ν on $[0, 1]$ satisfying $\nu(\{x\}) = 0$ for every x in $[0, 1]$.

The Poisson–Dirichlet distribution was introduced by Kingman [2] to describe the equilibrium distribution of gene frequencies in a large neutral population at a particular locus under the influence of mutation and genetic drift. The component $P_k(\theta)$ represents the proportion of the kth most frequent allele.

A different way of describing $PD(\theta)$ is through the size-biased permutation $(\tilde{P}_1(\theta), \tilde{P}_2(\theta), \ldots)$ of $\mathbf{P}(\theta)$, given by

$$P\{\tilde{P}_1(\theta) = P_i(\theta)|\mathbf{P}(\theta)\} = P_i(\theta), \quad i \geq 1,$$
$$P\{\tilde{P}_{n+1}(\theta) = P_j(\theta)|\tilde{P}_1(\theta), \ldots, \tilde{P}_n(\theta); \mathbf{P}(\theta)\} = \frac{P_j(\theta)\chi_B}{1 - \sum_{k=1}^{n} \tilde{P}_k(\theta)},$$

where $B = \{P_j(\theta) \neq \tilde{P}_k(\theta), 1 \leq k \leq n\}$ and χ_B is the corresponding indicator function. Clearly $PD(\theta)$ is the law of the descending order statistics of $(\tilde{P}_1(\theta), \tilde{P}_2(\theta), \ldots)$.

Let U_k, $k = 1, 2, \ldots$, be a sequence of independent, identically distributed random variables with common distribution $Beta(1, \theta)$ and set

$$X_1 = U_1, \; X_n = (1 - U_1) \cdots (1 - U_{n-1})U_n, \quad n \geq 2. \tag{1.3}$$

It is well known (cf. [3]) that the size-biased permutation $(\tilde{P}_1(\theta), \tilde{P}_2(\theta), \ldots)$ has the same law as $(X_1, X_2, \ldots)$. The representation (1.3) is called the GEM representation named after R.C. Griffiths, S. Engen and J.W. McCloskey for their contributions to the development of the structure. The $PD(\theta)$ also appears as the unique reversible measure (cf. [4]) of the infinitely-many-neutral-alleles diffusion process with state space ∇ and generator

$$A = \frac{1}{2} \sum_{i,j=1}^{\infty} p_i(\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} - \frac{\theta}{2} \sum_{i=1}^{\infty} p_i \frac{\partial}{\partial p_i}$$

defined on an appropriate domain. The word neutral refers to no selective advantages among the alleles.

The infinitely-many-neutral-alleles diffusion process can be derived as the limit in distribution of a sequence of the Wright–Fisher diffusions in population genetics as the population size goes to infinity. If u is the individual mutation rate and N_e is the effective population size, then the parameter $\theta = 4N_e u$ will be the scaled population mutation rate.

In this paper we will focus on the asymptotic behavior of $PD(\theta)$ when θ converges to zero. In terms of the diffusion models, this limit can be realized in two different ways: either the drift term or the scaled mutation rate goes to zero or the diffusion term goes to infinity. The latter corresponds to the extreme scenario when the population is overwhelmed by the force of genetic drift.

When θ is large, the proportions of different alleles under $PD(\theta)$ are evenly spread and approach zero. By direct calculation, $\lim_{\theta\to 0} X_1 = 1$. Since $X_1 \leq P_1(\theta) \leq 1$, it follows that $PD(\theta)$ concentrates around the point $(1, 0, \ldots)$ when θ is small. There are extensive studies of the asymptotic behavior of $PD(\theta)$ when θ goes to infinity [5–10]. Since the proportions of alleles are evenly spread and uniformly small, it is thus natural to see Gaussian structures [7,10] for large θ. For small mutation rates, the study is very limited. The author is aware of only results in [11] for Dirichlet process, and in [12,13] for the infinitely-many-neutral-alleles diffusion model.

The case of $\theta = 1$ is special. It appears as an asymptotic distribution in random number theory [14]. It is also a critical value in the boundary behavior of the infinitely-many-neutral-alleles model. By using techniques from the theory of Dirichlet forms, it was shown in [15] that for the infinitely-many-neutral-alleles model, with probability one, there will exist times at which the sample path will hit the boundary of a finite-dimensional sub-simplex of ∇ or, equivalently, the single point $(1, 0, \ldots)$ iff θ is less than one. The intuition here is that it is possible to have finite number of alleles in the population if mutation rate is small.

But in equilibrium, with $PD(\theta)$ probability one, the number of alleles is always infinity as long as θ is strictly positive. In other words, the critical value of θ between finite number of alleles and infinite number of alleles is zero for $PD(\theta)$. In physical terms this sudden change from one to infinity can be viewed as a phase transition. The objective of this paper is to investigate the microscopic structures during this phase transition. The limiting procedure involved will be θ going to zero. The tools we use are from the theory of large deviations. Our result will reveal a transition structure that can be viewed as a "ladder of energy".

The paper is organized as follows. In Section 2, we establish the large deviation principle for $PD(\theta)$ when θ goes to zero on space $\bar{\nabla}$. The rate function is identified explicitly. Since the rate function takes the value of infinity outside ∇, the large deviation principle also holds in ∇. When a sample of size r is selected from a population with distribution $PD(\theta)$, the probability that all samples are of the same type is called the population homozygosity of order r. In Section 3, the large deviation result is used to study the asymptotic behavior of the homozygosity and the impact of selection. It will be shown that, in contrast to the neutral case, the population under overdominant selection can preserve more than one alleles when θ goes to zero and the selection intensity goes to infinity in a particular way.

2. Large deviations

In this section, we establish the large deviation principle for $PD(\theta)$ when θ goes to zero. The result will be obtained through a series of lemmas and the main techniques in the proof are exponential approximation and the contraction principle [16].

Let $U = U(\theta)$ be a $Beta(1, \theta)$ random variable, $E = [0, 1]$, and $\lambda(\theta) = (-\log(\theta))^{-1}$.

Lemma 2.1. *The family of laws of $U(\theta)$ satisfies a large deviation principle on E as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$I(p) = \begin{cases} 0, & p = 1 \\ 1, & \text{else.} \end{cases} \tag{2.1}$$

Proof. For any $a < b$ in E, let $\mathbf{I}$ denote one of the intervals (a, b), $[a, b)$, $(a, b]$, and $[a, b]$. It follows from direct calculation that for $b < 1$

$$\lim_{\theta\to 0} \lambda(\theta) \log P\{U \in \mathbf{I}\} = -\lim_{\theta\to 0} \frac{\log(1 - c^\theta)}{\log(\theta)} = -1,$$

where $c = \frac{1-b}{1-a}$. If $b = 1$, then $\lim_{\theta\to 0} \lambda(\theta) \log P\{U \in \mathbf{I}\} = 0$. These, combined with compactness of E, implies the result. □

Lemma 2.2. *For $(X_1, X_2, \ldots)$ defined in (1.3) and any $n \geq 1$, the family of laws of $P_{1,n}(\theta) = \max\{X_1, \ldots, X_n\}$ satisfies a large deviation principle on E as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$I_n(p) = \begin{cases} 0, & p = 1 \\ k, & p \in \left[\dfrac{1}{k+1}, \dfrac{1}{k}\right), k = 1, 2, \ldots, n-1 \\ n, & \text{else.} \end{cases} \tag{2.2}$$

Proof. Noting that $P_{1,n}(\theta)$ is a continuous function of $(U_1, \ldots, U_n)$, it follows from Lemma 2.1, the independence, and the contraction principle that the family of the laws of $P_{1,n}(\theta)$ satisfies a large deviation principle on E with speed $\lambda(\theta)$ and rate function

$$I'(p) = \inf\left\{\sum_{i=1}^{n} I(u_i) : u_i \in E, 1 \leq i \leq n; \right.$$
$$\left. \max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_{n-1})u_n\} = p\right\}.$$

For $p = 1$, one has $I'(1) = 0$ by choosing $u_i = 1$ for $i = 1, \ldots, n$. If p is in $[1/2, 1)$, then at least one of the u_i is not one. By choosing $u_1 = p, u_i = 1, i = 2, \ldots, n$, it follows that $I'(p) = 1$ for p in $[1/2, 1)$.

For each $m \geq 2$, we have

$$\begin{aligned} &\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_m)\} \\ &\quad = \max\{u_1, (1-u_1)\max\{u_2, \ldots, (1-u_2)\cdots(1-u_m)\}\}. \end{aligned} \tag{2.3}$$

Noting that

$$\max\{u_1, 1-u_1\} \geq \frac{1}{2}, \quad u_1 \in E,$$

it follows from (2.3) and induction that

$$\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_m)\} \geq \frac{1}{m+1}, \quad u_i \in E, i = 1, \ldots, m. \tag{2.4}$$

Thus, for $2 \le k \le n-1$, and p in $[\frac{1}{k+1}, \frac{1}{k})$, in order for the equality

$$\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_{n-1})u_n\} = p$$

to hold, it is necessary that $u_1, u_2, \ldots, u_k$ are all less than one. In other words, $I'(p) \ge k$. Since the function $\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_k)\}$ is a surjection from E^k into $[\frac{1}{k+1}, 1]$, there exists $u_1 < 1, \ldots, u_k < 1$ such that

$$\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_k)\} = p.$$

By choosing $u_j = 1$ for $j = k+1, \ldots, n$, it follows that $I'(p) = k$.

Finally for p in $[0, \frac{1}{n})$, in order for

$$\max\{u_1, (1-u_1)u_2, \ldots, (1-u_1)\cdots(1-u_{n-1})u_n\} = p$$

to have solutions, each u_i has to be less than one and, thus, $I'(p) = n$. Therefore, $I'(p) = I_n(p)$ for all p in E. □

Lemma 2.3. *The laws of $P_1(\theta)$ under $PD(\theta)$ satisfy a large deviation principle on E as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$S_1(p) = \begin{cases} 0, & p = 1 \\ k, & p \in \left[\dfrac{1}{k+1}, \dfrac{1}{k}\right), k = 1, 2, \ldots \\ \infty, & p = 0. \end{cases} \tag{2.5}$$

Proof. First note that $P_1(\theta)$ has the same distribution as $\hat{P}_1(\theta) = \max\{X_i : i \ge 1\}$. For any $\delta > 0$, it follows from direct calculation that for any $n \ge 1$

$$\begin{aligned} P\{\hat{P}_1(\theta) - P_{1,n}(\theta) > \delta\} &\le P\{(1-U_1)\cdots(1-U_n) > \delta\} \\ &\le \delta^{-1}\left(\frac{\theta}{1+\theta}\right)^n, \end{aligned}$$

which implies that

$$\limsup_{\theta\to 0} \lambda(\theta) \log P\{\hat{P}_1(\theta) - P_{1,n}(\theta) > \delta\} \le -n. \tag{2.6}$$

Hence $\{P_{1,n}(\theta) : \theta > 0\}$ are exponentially good approximations of $\{\hat{P}_1(\theta) : \theta > 0\}$. By direct calculation, for every closed subset F of E

$$\inf_{q\in F} S_1(q) = \limsup_{n\to\infty} \inf_{q\in F} I_n(q).$$

This, combined with Theorem 4.2.16 in [16] and the fact that $S_1(p)$ is a good rate function, implies that a large deviation principle holds for the laws of $\tilde{P}_1$ with speed $\lambda(\theta)$ and rate function

$$\sup_{\delta>0} \liminf_{n\to\infty} \inf_{|q-p|<\delta} I_n(q),$$

which is clearly equal to $S_1(p)$. □

For any $m \ge 1$, let

$$\nabla_m = \left\{(p_1, \ldots, p_m) : 0 \le p_m \le \cdots \le p_1, \sum_{k=1}^m p_k \le 1\right\}, \tag{2.7}$$

and set $Q_{m,\theta}$ to be the law of $(P_1(\theta), \ldots, P_m(\theta))$ under $PD(\theta)$ on space ∇_m.

For any $\delta > 0$, and any $(p_1, \ldots, p_m) \in \nabla_m$, let

$$G((p_1, \ldots, p_m); \delta) = \{(q_1, \ldots, q_m) \in \nabla_m : |q_k - p_k| < \delta, k = 1, \ldots, m\},$$
$$F((p_1, \ldots, p_m); \delta) = \{(q_1, \ldots, q_m) \in \nabla_m : |q_k - p_k| \le \delta, k = 1, \ldots, m\}.$$

Lemma 2.4. *For fixed $m \ge 2$, the family $\{Q_{m,\theta} : \theta > 0\}$ satisfies a large deviation principle on the space ∇_m as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$S_m(p_1, \ldots, p_m) = \begin{cases} 0, & (p_1, p_2, \ldots, p_m) = (1, 0 \ldots, 0) \\ l - 1, & 2 \le l \le m, \sum_{k=1}^{l} p_k = 1, p_l > 0 \\ m + S_1\left(\dfrac{p_m}{1 - \sum_{i=1}^{m} p_i} \wedge 1\right), & \sum_{k=1}^{m} p_k < 1, p_m > 0 \\ \infty, & \text{else.} \end{cases} \tag{2.8}$$

Proof. Let $m \ge 2$ be fixed, and g_1^θ denotes the density function of $P_1(\theta)$. Then for any $p \in (0, 1)$

$$g_1^\theta(p) p(1-p)^{1-\theta} = \theta \int_0^{(p/(1-p))\wedge 1} g_1^\theta(x) \mathrm{d}x. \tag{2.9}$$

The joint density function g_m^θ of $(P_1(\theta), \ldots, P_m(\theta))$ is given by (cf. [17])

$$g_m^\theta(p_1, \ldots, p_m) = \frac{\theta^{m-1}\left(1 - \sum_{k=1}^{m-1} p_k\right)^{\theta-2}}{p_1 \cdots p_{m-1}} g_1^\theta\left(\frac{p_m}{1 - \sum_{k=1}^{m-1} p_k}\right),$$

for

$$(p_1, \ldots, p_m) \in \nabla_m^\circ = \left\{(p_1, \ldots, p_m) \in \nabla_m : 0 < p_m < \cdots < p_1 < 1, \sum_{k=1}^{m} p_k < 1\right\},$$

and is zero otherwise. Thus for any fixed $(p_1, \ldots, p_m) \in \nabla_m^\circ$ we have

$$g_m^\theta(p_1, \ldots, p_m) = \frac{\theta^m\left(1 - \sum_{k=1}^{m} p_k\right)^{\theta-1}}{p_1 \cdots p_m} \int_0^{(p_m/(1-\sum_{k=1}^{m} p_k))\wedge 1} g_1^\theta(u) \mathrm{d}u. \tag{2.10}$$

The key step in the proof is to show that for every $(p_1, \ldots, p_m)$ in ∇_m,

$$\begin{aligned} &\lim_{\delta\to 0} \liminf_{\theta\to 0} \lambda(\theta) \log Q_{m,\theta}(F((p_1, \ldots, p_m); \delta)) \\ &\quad = \lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log Q_{m,\theta}(G((p_1, \ldots, p_m); \delta)) \\ &\quad = -S_m(p_1, \ldots, p_m). \end{aligned} \tag{2.11}$$

For any $(p_1, \ldots, p_m)$ in ∇_m satisfying $\sum_{i=1}^m p_i > 0$, define

$$r = r(p_1, \ldots, p_m) = \max\{i : 1 \le i \le m,\ p_i > 0\}. \tag{2.12}$$

We divide the proof into several disjoint cases.

Case I: $r = 1$, i.e., $(p_1, \ldots, p_m) = (1, \ldots, 0)$.

For any $\delta > 0$,

$$F((1, \ldots, 0); \delta) \subset \{(q_1, \ldots, q_m) \in \nabla_m : |q_1 - 1| \le \delta\},$$

and one can choose $\delta' < \delta$ such that

$$\{(q_1, \ldots, q_m) \in \nabla_m : |q_1 - 1| < \delta'\} \subset G((1, \ldots, 0); \delta).$$

These combined with Lemma 2.3 implies (2.11) in this case.

Case II: $r = m$, $\sum_{k=1}^m p_k < 1$.

Choose $\delta > 0$ so that

$$\delta < \min\left\{ p_m, \frac{1 - \sum_{i=1}^m p_i}{m} \right\}.$$

By (2.10), we have that for any $(q_1, \ldots, q_m)$ in $F((p_1, \ldots, p_m), \delta) \cap \nabla_m^\circ$

$$g_m^\theta(q_1, \ldots, q_m) \le \frac{\theta^m \left(1 - \sum_{k=1}^m (p_k + \delta)\right)^{\theta-1}}{(p_1 - \delta) \cdots (p_m - \delta)} \int_0^{\frac{p_m+\delta}{1-\sum_{k=1}^m (p_k+\delta)} \wedge 1} g_1^\theta(u)\mathrm{d}u,$$

which, combined with Lemma 2.3, implies

$$\begin{aligned}
&\lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log Q_{m,\theta}\{F((p_1, \ldots, p_m); \delta)\} \\
&\le -m + \lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log P\left\{ P_1(\theta) \le \frac{p_m + \delta}{1 - \sum_{k=1}^m (p_k + \delta)} \wedge 1 \right\} \\
&\le -\left[m + S_1\left(\frac{p_m}{1 - \sum_{i=1}^m p_i} \wedge 1 \right) \right],
\end{aligned} \tag{2.13}$$

where the right continuity of $S_1(\cdot)$ is used in the last inequality.

On the other hand, let

$$\tilde{G}((p_1, \ldots, p_m), \delta) = \prod_{i=1}^m \left(p_i + \frac{\delta}{2}, p_i + \delta \right) \cap \nabla_m^\circ,$$

which is clearly a subset of $G((p_1, \ldots, p_m), \delta)$. Using (2.10) again it follows that for any $(q_1, \ldots, q_m)$ in $\tilde{G}((p_1, \ldots, p_m), \delta)$

$$g_m^\theta(q_1, \ldots, q_m) \ge \theta^m \frac{\left(1 - \sum_{k=1}^m (p_k + \delta/2)\right)^{\theta-1}}{(p_1 + \delta) \cdots (p_m + \delta)} \int_0^{((p_m+\delta/2)/(1-\sum_{k=1}^m (p_k+\delta/2)))\wedge 1} g_1^\theta(u)\mathrm{d}u,$$

which, combined with Lemma 2.3, implies

$$\liminf_{\theta\to 0}\lambda(\theta)\log Q_{m,\theta}\{G((p_1,\ldots,p_m);\delta)\} \geq \liminf_{\theta\to 0}\lambda(\theta)\log Q_{m,\theta}\{\tilde{G}((p_1,\ldots,p_m);\delta)\}$$
$$\geq -m - S_1\left(\frac{p_m+\delta/2}{1-\sum_{i=1}^{m}(p_i+\delta/2)}\wedge 1\right).$$

It follows, by letting δ go to zero, that

$$\liminf_{\delta\to 0}\liminf_{\theta\to\infty}\lambda(\theta)\log Q_{m,\theta}\{G((p_1,\ldots,p_m);\delta)\} \geq -S_m(p_1,\ldots,p_m). \tag{2.14}$$

Case III: $2\leq r\leq m-1$, $\sum_{i=1}^{r}p_i<1$ or $p_1=0$.

This case follows from estimate (2.13) and the fact that $S_1(0)=-\infty$.

Case IV: $r=m$, $\sum_{k=1}^{m}p_k=1$.

Noting that for any $\delta>0$

$$F((p_1,\ldots,p_m);\delta)\cap\nabla_m^\circ\subset\{(q_1,\ldots,q_m)\in\nabla_m^\circ:|q_i-p_i|\leq\delta, i=1,\ldots,m-1\}.$$

By applying **Case II** to $(P_1(\theta),\ldots,P_{m-1}(\theta))$ at the point $(p_1,\ldots,p_{m-1})$, we get

$$\lim_{\delta\to 0}\limsup_{\theta\to 0}\lambda(\theta)\log Q_{m,\theta}\{F((p_1,\ldots,p_m);\delta)\} \leq -[m-1+S_1(1)]$$
$$= -(m-1). \tag{2.15}$$

On the other hand, one can choose $\delta>0$ small so that $\frac{q_m}{1-\sum_{i=1}^{m}q_i}>1$ for any $(q_1,\ldots,q_m)$ in $G((p_1,\ldots,p_m);\delta)\cap\nabla_m^\circ$.

Set

$$\tilde{G}=\{(q_1,\ldots,q_m)\in\nabla_m^\circ: p_i<q_i<p_i+\delta/(m-1),$$
$$i=1,\ldots,m-1;\ p_m-\delta<q_m<p_m\}.$$

Clearly $\tilde{G}$ is a subset of $G((p_1,\ldots,p_m);\delta)$. It follows from (2.10) that for any $(q_1,\ldots,q_m)$ in $\tilde{G}$,

$$g_m^\theta(q_1,\ldots,q_m)\geq\frac{\theta^{m-1}\left[\theta\left(1-\sum_{i=1}^{m}q_i\right)^{\theta-1}\right]}{(p_1+\delta/(m-1))\cdots(p_{m-1}+\delta/(m-1))p_m}.$$

For $m\geq 2$, let

$$A_m=\left\{(q_1,\ldots,q_{m-1})\in\nabla_{m-1}: p_i<q_i<p_i+\delta/(m-1),\right.$$
$$\left. i=1,\ldots,m-1,\sum_{j=1}^{m-1}q_j<1\right\}.$$

Then

$$\int_{\tilde{G}}\theta\left(1-\sum_{i=1}^{m}q_i\right)^{\theta-1}\mathrm{d}q_1\cdots\mathrm{d}q_m$$

$$= \int_{A_m} \mathrm{d}q_1 \cdots \mathrm{d}q_{m-1} \int_{p_m-\delta}^{p_m \wedge \left(1-\sum_{i=1}^{m-1} q_i\right)} \theta \left(1 - \sum_{i=1}^{m} q_i\right)^{\theta-1} \mathrm{d}q_m$$

$$= \int_{A_m} \left(1 + \delta - p_m - \sum_{i=1}^{m-1} q_i\right)^{\theta} \mathrm{d}q_1 \cdots \mathrm{d}q_{m-1},$$

which converges to a strictly positive number depending only on δ and $(p_1, \ldots, p_m)$ as θ goes to zero. Hence

$$\begin{aligned} &\lim_{\delta\to 0} \liminf_{\theta\to 0} \lambda(\theta) \log Q_{m,\theta}\{G((p_1, \ldots, p_m); \delta)\} \\ &\geq \lim_{\delta\to 0} \liminf_{\theta\to 0} \lambda(\theta) \log Q_{m,\theta}\{\tilde{G}\} \geq -(m-1). \end{aligned} \tag{2.16}$$

Case V: $2 \leq r \leq m-1$, $\sum_{i=1}^{r} p_i = 1$.

First note that for any $\delta > 0$, $F((p_1, \ldots, p_m); \delta)$ is a subset of

$$\{(q_1, \ldots, q_m) \in \nabla_m : |q_i - p_i| \leq \delta, i = 1, \ldots, r\}.$$

On the other hand, for each $\delta > 0$ one can choose $\delta_0 < \delta$ such that for any $\delta' \leq \delta_0$

$$G((p_1, \ldots, p_m); \delta) \supset \{(q_1, \ldots, q_m) \in \nabla_m^{\circ}; |q_i - p_i| < \delta', i = 1, \ldots, r\}.$$

Thus the result now follows from **Case IV** for $(P_1(\theta), \ldots, P_r(\theta))$.

The lemma now follows from (2.11) and the fact that ∇_m is compact. □

For any $n \geq 1$, set

$$L_n = \left\{(p_1, \ldots, p_n, 0, 0, \ldots) \in \bar{\nabla} : \sum_{i=1}^{n} p_i = 1\right\},$$

and

$$L = \bigcup_{i=1}^{\infty} L_i.$$

Now we are ready to state and prove the main result of this section.

Theorem 2.5. *The family $\{P\,D(\theta) : \theta > 0\}$ satisfies a large deviation principle on $\bar{\nabla}$ as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$S(\mathbf{p}) = \begin{cases} 0, & \mathbf{p} \in L_1 \\ n-1, & \mathbf{p} \in L_n,\ p_n > 0, n \geq 2 \\ \infty, & \mathbf{p} \notin L. \end{cases} \tag{2.17}$$

Remark. Since $\{S(\mathbf{p}) < \infty\}$ is a subset of ∇, the large deviation principle also holds in ∇.

Proof. First note that the topology of the space $\bar{\nabla}$ can be generated by the following metric

$$d(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{\infty} \frac{|p_k - q_k|}{2^k},$$

where $\mathbf{p} = (p_1, p_2, \ldots)$, $\mathbf{q} = (q_1, q_2, \ldots)$. For any fixed $\delta > 0$, let $B(\mathbf{p}, \delta)$ and $\bar{B}(\mathbf{p}, \delta)$ denote the respective open and closed balls centered at $\mathbf{p}$ with radius $\delta > 0$.

We start with the case that $\mathbf{p}$ is not in L.

For any $k \geq 1$, $\delta' > 0$, set

$$\bar{B}_{k,\delta'}(\mathbf{p}) = \{(q_1, q_2, \ldots) \in \bar{\nabla} : |q_i - p_i| \leq \delta', i = 1, \ldots, k\}.$$

Choose $\delta > 0$ so that $2^k \delta < \delta'$. Then

$$\bar{B}(\mathbf{p}, \delta) \subset \bar{B}_{k,\delta'}(\mathbf{p}),$$

and

$$\begin{aligned}
&\lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log PD(\theta)\{\bar{B}(\mathbf{p}, \delta)\} \leq \limsup_{\theta\to 0} \lambda(\theta) \log PD(\theta)\{\bar{B}_{k,\delta'}(\mathbf{p})\} \\
&\quad \leq \limsup_{\theta\to 0} \lambda(\theta) \log Q_{k,\theta}\{F((p_1, \ldots, p_k), \delta')\} \\
&\quad \leq -\inf\{S_k(q_1, \ldots, q_k) : (q_1, \ldots, q_k) \in F((p_1, \ldots, p_k), \delta')\}. \qquad (2.18)
\end{aligned}$$

Letting δ' go to zero, and then k go to infinity, we get

$$\begin{aligned}
&\lim_{\delta\to 0} \liminf_{\theta\to 0} \lambda(\theta) \log PD(\theta)\{B(\mathbf{p}, \delta)\} \\
&\quad = \lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log PD(\theta)\{\bar{B}(\mathbf{p}, \delta)\} = -\infty. \qquad (2.19)
\end{aligned}$$

Next consider the case of $\mathbf{p}$ belonging to L. Without loss of generality, we assume that $\mathbf{p}$ belongs to L_n with $p_n > 0$.

For any $\delta > 0$, let

$$\tilde{G}(\mathbf{p}; \delta) = \{\mathbf{q} \in \bar{\nabla} : |q_k - p_k| < \delta, k = 1, \ldots, n\},$$
$$\tilde{F}(\mathbf{p}; \delta) = \{\mathbf{q} \in \bar{\nabla} : |q_k - p_k| \leq \delta, k = 1, \ldots, n\}.$$

Clearly, $\bar{B}(\mathbf{p}, \delta)$ is a subset of $\tilde{F}(\mathbf{p}; 2^n\delta)$. Since $\sum_{i=1}^n p_i = 1$, it follows that, for any $\delta > 0$, one can find $\delta' < \delta$ such that

$$B(\mathbf{p}, \delta) \supset \tilde{G}(\mathbf{p}; \delta').$$

Using results on $(P_1(\theta), \ldots, P_n(\theta))$ in **Case V** in the proof of Lemma 2.4, we get

$$\begin{aligned}
&\lim_{\delta\to 0} \liminf_{\theta\to 0} \lambda(\theta) \log PD(\theta)(B(\mathbf{p}, \delta)) \\
&\quad = \lim_{\delta\to 0} \limsup_{\theta\to 0} \lambda(\theta) \log PD(\theta)(\bar{B}(\mathbf{p}, \delta)) = -(n-1). \qquad (2.20)
\end{aligned}$$

Finally, the theorem follows from the compactness of $\bar{\nabla}$. □

Remarks. 1. Consider the rate function $S(\cdot)$ as an "energy" function, then the energy needed to get $n \geq 2$ different alleles is $n - 1$. The values of $S(\cdot)$ form a "ladder of energy". The energy needed to get infinite number of alleles is infinity.

2. The effective domain of $S(\cdot)$, defined as $\{\mathbf{p} \in \bar{\nabla} : S(\mathbf{p}) < \infty\}$, is clearly L. This is in sharp contrast to the result in [8] where the rate function associated with large mutation rate has an effective domain of $\{\mathbf{p} \in \bar{\nabla} : \sum_{i=1}^\infty p_i < 1\}$. The two effective domains are disjoint. One is part of the boundary of $\bar{\nabla}$ and the other is the interior of $\bar{\nabla}$, and both have no intersections with the set $\{\mathbf{p} \in \bar{\nabla} : p_1 > p_2 \cdots > 0, \sum_{i=1}^\infty p_i = 1\}$.

3. Applications

In this section we will discuss two applications of Theorem 2.5. The first one is concerned with the large deviation principle for the homozygosity.

A population of diploid individuals, where the chromosomes occur in homologous pairs, can be divided into two groups: the homozygote and the heterozygote. The frequencies of the homozygote and the heterozygote in the population are called the homozygosity and heterozygosity, respectively. In a randomly mating population with allele frequencies $(p_1, p_2, \ldots)$, the homozygosity is given by

$$H_2(p_1, p_2, \ldots) = \sum_{i=1}^{\infty} p_i^2.$$

The heterozygosity is thus given by $1 - H_2(p_1, p_2, \ldots)$ and has been used to describe levels of variation in populations that fail to satisfy the random mating assumption. More information on the homozygosity and the heterozygosity can be found in [18,19].

For $r \geq 2$, select a random sample of size r from a population whose allelic types have distribution $PD(\theta)$. The probability that all samples are of the same type is given by

$$H_r(P_1(\theta), \ldots) = \sum_{i=1}^{\infty} P_i^r(\theta). \tag{3.1}$$

For $r = 2$, this is the homozygosity. Following [7], we call $H_r(\cdot)$ the rth order population homozygosity. It is clear that $H_r(P_1(\theta), \ldots)$ converges to one as θ approaches zero. Our next theorem describes the large deviations of $H_r(\theta)$ from one.

Theorem 3.1. *For any integer $r \geq 2$, the family of laws of $H_r(P_1(\theta), \ldots)$ satisfies a large deviation principle on E as θ goes to zero with speed $\lambda(\theta)$ and rate function*

$$J(p) = \begin{cases} 0, & p = 1 \\ n - 1, & p \in \left[\dfrac{1}{n^{r-1}}, \dfrac{1}{(n-1)^{r-1}}\right), n = 2, \ldots \\ \infty, & p = 0. \end{cases} \tag{3.2}$$

Thus in terms of large deviations, $H_r(P_1(\theta), \ldots)$ behaves the same as $P_1^{r-1}(\theta)$.

Proof. For any integer $r > 1$, $H_r(\mathbf{p})$ is clearly continuous on $\bar{\nabla}$. By Theorem 2.5 and the contraction principle, the family of the laws of $H_r(P_1(\theta), \ldots)$ satisfies a large deviation principle with speed $\lambda(\theta)$ and rate function

$$\inf\{S(\mathbf{q}) : \mathbf{q} \in \bar{\nabla}, H_r(\mathbf{q}) = p\} = \inf\{S(\mathbf{q}) : \mathbf{q} \in L, H_r(\mathbf{q}) = p\}.$$

For $p = 1$, it follows by choosing $\mathbf{q} = (1, 0, \ldots)$ that $\inf\{S(\mathbf{q}) : \mathbf{q} \in \bar{\nabla}, H_r(\mathbf{q}) = p\} = 0$. For $p = 0$, there does not exist $\mathbf{q}$ in L such that $H_r(\mathbf{q}) = p$. Hence $\inf\{S(\mathbf{q}) : \mathbf{q} \in L, H_r(\mathbf{q}) = p\} = \infty$.

For any $n \geq 2$, the minimum of $\sum_{i=1}^{n} q_i^r$ over L_n is $n^{-(r-1)}$ which is achieved when all $q_i's$ are equal. Hence for

$$p \in [n^{-(r-1)}, (n-1)^{-(r-1)}),$$

we have

$$\inf\{S(\mathbf{q}) : \mathbf{q} \in \bar{\nabla}, H_r(\mathbf{q}) = p\} = n - 1 = J(p). \quad \square$$

For any $\alpha(\theta) > 0$ and any nonzero constant s, the Poisson–Dirichlet distribution with selection considered here is a probability measure on $\bar{\nabla}$ given by

$$P_{\alpha(\theta),sH_r,\theta}(\mathrm{d}\mathbf{p}) = \left(\int_{\bar{\nabla}} \mathrm{e}^{s\alpha(\theta)H_r(\mathbf{q})} PD(\theta)(\mathrm{d}\mathbf{q})\right)^{-1} \mathrm{e}^{s\alpha(\theta)H_r(\mathbf{p})} PD(\theta)(\mathrm{d}\mathbf{p}),$$

where $\alpha(\theta)$ is the selection intensity. The case of $r = 2$ corresponds to Theorem 4.4 in [20] with the fitness function

$$\sigma(i, j) = s\alpha(\theta)\delta_{ij},$$

and $s > 0 (< 0)$ corresponds to underdominant (overdominant) selection. The case of $r > 2$ can be rightfully viewed as a mathematical generalization.

In our second application, Theorem 2.5 is used to derive the large deviation principle for $P_{\alpha(\theta),sH_r,\theta}(\mathrm{d}\mathbf{p})$.

Theorem 3.2. *The family* $\{P_{\alpha(\theta),sH_r,\theta} : \theta > 0\}$ *satisfies a large deviation principle on* $\bar{\nabla}$ *as* θ *goes to zero with speed* $\lambda(\theta)$ *and rate function*

$$S'(\mathbf{p}) = \begin{cases} S(\mathbf{p}), & \lim\limits_{\theta\to 0} \alpha(\theta)\lambda(\theta) = 0 \\ S(\mathbf{p}) + sc(1 - H_r(\mathbf{p})), & \lim\limits_{\theta\to 0} \alpha(\theta)\lambda(\theta) = c > 0, s > 0 \\ S(\mathbf{p}) + |s|cH_r(\mathbf{p}) - \inf\left\{\dfrac{|s|c}{n^{r-1}} + n - 1 : n \geq 1\right\}, & \\ \quad \lim\limits_{\theta\to 0} \alpha(\theta)\lambda(\theta) = c > 0, s < 0. & \end{cases} \tag{3.3}$$

Proof. By putting c and s together, we can assume, without loss of generality, that $c = 1$. Theorem 2.5 combined with Varadhan's lemma and the Laplace method implies that the family $\{P_{\alpha(\theta),sH_r,\theta} : \theta > 0\}$ satisfies a large deviation principle on $\bar{\nabla}$ with speed $\lambda(\theta)$ and rate function

$$\sup\{sH_r(\mathbf{q}) - S(\mathbf{q}) : \mathbf{q} \in \bar{\nabla}\} - (sH_r(\mathbf{p}) - S(\mathbf{p})).$$

The theorem then follows from the fact that

$$\sup\{sH_r(\mathbf{q}) - S(\mathbf{q}) : \mathbf{q} \in \bar{\nabla}\} = \begin{cases} s, & s > 0 \\ -\inf\left\{\dfrac{|s|}{n^{r-1}} + n - 1 : n \geq 1\right\}, & s < 0. \end{cases}$$ □

Remarks. 1. The selection has an impact on the rate function only when the selection intensity $\alpha(\theta)$ is proportional to $\lambda(\theta)^{-1}$.

2. The neutral case corresponds to $s = 0$. Assume that $\alpha(\theta) = (\lambda(\theta))^{-1}$. Then for $s > 0$ the homozygote has selective advantage, and the small mutation rate limit is $(1, 0, \ldots)$. The energy $S'(\mathbf{p})$ needed for a large deviation from $(1, 0, \ldots)$ is larger than the neutral energy $S(\mathbf{p})$. For $s < 0$, the heterozygote has selection advantage. Since $S'(\cdot)$ may reach zero at a point that is different from $(1, 0, \ldots)$, several alleles can coexist in the population when the selection intensity goes to infinity and θ approaches zero.

3. Let $r = 2, \lambda_k = \frac{k(k+1)}{2}, k \geq 1$. Then for $-2 < s < 0$, $(1, 0, \ldots)$ is the unique zero point of $S'(\cdot)$; for $-2\lambda_{k+1} < s < -2\lambda_k$, the unique zero point of $S'(\cdot)$ is $(\frac{1}{k+1}, \ldots, \frac{1}{k+1}, 0, \ldots)$; for $s = -2\lambda_k$, $S'(\cdot)$ has two zero points $(\frac{1}{k}, \ldots, \frac{1}{k}, 0, \ldots)$ and $(\frac{1}{k+1}, \ldots, \frac{1}{k+1}, 0, \ldots)$. It is worth noting that $\{\lambda_k : k \geq 1\}$ are the death rates of Kingman's coalescent.

Acknowledgments

The author is grateful to an anonymous referee and an associate editor for their detailed comments and suggestions. The author's research was supported by the Natural Science and Engineering Research Council of Canada.

References

[1] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Statist. 1 (1973) 209–230.
[2] J.F.C. Kingman, Random discrete distributions, J. Roy. Statist. Soc. 37 (1975) 1–22.
[3] J.W. McCloskey, A model for the distribution of individuals by species in an environment, Ph.D. Thesis, Michigan State University, 1965.
[4] S.N. Ethier, T.G. Kurtz, The infinitely-many-neutral-alleles diffusion model, Adv. Appl. Probab. 13 (1981) 429–452.
[5] G.A. Watterson, H.A. Guess, Is the most frequent allele the oldest? Theoret. Popul. Biol. 11 (1977) 141–160.
[6] R.C. Griffiths, On the distribution of allele frequencies in a diffusion model, Theoret. Popul. Biol. 15 (1979) 140–158.
[7] P. Joyce, S.M. Krone, T.G. Kurtz, Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula, Ann. Appl. Probab. 12 (2002) 101–124.
[8] D. Dawson, S. Feng, Asymptotic behavior of Poisson–Dirichlet distribution for large mutation rate, Ann. Appl. Probab. 16 (2) (2006) 562–582.
[9] S. Feng, Large deviations associated with Poisson–Dirichlet distribution and Ewens sampling formula, Ann. Appl. Probab. 17 (5–6) (2007) 1570–1595.
[10] S. Feng, F. Gao, Moderate deviations for Poisson–Dirichlet distribution, Ann. Appl. Probab. 18 (5) (2008) 1794–1824.
[11] J. Sethraman, R.C. Tiwari, Convergence of Dirichlet measures and the interpretation of their parameters, in: S.S. Gupta, J.O. Berger (Eds.), in: Statistical Decision Theory and Related Topics III, vol. 2, Academic, New York, 1982, pp. 305–315.
[12] S.N. Ethier, A class of infinite-dimensional diffusions occurring in population genetics, Indiana Univ. Math. J. 30 (1981) 925–935.
[13] S.N. Ethier, R.C. Griffiths, The transition function of a Fleming-Viot process, Ann. Probab. 21 (3) (1993) 1571–1590.
[14] R. Arratia, A.D. Barbour, S. Tavaré, Logarithmic combinatorial structures: A probabilistic approach, in: EMS Monographs in Mathematics, European Mathematical Society(EMS), Zürich, 2003.
[15] B. Schmuland, A result on the infinitely many neutral alleles diffusion model, J. Appl. Probab. 28 (1991) 253–267.
[16] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, 1st ed., Jones and Bartlett Publishers, Boston, 1992.
[17] G.A. Watterson, The stationary distribution of the infinitely-many neutral alleles diffusion model, J. Appl. Probab. 13 (1976) 639–651.
[18] W.J. Ewens, Mathematical Population Genetics, vol. I, Springer-Verlag, New York, 2004.
[19] G.A. Watterson, Heterosis or neutrality, Genetics 85 (1977) 789–814.
[20] S.N. Ethier, T.G. Kurtz, Convergence to Fleming-Viot processes in the weak atomic topology, Stochastic Process. Appl. 54 (1994) 1–27.