ISSN 2279-9362

Collegio Carlo Alberto

Moderate deviations for Ewens-Pitman exchangeable random partitions

Stefano Favaro
Shui Feng
Fuqing Gao

No. 527
December 2017

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Moderate deviations for Ewens-Pitman exchangeable random partitions

Stefano Favaro[1], Shui Feng[2] and Fuqing Gao[3]

[1] University of Torino and Collegio Carlo Alberto, Torino, Italy.
E-mail: stefano.favaro@unito.it

[2] McMaster University, Hamilton, Canada.
E-mail: shuifeng@univmail.cis.mcmaster.ca

[3] Wuhan University, Hubei, China
E-mail: fqgao@whu.edu.cn

Abstract

Consider a population of individuals belonging to an infinity number of types, and assume that type proportions follow the two-parameter Poisson-Dirichlet distribution. A sample of size n is selected from the population. The total number of different types and the number of types appearing in the sample with a fixed frequency are important statistics. In this paper we establish the moderate deviation principles for these quantities. The corresponding rate functions are explicitly identified, which help revealing a critical scale and understanding the exact role of the parameters. Conditional, or posterior, counterparts of moderate deviation principles are also established.

Key words and phrases: α-diversity; exchangeable random partition; Dirichlet process; large and moderate deviation; random probability measure; two parameter Poisson-Dirichlet distribution

1 Introduction

Consider a population of countable number of individuals belonging to an infinite number of types. The type of each individual is labelled by a point in a Polish space S. The type proportions in the population are thus a point $\mathbf{p} = (p_1, p_2, \ldots)$ in the space $\triangle := \{\mathbf{q} = (q_1, q_2, \ldots) : q_i \geq 0, \sum_{j=1}^{\infty} q_j = 1\}$. For each $n \geq 1$, let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population with X_i denoting the type of the ith sample. The sample diversity is defined as

$$K_n := \text{total number of different types in the sample.}$$

For any $1 \leq l \leq n$, set

$$M_{l,n} := \text{total number of types that appear in the sample } l \text{ times.}$$

The quantity $M_{l,n}$ is typically referred to as the sample diversity with frequency l. Both the random variables K_n and $M_{l,n}$, as well as related functions, provide important statistics for inference about the population diversity.

A natural scheme arises in the occupancy problem. Consider a countable numbers of urns. Balls are put into the urns independently and each ball lands in urn i with probability p_i. After n balls are put into the urns, the total number of occupied urns is K_n, and $M_{l,n}$ is the numbers of urns with l balls inside. Assuming that $p_1 \geq p_2 \geq \ldots$, a comprehensive study of K_n and $M_{l,n}$ was carried out in [15]. See also [14], [1], [2] for some recent contributions. A comprehensive survey of recents progresses in this context is found in [11].

Adding randomness to the type proportions $\mathbf{p}$, the population will have random type proportions with the law $\mathcal{P}$ being a probability on $\triangle$. Note that, instead of being independent and identically distributed (iid), the random sample $X_1, X_2, \ldots, X_n$ becomes exchangeable. In particular, following the de Finetti theorem, the random type proportions are recovered from the masses of the limit of empirical distributions of the random sample as n tends to infinity. This framework fits naturally in the context of Bayesian nonparametric inference. See, e.g., [7]. In particular the law $\mathcal{P}$ can be viewed as the prior distribution on the unknown species composition $(p_i)_{i\geq 1}$ of the population. The main interests in Bayesian nonparametrics are the posterior distribution of $\mathcal{P}$ given an initial sample $(X_1, \ldots, X_n)$ and associated statistical inferences. More specifically, given an initial sample $(X_1, \ldots, X_n)$, interest lies in making inference based on certain statistics induced by an additional unobserved sample of size m. These include, among others, the sample diversity $K_m^{(n)}$ and the sample diversity $M_{l,m}^{(n)}$ with frequency l to be observed in the additional sample of size m. We call $K_m^{(n)}$ and $M_{l,m}^{(n)}$ the posterior sample diversity and the posterior sample diversity with frequency l, respectively.

The most studied family of probabilities on $\triangle$ is Kingman's Poisson-Dirichlet distribution ([16]) describing in the genetics context the distribution of allele frequencies in a neutral population. This is followed by the study of the two-parameter Poisson-Dirichlet distribution ([18]). Various generalizations of these models can be found in [3], [19] and the references therein.

The focus of this paper is on the asymptotic behaviour of all these sample diversities when the random proportions in the population follow Kingman's Poisson-Dirichlet distribution and its two-parameter generalization. Specifically, for any α in $[0,1)$ and $\theta > -\alpha$, let U_k, $k = 1, 2, \cdots$, be a sequence of independent random variables such that U_k has $Beta(1-\alpha, \theta+k\alpha)$ distribution. If

$$V_1(\alpha,\theta) = U_1,\ V_n(\alpha,\theta) = (1-U_1)\cdots(1-U_{n-1})U_n,\ n \geq 2.$$

then

$$\mathbf{V}(\alpha,\theta) = (V_1(\alpha,\theta), V_2(\alpha,\theta), \cdots) \in \triangle$$

with probability 1. The law of the descending order statistic $\mathbf{P}(\alpha,\theta) = (P_1(\alpha,\theta), P_2(\alpha,\theta), \cdots)$ of $\mathbf{V}(\alpha,\theta)$ is the so-called the two-parameter Poisson-Dirichlet distribution and is denoted by $PD(\alpha,\theta)$. Kingman's Poisson-Dirichlet distribution which corresponds to $\alpha = 0$. The sample diversities $K_n, K_m^{(n)}$, $M_{l,n}$ and $M_{l,m}^{(n)}$ depend on the parameters θ and α. For notational convenience we will not indicate the dependence explicitly. When $\alpha = 0$, the parameter θ corresponds to the scaled population mutation rate. The sample diversity K_n turns out to be a sufficient statistic for the estimation of θ.

There have been many studies on the behaviour of K_n and $M_{l,n}$, as n goes to infinity, and of $K_m^{(n)}$ and $M_{l,m}^{(n)}$, as m goes to infinity. In the case $\alpha = 0$, one can represent K_n as the summation of independent Bernoulli random variables and show that $\frac{K_n}{\ln n}$ converges to θ almost surely. In [12] ($\alpha = 0, \theta = 1$) and [13]($\alpha = 0$, general θ) the following central limit theorem was obtained

$$\frac{K_n - \theta \ln n}{\sqrt{\ln n}} \Rightarrow N(0,1),$$

as n goes to infinity, with $\Rightarrow$ denoting the weak convergence. When the parameter α is positive, the Gaussian limit no longer holds. In particular, it was shown in [17] that one has

$$\lim_{n\to\infty} \frac{K_n}{n^\alpha} = S_{\alpha,\theta}, \qquad a.s.$$

where $S_{\alpha,\theta}$ is related to the Mittag-Leffler distribution. For any $l \geq 1$, the following holds ([19]):

$$\lim_{n\to\infty} \frac{M_{l,n}}{n^\alpha} = (-1)^{l-1} \binom{\alpha}{l} S_{\alpha,\theta}, \qquad a.s.$$

The random variable $S_{\alpha,\theta}$ is referred to as the α-diversity of the $PD(\alpha,\theta)$ distribution. Large deviation principles for K_n were established in [10]. The fluctuation behaviour of $K_m^{(n)}$ and $M_{l,m}^{(n)}$, as m goes to infinity, were studied in [6], where the notion of posterior α-diversity were introduced. Moreover, the associated large deviation principles have been recently established in [8] and [9].

The main results of the present paper are the moderate deviation principles (henceforth MDPs) for the sample diversities K_n, $K_m^{(n)}$, $M_{l,n}$ and $M_{l,m}^{(n)}$ under $PD(\alpha,\theta)$ with $\alpha > 0$. Our study is motivated by a better understanding of the non-Gaussian moderate deviation behaviour and a refined analysis about the role of the parameters α and θ involved. Interestingly, our results identify a critical scale and reveal the role of the parameters θ and α explicitly. The paper is organized as follows. Section 2 contains the study of MDPs for the sample diversities K_n and $M_{l,n}$. The corresponding results for the posterior sample diversities are then presented in Section 3. A key step here is a Bernoulli representation of $K_m^{(n)}$ and $M_{l,m}^{(n)}$. All terminologies and theorems on large and moderate deviations are based on the reference [5].

2 Moderate deviations for K_n and $M_{l,n}$

In the case $\alpha = 0$ and $\theta > 0$, K_n is the summation of independent Bernoulli random variables, and for each $1 \leq l \leq n$ $M_{l,n}$ is approximately a Poisson random variable. Accordingly, the corresponding moderate deviations are standard. Hence we assume in the sequel that $0 < \alpha < 1$ and $\theta + \alpha > 0$.

Moderate deviations in these cases lie between the fluctuation limit results for $\frac{K_n}{n^\alpha}$ and $\frac{M_{l,n}}{n^\alpha}$, and the large deviation results for $\frac{K_n}{n}$ and $\frac{M_{l,n}}{n}$, respectively. In particular our objectives consist of establishing large deviation principles for $\frac{K_n}{n^\alpha \beta_n}$ and $\frac{M_{l,n}}{n^\alpha \beta_n}$ where β_n converges to infinity at a slower pace than $n^{1-\alpha}$ as n tends to infinity. More specifically, we assume that β_n satisfies

$$\lim_{n\to\infty} \frac{\beta_n}{n^{1-\alpha}} = 0, \quad \lim_{n\to\infty} \frac{\beta_n}{(\ln n)^{1-\alpha}} = \infty. \tag{1}$$

The assumption that β_n grows faster that $(\ln n)^{1-\alpha}$ is crucial for establishing the following MDP.

Theorem 2.1 *For any $\alpha \in (0,1)$ and for any $\theta > -\alpha$, $\frac{K_n}{n^\alpha \beta_n}$ satisfies a large deviation principle on $\mathbb{R}$ with speed $\beta_n^{1/(1-\alpha)}$ and rate function $I_\alpha(\cdot)$ defined by*

$$I_\alpha(x) = \begin{cases} (1-\alpha)\alpha^{\alpha/(1-\alpha)} x^{1/(1-\alpha)} & \text{if } x > 0, \\ \\ +\infty & \text{if } x \leq 0. \end{cases}$$

Proof. Let us define $\tilde{K}_n = \frac{K_n}{n^\alpha \beta_n}$. First, by a direct calculation, one has that for any $\lambda \leq 0$

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \mathbb{E}\left[\exp\{\lambda \beta_n^{1/(1-\alpha)} \tilde{K}_n\}\right] = 0.$$

For any $\lambda > 0$, set $y_n = 1 - \exp\{-\lambda n^{-\alpha} \beta_n^{\alpha/(1-\alpha)}\}$. First assume $\theta = 0$. Then by equation (3.5) in [10], we have

$$\begin{aligned} \mathbb{E}\left[\exp\{\lambda \beta_n^{1/(1-\alpha)} \tilde{K}_n\}\right] &= \mathbb{E}\left[(1-y_n)^{-K_n}\right] \\ &= \sum_{i=0}^{\infty} y_n^i \binom{i\alpha + n - 1}{n-1}. \end{aligned}$$

Let $\lfloor i\alpha \rfloor$ denote the integer part of $i\alpha$. It follows from direct calculation that

$$\sum_{i=0}^{\infty} y_n^i \binom{i\alpha + n - 1}{n-1}$$

$$\geq \sum_{i=0}^{\infty} y_n^i \binom{\lfloor i\alpha \rfloor + n - 1}{n-1} = \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} \sum_{\lfloor i\alpha \rfloor = k} y_n^i$$
$$\geq y_n^{1/\alpha} \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} (y_n^{1/\alpha})^k = \frac{y_n^{1/\alpha}}{(1-y_n^{1/\alpha})^n}.$$

On the other hand,

$$\sum_{i=0}^{\infty} y_n^i \binom{i\alpha + n - 1}{n-1}$$
$$\leq \sum_{i=0}^{\infty} y_n^i \binom{\lfloor i\alpha \rfloor + n}{n-1} = \sum_{i=0}^{\infty} y_n^i \frac{\lfloor i\alpha \rfloor + n}{\lfloor i\alpha \rfloor + 1} \binom{\lfloor i\alpha \rfloor + n - 1}{n-1}$$
$$\leq n \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} \sum_{\lfloor i\alpha \rfloor = k} (y_n^{1/\alpha})^{i\alpha} \leq \frac{n}{\alpha} \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} (y_n^{1/\alpha})^k$$
$$= \frac{n}{\alpha} \frac{1}{(1-y_n^{1/\alpha})^n}.$$

Putting these together and applying assumption (1) one gets

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \mathbb{E}\left[\exp\{\lambda n^{-\alpha} \beta_n^{\alpha/(1-\alpha)} K_n\}\right]$$
$$= \lim_{n\to\infty} \ln \left[1 - \left(1 - \exp\{-\lambda n^{-\alpha} \beta_n^{\alpha/(1-\alpha)}\}\right)^{1/\alpha}\right]^{-n\beta_n^{-1/(1-\alpha)}}$$
$$= \lambda^{1/\alpha}.$$

Since the law of K_n under $PD(\alpha, \theta)$ is equivalent to the law of K_n under $PD(\alpha, 0)$, the above limit holds for $\lambda \geq 0$,

Set

$$\Lambda(\lambda) = \begin{cases} \lambda^{1/\alpha} & \text{if } \lambda > 0, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Noting that $I_\alpha(x) = \sup_{\lambda \in \mathbb{R}}\{\lambda x - \Lambda(\lambda)\}$, the conclusion holds following Gärtner-Ellis theorem ([5]).

□

Theorem 2.1 introduces a moderate deviation principle for K_n. Rewrite the rate function as

$$I_\alpha(x) = \exp\{\frac{1}{1-\alpha}[H_\alpha + \ln x]\}$$

with $H_\alpha = (1-\alpha)\ln(1-\alpha) + \alpha \ln \alpha$ being the entropy function, it follows that $\alpha x = 1$ is a critical curve. For $0 < x \leq 1$, $I_\alpha(x)$ is decreasing in α. For $x > 1$ $I_\alpha(x)$ decreases for α in

$(0, 1/x)$, increases for α in $(1/x, 1)$. The minimum is achieved at the point $1/x$. Discounting the scale differences, these results provide a refined comparison between different models in terms of deviation manners.

In the next theorem we establish the MDP for $M_{l,n}$ for any $l \geq 1$.

Theorem 2.2 *For any $\alpha \in (0,1)$ and for any $\theta > -\alpha$, $\frac{M_{l,n}}{n^\alpha \beta_n}$ satisfies a large deviation principle on $\mathbb{R}$ with speed $\beta_n^{1/(1-\alpha)}$ and rate function $I_{\alpha,l}(\cdot)$ defined by*

$$I_{\alpha,l}(x) = \begin{cases} (1-\alpha)\left(\frac{l!}{(1-\alpha)_{(l-1)\uparrow 1}}\right)^{\alpha/(1-\alpha)} x^{1/(1-\alpha)} & \text{if } x > 0, \\ \\ +\infty & \text{if } x \leq 0, \end{cases}$$

where $(a)_{j\uparrow b} = a(a+b)\cdots(a+(j-1)b)$ with the proviso $(a)_{0\uparrow b} = 1$.

Proof. Let y_n be as in Theorem 2.1. Set

$$y_{n,l} = \frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!} \frac{y_n}{1-y_n}.$$

By an argument similar to the proof of Lemma 2.1 in [8], we obtain that for any $\lambda > 0$

$$\begin{aligned} \mathbb{E}\left[\exp\{\lambda n^{-\alpha}\beta_n^{\alpha/(1-\alpha)} M_{l,n}\}\right] &= \mathbb{E}\left[\left(\frac{1}{1-y_n}\right)^{M_{l,n}}\right] \\ &= \sum_{i=0}^{\lfloor n/l \rfloor} y_{n,l}^i \frac{n}{n-il+\alpha i}\binom{n-il+i\alpha}{n-il}. \end{aligned}$$

Note that, since $1 \leq \frac{n}{n-il+\alpha i} \leq \frac{l}{\alpha}$ for $i = 0, \ldots, \lfloor n/l \rfloor$, it follows that the large n approximation of

$$\mathbb{E}\left[\exp\{\lambda n^{-\alpha}\beta_n^{\alpha/(1-\alpha)} M_{l,n}\}\right]$$

is equivalent to that of

$$H_{n,l} = \sum_{i=0}^{\lfloor n/l \rfloor} y_{n,l}^i \binom{n-il+i\alpha}{n-il}.$$

Set

$$H_{n,l}^- = \sum_{i=0}^{\lfloor n/l \rfloor} y_{n,l}^i \binom{n-il+\lfloor i\alpha \rfloor}{n-il}$$

and

$$H_{n,l}^+ = \sum_{i=0}^{\lfloor n/l \rfloor} y_{n,l}^i \binom{n-il+\lfloor i\alpha \rfloor + 1}{n-il}.$$

It is clear that

$$H_{n,l}^- \le H_{n,l} \le H_{n,l}^+ \le (n+1)H_{n,l}^-.$$

The assumption for β_n guarantees that the factor $n+1$ in the upper bound does not contribute to the scaled logarithmic limit. Accordingly, we can write

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \mathbb{E}\left[\exp\{\lambda n^{-\alpha}\beta_n^{\alpha/(1-\alpha)} M_{l,n}\}\right] = \lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln H_{n,l}^-. \tag{2}$$

To estimate $H_{n,l}^-$, we write

$$\begin{aligned} H_{n,l}^- &= \sum_{i=0}^{\lfloor n/l \rfloor} (y_{n,l}^{1/\alpha})^{i\alpha} \frac{(n-il+1)\cdots(n-il+\lfloor i\alpha \rfloor)}{(\lfloor i\alpha \rfloor)!} \\ &= \sum_{i=0}^{\lfloor n/l \rfloor} (y_{n,l}^{1/\alpha})^{i\alpha - \lfloor i\alpha \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(1-il)/n)\cdots(1+(\lfloor i\alpha \rfloor - il)/n)}{(\lfloor i\alpha \rfloor)!} \end{aligned}$$

which is controlled from below by

$$\sum_{i=0}^{\lfloor n/l \rfloor} (y_{n,l}^{1/\alpha})^{i\alpha - \lfloor i\alpha \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(1-il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}$$

and from above by

$$\sum_{i=0}^{\lfloor n/l \rfloor} (y_{n,l}^{1/\alpha})^{i\alpha - \lfloor i\alpha \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(\lfloor i\alpha \rfloor - il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}.$$

Since $(y_{n,l}^{1/\alpha})^{i\alpha - \lfloor i\alpha \rfloor}$ does not affect the scaled logarithmic limit in (2), it suffices to focus on

$$D_{n,l} = \sum_{i=0}^{\lfloor n/l \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(1-il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}$$

and

$$J_{n,l} = \sum_{i=0}^{\lfloor n/l \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(\lfloor i\alpha \rfloor - il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}$$

Set $\gamma_n = \lfloor \beta_n^{1/(1-\alpha)} \rfloor$ and write

$$D_{n,l} = D_{n,l}^1 + D_{n,l}^2$$

with

$$D_{n,l}^1 = \sum_{i=0}^{\gamma_n} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(1-il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}.$$

It follows that

$$D_{n,l}^2 = \sum_{i=\gamma_n+1}^{\lfloor n/l \rfloor} (n y_{n,l}^{1/\alpha})^{\lfloor i\alpha \rfloor} \frac{(1+(1-il)/n)^{\lfloor i\alpha \rfloor}}{(\lfloor i\alpha \rfloor)!}$$

$$\leq \sum_{i=\gamma_n+1}^{\lfloor n/l\rfloor} \frac{(ny_{n,l}^{1/\alpha})^{\lfloor i\alpha\rfloor}}{(\lfloor i\alpha\rfloor)!} \leq \frac{1}{\alpha} \sum_{k=\lfloor(\gamma_n+1)\alpha\rfloor}^{\infty} \frac{(ny_{n,l}^{1/\alpha})^k}{k!} \tag{3}$$

$$\leq \frac{1}{\alpha} \frac{(ny_{n,l}^{1/\alpha})^{\lfloor(\gamma_n+1)\alpha\rfloor}}{\lfloor(\gamma_n+1)\alpha\rfloor!} \exp\{ny_{n,l}^{1/\alpha}\}.$$

By direct calculation, we have

$$\lim_{n\to\infty} \frac{ny_{n,l}^{1/\alpha}}{\beta_n^{1/(1-\alpha)}} = \left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\lambda\right)^{1/\alpha} \tag{4}$$

and

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln\lfloor(\gamma_n+1)\alpha\rfloor! = \infty. \tag{5}$$

Hence

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln D_{n,l}^2 = -\infty.$$

This implies that

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln D_{n,l} = \lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln D_{n,l}^1.$$

Noting that $\lim_{n\to\infty} \max_{10\leq i\leq\gamma_n}\{|(1-il)/n|\} = 0$, we obtain

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln D_{n,l}^1 = \lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \sum_{i=0}^{\gamma_n} \frac{(ny_{n,l}^{1/\alpha})^{\lfloor i\alpha\rfloor}}{(\lfloor i\alpha\rfloor)!}.$$

By an argument similar to that used in deriving the estimation (3), and taking into account of (4), we obtain that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln D_{n,l} \\ &= \lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \sum_{i=0}^{\gamma_n} \frac{(ny_{n,l}^{1/\alpha})^{\lfloor i\alpha\rfloor}}{(\lfloor i\alpha\rfloor)!} \\ &= \lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln \exp\{ny_{n,l}^{1/\alpha}\} \\ &= \left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\lambda\right)^{1/\alpha}, \end{aligned} \tag{6}$$

Similarly we can prove that

$$\lim_{n\to\infty} \frac{1}{\beta_n^{1/(1-\alpha)}} \ln J_{n,l} = \left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\lambda\right)^{1/\alpha}. \tag{7}$$

The result now follows from (2), (6), (7) and Gärtner-Ellis theorem.

□

3 Moderate deviations for $K_m^{(n)}$ and $M_{l,m}^{(n)}$

Given $n \geq 1$, let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be a sample from the population with type proportions following two parameter Poisson-Dirichlet distribution $PD(\alpha, \theta)$. Let the sample $\mathbf{X}_n$ featuring $K_n = j \leq n$ distinct types with corresponding frequencies $\mathbf{N}_n = (N_{1,1}, \ldots, N_{1,K_n}) = (n_1, \ldots, n_j)$, and let $M_{l,n}$ be the number of distinct types with frequency $1 \leq l \leq n$. Now consider an additional sample $\mathbf{X}_m^{(n)} = (X_{n+1}, \ldots, X_{n+m})$ of size m, and let $K_m^{(n)}$ and $M_{l,m}^{(n)}$ be the sample diversity and sample diversity with frequency $1 \leq l \leq m$ in $\mathbf{X}_m^{(n)}$. In this section we derive the MDPs for $K_m^{(n)}$ and $M_{l,m}^{(n)}$ as m tends to infinity given $\mathbf{X}_n$, K_n and $\mathbf{N}_n$. The law of the type proportions of the population is now the posterior distribution of $PD(\alpha, \theta)$ given $\mathbf{X}_n$. Structurally we can divide the type into two groups: types appeared in the sample $\mathbf{X}_n$ and brand new types.

Let $L_m^{(n)}$ be the number of X_{n+i}'s, for $i = 1, \ldots, m$, that do not coincide with X_i's, for $i = 1, \ldots, n$. Also, let

i) $\tilde{K}_m^{(n)}$ be the number of new distinct types in the additional sample $\mathbf{X}_m$, i.e. the number of types in $\mathbf{X}_m^{(n)}$ which do not coincide with any of the types that appear in the initial sample $\mathbf{X}_n$;

ii) $\tilde{M}_{l,m}^{(n)}$ be the number of new distinct types with frequency l in the additional sample $\mathbf{X}_m$, i.e., the number of types with frequency l among the new types that appear in $\mathbf{X}_m^{(n)}$, such that

$$\sum_{l=1}^{m} \tilde{M}_{l,m}^{(n)} = \tilde{K}_m^{(n)} \quad \text{and} \quad \sum_{l=1}^{n} l\tilde{M}_{l,m}^{(n)} = L_m^{(n)}.$$

Since the sample $\mathbf{X}_n$ is fixed, the moderate deviations for $K_m^{(n)}$ and $M_{l,m}^{(n)}$ are equivalent to the corresponding moderate deviations for $\tilde{K}_m^{(n)}$ and $\tilde{M}_{m,l}^{(n)}$. Thus we will focus on $\tilde{K}_m^{(n)}$ and $\tilde{M}_{m,l}^{(n)}$ in the sequel. The key step in the proof is the following representation for the conditional, or posterior, distributions of $\tilde{K}_m^{(n)}$ given $(K_n, \mathbf{N}_n)$ and of $\tilde{M}_{l,m}^{(n)}$ given $(K_n, \mathbf{N}_n)$, for any $l = 1, \ldots, m$. With a slight abuse of notation, throughout this section we write $X \,|\, Y$ to denote a random variable whose distribution coincides with the conditional distribution of X given Y.

Theorem 3.1 *For any $k \geq 1$ and $p \in [0, 1]$, let $Z_{k,p}$ be Binomial random variable with parameter (k, p), and for any $a, b > 0$ let $B_{a,b}$ be a Beta random variable with parameter (a, b). If K_m^* and $M_{l,m}^*$ denote the number of distinct types and the number of distinct types with frequency $1 \leq l \leq m$, respectively, in a sample of size m from $PD(\alpha, \theta + n)$, then we have*

$$\tilde{K}_m^{(n)} \,|\, (K_n = j, \mathbf{N}_n = (n_1, \ldots, n_j)) \stackrel{d}{=} \tilde{K}_m^{(n)} \,|\, (K_n = j) \stackrel{d}{=} Z_{K_m^*, B_{\frac{\theta}{\alpha}+j, \frac{n}{\alpha}-j}} \tag{8}$$

and

$$\tilde{M}^{(n)}_{l,m} \,|\, (K_n = j, \mathbf{N}_n = (n_1, \ldots, n_j)) \stackrel{d}{=} \tilde{M}^{(n)}_{l,m} \,|\, (K_n = j) \stackrel{d}{=} Z_{M^*_{l,m}, B_{\frac{\theta}{\alpha}+j, \frac{n}{\alpha}-j}} \tag{9}$$

where $\stackrel{d}{=}$ *denotes the equality in distribution, and* $B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j}$ *is independent of* K^*_m *and of* $M^*_{l,m}$.

Proof. Since all random variables involved are bounded, it suffices to verify the equality of all moments. We start by recalling some moment formulate for K^*_m and $M^*_{l,m}$ (cf. [20] and [6]). In particular one has

$$\mathbb{E}[(K^*_m)_{r\downarrow 1}] = \left(\frac{\theta+n}{\alpha}\right)_{r\uparrow 1} \sum_{i=0}^{r} (-1)^{r-i} \binom{r}{i} \frac{(\theta+n+i\alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} \tag{10}$$

and

$$\begin{aligned} &\mathbb{E}[(M^*_{l,m})_{r\downarrow 1}] \\ &\quad = (m)_{rl\downarrow 1} \left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\right)^r \left(\frac{\theta+n}{\alpha}\right)_{r\uparrow 1} \frac{(\theta+n+r\alpha)_{(m-rl)\uparrow 1}}{(\theta+n)_{m\uparrow 1}}, \end{aligned} \tag{11}$$

where $(c)_{j\downarrow 1} = (c)_{j\uparrow -1}$ Moreover, let us recall the factorial moment of order r of the Binomial random variable $Z_{n,p}$, i.e.,

$$\mathbb{E}[(Z_{n,p})^r] = \sum_{t=0}^{r} S(r,t)(n)_{t\downarrow 1} p^t, \tag{12}$$

with $S(n,k)$ being the Stirling number of the second kind. If $S(n,k;a)$ denotes the non-central Stirling number of the second kind, see [4], then by means of Proposition 1 in [7] we have

$$\begin{aligned} &\mathbb{E}[(\tilde{K}^{(n)}_m)^r \,|\, K_n = j] \\ &\quad = \sum_{i=0}^{r} (-1)^{r-i} \left(j+\frac{\theta}{\alpha}\right)_{i\uparrow 1} S\left(r,i;j+\frac{\theta}{\alpha}\right) \frac{(\theta+n+i\alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} \\ &\text{(by expanding } S(r,i;j+\theta/\alpha) \text{ as a finite sum)} \\ &\quad = \sum_{i=0}^{r} (-1)^{-i} \frac{(\theta+n+i\alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} \sum_{t=i}^{r} (-1)^t \binom{t}{i} S(r,t) \left(j+\frac{\theta}{\alpha}\right)_{t\uparrow 1} \\ &\quad = \sum_{t=0}^{r} S(r,t) \frac{\left(j+\frac{\theta}{\alpha}\right)_{t\uparrow 1}}{\left(\frac{\theta+n}{\alpha}\right)_{t\uparrow 1}} \left(\frac{\theta+n}{\alpha}\right)_{t\uparrow 1} \sum_{i=0}^{t} (-1)^{t-i} \binom{t}{i} \frac{(\theta+n+i\alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} \\ &\text{(by Equation (10))} \\ &\quad = \sum_{t=0}^{r} S(r,t) \frac{\left(j+\frac{\theta}{\alpha}\right)_{t\uparrow 1}}{\left(\frac{\theta+n}{\alpha}\right)_{t\uparrow 1}} \mathbb{E}[(K^*_m)_{t\downarrow 1}] \\ &\text{(by expanding } (j+\theta/\alpha)_{t\uparrow 1}/((\theta+n)/\alpha)_{t\uparrow 1} \text{ as an Euler integral)} \\ &\quad = \sum_{t=0}^{r} S(r,t) \mathbb{E}[(K^*_m)_{t\downarrow 1}] \frac{\Gamma\left(\frac{\theta+n}{\alpha}\right)}{\Gamma\left(\frac{\theta}{\alpha}+j\right)\Gamma\left(\frac{n}{\alpha}-j\right)} \int_0^1 x^{t+\frac{\theta}{\alpha}+j-1}(1-x)^{\frac{n}{\alpha}-j-1} dx \end{aligned}$$

$$= \sum_{t=0}^{r} S(r,t)\mathbb{E}[(K_m^*)_{t\downarrow 1}]\mathbb{E}[(B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j})^t]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{r} S(r,t)(K_m^*)_{t\downarrow 1}(B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j})^t\right]\right]$$

(by Equation (12))

$$= \mathbb{E}\left[\left(Z_{K_m^*,B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j}}\right)^r\right]$$

and the proof of the representation (8) is completed. Similarly, by Theorem 2 in [6] we can write

$$\mathbb{E}[(\tilde{M}_{l,m}^{(n)})^r \mid K_n = j]$$

$$= \sum_{t=0}^{r} S(r,t)(m)_{tl\downarrow 1}\left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\right)^t\left(j+\frac{\theta}{\alpha}\right)_{t\uparrow 1}\frac{(\theta+n+t\alpha)_{(m-tl)\uparrow 1}}{(\theta+n)_{m\uparrow 1}}$$

(by Equation (11))

$$= \sum_{t=0}^{r} S(r,t)\frac{(j+\frac{\theta}{\alpha})_{t\uparrow 1}}{(\frac{\theta+n}{\alpha})_{t\uparrow 1}}\mathbb{E}[(M_{l,m}^*)_{t\downarrow 1}]$$

(by expanding $(j+\theta/\alpha)_{t\uparrow 1}/((\theta+n)/\alpha)_{t\uparrow 1}$ as an Euler integral)

$$= \sum_{t=0}^{r} S(r,t)\mathbb{E}[(M_{l,m}^*)_{t\downarrow 1}]\frac{\Gamma\left(\frac{\theta+n}{\alpha}\right)}{\Gamma\left(\frac{\theta}{\alpha}+j\right)\Gamma\left(\frac{n}{\alpha}-j\right)}\int_0^1 x^{t+\frac{\theta}{\alpha}+j-1}(1-x)^{\frac{n}{\alpha}-j-1}d\,x$$

$$= \sum_{t=0}^{r} S(r,t)\mathbb{E}[(M_{l,m}^*)_{t\downarrow 1}]\mathbb{E}[(B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j})^t]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sum_{t=0}^{r} S(r,t)(M_{l,m}^*)_{t\downarrow 1}(B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j})^t\right]\right]$$

(by Equation (12))

$$= \mathbb{E}\left[\left(Z_{M_{l,m}^*,B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j}}\right)^r\right]$$

and the proof of the representation (9) is completed.

□

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

Theorem 3.2 *For any $\alpha \in (0,1)$ and $\theta > -\alpha$, the conditional laws of $\frac{\tilde{K}_m^{(n)}}{m^\alpha \beta_m}$ and $\frac{\tilde{M}_{m,l}^{(n)}}{m^\alpha \beta_m}$ satisfy MDPs that are the same as $\frac{K_m}{m^\alpha \beta_m}$ and $\frac{M_{l,m}}{m^\alpha \beta_m}$, respectively, as m tends to infinity.*

Proof. First observe that the MDPs for $\frac{K_m^*}{m^\alpha \beta_m}$ and $\frac{M_{m,l}^*}{m^\alpha \beta_m}$ are the same as the corresponding MDPs for $\frac{K_m}{m^\alpha \beta_m}$ and $\frac{M_{l,m}}{m^\alpha \beta_m}$, respectively. Furthermore, for any $\lambda \leq 0$ it is not difficult to see that

$$\lim_{m\to\infty}\frac{1}{\beta_m^{1/(1-\alpha)}}\ln \mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{K}_m^{(n)}}|K_n = j]$$

$$
= \lim_{m\to\infty} \frac{1}{\beta_m^{1/(1-\alpha)}} \ln \mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{M}_{m,l}^{(n)}}|K_n = j]
$$

$$
= 0.
$$

Let $\{Y_i : i \geq 1\}$ be iid Bernoulli with parameter $\eta = B_{\frac{\theta}{\alpha}+j,\frac{n}{\alpha}-j}$. it follows from Theorem 3.1 that

$$
\tilde{K}_m^{(n)} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{K_m^*} Y_i, \quad \tilde{M}_{m.l}^{(n)} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{M_{l,m}^*} Y_i.
$$

Hence for $\lambda > 0$,

$$
\mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{K}_m^{(n)}}|K_n = j] \quad \leq \quad \mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}K_m^*}]
$$

and

$$
\begin{aligned}
&\mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{K}_m^{(n)}}|K_n = j] \\
&\quad \mathbb{E}\left[\mathbb{E}[\left(1-\eta+\eta e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}}\right)^{K_m^*}]\right] \\
&\quad \geq \mathbb{E}\left[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}K_m^*}\mathbb{E}[\eta^{K_m^*}]\right] \\
&\quad \geq \mathbb{E}\left[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}K_m^*}\frac{\Gamma(\frac{\theta+n}{\alpha})}{\Gamma(\frac{\theta}{\alpha})}\frac{\Gamma(K_m^*+\frac{\theta}{\alpha})}{\Gamma(K_m^*+\frac{\theta+n}{\alpha})}\right] \\
&\quad \geq \frac{1}{m^{\gamma(m,\alpha,\theta,n,j)}}\mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}K_m^*}]
\end{aligned}
$$

where $\gamma(m, \alpha, \theta, n, j)$ is sequence of positive numbers converging to $\frac{n}{\alpha} - j$ for large m. Thus we have

$$
\lim_{m\to\infty} \frac{1}{\beta_m^{1/(1-\alpha)}} \ln \mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{K}_m^{(n)}}|K_n = j] = \lambda^{1/\alpha}. \tag{13}
$$

Similarly we can show that

$$
\lim_{m\to\infty} \frac{1}{\beta_m^{1/(1-\alpha)}} \ln \mathbb{E}[e^{\lambda m^{-\alpha}\beta_m^{\alpha/(1-\alpha)}\tilde{M}_m^{(n)}}|K_n = j] = \left(\frac{\alpha(1-\alpha)_{(l-1)\uparrow 1}}{l!}\lambda\right)^{1/\alpha}
$$

which combined with (13) led to the theorem.

□

The MDP results in Theorems 2.1, 2.2 and 3.2 identify a critical scale at $(\ln m)^{1-\alpha}$. It is not clear whether MDP holds when β_m is at or has a slower growth rate than $(\ln m)^{1-\alpha}$. Our calculations indicate that if such MDPs hold true, then the posterior MDP and the unconditional MDP may be different.

References

[1] A. Barbour and A. Gnedin (2009). Small counts in the infinite occupancy scheme, *Electron. J. Probab.*, **14**, 365–384.

[2] A. Ben-Hamou, S. Boucheron and M.I. Ohannessian (2016). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, in press.

[3] J. Bertoin, *Random fragmentation and coagulation processes*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2006.

[4] C.A. Charalambides, *Enumerative combinatorics*, Chapman and Hall/CRC, 2002.

[5] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer, New York, 1998.

[6] S. Favaro, A. Lijoi, and I. Prünster (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754.

[7] S. Favaro, A. Lijoi, R.H. Mena, and I. Prünster (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993–1008.

[8] S. Favaro and S. Feng (2014). Asymptotics for the number of blocks in a conditional Ewens-Pitman sampling model, *Electron. J. Probab.*, **19**, 1–15.

[9] S. Favaro and S. Feng (2015). Large deviation principles for the Ewens-Pitman sampling model. *Electron. J. Probab.*, **20**, 1–27.

[10] S. Feng and F.M. Hoppe (1998). Large deviation principles for some random combinatorial structures in population genetics and Brownian motion. *Ann. Appl. Probab.*, **8**, 975–994.

[11] A. Gnedin, B. Hansen, and J. Pitman (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotic and power laws. *Probability Surveys*, 4:146–171.

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

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

[14] H. Hwang and S. Janson (2008) Local limit theorems for finite and infinite urn models *Ann. Probab*, **36**:992–1022

[15] S. Karlin (1967). Central limit theorems for certain infinite urn schemes. *J. Math. and Mech.* , **17**, No.4:373–401.

[16] J.F.C. Kingman (1975). Random discrete distributions. *J. Roy. Stat. Soc. Ser. B,* **37**, 1-22.

[17] J. Pitman (1992). Notes on the two parameter generalization of the Ewens random partition structure. Unpublished notes

[18] J. Pitman and M. Yor (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**, 855–900.

[19] J. Pitman. *Combinatorial stochastic processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer-Verlag, New York, 2006.

[20] H. Yamato and M. Sibuya (2000). Moments of some statistics of Pitman sampling formula. *Bull. Inform. Cybernet.*, **32**, 1–10.