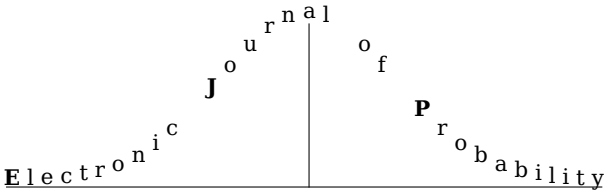


Electron. J. Probab. **20** (2015), no. 40, 1–26.
ISSN: 1083-6489 DOI: 10.1214/EJP.v20-3668

Large deviation principles for the Ewens-Pitman sampling model

Stefano Favaro* Shui Feng†

Abstract

Let $M_{l,n}$ be the number of blocks with frequency l in the exchangeable random partition induced by a sample of size n from the Ewens-Pitman sampling model. We show that as n tends to infinity $n^{-1}M_{l,n}$ satisfies a large deviation principle and we characterize the corresponding rate function. A conditional counterpart of this large deviation principle is also presented. Specifically, given an initial observed sample of size n from the Ewens-Pitman sampling model, we consider an additional unobserved sample of size m thus giving rise to an enlarged sample of size $n+m$. As m tends to infinity, and for any fixed n, we establish a large deviation principle for the conditional number of blocks with frequency l in the enlarged sample, given the initial sample. Interestingly this conditional large deviation principle coincides with the unconditional large deviation principle for $M_{l,n}$, namely there is no long lasting impact of the given initial sample. Applications of the conditional large deviation principle are discussed in the context of Bayesian nonparametric inference for species sampling problems.

Keywords: Bayesian nonparametrics; discovery probability; Ewens-Pitman sampling model; exchangeable random partition; large deviations; population genetics; species sampling problems.
AMS MSC 2010: Primary 60F10, Secondary 92D10.
Submitted to EJP on July 13, 2014, final version accepted on April 8, 2015.

1 Introduction

The Ewens-Pitman sampling model was introduced by Pitman [25] as a generalization of the celebrated sampling model by Ewens [9]. See Pitman [29] for a comprehensive review. In order to define the Ewens-Pitman sampling model, let $\mathbb{X}$ be a Polish space and let ν be a nonatomic probability measure on $\mathbb{X}$. For any $\alpha \in (0,1)$ and $\theta > -\alpha$ let $(X_i)_{i\geq 1}$ be a sequence of $\mathbb{X}$-valued random variables such that $\mathbb{P}[X_1 \in \cdot] = \nu(\cdot)$, and for any $i \geq 1$

$$\mathbb{P}[X_{i+1} \in \cdot \,|\, X_1, \ldots, X_i] = \frac{\theta + j\alpha}{\theta + i}\nu(\cdot) + \frac{1}{\theta + i}\sum_{l=1}^{j}(n_l - \alpha)\delta_{X_l^*}(\cdot) \tag{1.1}$$

with $X_1^*, \ldots, X_j^*$ being the j distinct values in $(X_1, \ldots, X_i)$ with corresponding frequencies $(n_1, \ldots, n_j)$. The conditional probability (1.1) is referred to as the Ewens-Pitman

*University of Torino and Collegio Carlo Alberto, Italy. E-mail: stefano.favaro@unito.it
†McMaster University, Canada. E-mail: shuifeng@univmail.cis.mcmaster.ca

sampling model. The Ewens sampling model is recovered as special case of (1.1) by letting $\alpha \to 0$. Pitman [25] showed that $(X_i)_{i\geq 1}$ is exchangeable and its de Finetti measure Π is the distribution of the two parameter Poisson-Dirichlet process $\tilde{P}_{\alpha,\theta,\nu}$ in Perman [24], namely

$$\begin{aligned} X_i \,|\, \tilde{P}_{\alpha,\theta,\nu} &\overset{\text{iid}}{\sim} \tilde{P}_{\alpha,\theta,\nu} \qquad i = 1,\dots,n \\ \tilde{P}_{\alpha,\theta,\nu} &\sim \Pi, \end{aligned} \tag{1.2}$$

for any $n \geq 1$. In particular $\tilde{P}_{\alpha,\theta,\nu} \overset{\text{d}}{=} \sum_{i\geq 1} V_i \prod_{l\leq i-1}(1-V_l)\delta_{Z_i}$ where $(V_i)_{i\geq 1}$ are independent random variables such that V_i is distributed according to a Beta distribution with parameter $(1-\alpha, \theta+i\alpha)$, and $(Z_i)_{i\geq 1}$ are random variables independent of $(V_i)_{i\geq 1}$ and independent and identically distributed according to ν. See, e.g., Perman [24] and Pitman and Yor [28] for further details on the discrete random probability measure $\tilde{P}_{\alpha,\theta,\nu}$.

According to (1.1) a sample $(X_1,\dots,X_n)$ from $\tilde{P}_{\alpha,\theta,\nu}$ induces an exchangeable random partition of the set $\{1,\dots,n\}$ into $K_n \leq n$ blocks with corresponding frequencies $\mathbf{N}_n = (N_1,\dots,N_{K_n})$ such that $\sum_{1\leq i\leq K_n} N_{i,n} = n$. As shown in Pitman [25] this exchangeable random partition leads to the following generalization of the celebrated Ewens sampling formula: if $M_{l,n}$ denotes the number of blocks with frequency $1 \leq l \leq n$, namely $M_{l,n} = \sum_{1\leq i\leq K_n} \mathbb{1}_{\{N_{i,n}=l\}}$ such that $K_n = \sum_{1\leq l\leq n} M_{l,n}$ and $n = \sum_{1\leq l\leq n} l M_{l,n}$, then

$$\mathbb{P}[(M_{1,n},\dots,M_{n,n}) = (m_1,\dots,m_n)] = \frac{\prod_{i=0}^{j-1}(\theta+i\alpha)}{(\theta)_{(n)}} n! \prod_{i=1}^{n}\left(\frac{(1-\alpha)_{(i-1)}}{i!}\right)^{m_i}\frac{1}{m_i!}, \tag{1.3}$$

where $(x)_{(n)} = x(x+1)\cdots(x+n-1)$ with the proviso $(x)_{(0)} = 1$. The distribution (1.3) is known as the Ewens-Pitman sampling formula and it has been the subject of a rich and active literature. In particular there have been several studies on the large n asymptotic behaviour of K_n in terms of fluctuation limits and large deviations. See, e.g., Pitman [27], Feng and Hoppe [15] and Favaro and Feng [12]. In this paper we focus on large deviations for $M_{l,n}$: for any $\alpha \in (0,1)$ and $\theta > -\alpha$ we show that $n^{-1}M_{l,n}$ satisfies a large deviation principle with speed n and we characterize the corresponding rate function. We also present a conditional counterpart of this large deviation principle. These results complete the study initiated in Feng and Hoppe [15] and Favaro and Feng [12].

1.1 Notation and background

We start by recalling the distribution of the random variables K_n and $M_{l,n}$ induced by a sample $(X_1,\dots,X_n)$ from $\tilde{P}_{\alpha,\theta,\nu}$. For any nonnegative integers n and $j \leq n$ we denote by $\mathscr{C}(n,j;\alpha)$ the so-called generalized factorial coefficient, namely $\mathscr{C}(n,j;\alpha) = (j!)^{-1}\sum_{i=1}^{j}(-1)^i\binom{j}{i}(-i\alpha)_{(n)}$. See Charalambides [6] for a detailed account on this combinatorial coefficient. According to results in Pitman [25] and Favaro et al. [11], one has

$$\mathbb{P}[K_n = j] = \frac{\prod_{i=0}^{j-1}(\theta+i\alpha)}{(\theta)_n}\mathscr{C}(n,j;\alpha) \tag{1.4}$$

and

$$\begin{aligned} \mathbb{P}[M_{l,n} = m_l] = \sum_{i=0}^{n-m_l}(-1)^i \binom{n}{i, m_l, n-lm_l-li} \alpha^{m_l+i}\left(\frac{\theta}{\alpha}\right)_{(m_l+i)} \\ \times \left(\frac{(1-\alpha)_{(l-1)}}{i!}\right)^{m_l+i}\frac{(\theta+(m_l+i)\alpha)_{(n-lm_l-li)}}{(\theta)_{(n)}}, \end{aligned} \tag{1.5}$$

respectively. For $\alpha \to 0$ the distributions (1.4) and (1.5) reduce to distributions originally obtained in Ewens [9], Watterson [32] and Watterson [33]. Indeed one has $\lim_{\alpha\to 0} \alpha^{-j}\mathscr{C}(n,j;\alpha) = |s(n,j)|$, where $|s(n,j)|$ denotes the signless Stirling number of the first type.

Conditional counterparts of (1.4) and (1.5) have been proposed in Lijoi et al. [22] and Favaro et al. [11]. See also Griffiths and Spanò [19], Lijoi et al. [23] and Bacallado et al. [4]. Specifically, let $(X_1,\ldots,X_n)$ be an initial sample from $\tilde{P}_{\alpha,\theta,\nu}$ and featuring $K_n = j$ blocks with frequencies $\mathbf{N}_n = \mathbf{n} = (n_1,\ldots,n_j)$, and let $(X_{n+1},\ldots,X_{n+m})$ be an additional unobserved sample. This is equivalent to say that $(X_{n+1},\ldots,X_{n+m})$ is a sample from a discrete random probability measure $\tilde{P}^{(n)}_{\alpha,\theta,\nu}$ whose distribution is the conditional distribution of $\tilde{P}_{\alpha,\theta,\nu}$ given $(X_1,\ldots,X_n)$. This conditional distribution is characterized in Corollary 20 of Pitman [26]. Specifically, if $(W_1,\ldots,W_{j+1})$ a random variable distributed according to a Dirichlet distribution with parameter $(n_1-\alpha,\ldots,n_j-\alpha,\theta+j\alpha)$, then

$$\tilde{P}^{(n)}_{\alpha,\theta,\nu} \overset{\mathrm{d}}{=} \sum_{i=1}^{j} W_i \delta_{X^*_i} + W_{j+1}\tilde{P}_{\alpha,\theta+j\alpha,\nu}$$

where $(W_1,\ldots,W_{j+1})$ is independent of $\tilde{P}_{\alpha,\theta+j\alpha,\nu}$. Proposition 1 in Lijoi et al. [22] provides the conditional distribution, given $(K_n,\mathbf{N}_n)$, of the number $K^{(n)}_m$ of new blocks in $(X_{n+1},\ldots,X_{n+m})$, whereas Proposition 5 and Proposition 6 in Favaro et al. [11] provide the conditional distribution, given $(K_n,\mathbf{N}_n)$, of the number $M^{(n)}_{l,m}$ of blocks with frequency $l \geq 1$ in $(X_1,\ldots,X_{n+m})$. These conditional distributions have found applications in Bayesian nonparametric inference for species sampling problems arising from ecology, bioinformatic, genetic, etc. Indeed from a Bayesian perspective (1.2) is a nonparametric model for the individuals X_i's of a population with infinite species X^*_i's, with Π being the prior distribution on the composition of such a population. Within this Bayesian framework $\mathbb{P}[K^{(n)}_m = k \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ and $\mathbb{P}[M^{(n)}_{l,m} = m_l \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ are the posterior distributions of the number of new species in $(X_{n+1},\ldots,X_{n+m})$ and the number of species with frequency l in the enlarged sample $(X_1,\ldots,X_{n+m})$, respectively, given $(X_1,\ldots,X_n)$. Hence $\mathbb{E}_{\alpha,\theta}[K^{(n)}_m \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ and $\mathbb{E}_{\alpha,\theta}[M^{(n)}_{l,m} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ are the corresponding Bayesian nonparametric estimators under a squared loss function.

For any $\alpha \in (0,1)$ and $q > -1$, let $S_{\alpha,q\alpha}$ be a positive random variable such that $\mathbb{P}[S_{\alpha,q\alpha} \in dy] = \alpha^{-1}\Gamma^{-1}(q+1)\Gamma(q\alpha+1)y^{q-1-1/\alpha}f_\alpha(y^{-1/\alpha})dy$ where f_α denotes the density function of the positive α-stable random variable and $\Gamma(\cdot)$ denotes the Gamma function. Specifically, $S^{-1/\alpha}_{\alpha,q\alpha}$ is the so-called polynomially tilted positive α-stable random variable. Pitman [27] and Pitman [29] established a large n fluctuation limit for K_n and $M_{l,n}$, namely

$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_{\alpha,\theta} \qquad \text{a.s.} \tag{1.6}$$

and

$$\lim_{n\to+\infty} \frac{M_{l,n}}{n^\alpha} = \frac{\alpha(1-\alpha)_{(l-1)}}{l!} S_{\alpha,\theta} \qquad \text{a.s.} \tag{1.7}$$

In contrast, for $\alpha \to 0$ Korwar and Hollander [21] and Arratia et al. [2] showed that $\lim_{n\to+\infty} K_n/\log n = \theta$ and $\lim_{n\to+\infty} M_{l,n} = P_{\theta/l}$ almost surely, where $P_{\theta/l}$ is a Poisson random variable with parameter θ/l. A weak convergence version of (1.6) and (1.7) can also be derived from asymptotic results for urn model with weighted balls. See Proposition 16 in Flajolet et al. [16] and Theorem 5 in Janson [20] for details. Let $X \,|\, Y$ be a random variable whose distribution coincides with the conditional distribution of X given Y. A conditional counterpart of (1.6) and (1.7) has been established in Favaro et al. [10]. Specifically, if $(K_n,\mathbf{N}_n)$ is the random partition induced by a sample of size n

from $\tilde{P}_{\alpha,\theta,\nu}$, then

$$\lim_{m\to+\infty} \frac{K_m^{(n)}}{m^\alpha} \,|\, (K_n = j, \mathbf{N}_n = \mathbf{n}) = B_{j+\theta/\alpha,n/\alpha-j} S_{\alpha,\theta+n} \qquad \text{a.s.} \tag{1.8}$$

and

$$\lim_{m\to+\infty} \frac{M_{l,m}^{(n)}}{m^\alpha} \,|\, (K_n = j, \mathbf{N}_n = \mathbf{n}) = \frac{\alpha(1-\alpha)_{(l-1)}}{l!} B_{j+\theta/\alpha,n/\alpha-j} S_{\alpha,\theta+n} \qquad \text{a.s.} \tag{1.9}$$

where $B_{j+\theta/\alpha,n/\alpha-j}$ is a random variable independent of $S_{\alpha,\theta+n}$ and distributed according to a Beta distribution with parameter $(j+\theta/\alpha, n/\alpha - j)$. Moreover, for $\alpha \to 0$ one has $\lim_{m\to+\infty} K_m^{(n)}/\log m = \theta$ and $\lim_{m\to+\infty} M_{l,m}^{(n)} = P_{\theta/l}$ almost surely. The reader is referred to Arratia et al. [3], Barbour and Gnedin [5], Schweinsberg [30] and Favaro and Feng [12] for recent generalizations and refinements of the fluctuation limits (1.6), (1.7), (1.9) and (1.8).

Feng and Hoppe [15] further investigated the large n asymptotic behaviour of the random variable K_n and, in particular, they established a large deviation principle for K_n. Interestingly the large deviation principle is characterized by a rate function depending only on the parameter α, which displays the different roles of the two parameters, $\alpha \in (0,1)$ and $\theta > -\alpha$, at different scales. Specifically, Feng and Hoppe [15] showed that $n^{-1}K_n$ satisfies a large deviation principle with speed n and rate function of the form

$$I^\alpha(x) = \begin{cases} \sup_\lambda\{\lambda x - \Lambda_\alpha(\lambda)\} & x \in [0,1] \\ \\ +\infty & \text{otherwise,} \end{cases} \tag{1.10}$$

where $\Lambda_\alpha(\lambda) = -\log(1-(1-e^{-\lambda})^{1/\alpha})\mathbb{1}_{(0,+\infty)}(\lambda)$. As $\alpha \to 0$ it was shown by Feng and Hoppe [15] that $(\log n)^{-1}K_n$ satisfies a large deviation principle with speed $\log n$ and rate function

$$I_\theta(x) = \begin{cases} x\log\frac{x}{\theta} - x + \theta & x > 0 \\ \\ \theta & x = 0 \\ \\ +\infty & x < 0. \end{cases} \tag{1.11}$$

As suggested by (1.6) and (1.8), one may expect that K_n and $K_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ have different asymptotic behaviours also in terms of large deviations, as n and m tend to infinity, respectively. Favaro and Feng [12] showed that for any fixed n and as m tends to infinity, $m^{-1}K_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ satisfies a large deviation principle with speed m and rate function (1.10), for $\alpha \in (0,1)$, and rate function (1.11), for $\alpha \to 0$. In other words, there is no long lasting impact of the given initial sample to the large deviation principle for $K_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$.

1.2 Main results and outline of the paper

Under the Ewens-Pitman sampling model we establish a large deviation principle for $M_{l,n}$, for any $l \geq 1$. Specifically, for any $\lambda > 0$ and $l \geq 1$ let us define $x = 1 - \mathrm{e}^{-\lambda}$ and $\tilde{x} = \alpha x(1-\alpha)_{(l-1)}/(1-x)l!$. Moreover, let $\epsilon_0(\lambda)$ be the unique solution of the following equation

$$(l-\alpha)\log(1-(l-\alpha)\epsilon) - l\log(1-l\epsilon) - \alpha\log\alpha\epsilon - \log\tilde{x} = 0,$$

and let

$$\Lambda_{\alpha,l}(\lambda) = \begin{cases} \log\left(1 + \frac{\alpha\epsilon_0(\lambda)}{1-l\epsilon_0(\lambda)}\right) & \text{if } \lambda > 0 \\ \\ 0 & \text{if } \lambda \leq 0. \end{cases}$$

We show that

$$-I_l^\alpha(A^\circ) \leq \liminf_{n\to+\infty} \frac{1}{n}\log \mathbb{P}\left[\frac{M_{l,n}}{n} \in A\right] \leq \limsup_{n\to+\infty} \frac{1}{n}\log \mathbb{P}\left[\frac{M_{l,n}}{n} \in A\right] \leq -I_l^\alpha(\bar{A})$$

where we set $I_l^\alpha(A) = \inf_{y\in A}\{I_l^\alpha(y)\}$ and $I_l^\alpha(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \Lambda_{\alpha,l}(\lambda)\}$ for any measurable set $A \subset \mathbb{R}$, and where A° and $\bar{A}$ denote the interior and the closure of A, respectively. In other words, we can state the following large deviation principle for $M_{l,n}$.

Theorem 1.1. *Under the Ewens-Pitman sampling model $n^{-1}M_{l,n}$ satisfies a large deviation principle with speed n and rate function I_l^α, for any $\alpha \in (0,1)$ and $\theta > -\alpha$. In particular,*

$$\lim_{n\to+\infty} \frac{1}{n}\log \mathbb{P}\left[\frac{M_{l,n}}{n} > x\right] = -I_l^\alpha(x)$$

for almost all $x > 0$, where $I_l^\alpha(0) = 0$, $I_l^\alpha(x) < +\infty$ for any $x \in (0, 1/l]$, and $I_l^\alpha(x) = +\infty$ for any $x \notin [0, 1/l]$.

Like the large deviation principle for K_n in Feng and Hoppe [15], the large deviation principle for $M_{l,n}$, for any $l \geq 1$, is characterized by a rate function depending only on α. In other terms, for any $\theta > -\alpha$ the rate function I_l^α coincides with the rate function of the special case $\alpha \in (0,1)$ and $\theta = 0$. An intuitive explanation for this comes from the representation of $\tilde{P}_{\alpha,\theta,\nu}$ through a collection of independent Beta random variables $(V_i)_{i\geq 1}$. Large deviations are determined mostly by the tail of $(V_i)_{i\geq 1}$ where α plays the dominant role. We derive an explicit expression for I_l^α under the assumption $\alpha = 1/2$ and $l = 1$. The speciality of the case $\alpha = 1/2$ is determined by a well-known interplay between $\tilde{P}_{1/2,0,\nu}$ and the Brownian motion: the distribution of the decreasing ordered random masses of $\tilde{P}_{1/2,0,\nu}$ coincides with the distribution of the decreasing ordered length of excursions of a Brownian motion, away from 0, in the time interval $[0,1]$. In particular Pitman [27] showed that the distribution of K_n can be derived from the zeros of a Brownian motion. For $\theta = 1/2$ a similar relationship follows with respect to the Brownian bridge. We refer to Pitman [29] for a detailed account on $\tilde{P}_{1/2,0,\nu}$ and $\tilde{P}_{1/2,1/2,\nu}$.

We also present a conditional counterpart of Theorem 1.1, in the same spirit as the fluctuation limit (1.9) represents a conditional counterpart of (1.7). In particular we establish a large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$, as $m \to +\infty$, where $(K_n, \mathbf{N}_n)$ is the random partition induced by a sample of size n from $\tilde{P}_{\alpha,\theta,\nu}$. We can state the following theorem.

Theorem 1.2. *Under the Ewens-Pitman sampling model $m^{-1}M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ satisfies a large deviation principle with speed m and rate function I_l^α, for any $\alpha \in (0,1)$ and $\theta > -\alpha$. In particular,*

$$\lim_{m\to+\infty} \frac{1}{m}\log \mathbb{P}\left[\frac{M_{l,m}^{(n)}}{m} > x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] = -I_l^\alpha(x)$$

for almost all $x > 0$, where $I_l^\alpha(0) = 0$, $I_l^\alpha(x) < +\infty$ for any $x \in (0, 1/l]$, and $I_l^\alpha(x) = +\infty$ for any $x \notin [0, 1/l]$.

Note that, similarly to the large deviation principle for $K_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ obtained in Favaro and Feng [12], the large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$, as $m \to +\infty$, coincides with the large deviation principle for $M_{l,n}$, as $n \to +\infty$. In other words, there is no long lasting impact of the given initial sample to the large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$.

A closer inspection of (1.8) and (1.9) reveals that for $l = 1$ the large deviation principle in Theorem 1.2 has a natural interpretation in the context of Bayesian nonparametric inference for discovery probabilities. Specifically, let $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ be the

conditional, or posterior, distribution of the probability of discovering a new species at the $(n+m+1)$-th draw, given the random partition $(K_n, \mathbf{N}_n)$ induced by $(X_1, \ldots, X_n)$. We show that $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n]$ and $\mathbb{P}[m^{-1}M_{1,m}^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ are approximately equal for large m. Accordingly Theorem 1.2 provides a large m approximation of $\mathbb{P}[D_m^{(n)} \geq x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$, which is a Bayesian nonparametric estimator of the decay of the discovery probability. Similarly, $\mathbb{E}_{\alpha,\theta}[m^{-1}M_{1,m}^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ provides a large m approximation of the Bayesian nonparametric estimator of the discovery probability, namely $\mathbb{E}_{\alpha,\theta}[D_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$. An illustration of these asymptotic estimators is presented by using a genomic dataset. The interest in the estimators $\mathbb{E}_{\alpha,\theta}[D_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ and $\mathbb{P}[D_m^{(n)} \geq x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$, as well as in their large m approximations, is related to the classical problem of determining the optimal sample size in species sampling problems. Indeed this problem is typically faced by setting a threshold τ for an exact or approximate mean functional of $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$, and then making inference on the sample size m for which this mean functional falls below, or above, τ. This procedure naturally introduces a criterion for evaluating the effectiveness of further sampling.

The paper is structured as follows. In Section 2 we present the proof of Theorem 1.1 and we derive an explicit expression for the rate function I_l^α under the assumption $\alpha = 1/2$ and $l = 1$. Section 3 contains the proof of Theorem 1.2. In Section 4 we discuss potential applications of Theorem 1.2 in the context of Bayesian nonparametric inference for species sampling problems and, in particular, in the context of discovery probabilities.

2 Large deviations for $M_{l,n}$

The large deviation principle for $M_{l,n}$ is established through a detailed study of the moment generating function of the random variable $M_{l,n}$. This is in line with the approach originally adopted in Feng and Hoppe [15] for K_n. For any $\lambda > 0$ let $y = 1 - e^{-\lambda}$ and

$$G_{M_{l,n}}(y; \alpha, \theta) = \mathbb{E}_{\alpha,\theta}\left[\left(\frac{1}{1-y}\right)^{M_{l,n}}\right] = \sum_{i \geq 0} \frac{y^i}{i!}\mathbb{E}_{\alpha,\theta}[(M_{l,n})_{(i)}] \tag{2.1}$$

be the moment generating function of the random variable $M_{l,n}$. Let $(y)_{[n]} = y(y-1)\cdots(y-n+1)$ be the falling factorial of y of order n, with the proviso $(y)_{[0]} = 1$. Proposition 1 in Favaro et al. [11] provides an explicit expression for $\mathbb{E}_{\alpha,\theta}[(M_{l,n})_{[r]}]$. Recalling that $(y)_{(n)} = \sum_{0 \leq i \leq n} \sum_{0 \leq j \leq i} |s(n,i)| S(i,j) (y)_{[j]}$, where s and S denote the Stirling number of the first type and the second type, an explicit expression for $\mathbb{E}_{\alpha,\theta}[(M_{l,n})_{(r)}]$ is obtained. Specifically,

$$\mathbb{E}_{\alpha,\theta}[(M_{l,n})_{(r)}] = r! \sum_{i=0}^{r} \binom{r-1}{r-i} \frac{\left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i \left(\frac{\theta}{\alpha}\right)_{(i)} (n)_{[il]} (\theta + i\alpha)_{(n-il)}}{i! (\theta)_{(n)}} \tag{2.2}$$

and

$$\mathbb{E}_{\alpha,0}[(M_{l,n})_{(r)}] = (r-1)! \sum_{i=0}^{r} \binom{r}{i} \frac{\left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i (n)_{[il]} (i\alpha)_{(n-il)}}{\alpha \Gamma(n)}, \tag{2.3}$$

where the sums over i is nonnull for $0 \leq i \leq \min(r, \lfloor n/l \rfloor)$. The next lemma provides an explicit expression for $G_{M_{l,n}}(y; \alpha, 0)$. This result follows by combining (2.3) with the series expansion on the right-hand side of (2.1), and by means of standard combinatorial manipulations.

Lemma 2.1. *For any* $\alpha \in (0,1)$

$$G_{M_{l,n}}(y; \alpha, 0) \tag{2.4}$$

$$= \sum_{i=0}^{\lfloor n/l \rfloor} \left(\frac{y}{1-y}\right)^i \left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{n}{n-il} \binom{n-il+i\alpha-1}{n-il-1}.$$

Proof. The proof is obtained by simply combining the right-hand side of (2.1), with $\theta = 0$, with the rising factorial moment displayed in (2.3). This leads to write $G_{M_{l,n}}(y;\alpha,0)$ as follows

$$\begin{aligned} &G_{M_{l,n}}(y;\alpha,0) \\ &= \sum_{t\geq 0} \frac{y^t}{t!} \sum_{i=0}^{t} \binom{t}{i} (t-1)! \left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(n)_{[il]}(i\alpha)_{(n-il)}}{\alpha\Gamma(n)} \\ &= \sum_{i\geq 0} \frac{1}{i} \left(\frac{y}{1-y}\right)^i \left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(n)_{[il]}(i\alpha)_{(n-il)}}{\alpha\Gamma(n)}, \end{aligned}$$

where the last equality is obtained by interchanging the summations and by means of $\sum_{t\geq i} \binom{t}{i} y^t (t-1)!/t! = i^{-1}(y/(1-y))^i$. Since $(n)_{[il]} = 0$ for $i > \lfloor n/l \rfloor$ then we can write the last expression as

$$\begin{aligned} &G_{M_{l,n}}(y;\alpha,0) \\ &= \sum_{i=0}^{\lfloor n/l \rfloor} \frac{1}{i} \left(\frac{y}{1-y}\right)^i \left(\alpha \frac{(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(n)_{[il]}(i\alpha)_{(n-il)}}{\alpha\Gamma(n)}. \end{aligned}$$

The proof is completed by rearranging the rising and the falling factorial moments by means of the well-known identities $(i\alpha)_{(n-il)} = \Gamma(n-il+i\alpha)/\Gamma(i\alpha) = (n-il+i\alpha-1)!/(i\alpha-1)!$. □

We exploit (2.4) and (2.2) in order to establish the large deviation principle for $M_{l,n}$ stated in Theorem 1.1. The proof of this theorem is split into three main parts. The first two parts deal with the large deviation principle for $M_{l,n}$ under the assumption $\alpha \in (0,1)$ and $\theta = 0$, whereas the third part deals with the general case $\alpha \in (0,1)$ and $\theta > -\alpha$.

Proof of Theorem 1.1. In the first part of the proof we show that, assuming $\alpha \in (0,1)$ and $\theta = 0$, $n^{-1}M_{l,n}$ satisfies a large deviation principle with speed n and rate function I_l^α. By exploiting the rising factorial moment (2.3) we can write the large n approximation

$$\mathbb{E}_{\alpha,0}[M_{l,n}] = \frac{\alpha(1-\alpha)_{(l-1)}}{\alpha\Gamma(n)l!}(n)_{[l]}(\alpha)_{(n-l)} \approx n^\alpha$$

and

$$G_{M_{l,n}}(y;\alpha,0) = \sum_{i=1}^{\lfloor n/l \rfloor} \tilde{y}^i \frac{n}{n-il} \binom{n-il+\alpha i-1}{n-il-1}$$

where we defined $\tilde{y} = \alpha y(1-\alpha)_{(l-1)}/(1-y)l!$. Note that if n/l is an integer, then the final term in the above expression corresponds to $n\tilde{y}^{n/l}$. By direct calculation we can prove that $\lim_{n\to+\infty} n^{-1} \log \mathbb{E}_{\alpha,0}[e^{\lambda M_{l,n}}] = 0$ for any $\lambda \leq 0$. Furthermore, for any $\lambda > 0$ and $y = 1 - e^{-\lambda}$,

$$\begin{aligned} &\lim_{n\to+\infty} \frac{1}{n} \log G_{M_{l,n}}(y;\alpha,0) \\ &= \lim_{n\to+\infty} \frac{1}{n} \log \max \left\{ \tilde{y}^i \binom{n-il+\alpha i-1}{n-il-1}; i = 0,\ldots,\frac{n}{l} \right\} \\ &= \lim_{n\to+\infty} \max \left\{ \frac{1}{n} \log \tilde{y}^i \binom{n-il+\alpha i-1}{n-il-1}; i = 0,\ldots,\frac{n}{l} \right\}. \end{aligned}$$

For $\alpha i < 1$, it is clear that one has $\lim_{n\to+\infty} n^{-1}\log\tilde{y}^i\binom{n-(l-\alpha)i-1}{n-li-1} = 0$. For any i satisfying $0 \le n - il < 1$, there are two possibilities: either $n = il$ or $i = \lfloor n/l \rfloor < n/l$. In both cases one has

$$\lim_{n\to+\infty}\frac{1}{n}\log\tilde{y}^i\binom{n-(l-\alpha)i-1}{n-li-1} = \frac{1}{l}\log\tilde{y}.$$

Next we consider the case for which i satisfies $n - li \ge 1$ and $\alpha i \ge 1$. For $0 < \epsilon < 1/l$, set $\phi(\epsilon) = \epsilon\log\epsilon$ and $\varphi(\epsilon) = \phi(1-(l-\alpha)\epsilon) - \phi(1-l\epsilon) - \phi(\alpha\epsilon) + \epsilon\log\tilde{y}$. By means of the representation $\Gamma(z) = \sqrt{2\pi}z^{z-1/2}e^{-z}[1+r(z)]$, where we set $|r(z)| \le e^{1/12z} - 1$ for any $z > 0$, we can write

$$\begin{aligned}
&\binom{n-il+\alpha i-1}{n-il-1}\\
&\quad= \frac{\Gamma(n-(l-\alpha)i)}{\alpha i\Gamma(n-li)\Gamma(\alpha i)}\\
&\quad= \frac{(1+r(n-(l-\alpha)i))e}{\sqrt{2\pi}(1+r(n-li))(1+r(\alpha i))}\left(\frac{(n-li)}{\alpha i(n-(l-\alpha)i)}\right)^{1/2}\\
&\qquad\times\frac{(1-(l-\alpha)i/n)^{n-(l-\alpha)i}}{(1-li/n)^{n-li}(\alpha i/n)^{\alpha i}}\\
&\quad= \frac{(1+r(n-(l-\alpha)i))e}{\sqrt{2\pi}(1+r(n-li))(1+r(\alpha i))}\left(\frac{(n-li)(\alpha i+1)}{(n-(l-\alpha)i)}\right)^{1/2}\\
&\qquad\times\exp\left\{n\left[\phi\left(1-(l-\alpha)\frac{i}{n}\right)-\phi\left(1-l\frac{i}{n}\right)-\phi\left(\alpha\frac{i}{n}\right)\right]\right\}.
\end{aligned}$$

The fact that $\alpha^{-1} \le i \le (n-1)/l$ implies that $(1+r(n-(l-\alpha)i))e/\sqrt{2\pi}(1+r(n-li))(1+r(\alpha i))$ is uniformly bounded from above by some constant $d_1 > 0$. Accordingly, we can write

$$\begin{aligned}
&\binom{n-il+\alpha i-1}{n-il-1}\\
&\quad\le d_1\sqrt{n}\exp\left\{n\left[\phi\left(1-(l-\alpha)\frac{i}{n}\right)-\phi\left(1-l\frac{i}{n}\right)-\phi\left(\alpha\frac{i}{n}\right)\right]\right\},
\end{aligned}\tag{2.5}$$

and

$$\begin{aligned}
&\lim_{n\to+\infty}\max\left\{\frac{1}{n}\log\tilde{y}^i\binom{n-il+\alpha i-1}{n-il-1};\ \frac{1}{\alpha}\le i\le\frac{(n-1)}{l}\right\}\\
&\quad\le\max\left\{\varphi(\epsilon):0<\epsilon<\frac{1}{l}\right\}.
\end{aligned}\tag{2.6}$$

In particular, by combining the inequalities stated in (2.5) and (2.6), respectively, we have

$$\begin{aligned}
&\lim_{n\to+\infty}\frac{1}{n}\log G_{M_{l,n}}(y;\alpha,0)\\
&\quad=\lim_{n\to+\infty}\frac{1}{n}\log\max\left\{\tilde{y}^i\binom{n-il+\alpha i-1}{n-il-1};\ i=0,\ldots,\frac{n}{l}\right\}\\
&\quad=\max\left\{\max\left\{\lim_{n\to+\infty}\frac{1}{n}\log\tilde{y}^i\binom{n-il+\alpha i-1}{n-il-1};\ i<\frac{1}{\alpha}\right\},\right.\\
&\qquad\left.\max\left\{\lim_{n\to+\infty}\frac{1}{n}\log\tilde{y}^i\binom{n-il+\alpha i-1}{n-il-1};\ \frac{1}{\alpha}\le i\le\frac{n-1}{l}\right\},\ \frac{1}{l}\log\tilde{y}\right\}\\
&\quad=\max\left\{0,\max\left\{\varphi(\epsilon):0<\epsilon<\frac{1}{l}\right\},\ \frac{1}{l}\log\tilde{y}\right\}
\end{aligned}$$

$$\leq \max\left\{\varphi(\epsilon) : 0 < \epsilon < \frac{1}{l}\right\}.$$

On the other hand, for any ϵ in $(0, 1/l)$, there exists a sequence $(i_n)_{n\geq 1}$ such that $(i_n/n)_{n\geq 1}$ converges to ϵ as n tends to infinity. For this particular sequence we can write

$$\begin{aligned}\varphi(\epsilon) &= \lim_{n\to+\infty} \frac{1}{n} \log \tilde{y}^{i_n} \binom{n - i_n l + \alpha i_n - 1}{n - i_n l - 1} \\ &\leq \lim_{n\to+\infty} \frac{1}{n} \log G_{M_{l,n}}(y; \alpha, 0).\end{aligned}$$

Thus

$$\lim_{n\to+\infty} \frac{1}{n} \log G_{M_{l,n}}(y; 0, \alpha) = \max\left\{\varphi(\epsilon) : 0 \leq \epsilon \leq \frac{1}{l}\right\}.$$

Noting that

$$\varphi'(\epsilon) = -(l-\alpha)\log(1-(l-\alpha)\epsilon) + l\log(1-l\epsilon) - \alpha\log\alpha\epsilon + \log\tilde{y}, \tag{2.7}$$

one has

$$\varphi(\epsilon) = \log(1-(l-\alpha)\epsilon) - \log(1-l\epsilon) + \varphi'(\epsilon)\epsilon. \tag{2.8}$$

Since $\varphi'(0+) = +\infty$ and $\varphi'(1/l-) = -\infty$, then the function $\varphi(\epsilon)$ reaches a maximum at a point ϵ_0 in the set $(0, 1/l)$ where $\varphi'(\epsilon_0) = 0$. Clearly ϵ_0 depends on the parameter α, and it also depends on l and λ. Moreover note that $\varphi''(\epsilon) = -\alpha/\epsilon(1-(l-\alpha)\epsilon)(1-l\epsilon) < 0$, which implies that $\epsilon_0(\lambda)$ is unique and that $\Lambda_{\alpha,l}(\lambda) = \log[1 + \alpha\epsilon_0/(1-l\epsilon_0)]$. In particular, since

$$\log\tilde{y} = \lambda + \log\frac{e^\lambda - 1}{e^\lambda} + \log\frac{\alpha(1-\alpha)_{(l-1)}}{l!}$$

and $\varphi'(\epsilon_0) = -(l-\alpha)\log(1-(l-\alpha)\epsilon_0) + l\log(1-l\epsilon_0) - \alpha\log\alpha\epsilon_0 + \log\tilde{y} = 0$, then one obtains

$$\begin{aligned}&\lambda + \log\frac{e^\lambda - 1}{e^\lambda} + \log\frac{\alpha(1-\alpha)_{(l-1)}}{\alpha^\alpha l!} \\ &\quad = l\log\frac{1-(l-\alpha)\epsilon_0}{1-l\epsilon_0} + \alpha\log\frac{\epsilon_0}{1-(l-\alpha)\epsilon_0}.\end{aligned}$$

Set

$$h_1(\lambda) = \lambda + \log\frac{e^\lambda - 1}{e^\lambda} + \log\frac{\alpha(1-\alpha)_{(l-1)}}{\alpha^\alpha l!} \tag{2.9}$$

and

$$h_2(\epsilon_0) = l\log\frac{1-(l-\alpha)\epsilon_0}{1-l\epsilon_0} + \alpha\log\frac{\epsilon_0}{1-(l-\alpha)\epsilon_0}. \tag{2.10}$$

Since h_1 and h_2 are strictly increasing functions with differentiable inverses, then $\epsilon_0 = h_2^{-1} \circ h_1(\lambda)$ is a differentiable strictly increasing function and, in particular, we have $\lim_{\lambda\to 0}\epsilon_0 = 0$ and $\lim_{\lambda\to+\infty}\epsilon_0 = 1/l$. Now, if we set $\Lambda_{\alpha,l}(\lambda)$ to be zero for nonpositive λ, and for $\lambda > 0$

$$\Lambda_{\alpha,l}(\lambda) = \log\left(1 + \frac{\alpha h_2^{-1} \circ h_1(\lambda)}{1 - l h_2^{-1} \circ h_1(\lambda)}\right), \tag{2.11}$$

then it is clear that $\{\lambda : \Lambda_{\alpha,l}(\lambda) < +\infty\} = \mathbb{R}$ and $\Lambda_{\alpha,l}(\lambda)$ is differentiable for $\lambda \neq 0$. The left derivative of $\Lambda_{\alpha,l}(\lambda)$ at zero is clearly zero. On the other hand, for any $\lambda > 0$ we can write

$$\frac{d\Lambda_{\alpha,l}(\lambda)}{d\lambda} = \left[\frac{\alpha - l}{1 + (\alpha - l)\epsilon_0} + \frac{l}{1 - l\epsilon_0}\right]\frac{d\epsilon_0}{d\lambda}.$$

Since ϵ_0 converges to zero it follows from direct calculation that, as $\lambda \downarrow 0$ one obtains the following

$$\frac{d\epsilon_0}{d\lambda} = \frac{(e^{h_1(\lambda)})'}{(e^{h_2(\epsilon)})'|_{\epsilon=\epsilon_0}} \to 0.$$

Hence $\Lambda_{\alpha,l}(\lambda)$ is differentiable everywhere. By the Gärtner-Ellis theorem, see Dembo and Zeitouni [7] for details, a large deviation principle holds for $n^{-1}M_{l,n}$ on space $\mathbb{R}$ as n tends to infinity with speed n and good rate function $I_l^\alpha(x) = \sup_\lambda\{\lambda x - \Lambda_{\alpha,l}(\lambda)\}$. This completes the first part of the proof. In the second part of the proof we further specify the rate function I_l^α. In particular, let us rewrite $\Lambda_{\alpha,l}(\lambda)$ as $\Lambda_{\alpha,l}(\lambda) = \lambda/l + \tilde{\Lambda}_{\alpha,l}(\lambda)$, where we defined

$$\tilde{\Lambda}_{\alpha,l}(\lambda) = -\lambda/l,$$

for $\lambda \leq 0$, and

$$\tilde{\Lambda}_{\alpha,l}(\lambda) = \frac{1}{l}\log\frac{e^\lambda - 1}{e^\lambda} + \frac{1}{l}\log\frac{\alpha(1-\alpha)_{(l-1)}}{\alpha^\alpha l!} - \frac{\alpha}{l}\log\frac{\epsilon_0}{1-(l-\alpha)\epsilon_0}$$

for any $\lambda > 0$. We observe that, since there exists a strictly positive constant $d_2 > 0$ such that $\epsilon_0 \geq d_2$ for $\lambda \geq 1$, then the function $\tilde{\Lambda}_{\alpha,l}(\lambda)$ is uniformly bounded for $\lambda \geq 1$. This implies that the rate function $I_l^\alpha(x) = \sup_\lambda\{\lambda\,(x - 1/l) - \tilde{\Lambda}_{\alpha,l}(\lambda)\}$ is infinity for any point $x > 1/l$, which is consistent with the fact that $n^{-1}M_{l,n} \leq 1/l$. Additionally we have that

$$I_l^\alpha(x) = \begin{cases} 0 & \text{if } x = 0 \\ < +\infty & \text{if } x \in (0, 1/l] \\ +\infty & \text{otherwise .} \end{cases} \tag{2.12}$$

For this to hold, we need to verify that the rate function $I_l^\alpha(x)$ is finite for x in $(0, 1/l]$. By definition,

$$\sup_{0\leq\lambda\leq 1}\{\lambda x - \Lambda(\lambda)\} \leq \sup_{0\leq\lambda\leq 1}\{\lambda x\} = x < +\infty \tag{2.13}$$

for any x in $(0, 1/l]$. For any $\lambda \geq 1$, let d_2 be the value of ϵ_0 at $\lambda = 1$. Then $\epsilon_0 \geq d_2$ for any $\lambda \geq 1$ and this implies that $\tilde{\Lambda}(\lambda)$ is bounded for all $\lambda \geq 1$. Accordingly, we can write $\sup_{\lambda\geq 1}\{\lambda\,(y - 1/l) - \tilde{\Lambda}_{\alpha,l}(\lambda)\} \leq \sup_{\lambda\geq 1}\{|\tilde{\Lambda}_{\alpha,l}(\lambda)|\} < +\infty$, which combined with (2.13) implies (2.12). This completes the second part of the proof. Finally, in the third part of the proof we extend the large deviation principle to the case $\alpha \in (0, 1)$ and $\theta > -\alpha$. By combining the definition (2.1) with (2.2), and by means of combinatorial manipulations, one has

$$\begin{aligned} &G_{M_{l,n}}(y;\alpha,\theta) \\ &\quad = \sum_{i=0}^{\lfloor n/l\rfloor} D(\alpha,\theta,n,i)\left(y\alpha\frac{(1-\alpha)_{(1-y)(l-1)}}{l!}\right)^i \frac{n}{(n-il)}\binom{n-il+i\alpha-1}{n-il-1}, \end{aligned}$$

where the function $D(\alpha,\theta,n,i)$ is such that $D(\alpha,\theta,n,0) = 1$ and, in general, for any $1 \leq i \leq \lfloor n/l\rfloor$,

$$D(\alpha,\theta,n,i) = \frac{\Gamma(n)}{(\theta+1)_{(n-1)}}\frac{(\theta/\alpha+1)_{(i-1)}}{\Gamma(i)}\frac{(\theta+i\alpha)_{(n-il)}}{(i\alpha)_{(n-il)}}.$$

Note that, since $\theta/\alpha > -1$, it follows from standard basic algebra that one can find positive constants, say d_3 and d_4, that are independent of n and of i, and such that it follows

$$d_3 n^{-2} \leq D(\alpha,\theta,n,i) \leq d_4 n^k \tag{2.14}$$

where k is the smallest integer greater than $1+|\theta|+|\theta/\alpha|$. Accordingly, we have the identities

$$\begin{aligned}
&\lim_{n\to+\infty}\frac{1}{n}\log\mathbb{E}_{\alpha,\theta}[e^{\lambda M_{l,n}}]\\
&\quad=\lim_{n\to+\infty}\frac{1}{n}\log G_{M_{l,n}}(y;\alpha,\theta)\\
&\quad=\lim_{n\to+\infty}\frac{1}{n}\log G_{M_{l,n}}(y;\alpha,0)\\
&\quad=\lim_{n\to+\infty}\frac{1}{n}\log\mathbb{E}_{\alpha,0}[e^{\lambda M_{l,n}}]=\Lambda_{\alpha,l}(\lambda).
\end{aligned}$$

Then, for any $\alpha\in(0,1)$ and $\theta>-\alpha$, $n^{-1}M_{l,n}$ satisfies a large deviation principle with speed n and with rate function I_l^α. This completes the third part of the proof and the proof. □

In general it is difficult to get a more explicit expression for I_l^α. Indeed, $\Lambda_{\alpha,l}$ depends on λ in an implicit form, namely $\Lambda_{\alpha,l}$ is a function of λ in terms of $h_2^{-1}\circ h_1(\lambda)$, where h_1 and h_2 are in (2.9) and (2.10) respectively. However, under the assumption $\alpha=1/2$ and $l=1$, an explicit expression for I_l^α can be derived. For general $\alpha\in(0,1)$ and $\theta>-\alpha$, the rate function I_l^α displayed in (2.11) can be easily evaluated by means of numerical techniques.

Proposition 2.2. *For any* $x\in[0,1]$

$$I_1^{1/2}(x)=x\log[B_1(x)+1]+\log 2-\log\left(1+\sqrt{B_1^2(x)+1}\right),$$

where

$$B_1(x)=2\sqrt{\frac{-p}{3}}\cos\left(\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right)\right)-\frac{2}{3(1-x)}.$$

Proof. Let us consider the equation (2.7). Under the assumption $\alpha=1/2$ and $l=1$, the equation $-(l-\alpha)\log(1-(l-\alpha)\epsilon_0)+l\log(1-l\epsilon_0)-\alpha\log\alpha\epsilon_0+\log\tilde{y}=0$ becomes of the form

$$-\frac{1}{2}\log\left(1-\frac{\epsilon_0}{2}\right)+\log(1-\epsilon_0)-\frac{1}{2}\log\epsilon_0+\frac{1}{2}\log 2+\log(e^\lambda-1)-\log 2=0.$$

Equivalently we have $(e^\lambda-1)^2=(2-\epsilon_0)\epsilon_0/(1-\epsilon_0)^2$. By solving the equation we obtain $\epsilon_0=1-1/\sqrt{B^2+1}$ with $B=e^\lambda-1$. Going back to the rate function, we have the following identities

$$\begin{aligned}
I_1^{1/2}(x)&=\sup_{\lambda>0}\left\{\lambda x-\log\frac{1-\epsilon_0/2}{1-\epsilon_0}\right\}\\
&=\sup_{\lambda>0}\left\{\lambda x-\log\frac{2-\epsilon_0}{1-\epsilon_0}\right\}+\log 2\\
&=\sup_{\lambda>0}\left\{\lambda x-\log(1+\sqrt{B^2+1})\right\}+\log 2.
\end{aligned}$$

It is known that $I_1^{1/2}(0)=0$. Moreover, for $x=1$, we have the following expression for the rate function

$$\begin{aligned}
&\sup_{\lambda>0}\left\{\lambda-\log(1+\sqrt{B^2+1})\right\}\\
&\quad=\sup_{\lambda>0}\left\{\log\frac{B+1}{1+\sqrt{B^2+1}}\right\}
\end{aligned}$$

$$= \lim_{\lambda \to +\infty} \log \frac{B+1}{1+\sqrt{B^2+1}} = 0,$$

which implies that $I_1^{1/2}(1) = \log 2$. In general, for $0 < x < 1$, set $h(\lambda) = \lambda x - \log(1 + \sqrt{B^2+1})$. Then $h'(\lambda) = x - B(B+1)/(B^2+1+\sqrt{B^2+1})$ the solution of the equation $h'(\lambda) = 0$ satisfies

$$(1-x)^2 B^3 + 2(1-x)B^2 + (1-x)^2 B - 2x = 0, \tag{2.15}$$

and

$$\begin{aligned}\Delta &= 64x(1-x)^3 + 4(1-x)^6 - 36x(1-x)^5 - 4(1-x)^8 - 108x^2(1-x)^4 \\ &= 4(1-x)^6[1-(1-x)^2] + (1-x)^3 x[64 - 36(1-x)^2 - 108x(1-x)] \\ &\geq 4(1-x)^6[1-(1-x)^2] + (1-x)^3 x[64 - 36 - 27] > 0\end{aligned}$$

is the discriminant. Let $G(B)$ denote the left-hand side of the equation displayed in (2.15). By a direct calculation it follows that $G'(B) = 0$ has two negative roots. This, combined with the fact that $G(0) = -2x < 0$, implies that one and only one of the three roots of (2.15) is positive. Denote this root by $B_1(x)$. Then the rate function coincides with

$$I_1^{1/2}(x) = x \log[B_1(x) + 1] + \log 2 - \log\left(1 + \sqrt{B_1^2(x) + 1}\right). \tag{2.16}$$

By means of a change of variable in the equation (2.15), such that $C = B + 2/(3(1-x))$ we obtain the following depressed form of the equation $C^3 + pC + q = 0$ where we defined

$$p = 1 - \frac{4}{3(1-x)^2} < 0$$

and

$$q = \frac{16 - 18(1-x)^2 - 54x(1-x)}{27(1-x)^3}.$$

Then we can write

$$B_1(x) = 2\sqrt{\frac{-p}{3}} \cos\left(\frac{1}{3} \arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right)\right) - \frac{2}{3(1-x)} \tag{2.17}$$

follows by a direct application of the Viéte's trigonometric formula. The proof is completed by simply combining the rate function displayed in (2.16) with the function B_1 in (2.17). □

To some extent Theorem 1.1 provides a generalization of the large deviation principle for K_n established in Theorem 1.2 of Feng and Hoppe [15]. Indeed, recall that one has the following relations between K_n and $M_{l,n}$: $K_n = \sum_{1 \leq i \leq n} M_{l,n}$ and $n = \sum_{1 \leq i \leq n} l M_{l,n}$. So far it is not clear to us how to relate the large deviation principle for $M_{l,n}$, for any $l \geq 1$, with the large deviation principle for K_n. In this respect we retain that results introduced in Dinwoodie and Zabell [8] may be helpful in understanding such a relationship.

3 Large deviations for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$

Let $(X_1, \ldots, X_n)$ be an initial sample from $\tilde{P}_{\alpha,\theta,\nu}$ and let $(X_{n+1}, \ldots, X_{n+m})$ be an additional sample, for any $m \geq 1$. Furthermore, let $X_1^*, \ldots, X_{K_n}^*$ be the labels identifying the K_n blocks generated by $(X_1, \ldots, X_n)$ with corresponding frequencies $\mathbf{N}_n$, and let $L_m^{(n)} = \sum_{1 \leq i \leq m} \prod_{1 \leq k \leq K_n} \mathbb{1}_{\{X_k^*\}^c}(X_{n+i})$ be the number of elements in the additional sample that do not coincide with elements in the initial sample. If we denote by $K_m^{(n)}$ the

number of new blocks generated by these $L_m^{(n)}$ elements and by $X^*_{K_n+1},\dots,X^*_{K_n+K_m^{(n)}}$ their labels, then

$$S_i = \sum_{l=1}^{m} \mathbb{1}_{\{X^*_{K_n+i}\}}(X_{n+l}), \tag{3.1}$$

for $i = 1,\dots,K_m^{(n)}$, are the frequencies of the $K_m^{(n)}$ blocks. Finally, note that the frequencies of the blocks generated by the remaining $m - L_m^{(n)}$ elements of the additional sample are

$$R_i = \sum_{l=1}^{m} \mathbb{1}_{\{X_i^*\}}(X_{n+l}), \tag{3.2}$$

for $i = 1,\dots,K_n$. The blocks generated by the $m - L_m^{(n)}$ elements of the additional sample are termed “old" to be distinguished from the $K_m^{(n)}$ new blocks generated by the $L_m^{(n)}$ elements of the additional sample. The random variables (3.1) and (3.2), together with $L_m^{(n)}$ and $K_m^{(n)}$, completely describe the conditional random partition induced by $(X_{n+1},\dots,X_{n+m})$ given $(X_1,\dots,X_n)$. See Lijoi et al. [23] and Favaro et al. [11] for a comprehensive study on the conditional distributions of these random variables given the initial sample.

The random variables (3.1) and (3.2) lead to define the number $M_{l,m}^{(n)}$ of blocks with frequency l in $(X_1,\dots,X_{n+m})$. This is the number of new blocks with frequency l generated by $(X_{n+1},\dots,X_{n+m})$ plus the number of old blocks with frequency l that arise by updating, via $(X_{n+1},\dots,X_{n+m})$, the frequencies already induced by $(X_1,\dots,X_n)$. Specifically, let

$$N_{l,m}^{(n)} = \sum_{i=1}^{K_m^{(n)}} \mathbb{1}_{\{S_i=l\}}$$

be the number of new blocks with frequency l. Specifically, these new blocks are generated by the $L_m^{(n)}$ elements of the additional sample $(X_{n+1},\dots,X_{n+m})$. Furthermore, let

$$O_{l,m}^{(n)} = \sum_{i=1}^{K_n} \mathbb{1}_{\{N_i+R_i=l\}}$$

be the number of old blocks with frequency l. These old blocks are generated by updating, via the $m - L_m^{(n)}$ elements of the additional sample, the frequencies of random partition induced by the initial sample. Therefore, $M_{l,m}^{(n)} = O_{l,m}^{(n)} + N_{l,m}^{(n)}$. The conditional distribution of $M_{l,m}^{(n)}$, given the initial sample, has been recently derived and investigated in Favaro et al. [11].

The study of large deviations for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ reduces to the study of large deviations for the conditional number of new blocks with frequency l, namely $N_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$. Indeed $N_{l,m}^{(n)} \leq M_{l,m}^{(n)} \leq N_{l,m}^{(n)} + n$. Hence, by means of a direct application of Corollary B.9 in Feng [14], $m^{-1}M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ and $m^{-1}N_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ satisfy the same large deviation principle. As in Theorem 1.1, this large deviation principle is established through the study of the moment generating function of $N_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$. For any $\lambda > 0$ and $y = 1 - \mathrm{e}^{-\lambda}$, let

$$\begin{aligned} &G_{N_{l,m}^{(n)}}(y;\alpha,\theta) \\ &\quad = \mathbb{E}_{\alpha,\theta}\left[\left(\frac{1}{1-y}\right)^{N_{l,m}^{(n)}} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] \end{aligned} \tag{3.3}$$

$$= \sum_{i\geq 0} \frac{y^i}{i!} \mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(i)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}].$$

Theorem 1 in Favaro et al. [11] provides an expression for the falling factorial moment $\mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{[r]} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$. By exploiting the relation between falling factorials and rising factorials in terms of signless Stirling numbers of the first type and Stirling numbers of the second type, an explicit expression for $\mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ is obtained. Specifically,

$$\begin{aligned} &\mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &= r! \sum_{i=0}^{r} \binom{r-1}{r-i} \frac{\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i \left(\frac{\theta}{\alpha}\right)_{(j+i)} (m)_{[il]} (\theta + i\alpha + n)_{(m-il)}}{i!(\theta+n)_{(m)}(\theta/\alpha)_{(j)}} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} &\mathbb{E}_{\alpha,0}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &= j(r-1)! \sum_{i=0}^{r} \binom{r}{i} \binom{j+i-1}{i-1} \frac{\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i (m)_{[il]} (i\alpha + n)_{(m-il)}}{(n)_{(m)}} \end{aligned} \tag{3.5}$$

where the sums over i is nonnull for $0 \leq i \leq \min(r, \lfloor m/l \rfloor)$. Note that $\mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] = \mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j]$. In other words the number K_n of blocks in the initial sample is a sufficient statistics for $\mathbb{E}_{\alpha,\theta}[(N^{(n)}_{l,m})_{(r)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$. This property of sufficiency was pointed out in Favaro et al. [11]. Along lines similar to Lemma 2.1, in the next lemma we provide an explicit expression for the moment generating function $G_{N^{(n)}_{l,m}}(y; \alpha, 0)$.

Lemma 3.1. *For any* $\alpha \in (0,1)$

$$\begin{aligned} &G_{N^{(n)}_{l,m}}(y; \alpha, 0) \\ &= \frac{m!}{(n)_{(m)}} \sum_{i=0}^{\lfloor m/l \rfloor} \left(\frac{y}{1-y}\right)^i \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i \\ &\quad \times \binom{j+i-1}{i} \frac{(i\alpha+n)}{(m-il)} \binom{n+m+i\alpha-il-1}{m-il-1}. \end{aligned} \tag{3.6}$$

Proof. The proof is obtained by simply combining the right-hand side of (3.3), with $\theta = 0$, with the rising factorial moment displayed in (3.5). This leads to write $G_{N^{(n)}_{l,m}}(y; \alpha, 0)$ as follows

$$\begin{aligned} &G_{N^{(n)}_{l,m}}(y; \alpha, 0) \\ &= j \sum_{t\geq 0} \frac{y^t}{t!} (t-1)! \sum_{i=0}^{t} \binom{t}{i} \binom{j+i-1}{i-1} \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(m)_{[il]}(i\alpha+n)_{(m-il)}}{(n)_{(m)}} \\ &= \sum_{i\geq 0} \binom{j+i-1}{i} \left(\frac{y}{1-y}\right)^i \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(m)_{[il]}(i\alpha+n)_{(m-il)}}{(n)_{(m)}}, \end{aligned}$$

where the last equality is obtained by interchanging the summations and by means of $\sum_{t\geq i} \binom{t}{i} y^t (t-1)!/t! = i^{-1}(y/(1-y))^i$. Since $(m)_{[il]} = 0$ for $i > \lfloor m/l \rfloor$ then we can write the last expression as

$$G_{N^{(n)}_{l,m}}(y; \alpha, 0)$$

$$= \sum_{i=0}^{\lfloor m/l \rfloor} \binom{j+i-1}{i} \left(\frac{y}{1-y}\right)^i \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i \frac{(m)_{[il]}(i\alpha+n)_{(m-il)}}{(n)_{(m)}}.$$

The proof is completed by rearranging the rising and the falling factorial moments by means of the identities $(i\alpha+n)_{(m-il)} = \Gamma(n+m-il+i\alpha)/\Gamma(i\alpha+n) = (n+m-il+i\alpha-1)!/(i\alpha+n-1)!$. □

We exploit the moment generating function (3.6) and rising factorial moment (2.2) in order to establish the large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ stated in Theorem 1.2. We also make use of the unconditional large deviation principle established in Theorem 1.1.

Proof of Theorem 1.2. As we anticipated, in order to prove the theorem, it is sufficient to prove the large deviation principle for $N_{l,m}^{(n)} \,|\, K_n$, for any $\alpha \in (0,1)$ and $\theta > -\alpha$. We start with the assumption $\alpha \in (0,1)$ and $\theta = 0$ and then we consider the general case. According to (3.6),

$$G_{N_{l,m}^{(n)}}(y;\alpha,0) = \sum_{i=0}^{\lfloor m/l \rfloor} \tilde{y}^i C(i,m;n,j,\alpha,l)$$

where

$$\begin{aligned}
&C(i,m;n,j,\alpha,l) \\
&\quad = \frac{m!}{(n)_{(m)}} \binom{j+i-1}{i} \frac{i\alpha+n}{m-il} \binom{n+m+i\alpha-il-1}{m-il-1} \\
&\quad = \binom{n+m+i\alpha-il-1}{n+m-il-1} \frac{m!}{(n)_{(m)}} \binom{j+i-1}{i} \frac{(m-il+1)_{(n-2)}}{(i\alpha+1)_{(n-2)}} \\
&\quad = \binom{n+m+i\alpha-il-1}{n+m-il-1} \\
&\qquad \times \frac{(n-1)!}{(m+1)\cdots(m+n-1)} \frac{(m-il+1)_{(n-2)}}{(i\alpha+1)_{(n-2)}} \binom{j+i-1}{i}
\end{aligned}$$

which is bounded below by $((n-1)!/(m+n)^{n-1})^2$, and from above by $(m+n)^{n+j-1}$. Hence,

$$G_{N_{l,m}^{(n)}}(y;\alpha,0) \leq (m+n)^{n+j-1} G_{M_{l,n+m}}(y;\alpha,0) \tag{3.7}$$

and

$$G_{N_{l,m}^{(n)}}(y;\alpha,0) \geq \frac{\left(G_{M_{l,n+m}}(y;\alpha,0) - \sum_{i=\lfloor m/l \rfloor+1}^{\lfloor (n+m)/l \rfloor} \tilde{y}^i \binom{n+m+i\alpha-il-1}{n+m-il-1}\right)}{\left(\frac{(n-1)!}{(m+n)^{n-1}}\right)^{-2}}. \tag{3.8}$$

Note that for any index i such that $\lfloor m/l \rfloor + 1 \leq i \leq \lfloor (m+n)/l \rfloor$, we can write the following inequalities $1 \leq \binom{n+m+i\alpha-il-1}{n+m-il-1} = (n+m-il)\cdots(n+m-il-1+i\alpha)/(i\alpha)! \leq (n+1)\cdots(n+i\alpha)/(i\alpha)! \leq (2n+m)^n$. Accordingly, the following limit can be easily verified

$$\lim_{m\to+\infty} \frac{1}{m} \log \sum_{i=\lfloor m/l \rfloor+1}^{\lfloor (n+m)/l \rfloor} \tilde{y}^i \binom{n+m+i\alpha-il-1}{n+m-il-1} = 0. \tag{3.9}$$

Accordingly, putting together equations displayed in (3.7), (3.8) and (3.9), we obtain the following identity

$$\lim_{m\to+\infty} \frac{1}{m} \log G_{N_{l,m}^{(n)}}(y;\alpha,0) = \lim_{m\to+\infty} \frac{1}{n+m} \log G_{M_{l,n+m}}(y;\alpha,0)$$

which, once combined with Theorem 1.1, implies that $m^{-1}N_{l,m}^{(n)} \,|\, K_n$ satisfies a large deviation principle with speed m and rate function I_l^α. In order to deal with the general case $\alpha \in (0,1)$ and $\theta > -\alpha$, we need a termwise comparison between (3.4) and (3.5). For any $i \leq m/l$ let

$$D(m,i;\alpha,\theta,n,j) = \frac{(n)_{(m)}}{(\theta+n)_{(m)}} \frac{(j-1)!(\frac{\theta}{\alpha})_{(j+i)}}{(j+i-1)!(\frac{\theta}{\alpha})_{(j)}} \frac{(\theta+n+i\alpha)_{(m-il)}}{(n+i\alpha)_{(m-il)1}}.$$

Then, one has

$$\begin{aligned}
&\mathbb{E}_{\alpha,\theta}[(N_{l,m}^{(n)})_{(r)} \,|\, K_n = j] \\
&\quad = \frac{j}{(n)_{(m)}}(r-1)! \sum_{i=0}^{r} D(m,i;\alpha,\theta,n,j) \binom{r}{i} \binom{j+i-1}{i-1} (m)_{[il]} \\
&\qquad \times \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i (i\alpha+n)_{(m-il)}.
\end{aligned}$$

By means of arguments similar to those used for deriving the inequalities in (2.14), it follows that one can find constants $d_5 > 0$ and $d_6 > 0$ and positive integers k_1 and k_2 independent of m and i such that $d_5(n+m)^{-k_1} \leq D(m,i;\alpha,\theta,n,j) \leq d_6(n+m)^{k_2}$. This leads to

$$\begin{aligned}
&d_5 \left(\frac{1}{n+m}\right)^{k_1} G_{N_{l,m}^{(n)}}(y;\alpha,0) \\
&\quad \leq G_{N_{l,m}^{(n)}}(y;\alpha,\theta) \\
&\quad \leq G_{N_{l,m}^{(n)}}(y;\alpha,0) d_6(n+m)^{k_2}.
\end{aligned}$$

Such a result, combined with the large deviation principle stated in Theorem 1.1, implies that $m^{-1}N_{l,m}^{(n)} \,|\, K_n$ satisfies a large deviation principle with speed m and rate function I_l^α. Hence, by a direct application of Corollary B.9 in Feng [14], $m^{-1}M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ satisfies a large deviation principle with speed m and rate function I_l^α, and the proof is completed. □

In contrast with the fluctuation limits (1.7) and (1.9), Theorem 1.1 and Theorem 1.2 show that in terms of large deviations the given initial sample $(X_1,\ldots,X_n)$ have no long lasting impact. Specifically the large deviation principle for $M_{l,n}$, as $n \to +\infty$, coincides with the large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$, as $m \to +\infty$ and fixed n. This is caused by the two different scalings involved, namely m^{-1} for large deviations and $m^{-\alpha}$ for the fluctuations. According to Corollary 20 in Pitman [26], the initial sample $(X_1,\ldots,X_n)$ leads to modify the parameter θ in the conditional distribution of $\tilde{P}_{\alpha,\theta,\nu}$ given $(X_1,\ldots,X_n)$. Hence we conjecture that the conditional and the unconditional large deviation results will be different if n is allowed to grow and it leads to a larger parameter θ. In the unconditional setting this kind of asymptotic behaviour is thoroughly discussed in Feng [13], where the parameter θ and the sample size n grow together and the large deviation result will depend on the relative growth rate between n and θ.

If m depends on n and both approach infinity, then one can expect very different behaviours in terms of law of large numbers and fluctuations. The large deviation principle for $M_{l,m}^{(n)} \,|\, (K_n, \mathbf{N}_n)$ may not be easily derived, by means of a direct comparison argument, from the large deviation principle of $N_{l,m}^{(n)} \,|\, K_n$. In this respect, it is helpful to study directly

$$G_{M_{l,m}^{(n)}}(y;\alpha,\theta) \tag{3.10}$$

$$= \mathbb{E}_{\alpha,\theta}\left[\left(\frac{1}{1-y}\right)^{M_{l,m}^{(n)}} \mid K_n = j, \mathbf{N}_n = \mathbf{n}\right]$$
$$= \sum_{i\geq 0}\frac{y^i}{i!}\mathbb{E}_{\alpha,\theta}[(M_{l,m}^{(n)})_{(i)} \mid K_n = j, \mathbf{N}_n = \mathbf{n}].$$

We intend to pursue this study further in a subsequent project. Here we conclude by providing an explicit expression for (3.10) with $\theta = 0$. An expression for any $\theta > -\alpha$ follows by means of similar arguments. The rising factorial moments of $M_{l,m}^{(n)} \mid (K_n, \mathbf{N}_n)$ are obtained from Theorem 3 in Favaro et al. [11]. Specifically, one has the following expressions

$$\begin{aligned}
&\mathbb{E}_{\alpha,\theta}[(M_{l,m}^{(n)})_{(r)} \mid (K_n = j, \mathbf{N}_n = \mathbf{n})] \\
&\quad = r!\sum_{t=0}^{r}\sum_{v=0}^{r-t}\binom{r-1}{v+t-1}\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^t \\
&\qquad \times \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\frac{m!}{(m-tl-vl+\sum_{h=1}^{v}n_{c_h})!} \\
&\qquad\quad \times \frac{\left(\frac{\theta}{\alpha}+j\right)_{(t)}(\theta+n+t\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-tl-vl+\sum_{h=1}^{v}n_{c_h})}}{t!(\theta+n)_{(m)}}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_{\alpha,0}[(M_{l,m}^{(n)})_{(r)} \mid (K_n = j, \mathbf{N}_n = \mathbf{n})] \qquad (3.11)\\
&\quad = r!\sum_{t=0}^{r}\sum_{v=0}^{r-t}\binom{r-1}{v+t-1}\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^t \\
&\qquad \times \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\frac{m!}{(m-tl-vl+\sum_{h=1}^{v}n_{c_h})!} \\
&\qquad\quad \times \frac{(j)_{(t)}(n+t\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-tl-vl+\sum_{h=1}^{v}n_{c_h})}}{t!(n)_{(m)}},
\end{aligned}$$

where the sum over the indexes t, v and $(c_1,\ldots,c_v)$ is non-null for any $t = 0,\ldots,r$, $v = 0,\ldots,r-t$ and $(c_1,\ldots,c_v) \in \mathcal{C}_{j,v}$ such that $(m-tl-vl+\sum_{1\leq h\leq v}n_{c_h}) \geq 0$. An explicit expression for (3.10) follows by combining the rising factorial moments of $M_{l,m}^{(n)} \mid (K_n, \mathbf{N}_n)$ with the right-hand side of (3.3), and by means of standard combinatorial manipulations.

Proposition 3.2. *For any* $\alpha \in (0,1)$

$$\begin{aligned}
&G_{M_{l,m}^{(n)}}(x;\alpha,0) \\
&\quad = \frac{m!}{(n)_{(m)}}\sum_{v=0}^{j}\left(\frac{x}{1-x}\right)^v \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!} \\
&\qquad \times \sum_{i=0}^{\lfloor\frac{m}{l}-v+\sum_{h=1}^{v}\frac{n_{c_h}}{l}\rfloor}\frac{1}{i!}\left(\frac{x}{1-x}\right)^i(j)_{(i)}\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i\frac{(n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})}{(m-il-vl+\sum_{h=1}^{v}n_{c_h})} \\
&\qquad\quad \times \binom{n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h}+m-il-vl+\sum_{h=1}^{v}n_{c_h}-1}{m-il-vl+\sum_{h=1}^{v}n_{c_b}-1}.
\end{aligned}$$

Proof. The proof is obtained by simply combining the right-hand side of (3.10), with $\theta = 0$, with the rising factorial moment displayed in (3.11). This leads to write $G_{M_{l,m}^{(n)}}(y;\alpha,0)$ as follows

$$
\begin{aligned}
&G_{M_{l,m}^{(n)}}(y;\alpha,0)\\
&\quad=\sum_{t\geq 0}\frac{y^t}{t!}\sum_{i=0}^{t}\binom{t}{i}\sum_{v=0}^{t-i}\binom{t-i}{v}\frac{(t-1)!}{(v+i-1)!}v!\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i\\
&\qquad\times\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\frac{m!}{(m-tl-vl+\sum_{h=1}^{v}n_{c_h})!}\\
&\qquad\quad\times(j)_{(i)}\frac{(n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-il-vl+\sum_{h=1}^{v}n_{c_h})}}{(n)_{(m)}}\\
&\quad=\sum_{i\geq 0}\sum_{v\geq 0}\frac{1}{(v+i-1)!}v!\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i\\
&\qquad\times\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\frac{m!}{(m-il-vl+\sum_{h=1}^{v}n_{c_h})!}\\
&\qquad\quad\times(j)_{(i)}\frac{(n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-il-vl+\sum_{h=1}^{v}n_{c_h})}}{(n)_{(m)}}\sum_{t\geq i+v}\frac{y^t}{t!}\binom{t}{i}\binom{t-i}{v}(t-1)!\\
&\quad=\sum_{i\geq 0}\sum_{v\geq 0}\frac{v!}{(v+i)!}\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i\left(\frac{y}{1-y}\right)^{i+v}\binom{i+v}{i}\\
&\qquad\times\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\frac{m!}{(m-il-vl+\sum_{h=1}^{v}n_{c_h})!}\\
&\qquad\quad\times(j)_{(i)}\frac{(n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-il-vl+\sum_{h=1}^{v}n_{c_h})}}{(n)_{(m)}},
\end{aligned}
$$

where in the last identity we used the fact that $\sum_{t\geq i+v}y^t(t!)^{-1}\binom{t}{i}\binom{t-i}{v}(t-1)! = ((v+i-1)!/i!v!)(y/(1-y))^{i+v}$. The sum over i and v are bounded by j and $\lfloor m/l - v + \sum_{h=1}^{v}n_{c_h}/l\rfloor$, respectively. Hence

$$
\begin{aligned}
&G_{M_{l,m}^{(n)}}(y;\alpha,0)\\
&\quad=\sum_{v=0}^{j}\left(\frac{y}{1-y}\right)^v\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}\prod_{h=1}^{v}\frac{(n_{c_h}-\alpha)_{(l-n_{c_h})}}{(l-n_{c_h})!}\\
&\qquad\times\sum_{i=0}^{\lfloor\frac{m}{l}-v+\sum_{h=1}^{v}\frac{n_{c_h}}{l}\rfloor}\frac{1}{i!}\left(\frac{y}{1-y}\right)^i(j)_{(i)}\left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!}\right)^i\frac{m!}{(m-il-vl+\sum_{h=1}^{v}n_{c_h})!}\\
&\qquad\quad\times\frac{(n+i\alpha+v\alpha-\sum_{h=1}^{v}n_{c_h})_{(m-il-vl+\sum_{h=1}^{v}n_{c_h})}}{(n)_{(m)}}.
\end{aligned}
$$

The proof is completed by rearranging the factorial moments by means of $(n+i\alpha+v\alpha-\sum_{1\leq h\leq v}n_{c_h})_{(m-il-vl+\sum_{1\leq h\leq v}n_{c_h})} = (n+i\alpha+v\alpha+m-il-vl-1)!/(n+i\alpha+v\alpha-\sum_{1\leq h\leq v}n_{c_h}-1)!$.

□

4 Discussion and numerical illustrations

Our large deviation results contribute to the study of conditional and unconditional properties of the Ewens-Pitman sampling model. Theorem 1.2 has potential applications in the context of Bayesian nonparametric inference for species sampling problems. Indeed, as we pointed out in the Introduction, in such a context $\mathbb{P}[M_{l,m}^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ takes on the interpretation of the posterior distribution of the number of species with frequency l in a sample $(X_1, \ldots, X_{n+m})$ from $\tilde{P}_{\alpha,\theta,\nu}$, given the initial observed sample $(X_1, \ldots, X_n)$ featuring $K_n = j$ species with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$. The reader is referred to Favaro et al. [11] for a comprehensive account on this posterior distribution with applications to Bayesian nonparametric inference for the so-called rare, or local, species variety.

For a large additional sample size m, $m^{-1}M_{l,m}^{(n)}$ represents the random proportion of species with frequency l in the enlarged sample $(X_1, \ldots, X_{n+m})$. In Theorem 1.2 we characterized the rate function I_l^α of a conditional large deviation principle associated to such a random proportion. The rate function I_l^α is nondecreasing over the set $[0, 1/l]$. Then the number of discontinuous points of I_l^α is at most countable and therefore $\inf_{z \geq x} I_l^\alpha(z) = \inf_{z > x} I_l^\alpha(z)$ for almost all $x \in [0, 1/l]$. Accordingly, for almost all $x > 0$ we can write

$$\begin{aligned} &\lim_{m \to +\infty} \frac{1}{m} \log \mathbb{P}\left[\frac{M_{l,m}^{(n)}}{m} \geq x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] \\ &= \lim_{m \to +\infty} \frac{1}{m} \log \mathbb{P}\left[\frac{M_{l,m}^{(n)}}{m} > x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] = -I_l^\alpha(x). \end{aligned} \tag{4.1}$$

Therefore the identity (4.1) provides a large m approximation of the Bayesian nonparametric estimator $\mathbb{P}[m^{-1}M_{l,m}^{(n)} \geq x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$, for any $x \geq 0$. In other words, we can write

$$\mathcal{T}_{l,m}^{(n)}(x) = \mathbb{P}\left[\frac{M_{l,m}^{(n)}}{m} \geq x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] \approx \exp\{-mI_l^\alpha(x)\}, \tag{4.2}$$

for any $x \geq 0$. Hereafter we thoroughly discuss the tail probability $\mathcal{T}_{1,m}^{(n)}$ within the context of Bayesian nonparametric inference for discovery probabilities. In particular we introduce a novel approximation, for large m, of the posterior distribution of the probability of discovering a new species at the $(n+m+1)$-th draw. Such an approximation, then, induces a natural interpretation of $\mathcal{T}_{1,m}^{(n)}$ within the context of Bayesian nonparametric inference for the probability of discovering a new species at the $(n+m+1)$-th draw.

4.1 Discovery probabilities and large deviations

Let $D_m^{(n)}$ be the probability of discovering a new species at the $(n+m+1)$-th draw. Since the additional sample $(X_{n+1} \ldots, X_{n+m})$ is assumed to be not observed, $D_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ is a random probability. The randomness $D_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ is determined by $(X_{n+1} \ldots, X_{n+m})$. In particular, by means of the predictive distribution (1.1), we observe that $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ is related to $\mathbb{P}[K_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ as follows

$$\begin{aligned} &\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &\quad = \mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j] \\ &\quad = \mathbb{P}\left[\frac{\theta + j\alpha + K_m^{(n)}\alpha}{\theta + n + m} \in \cdot \,|\, K_n = j\right] \end{aligned} \tag{4.3}$$

$$= \mathbb{P}\left[\frac{\theta + j\alpha + K_m^{(n)}\alpha}{\theta + n + m} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right],$$

where the conditional, or posterior, distribution $\mathbb{P}[K_m^{(n)} \in \cdot \,|\, K_n = j]$ was first obtained in Lijoi et al. [22] and then further investigated in Favaro et al. [10]. Specifically, let $\mathscr{C}(n, x, a, b) = (x!)^{-1} \sum_{0 \le i \le x} (-1)^i \binom{x}{i} (-ia - b)_{(n)}$ be the noncentral generalized factorial coefficient. See, e.g., Charalambides [6] for details. Then, for any $k = 0, 1, \ldots, m$, one has

$$\mathbb{P}[K_m^{(n)} = k \,|\, K_n = j] = \frac{(\theta/\alpha + j)_{(k)}}{(\theta + n)_{(m)}} \mathscr{C}(m, k; \alpha, -n + \alpha j), \tag{4.4}$$

and

$$\mathbb{E}_{\alpha,\theta}[K_m^{(n)} \,|\, K_n = j] = \left(\frac{\theta}{\alpha} + j\right)\left(\frac{(\theta + n + \alpha)_{(m)}}{(\theta + n)_{(m)}} - 1\right). \tag{4.5}$$

The distribution (4.3) takes on the interpretation of the posterior distribution of the probability of discovering a new species at the $(n+m+1)$-th draw. An explicit expression for this distribution is obtained by means of the distribution (4.4). Furthermore, $\mathcal{D}_m^{(n)} = \mathbb{E}_{\alpha,\theta}[D_m^{(n)} \,|\, K_n = j]$ provides the Bayesian nonparametric estimator, with respect to a squared loss function, of the probability of discovering a new species at the $(n+m+1)$-th draw. Of course an explicit expression of this estimator is obtained by combining (4.3) with (4.5).

We introduce a large m approximation of $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j]$ and a corresponding large m approximation of the Bayesian nonparametric estimator $\mathcal{D}_m^{(n)}$. Such an approximation sets a novel connection between the posterior distribution of the proportion of species with frequency 1 in the enlarged sample and the posterior distribution $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j]$. Specifically, by combining the fluctuation limit (1.8) with (4.3), one obtains

$$\lim_{m \to +\infty} \frac{D_m^{(n)}}{m^{\alpha-1}} \,|\, (K_n = j) = \alpha S_{\alpha,\theta}^{(n,j)} \qquad \text{a.s.} \tag{4.6}$$

where $S_{\alpha,\theta}^{(n,j)}$ has been defined in (1.8) and (1.9). In particular $\mathbb{E}[S_{\alpha,\theta}^{(n,j)}] = (j + \theta/\alpha)\Gamma(\theta + n)/\Gamma(\theta + n + \alpha)$. Then, for large m, the fluctuations (4.6) and (1.9) lead to the following approximation

$$\begin{aligned} &\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j] \\ &\quad \approx \mathbb{P}[m^{\alpha-1}\alpha S_{\alpha,\theta}^{(n,j)} \in \cdot \,|\, K_n = j] \\ &\quad \approx \mathbb{P}\left[\frac{M_{1,m}^{(n)}}{m} \in \cdot \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} \mathcal{D}_m^{(n)} &= \frac{\theta + j\alpha}{\theta + n} \frac{(\theta + n + \alpha)_m}{(\theta + n + 1)_m} \\ &\approx m^{\alpha-1}(j\alpha + \theta)\frac{\Gamma(\theta + n)}{\Gamma(\theta + n + \alpha)} \\ &\approx \mathbb{E}_{\alpha,\theta}\left[\frac{M_{1,m}^{(n)}}{m} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}\right] \\ &= \frac{m_1}{m}\frac{(\theta + n - 1 + \alpha)_m}{(\theta + n)_m} + (\theta + j\alpha)\frac{(\theta + n + \alpha)_{m-1}}{(\theta + n)_m} \end{aligned} \tag{4.8}$$

where the last identity of (4.8) is obtained by means of Theorem 3 in Favaro et al. [11]. The second approximation of (4.8) is somehow reminiscent of the celebrated Good-Turing

estimators introduced in Good [17] and Good and Toulmin [18]. Indeed, it shows that the estimator of the probability of discovering a new species at the $(n+m+1)$-th draw is related to the estimator of the number of species with frequency 1 in the enlarged sample.

Intuitively, when the parameter θ and the sample size n are moderately large and not overwhelmingly smaller than m, the exact value of $\mathcal{D}_m^{(n)}$ given in (4.8) is much smaller than its large m approximation, which is much smaller than the exact value of $m^{-1}\mathcal{M}_{1,m}^{(n)} = \mathbb{E}_{\alpha,\theta}[m^{-1}M_{1,m}^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$. This behaviour suggests that a finer normalization constant than m^α is to be used in the fluctuation limits (1.9) and (4.6), respectively. Equivalent, though less rough, normalization rates for (1.9) and (4.6) are given by

$$\begin{aligned} &r_M(m;\alpha,\theta,n,j,m_1) \\ &\quad = \frac{\Gamma(\theta+\alpha+n+m-1)}{\Gamma(\theta+n+m)}\left(m_1\frac{\theta+\alpha+n-1}{\theta+j\alpha}+m\right), \end{aligned} \tag{4.9}$$

and

$$r_D(m;\alpha,\theta,n,j) = \frac{\Gamma(\theta+\alpha+n+m)}{\Gamma(\theta+n+m+1)} \tag{4.10}$$

respectively. Obviously $r_M(m;\alpha,\theta,n,j,m_1)/m^\alpha \to 1$ and $r_D(m;\alpha,\theta,n,j)/m^{\alpha-1} \to 1$ as m tends to infinity. These corrected normalization rates are determined in such a way that $\mathcal{D}_m^{(n)}$ and $m^{-1}\mathcal{M}_{1,m}^{(n)}$ coincide with the corresponding asymptotic moments. Of course different procedures may be considered. Note that the number j of species and the number m_1 of species with frequency 1 affect the corrected normalization rate displayed in (4.9).

Besides being a novel large m approximation of the posterior distribution $\mathbb{P}[D_m^{(n)} \in \cdot \,|\, K_n = j]$, the result displayed in (4.7) induces a natural interpretation of Theorem 1.2, with $l = 1$, in the context of Bayesian nonparametric inference for discovery probabilities. Indeed by combining the approximations in (4.2) and (4.7) we can write the large m approximation

$$\begin{aligned} \mathcal{D}_m^{(n)}(x) &= \mathbb{P}[D_m^{(n)} \geq x \,|\, K_n = j] \\ &\approx \mathcal{T}_{1,m}^{(n)}(x) \\ &\approx \exp\{-mI_1^\alpha(x)\}. \end{aligned} \tag{4.11}$$

By exploiting the corrected normalization rates (4.9) and (4.10), a corrected version of (4.11) is

$$\begin{aligned} \mathcal{D}_m^{(n)}(x) &= \mathbb{P}[D_m^{(n)} \geq x \,|\, K_n = j] \\ &\approx \mathcal{T}_{1,m}^{(n)}\left(x\frac{r_M(m;\alpha,\theta,n,j,m_1)}{mr_D(m;\alpha,\theta,n,j)}\right) \\ &\approx \exp\left\{-mI_1^\alpha\left(x\frac{r_M(m;\alpha,\theta,n,j,m_1)}{mr_D(m;\alpha,\theta,n,j)}\right)\right\}. \end{aligned} \tag{4.12}$$

In other words Theorem 1.2 with $l = 1$ provides a large m approximation of the Bayesian nonparametric estimator of the right tail of the probability of discovering a new species at the $(n+m+1)$-th draw, without observing $(X_{n+1},\ldots,X_{n+m})$. If $\alpha = 1/2$ then the rate function in the approximations (4.11) and (4.12) can be exactly computed by means of Proposition 2.2.

4.2 An application to EST data

We conclude by presenting an brief illustration of our results to the analysis of Expressed Sequence Tag (EST) data. ESTs data, which have been first introduced and investigated in Adams et al. [1], are generated by partially sequencing randomly isolated gene transcripts that have been converted into complementary DNA (cDNA). ESTs play an fundamental role in the identification and discovery of organisms as they provide an attractive and efficient alternative to full genome sequencing. The resulting transcript sequences and their corresponding abundances are the main focus of interest providing the identification and level of expression of genes (species). Within the context of ESTs data an important issue to be addressed in terms of design of a future study is the determination of an "optimal" number of genes to be sequenced by the experimenter. Indeed, despite the novel advances in technology, sequencing is still expensive and therefore suitable cost-effectiveness thresholds must be established. This suggests that there is the need for assessing the relative redundancy of various cDNA libraries prepared from the same organism in order to detect which one yields new genes at a higher rate. Indeed, there are "normalization" protocols which aim at making the "frequencies" of genes in the library more uniform thus typically improving the discovery rate. However, performing such protocols is also expensive. The decision whether to proceed with sequencing of a non-normalized library or to resort to a normalization procedure, has to balance the involved costs: such a decision is necessarily based on the future discovery rate.

This practical issue naturally translates in the statistical problem in which, given an initial sample of ESTs, we are interest in making inference on the probability of discovering a new species at the $(n+m+1)$-th draw, namely $D_m^{(n)}$. Under the Bayesian nonparametric model (1.2), where the genes composition of the population is assumed to be modeled according to $\tilde{P}_{\alpha,\theta,\nu}$, $\mathcal{D}_m^{(n)}$ and $\mathcal{D}_m^{(n)}(x)$, as well as their large m approximations, provides useful pointwise predictive measures of the evolution of redundancy as the sequencing ideally proceeds. Hereafter we present an application of these measures to an EST dataset which is obtained by sequencing two cDNA libraries of the amitochondriate protist Mastigamoeba balamuthi: the first library is non-normalized, whereas the second library is normalized, namely it undergoes a normalization protocol which aims at making the frequencies of genes in the library more uniform so to increase the discovery rate. See Susko and Roger [31] for comprehensive account on the Mastigamoeba cDNA library. For the Mastigamoeba non-normalized the observed sample consists of $n = 715$ ESTs with $j = 460$ distinct genes whose frequencies are $m_{i,715} = 378, 33, 21, 9, 6, 1, 3, 1, 1, 1, 1, 5$ with $i \in \{1, 2, \ldots, 10\} \cup \{13, 15\}$. For the the Mastigamoeba normalized the observed sample consists of $n = 363$ with $j = 248$ distinct genes whose frequencies are $m_{i,363} = 200, 21, 14, 4, 3, 3, 1, 0, 1, 1$ with $i \in \{1, 2, \ldots, 9\} \cup \{14\}$. This means that we are observing $m_{1,n}$ genes which appear once, $m_{2,n}$ genes which appear twice, etc.

Under the Bayesian nonparametric model (1.2), the first issue to face is represented by the specification of the parameter (α, θ). This is typically achieved by adopting an empirical Bayes procedure in order to obtain an estimate $(\hat{\alpha}, \hat{\theta})$ of (α, θ). Specifically we fix (α, θ) so to maximize the likelihood function of the model (1.2) under the observed sample, namely

$$(\hat{\alpha}, \hat{\theta}) = \arg\max_{(\alpha,\theta)} \left\{ \frac{\prod_{i=0}^{j-1}(\theta + i\alpha)}{(\theta)_n} \prod_{i=1}^{j} (1-\alpha)_{(n_i-1)} \right\}.$$

Alternatively, one could specify a prior distribution for (α, θ). Here we adopt a less elaborate specification of the parameter (α, θ). We choice $\alpha = 1/2$ and then we set θ such

that $\mathbb{E}_{1/2,\theta}[K_n] = (2\theta)(((\theta+2^{-1})_n/(\theta)_n) - 1) = j$. Empirical investigations with simulated data suggests that $\alpha = 1/2$ is always a good choice when no precise prior information is available. See Lijoi et al. [22] for details. This approach gives $(\alpha, \theta) = (1/2, 206.75)$ for the Mastigamoeba non-normalized and $(\alpha, \theta) = (1/2, 132.92)$ for the Mastigamoeba normalized.

For the Mastigamoeba non-normalized and normalized cDNA libraries, Table 1 reports the exact estimate $\mathcal{D}_m^{(n)}$ and the corresponding large m approximate estimates under the uncorrected normalization rate $m^{\alpha-1}$ and the corrected normalization rate (4.9). These are denoted by $\bar{\mathcal{D}}_m^{(n)}$ and $\tilde{\mathcal{D}}_m^{(n)}$, respectively. In a similar fashion, Table 2 reports the exact estimate $m^{-1}\mathcal{M}_{1,m}^{(n)}$ and the corresponding large m approximate estimates under the uncorrected normalization rate m^{α} and the corrected normalization rate (4.10), respectively. These are denote by $m^{-1}\bar{\mathcal{M}}_{1,m}^{(n)}$ and $m^{-1}\tilde{\mathcal{M}}_{1,m}^{(n)}$, respectively. See (4.8) for details. Table 1 and Table 2 clearly show that the corrected normalization rates displayed in (4.9) and (4.10) are of fundamental importance when the additional sample size m is not much larger than the sample size n and the parameter θ. Figure 1 and Figure 2 show the large deviation approximations (4.11) and (4.12) of the pointwise estimate $\mathcal{D}_m^{(n)}(x)$.

Table 1. *Exact estimate and corresponding asymptotic estimates under the uncorrected and corrected normalization rate.*

cDNA Library	m	$\mathcal{D}_m^{(n)}$	$\bar{\mathcal{D}}_m^{(n)}$	$\tilde{\mathcal{D}}_m^{(n)}$
Mastigamoeba non-normalized	$\lfloor 100^{-1}n \rfloor$	0.472	5.438	0.472
	$\lfloor 10^{-1}n \rfloor$	0.456	1.696	0.456
	n	0.357	0.538	0.357
$(n = 715)$	$10n$	0.160	0.314	0.160
	$100n$	0.054	0.054	0.054
Mastigamoeba normalized	$\lfloor 100^{-1}n \rfloor$	0.516	5.770	0.516
	$\lfloor 10^{-1}n \rfloor$	0.500	1.923	0.500
	n	0.397	0.606	0.397
$(n = 363)$	$10n$	0.180	0.288	0.180
	$100n$	0.060	0.061	0.060

Table 2. *Exact estimate and corresponding asymptotic estimates under the uncorrected and corrected normalization rate.*

cDNA Library	m	$m^{-1}\mathcal{M}_{1,m}^{(n)}$	$m^{-1}\bar{\mathcal{M}}_{1,m}^{(n)}$	$m^{-1}\tilde{\mathcal{M}}_{1,m}^{(n)}$
Mastigamoeba non-normalized	$\lfloor 100^{-1}n \rfloor$	54.268	5.438	54.268
	$\lfloor 10^{-1}n \rfloor$	5.213	1.696	5.213
	n	0.752	0.538	0.752
$(n = 715)$	$10n$	0.178	0.314	0.178
	$100n$	0.054	0.054	0.054
Mastigamoeba normalized	$\lfloor 100^{-1}n \rfloor$	50.316	5.770	50.316
	$\lfloor 10^{-1}n \rfloor$	5.865	1.923	5.865
	n	0.812	0.606	0.812
$(n = 363)$	$10n$	0.199	0.288	0.199
	$100n$	0.061	0.061	0.061

Results displayed in Table 1 and Table 2, as well as Figure 1 and Figure 2, provide a natural guideline for selecting the size m of a future sample. Specifically, in the case the

estimated discovery rate $\tilde{\mathcal{D}}_m^{(n)}$ falls below a certain threshold τ suggested by the problem at issue, then it may be convenient to reduce the size of the future sample in a way that $\tilde{D}_m^{(n)}$ does not fall below τ. On the other hand, if $\tilde{\mathcal{D}}_m^{(n)}$ is still relatively high with respect to τ, one may decide to enlarge the size of the future sample. In both these situations the large deviation rates in Figure 1 and Figure 2 provide an approximate estimate of the decay of the discovery rate. As expected the Mastigamoeba normalized data exhibit a higher discovery rate, with respect to the Mastigamoeba non-normalized data. Of course this is the effect of the normalization protocol applied to the Mastigamoeba non-normalized data. However, since the discovery rate has a faster decay for the Mastigamoeba normalized data, it appears that, already for moderately large m, the effect of the normalization protocol is exhausted producing fewer number of new genes.

Figure 1. *Mastigamoeba non-normalized. Large deviation approximations of the estimate* $\mathcal{D}_m^{(715)}(x)$ *under the uncorrected (blue line) and corrected (red line) normalization rate.*

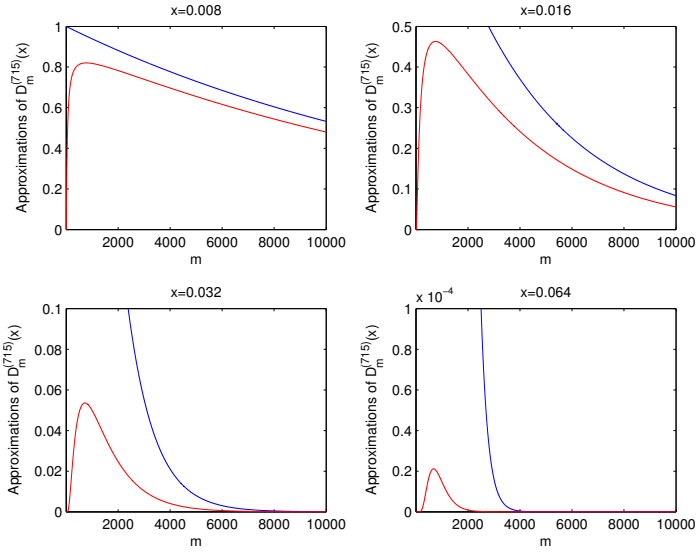


Figure 2. *Mastigamoeba normalized. Large deviation approximations of the estimate* $\mathcal{D}_m^{(363)}(x)$ *under the uncorrected (blue line) and corrected (red line) normalization rate.*

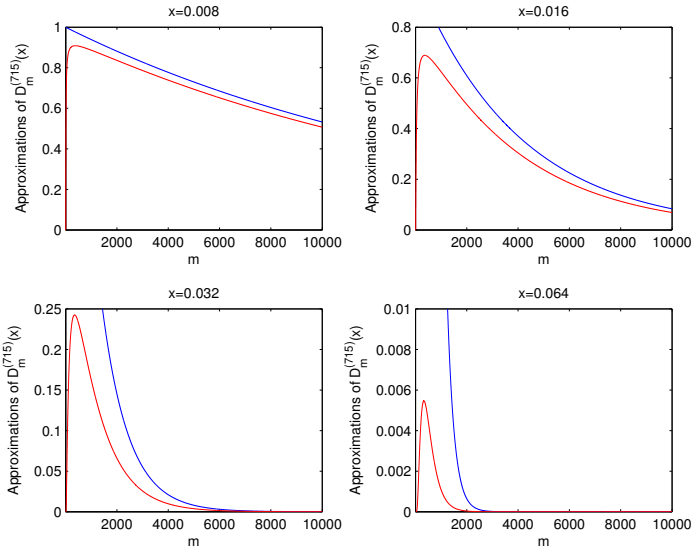
Acknowledgments. The authors are grateful to an anonymous Referee for valuable remarks and suggestions that have led to a substantial improvement of the paper. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406. Shui Feng is supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Adams, M., Kelley, J., Gocayne, J., Mark, D., Polymeropoulos, M., Xiao, H., Merril, C., Wu, A., Olde, B., Moreno, R., Kerlavage, A., McCombe, W. and Venter, J. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656.

[2] Arratia, R., Barbour, A.D. and Tavaré, S. (1992). Poisson process approximations for the Ewens sampling formula. *Ann. Appl. Probab.* **2**, 519–535. MR-1177897

[3] Arratia, R., Barbour, A.D. and Tavaré, S. (2003). *Logarithmic combinatorial structures: a probabilistic approach*. EMS Monograph in Mathematics. MR-2032426



[4] Bacallado, S., Favaro, S. and Trippa, L. (2013). Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli*, **21**, 1–37. MR-3322311

[5] Barbour, A.D. and Gnedin, A.V. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Probab.*, **13**, 365–384. MR-2480545

[6] Charalambides, C.A. (2005). *Combinatorial methods in discrete distributions*. Wiley Interscience. MR-2131068

[7] Dembo, A. and Zeitouni, O. (1998) *Large deviations techniques and applications*. Springer, New York. MR-1619036

[8] Dinwoodie, I.H. and Zabell, S.L. (1992). Large deviations for exchangeable random vectors. *Ann. Probab.*, **20**, 1147-1166 MR-1175254

[9] Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.*, **3**, 87–112. MR-0325177

[10] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B*, **71**, 993–1008. MR-2750254

[11] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754. MR-3114915

[12] Favaro, S. and Feng, S. (2013). Asymptotics for the conditional number of blocks in the Ewens-Pitman sampling model. *Electron. J. Probab.*, **19**, 1–15.

[13] Feng, S. (2007). Large deviations associated with Poisson-Dirichlet distribution and Ewens sampling formula. *Ann. Appl. Probab.*, **17**,1570–1595. MR-2358634

[14] Feng, S. (2010). *The Poisson-Dirichlet distribution and related topics: models and asymptotic behaviors*, Springer, Heidelberg. MR-2663265

[15] Feng, S. and Hoppe, F.M. (1998). Large deviation principles for some random combinatorial structures in population genetics and Brownian motion. *Ann. Appl. Probab.*, **8**, 975–994. MR-1661315

[16] Flajolet, P., Dumas, P. and Puyhaubert, V. (2006). Some exactly solvable models of urn process theory. *Discrete Math. Theor. Comput. Sci. Proceedings of the fourth colloquium on Mathematics and Computer Science*, 59–118. MR-2509623

[17] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264. MR-0061330

[18] Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63. MR-0077039

[19] Griffiths, R.C. and Spanò, D. (2007). Record indices and age-ordered frequencies in exchangeable Gibbs partitions. *Electron. J. Probab.*, **12**, 1101–1130. MR-2336601

[20] Janson, S. (2006) Limit theorems for triangular urn schemes. *Probab. Theory Related Fields*, **134**, 417–452. MR-2226887

[21] Korwar, R.M. and Hollander, M. (1973). Contribution to the theory of Dirichlet processes. *Ann. Probab.*, **1**,705–711. MR-0350950

[22] Lijoi, A., Mena, R.H. and Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering a new species *Biometrika*, **94**, 769–786. MR-2416792

[23] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.*, **18**, 1519–1547. MR-2434179

[24] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields*, **92**, 21–39. MR-1156448

[25] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, **102**, 145–158. MR-1337249

[26] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T.S. Ferguson, L.S. Shapley and J.B. MacQueen Eds.), Hayward: Institute of Mathematical Statistics, 245–267. MR-1481784

[27] Pitman, J. (1997). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, **3**, 79–66. MR-1466546

[28] Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, **25**, 855–900. MR-1434129

[29] Pitman, J. (2006). *Combinatorial stochastic processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer-Verlag, New York. MR-2245368

[30] Schweinsberg, J. (2010). The number of small blocks in exchangeable random partitions. *ALEA Lat. Am. J. Probab. Math. Stat.* **7**, 217–242. MR-2672786

[31] Susko, E. and Roger, A.J. (2004). Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics*, **20**, 2279–2287.

[32] Watterson G.A. (1974). The sampling theory of selectively neutral alleles. *Adv. in Appl. Probab.* **6**, 463–488. MR-0351504

[33] Watterson G.A. (1983). Lines of descent and the coalescent. *Theor. Pop. Biol.* **26**, 77–92. MR-0760232