Electron. J. Probab. **19** (2014), no. 21, 1–15.
ISSN: 1083-6489 DOI: 10.1214/EJP.v19-2881

Asymptotics for the number of blocks in a conditional Ewens-Pitman sampling model

Stefano Favaro* Shui Feng†

Abstract

The study of random partitions has been an active research area in probability over the last twenty years. A quantity that has attracted a lot of attention is the number of blocks in the random partition. Depending on the area of applications this quantity could represent the number of species in a sample from a population of individuals or the number of cycles in a random permutation, etc. In the context of Bayesian nonparametric inference such a quantity is associated with the exchangeable random partition induced by sampling from certain prior models, for instance the Dirichlet process and the two parameter Poisson-Dirichlet process. In this paper we generalize some existing asymptotic results from this prior setting to the so-called posterior, or conditional, setting. Specifically, given an initial sample from a two parameter Poisson-Dirichlet process, we establish conditional fluctuation limits and conditional large deviation principles for the number of blocks generated by a large additional sample.

Keywords: Bayesian nonparametrics; Dirichlet process; Ewens-Pitman sampling model; exchangeable random partition; fluctuation limit; large deviations; two parameter Poisson-Dirichlet process.
AMS MSC 2010: Primary 60F10, Secondary 92D10.
Submitted to EJP on June 19, 2013, final version accepted on February 12, 2014.

1 Introduction

Among various definitions of the Ewens-Pitman sampling model, a simple and intuitive one arises from Zabell [27] in terms of the following urn model. See also Feng and Hoppe [10]. Let $\mathbb{X}$ be a complete and separable metric space and let ν be a nonatomic probability measure on $\mathbb{X}$. Let $\alpha \in [0, 1)$ and consider an urn that initially contains a black ball with mass $\theta > 0$. Balls are drawn from the urn successively with probabilities proportional to their masses. When a black ball is drawn, it is returned to the urn together with a black ball of mass α and a ball of a new color, which is sampled from ν, with mass $(1 - \alpha)$. When a non-black ball is drawn, it is returned to the urn with

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

an additional ball of the same color with mass one. If $(X_i)_{i\geq 1}$ denotes the sequence of non-black colors, then

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

for any $i \geq 1$, with $X_1^*, \ldots, X_j^*$ being the j distinct colors in $(X_1, \ldots, X_i)$ with frequencies $\mathbf{n} = (n_1, \ldots, n_j)$. The predictive distribution (1.1) was first introduced in Pitman [21] for any $\alpha \in (0, 1)$ and $\theta > -\alpha$, and it is referred to as the Ewens-Pitman sampling model with parameter (α, θ, ν). In particular, Pitman [21] showed that the sequence $(X_i)_{i\geq 1}$ generated by (1.1) is exchangeable and its de Finetti measure Π is the distribution of the two parameter Poisson-Dirichlet process $\tilde{P}_{\alpha,\theta}$ in Perman et al. [20]. Accordingly, we can write

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

for any $n \geq 1$. See Pitman and Yor [23] for details on $\tilde{P}_{\alpha,\theta,\nu}$. For $\alpha = 0$ the urn model generating the X_i's reduces to the one in Hoppe [15], and the Ewens-Pitman sampling model reduces to the celebrated sampling model by Ewens [5]. Accordingly, for $\alpha = 0$ the two parameter Poisson-Dirichlet process reduces to the Dirichlet process by Ferguson [11]. The Ewens sampling model and its two parameter generalization play an important role in many research areas such as population genetics, machine learning, Bayesian nonparametrics, combinatorics and statistical physics. We refer to the monograph by Pitman [25] and references therein for a comprehensive account on these sampling models.

According to (1.1) and (1.2), a sample $(X_1, \ldots, X_n)$ from $\tilde{P}_{\alpha,\theta,\nu}$ induces a random partition of the set $\{1, \ldots, n\}$ into K_n blocks with corresponding frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$. The exchangeable sequence $(X_i)_{i\geq 1}$, then, induces an exchangeable random partition of the set of natural numbers $\mathbb{N}$. See Pitman [21] for details. Such a random partition has been the subject of a rich and active literature and, in particular, there have been several studies on the large n asymptotic behavior of K_n. For any $\alpha \in (0, 1)$ and $q > -1$, let $S_{\alpha,q\alpha}$ be a positive and finite random variable with density function of the form

$$g_{S_{\alpha,q\alpha}}(y) = \frac{\Gamma(q\alpha + 1)}{\alpha\Gamma(q + 1)} y^{q-1-1/\alpha} f_\alpha(y^{-1/\alpha}),$$

where f_α denotes the density function of a positive α-stable random variable. $S_{\alpha,q\alpha}$ is referred to as polynomially tilted α-stable. For any $\alpha \in (0, 1)$ and $\theta > -\alpha$ Pitman [22] showed that

$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_{\alpha,\theta} \qquad \text{a.s.} \tag{1.3}$$

See Pitman [25] and references therein for various generalizations and refinements of the fluctuation limit (1.3). In contrast, for $\alpha = 0$ and $\theta > 0$, Korwar and Hollander [17] showed that

$$\lim_{n\to+\infty} \frac{K_n}{\log n} = \theta \qquad \text{a.s.} \tag{1.4}$$

See Arratia et al. [1] for details. Weak convergence versions of (1.4) and (1.3) can also be derived from general asymptotic results for urn model with weighted balls. The reader is referred to Proposition 16 in Flajolet et al. [12] and Theorem 5 in Janson [16] for details. Fluctuation limits (1.4) and (1.3) display the crucial role of α in determining

both the clustering structure and the large n asymptotic behaviour of K_n. In general, $\alpha \in (0,1)$ controls the flatness of the distribution of K_n: the bigger α the flatter is the distribution of K_n.

Feng and Hoppe [10] further investigated the large n asymptotic behaviour of K_n and established a large deviation principle for $n^{-1}K_n$. Specifically, for any $\alpha \in (0,1)$ and $\theta > -\alpha$, they showed that $n^{-1}K_n$ satisfies a large deviation principle with speed n and rate function of the form

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

where $\Lambda_{\alpha}(\lambda) = -\log(1-(1-e^{-\lambda})^{1/\alpha})\mathbb{1}_{(0,+\infty)}(\lambda)$. Equation (1.3) shows that K_n fluctuates at the scale of n^{α}. This is analogous to a central limit type theorem where the fluctuation occurs at the scale of $\sqrt{n}$. Then the large deviation scaling of n can be understood through a comparison with the classical Cramér theorem where the law of large numbers is at the scale of n. In contrast, for $\alpha = 0$ and $\theta > 0$, Equation (1.4) is analogous to a law of large numbers type limit. In particular, it was shown in Feng and Hoppe [10] that $(\log n)^{-1}K_n$ satisfies a large deviation principle with speed $\log(n)$ and rate function of the form

$$I_{\theta}(u) = \begin{cases} u\log\frac{u}{\theta} - u + \theta & u > 0 \\ \theta & u = 0 \\ +\infty & u < 0. \end{cases} \tag{1.6}$$

It is worth pointing out that rate function (1.5) depends only on the parameter α which displays the different roles of the two parameters at different scales. We refer to Feng and Hoppe [10] for an intuitive explanation in terms of an embedding scheme for the Ewens-Pitman sampling model. See also Tavaré [26] for a similar embedding scheme in the Ewens sampling model.

In this paper we present conditional counterparts of the aforementioned asymptotic results. The problem of studying conditional properties of exchangeable random partitions has been first considered in Lijoi et al. [19]. This problem consists in evaluating, conditionally on an initial sample $(X_1,\dots,X_n)$ from $\tilde{P}_{\alpha,\theta,\nu}$, the distribution of statistics of an additional sample $(X_{n+1},\dots,X_{n+m})$, for any $m \geq 1$. Lijoi et al. [19] mainly focused on statistics of the so-called new X_{n+i}'s, namely X_{n+i}'s that do not coincide with observations in $(X_1,\dots,X_n)$. Note that, according to (1.1), for any $\alpha \in (0,1)$ and $\theta > -\alpha$ these statistics depend on $(X_1,\dots,X_n)$ via the sole K_n. For $\alpha = 0$ and $\theta > 0$ these statistics are independent of K_n. A representative example is given by the number $K_m^{(n)}$ of new blocks generated by $(X_{n+1},\dots,X_{n+m})$, given K_n. As discussed in Lijoi et al. [18], this statistic has direct applications in Bayesian nonparametric inference for species sampling problems arising from ecology, biology, genetics, linguistics, etc. In such a statistical context the distribution $\mathbb{P}[K_m^{(n)} \in \cdot \,|\, K_n = j]$ takes on the interpretation of the posterior distribution of the number of new species in the additional sample, given an observed sample featuring j species. Hence, the expected value with respect to $\mathbb{P}[K_m^{(n)} \in \cdot \,|\, K_n = j]$ provides the corresponding Bayesian nonparametric estimator. See, e.g., Griffiths and Spanò [14], Favaro et al. [6], Favaro et al. [7] and Bacallado et al. [2] for other contributions at the interface between Bayesian nonparametrics and species sampling problems.

For any $m \geq 1$, let $(X_1,\dots,X_n,X_{n+1},\dots,X_m)$ be a sample from $\tilde{P}_{\alpha,\theta,\nu}$. Within the conditional framework of Lijoi et al. [19], we investigate the large m asymptotic behaviour of the number $T_m^{(n)}$ of blocks generated by $(X_{n+1},\dots,X_{n+m})$, conditionally on the initial part $(X_1,\dots,X_n)$. With a slight abuse of notation, throughout the paper we write $X \,|\, Y$ to denote the random variable whose distribution corresponds to the conditional distribution of X given Y. The random variable $T_m^{(n)}$ consists of $K_m^{(n)}$ plus

the number $R_m^{(n)}$ of old blocks, namely blocks generated by the X_{n+i}'s that coincide with observations in $(X_1,\dots,X_n)$. Hence, differently from $K_m^{(n)}$, for any $\alpha\in[0,1)$ and $\theta>-\alpha$ the statistic $T_m^{(n)}$ depends on $(X_1,\dots,X_n)$ via the random partition $(K_n,\mathbf{N}_n)$. In other words, K_n does not provide a sufficient statistic for $T_m^{(n)}$. Intuitively K_n and $T_m^{(n)}\,|\,(K_n,\mathbf{N}_n)$ should have different asymptotic behaviour as n and m tends to infinity, respectively. This turns out to be the case in terms of fluctuation limits. But in terms of large deviations K_n and $T_m^{(n)}\,|\,(K_n,\mathbf{N}_n)$ have the same asymptotic behaviour. In order to detect the impact on large deviations of the given initial sample one may have to consider different limiting mechanisms. In Bayesian nonparametric inference for species sampling problems, large m conditional asymptotic analysis are typically motivated by the need of approximating quantities of interest from the posterior distribution. See Favaro et al. [6] for a thorough discussion. With this regards, our fluctuation limit provides a useful tools since, as we will see, computational burden for an exact evaluation of posterior distribution $\mathbb{P}[T_m^{(n)}\in\cdot\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}]$ becomes overwhelming for large j, n and m.

In Section 2 we introduce the random variable $T_m^{(n)}\,|\,(K_n,\mathbf{N}_n)$ and we present some distributional results for a finite sample size m. In Section 3 we study the large m asymptotic behaviour of $T_m^{(n)}\,|\,(K_n,\mathbf{N}_n)$ in terms of fluctuation limits and large deviation principles. In Section 4 we discuss our results with a view toward Bayesian nonparametric inference for species sampling problems. Some open problems are also discussed.

2 Preliminaries

Let $(X_1,\dots,X_n)$ be an initial sample from $\tilde{P}_{\alpha,\theta,\nu}$. In order to introduce the conditional framework of Lijoi et al. [19], one needs to consider an additional sample $(X_{n+1},\dots,X_{n+m})$ from $\tilde{P}_{\alpha,\theta,\nu}$. Let $X_1^*,\dots,X_{K_n}^*$ be the labels identifying the K_n blocks in $(X_1,\dots,X_n)$ and let

$$L_m^{(n)}=\sum_{i=1}^{m}\prod_{k=1}^{K_n}\mathbb{1}_{\{X_k^*\}^c}(X_{n+i}) \tag{2.1}$$

be the number of observations belonging to $(X_{n+1},\dots,X_{n+m})$ and not coinciding with observations in $(X_1,\dots,X_n)$. In particular, we denote by $K_m^{(n)}$ the number of new blocks generated by the $L_m^{(n)}$ observations and by $X_{K_n+1}^*,\dots,X_{K_n+K_m^{(n)}}^*$ their identifying labels. Moreover, let

$$\mathbf{M}_{L_m^{(n)}}=(M_1,\dots,M_{K_m^{(n)}}) \tag{2.2}$$

where

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

for any $i=1,\dots,K_m^{(n)}$, are the frequencies of the $K_m^{(n)}$ blocks detected among the $L_m^{(n)}$ observations in the additional sample. Specifically, (2.2) provides the random partitions induced by those observations in the additional sample generating new blocks. Analogously, let

$$\mathbf{S}_{m-L_m^{(n)}}=(S_1,\dots,S_{K_n}) \tag{2.3}$$

where

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

for any $i=1,\dots,K_n$, are the frequencies of the blocks detected among the $m-L_m^{(n)}$ observations in the additional sample. Specifically, (2.3) provides the updating for the

random partition induced by the initial sample. See Lijoi et al. [19] for the distribution of (2.1), (2.2) and (2.3).

The random variables in (2.1), (2.2) and (2.3), together with $K_m^{(n)}$, completely describe the conditional random partition induced by $(X_{n+1},\dots,X_{n+m})$, given $(X_1,\dots,X_n)$. In addition, let $R_m^{(n)}=\sum_{i=1}^{K_n}\mathbb{1}_{\{S_i>0\}}$ be the number of old blocks detected among the $m-L_m^{(n)}$ observations in the additional sample. These blocks are termed "old" to be distinguished from the new blocks detected among the $L_m^{(n)}$ observations. Hence, we introduce

$$T_m^{(n)}=R_m^{(n)}+K_m^{(n)},$$

which is the number of blocks generated by the additional sample. Hereafter we investigate the conditional distribution of $T_m^{(n)}$ given the random partition $(K_n,\mathbf{N}_n)$. We start by deriving falling factorial moments of $T_m^{(n)}\,|\,(K_n,\mathbf{N}_n)$. The resulting moment formulae are expressed in terms of noncentral generalized factorial coefficients $\mathscr{C}$ and noncentral Stirling numbers of the first kind s. Furthermore, we denote by S the noncentral Stirling numbers of the second kind. See Charalambides [3] for an account on these numbers.

Proposition 2.1. *Let $(X_1,\dots,X_n)$ be a sample from $\tilde{P}_{\alpha,\theta,\nu}$ featuring $K_n=j$ blocks with frequencies $\mathbf{N}_n=\mathbf{n}$. Then*

i) for any $\alpha\in(0,1)$ and $\theta>-\alpha$

$$\begin{aligned}&\mathbb{E}[(T_m^{(n)})_{r\downarrow 1}\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}] \qquad (2.4)\\ &=\frac{r!}{(\theta+n)_{m\uparrow 1}}\sum_{i=0}^{r}(-1)^{r-i}\left(\frac{\theta}{\alpha}+j\right)_{(r-i)\uparrow 1}\sum_{v=0}^{i}\binom{j-v}{i-v}(-1)^v\\ &\quad\times\sum_{(c_1,\dots,c_v)\in\mathcal{C}_{j,v}}\mathscr{C}(m,r-i;-\alpha,-\theta-n+\sum_{i=1}^{v}n_{c_i}-v\alpha);\end{aligned}$$

ii) for $\alpha=0$ and $\theta>0$

$$\begin{aligned}&\mathbb{E}[(T_m^{(n)})_{r\downarrow 1}\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}] \qquad (2.5)\\ &=\frac{r!}{(\theta+n)_{m\uparrow 1}}\sum_{i=0}^{r}(\theta)^{r-i}\sum_{v=0}^{i}\binom{j-v}{i-v}(-1)^v\\ &\quad\times\sum_{(c_1,\dots,c_v)\in\mathcal{C}_{j,v}}|s(m,r-i;\theta+n-\sum_{i=1}^{v}n_{c_i})|;\end{aligned}$$

where we defined $\mathcal{C}_{j,0}=\emptyset$ and $\mathcal{C}_{j,v}=\{(c_1,\dots,c_v):c_k\in\{1,\dots,j\},c_k\neq c_\ell,\ \text{if } k\neq\ell\}$ for any $v\geq 1$.

Proof. The random variables $K_m^{(n)}\,|\,(L_m^{(n)},K_n)$ and $R_m^{(n)}\,|\,(L_m^{(n)},K_n,\mathbf{N}_n)$ are independent. See Proposition 1 and Corollary 1 in Lijoi et al. [19] for details. Then, by a direct application of the Vandermonde identity, we can factorize the falling factorial moment as follows

$$\begin{aligned}&\mathbb{E}[(T_m^{(n)})_{r\downarrow 1}\,|\,L_m^{(n)}=s,K_n=j,\mathbf{N}_n=\mathbf{n}] \qquad (2.6)\\ &=\sum_{i=0}^{r}\binom{r}{i}\mathbb{E}[(K_m^{(n)})_{(r-i)\downarrow 1}\,|\,L_m^{(n)}=s,K_n=j]\\ &\qquad\times\mathbb{E}[(R_m^{(n)})_{i\downarrow 1}\,|\,L_m^{(n)}=s,K_n=j,\mathbf{N}_n=\mathbf{n}],\end{aligned}$$

and

$$\mathbb{P}[L_m^{(n)} = s \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] = \frac{1}{(\theta+n)_{m\uparrow 1}} \binom{m}{s} (n - j\alpha)_{(m-s)\uparrow 1} (\theta + j\alpha)_{s\uparrow 1}. \tag{2.7}$$

Let $\alpha \in (0,1)$ and $\theta > -\alpha$. We first consider the falling factorial moment of the number of new blocks. For any $r \geq 0$, Equation 25 in Lijoi et al. [19] and Proposition 1 in [6] lead to

$$\begin{aligned} &\mathbb{E}[(K_m^{(n)})_{r\downarrow 1} \,|\, L_m^{(n)} = s, K_n = j] \\ &= \sum_{l=0}^{r} (-1)^{r-l} \left(j + \frac{\theta}{\alpha}\right)_{l\uparrow 1} \frac{(\theta + j\alpha + l\alpha)_{s\uparrow 1}}{(\theta + j\alpha)_{s\uparrow 1}} \sum_{i=l}^{r} |s(r,i,0)| S\left(i, l; j + \frac{\theta}{\alpha}\right) \\ &= r!(-1)^r \left(\frac{\theta}{\alpha} + j\right)_{r\uparrow 1} \frac{\mathscr{C}(s, r; -\alpha, -\theta - j\alpha)}{(\theta + j\alpha)_{s\uparrow 1}}, \end{aligned} \tag{2.8}$$

where the last identity is obtained by means of Equation 2.57 and Equation 2.60 in Charalambides [3]. With regards to the falling factorial moment of the number of old blocks, for any $r \geq 0$, Equation 25 in Lijoi et al. [19] and Theorem 1 in Baccalado et al. [2] lead to

$$\begin{aligned} &\mathbb{E}[(R_m^{(n)})_{r\downarrow 1} \,|\, L_m^{(n)} = s, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &= r! \sum_{v=0}^{r} \binom{j-v}{r-v} (-1)^v \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \frac{(n - \sum_{i=1}^{v} n_{c_i} - (j-v)\alpha)_{(m-s)\uparrow 1}}{(n - j\alpha)_{(m-s)\uparrow 1}}. \end{aligned} \tag{2.9}$$

The proof of the part i) is completed by combining expressions (2.8) and (2.9) with (2.6) and then by integrating with respect to the distribution (2.7). Specifically, we can write

$$\begin{aligned} &\mathbb{E}[(T_m^{(n)})_{r\downarrow 1} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &= \frac{r!}{(\theta+n)_{m\uparrow 1}} \sum_{i=0}^{r} (-1)^{r-i} \left(\frac{\theta}{\alpha} + j\right)_{(r-i)\uparrow 1} \sum_{v=0}^{i} \binom{j-v}{i-v} (-1)^v \\ &\quad \times \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \sum_{s=0}^{m} \binom{m}{s} \mathscr{C}(s, r-i; -\alpha, -\theta - j\alpha)(n - \sum_{i=1}^{v} n_{c_i} - (j-v)\alpha)_{(m-s)\uparrow 1} \\ &= \frac{r!}{(\theta+n)_{m\uparrow 1}} \sum_{i=0}^{r} (-1)^{r-i} \left(\frac{\theta}{\alpha} + j\right)_{(r-i)\uparrow 1} \sum_{v=0}^{i} \binom{j-v}{i-v} (-1)^v \\ &\quad \times \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \mathscr{C}(m, r-i; -\alpha, -\theta - n + \sum_{i=1}^{v} n_{c_i} - v\alpha) \end{aligned} \tag{2.10}$$

where in the second equality the sum over the index $0 \leq s \leq m$ is obtained by exploiting the fact that $(n - \sum_{i=1}^{v} n_{c_i} - (j-v)\alpha)_{(m-s)\uparrow 1} = \mathscr{C}(m-s, 0; -\alpha, -n + \sum_{i=1}^{v} n_{c_i} + (j-v)\alpha)$ and noting

$$\binom{y+c}{y} \mathscr{C}(x, y+c; d, a+b) = \sum_{j=y}^{x-c} \binom{x}{j} \mathscr{C}(j, y; d, a) \mathscr{C}(x-j, c; d, b).$$

for any $x \geq 0$ and $0 \leq y \leq x$, for any $a > 0$, $b > 0$, $c > 0$ and for any real number d. For $\alpha = 0$ and $\theta > 0$ the result follows by taking the limit of (2.10) as $\alpha \to 0$. Specifically, in taking such a limit we make use of Equation 2.63 in Charalambides [3]. The proof is completed. □

A direct application of Proposition 2.1 lead to the distribution of $T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$. Indeed, by exploiting the relationship between probabilities and falling factorial moments in the case of discrete distributions, formulae (2.4) and (2.5) lead to the following expressions

$$\begin{aligned} &\mathbb{P}[T_m^{(n)} = x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \qquad (2.11)\\ &= \frac{(-1)^{x+j}}{(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{j} \sum_{y=0}^{x} (-1)^{y-v} \binom{v}{x-y-(j-v)} \left(\frac{\theta}{\alpha}+j\right)_y \\ &\quad \times \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \mathscr{C}(m, y; \alpha, j\alpha - n + \sum_{i=1}^{v} n_{c_i} - v\alpha) \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}[T_m^{(n)} = x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \qquad (2.12)\\ &= \frac{(-1)^{x+j}}{(\theta+n)_{m\uparrow 1}} \sum_{v=0}^{j} \sum_{y=0}^{x} (-1)^{y-x} \binom{v}{x-y-(j-v)} \theta^y \\ &\quad \times \sum_{(c_1,\ldots,c_j)\in\mathcal{C}_{j,v}} |s(m, y; n - \sum_{i=1}^{v} n_{c_i})|, \end{aligned}$$

respectively. Moment formulae for $T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ can be derived from Proposition 2.1 and by means of well-known relationships between falling factorial moments and moments.

3 Asymptotics for the conditional number of blocks

We start our conditional asymptotic analysis by establishing a fluctuation limit, as m tends to infinity, for $T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$. First, recall that $T_m^{(n)} = R_m^{(n)} + K_m^{(n)}$. For any $\alpha \in (0,1)$ and $\theta > -\alpha$, $\lim_{m\to+\infty} n^{-\alpha} R_m^{(n)} \,|\, (K_n, \mathbf{N}_n) = 0$ almost surely. Hence, the fluctuation limit for $T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ reduces to the fluctuation limit for $K_m^{(n)} \,|\, K_n$; such a fluctuation limit was established in Proposition 2 in Favaro et al. [6]. Similarly, for $\alpha = 0$ and $\theta > 0$ one has $\lim_{m\to+\infty} (\log m)^{-1} R_m^{(n)} \,|\, (K_n, \mathbf{N}_n) = 0$ almost surely and, furthermore, $K_m^{(n)}$ is independent of K_n. Hence, the fluctuation limit for $K_m^{(n)}$ coincides with the fluctuation limit for K_n in (1.4). For any $a, b > 0$ let $B_{a,b}$ a random variable distributed according to a Beta distribution with parameter (a, b). Then, we can state the following theorem.

Theorem 3.1. *Let $S_{\alpha,\theta}^{(n,j)}$ be the product of independent random variables $S_{\alpha,\theta+n}$ and $B_{j+\theta/\alpha, n/\alpha-j}$. Then*

- *for any $\alpha \in (0,1)$ and $\theta > -\alpha$*

$$\lim_{m\to+\infty} \frac{T_m^{(n)}}{m^\alpha} \,|\, (K_n = j, \mathbf{N}_n = \mathbf{n}) = S_{\alpha,\theta}^{(n,j)} \qquad \text{a.s.} \qquad (3.1)$$

- *for $\alpha = 0$ and $\theta > 0$*

$$\lim_{m\to+\infty} \frac{T_m^{(n)}}{\log m} \,|\, (K_n = j, \mathbf{N}_n = \mathbf{n}) = \theta \qquad \text{a.s.} \qquad (3.2)$$

As for the unconditional fluctuation limits in (1.3) and (1.4), weak convergence versions of (3.1) and (3.2) can alternatively be derived from general asymptotic results for

urn models. See Proposition 16 in Flajolet et al. [12] and Theorem 5 in Janson [16] for details. For any $\alpha \in (0,1)$ and $\theta > -\alpha$, if $n = j = 0$ then we recover (1.3) as special case of (3.1). Note that the dependence on n and j in the limiting random variable $S_{\alpha,\theta}^{(n,j)}$ indicates a long lasting impact of the given initial sample $(X_1, \ldots, X_n)$ to fluctuations. Furthermore, it is clear from Theorem 3.1 that one has $\lim_{m\to+\infty} m^{-1} T_m^{(n)} \,|\, (K_n, \mathbf{N}_n) = 0$ almost surely. Hereafter we establish a large deviation principle associated with this limiting procedure.

The study of large deviations for $m^{-1} T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ reduces to the study of large deviations for $m^{-1} K_m^{(n)} \,|\, K_n$. Indeed note that $K_m^{(n)} \leq T_m^{(n)} \leq K_m^{(n)} + n$. Then by Corollary B.9 in Feng [9], $m^{-1} T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ and $m^{-1} K_m^{(n)} \,|\, K_n$ satisfy the same large deviation principle. As in Feng and Hoppe [10], we establish a large deviation principle for $m^{-1} K_m^{(n)} \,|\, K_n$ by studying the moment generating of $K_m^{(n)} \,|\, K_n$. For any $\lambda > 0$ let $x = 1 - \mathrm{e}^{-\lambda}$ and let

$$G_{K_m^{(n)}}(x; \alpha, \theta) = \mathbb{E}\left[\left(\frac{1}{1-x} \right)^{K_m^{(n)}} \,|\, K_n = j \right] \tag{3.3}$$

be the moment generating function $K_m^{(n)} \,|\, K_n$. We focus on $\alpha \in (0,1)$ and $\theta > 0$. For $\alpha = 0$ and $\theta > 0$ the random variables $K_m^{(n)}$ and K_n are independent and, therefore, the large deviation principle for $m^{-1} K_m^{(n)}$ coincides with the large deviation principle for $n^{-1} K_n$ recalled in the Introduction. We start with two lemmas on the moment generating function $G_{K_m^{(n)}}$.

Lemma 3.2. *Let $(X_1, \ldots, X_n)$ be a sample from $\tilde{P}_{\alpha,\theta,\nu}$ featuring $K_n = j$ blocks. Then, for any $\alpha \in (0,1)$ and $\theta > -\alpha$*

$$G_{K_m^{(n)}}(x; \alpha, \theta) = (1-x)^{j+\frac{\theta}{\alpha}} \sum_{v \geq 0} \frac{x^v}{v!} \left(j + \frac{\theta}{\alpha} \right)_{v\uparrow 1} \frac{\binom{n+\theta+v\alpha+m-1}{n+\theta+m-1}}{\binom{n+\theta+v\alpha-1}{n+\theta-1}}.$$

Proof. The proof reduces to a straightforward application of Proposition 2.1. Indeed, the right-hand side of (3.3) can be expanded in terms of falling factorial moments of $K_m^{(n)} \,|\, K_n$, i.e.,

$$\begin{aligned} G_{K_m^{(n)}}(x; \alpha, \theta) &= \sum_{i \geq 0} x^i \mathbb{E}\left[\binom{i + K_m^{(n)} - 1}{K_m^{(n)} - 1} \,|\, K_n = j \right] \\ &= \sum_{i \geq 0} \frac{x^i}{i!} \sum_{l=0}^{i} |s(i,l,0)| \sum_{t=0}^{l} S(l,t,0) \mathbb{E}[(K_m^{(n)})_{t\downarrow 1} \,|\, K_n = j] \end{aligned} \tag{3.4}$$

where the falling factorial moment $\mathbb{E}[(K_m^{(n)})_{t\downarrow 1} \,|\, K_n = j]$ is read in (2.4) with $r = t$ at the index $i = 0$. Then, by Equation 2.60 and Equation 2.15 in Charalambides [3], we can write

$$\begin{aligned} &\sum_{l=0}^{i} |s(i,l,0)| \sum_{t=0}^{l} S(l,t,0) \mathbb{E}[(K_m^{(n)})_{t\downarrow 1} \,|\, K_n = j] \\ &\quad = \sum_{v=0}^{i} (-1)^{i-v} \left(j + \frac{\theta}{\alpha} \right)_{v\uparrow 1} \frac{(\theta + n + v\alpha)_{m\uparrow 1}}{(\theta + n)_{m\uparrow 1}} \\ &\qquad \times \sum_{l=v}^{i} (-1)^{i-l} |s(i,l,0)| S\left(l, v; j + \frac{\theta}{\alpha} \right) \end{aligned} \tag{3.5}$$

$$= \sum_{v=0}^{i} \binom{i}{v} \left(j + \frac{\theta}{\alpha}\right)_{v\uparrow 1} \left(-j - \frac{\theta}{\alpha}\right)_{(i-v)\uparrow 1} \frac{(\theta + n + v\alpha)_{m\uparrow 1}}{(\theta + n)_{m\uparrow 1}},$$

where the last equality is obtained by means of Equation (2.57) in Charalambides [3]. The proof is completed by combining (3.4) with (3.5) and by standard algebraic manipulations. □

Lemma 3.3. *For any $\alpha \in (0,1)$ and $\theta = 0$,*

$$\limsup_{m\to+\infty} \frac{1}{m} \log G_{K_m^{(n)}}(x;\alpha,0) \leq \Lambda_\alpha(\lambda) = \begin{cases} -\log(1-(1-e^{-\lambda})^{\frac{1}{\alpha}}) & \text{if } \lambda > 0 \\ 0 & \text{otherwise .} \end{cases}$$

Proof. Let $(a_n)_{n\geq 1}$ be a sequence of increasing positive numbers satisfying $a_n/n \to 1$ as $n \to +\infty$. Then we can find two increasing sequences of positive integers, say $(b_n)_{n\geq 1}$ and $(c_n)_{n\geq 1}$, such that $b_n \leq a_n \leq c_n$ and $\lim_{n\to+\infty} b_n/n = \lim_{n\to+\infty} c_n/n = 1$. Then, by combining Lemma 3.1 with Equation (3.5) in Feng and Hoppe [10], for any $0 < \alpha$ and $x < 1$ one obtains

$$\lim_{n\to+\infty} \frac{1}{n} \log \sum_{i\geq 0} x^i \frac{\Gamma(a_n + \alpha i)}{\Gamma(a_n)\Gamma(\alpha i + 1)} = \Lambda_\alpha(-\log(1-x)). \tag{3.6}$$

Consider the moment generating function $G_{K_m^{(n)}}(x;0,\alpha)$. Direct calculations one obtains the identity

$$\frac{(j)_{v\uparrow 1}}{v!\binom{n+v\alpha-1}{n-1}} = \frac{(n-1)!}{(j-1)!} \frac{(v+1)_{(j-1)\uparrow 1}}{(v\alpha+1)_{(n-1)\uparrow 1}} = C_0(n,j,\alpha,v) v^{j-n}, \tag{3.7}$$

where $C_0(n,j,\alpha,v)$ is uniformly bounded in v from above and below by positive constants. Then,

$$\begin{aligned} &\limsup_{m\to+\infty} \frac{1}{m} \log G_{K_m^{(n)}}(x;\alpha,0) \\ &= \limsup_{m\to+\infty} \frac{1}{m} \log \left((1-x)^j \sum_{v\geq 0} C_0(n,j,\alpha,v) v^{j-n} x^v \binom{n+v\alpha+m-1}{n+m-1} \right) \\ &\leq \limsup_{m\to+\infty} \frac{1}{n+m} \log \sum_{v\geq 0} x^v \binom{n+v\alpha+m-1}{n+m-1} = \Lambda_\alpha(\lambda) \end{aligned}$$

where the last equality is obtained by a direct application of (3.6). The proof is completed. □

Proposition 2.1, Lemma 3.2 and Lemma 3.3 are exploited in order to derive the large deviation principle for $m^{-1}K_m^{(n)} \,|\, K_n$ and, hence, for $m^{-1}T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$. We can state the following theorem.

Theorem 3.4. *For any $\alpha \in (0,1)$ and $\theta > -\alpha$, $m^{-1}T_m^{(n)} \,|\, (K_n, \mathbf{N}_n)$ satisfies a large deviation principle with speed m and rate function I^α in* (1.5)*. For $\alpha = 0$ and $\theta > 0$, $(\log m)^{-1}(T_m^{(n)} \,|\, K_n, \mathbf{N}_n)_{m\geq 1}$ satisfies a large deviation principle with speed $\log m$ and rate function I_θ in* (1.6)*.*

Proof. We only need to prove the large deviation principle for $m^{-1}K_m^{(n)} \,|\, K_n$ with $\alpha \in (0,1)$ and $\theta > -\alpha$. Let us consider the first moment of $K_m^{(n)} \,|\, K_n$. Such a moment is read in (2.6) with $r = 1$ at the index $i = 0$. Specifically one has $\mathbb{E}[K_m^{(n)} \,|\, K_n = j] =$

$(j+\theta/\alpha)\left((\theta+n+\alpha)_{m\uparrow 1}/(\theta+n)_{m\uparrow 1}-1\right)=O(m^{\alpha})$. For any $\lambda\leq 0$, Jensen's inequality leads to

$$\begin{aligned}0&\geq \lim_{m\to+\infty}\frac{1}{m}\log\mathbb{E}[e^{\lambda K_m^{(n)}}\,|\,K_n=j] \\ &\geq \lim_{m\to+\infty}\frac{1}{m}\mathbb{E}[\lambda K_m^{(n)}\,|\,K_n=j]\geq\lim_{m\to+\infty}\frac{\lambda m^{\alpha}}{m}=0.\end{aligned} \tag{3.8}$$

For any $\lambda>0$, we start by considering $G_{K_m^{(n)}}(x;\alpha,0)$ and then we move to the general case $\theta>-\alpha$. For any $n\geq 1$ and $1\leq j\leq n$ let $H_m(x;\alpha,0)=1+\sum_{v\geq 1}x^v v^{j-n}\binom{n+v\alpha+m-1}{n+m-1}$. If $n=j$, then $H_m(x;\alpha,0)$ can be estimated as in (3.6). On the other hand, for $n>j$ the $(n-j)$-th order derivative of $H_m(x;\alpha,0)$ with respect to x coincides with the following expression

$$\begin{aligned}H_m^{(n-j)}(x;\alpha,0)&=\frac{\mathrm{d}^{n-j}}{\mathrm{d}x^{n-j}}H_m(x;\alpha,0)\\ &=\sum_{v\geq n-j}x^{v-(n-j)}(n-j)!(v)_{(n-j)\downarrow 1}v^{j-n}\binom{n+v\alpha+m-1}{n+m-1}\\ &\geq g(n,j)\sum_{v\geq 0}x^v\binom{n+v\alpha+m-1}{n+m-1}-g(n,j)\sum_{v=0}^{n-j-1}\binom{n+v\alpha+m-1}{n+m-1}\end{aligned}$$

where $g(n,j)=(n-j)!/(n-j)^{n-j}$. For $x\in(0,1)$ and $\varepsilon\in(0,x)$, integrating $(n-j)$ times over $(0,x)$ lead to

$$\begin{aligned}H_m(x;\alpha,0)&\geq H_m(x;\alpha,0)-\sum_{i=0}^{n-j-1}\frac{H_m^{(i)}(0;\alpha,0)}{i!}x^i\\ &=\int_0^x\int_0^{x_1}\cdots\int_0^{x_{n-j-1}}H_m^{(n-j)}(y;\alpha,0)\mathrm{d}y\cdots\mathrm{d}x_1\\ &\geq\int_{x-\varepsilon}^x\int_{x-\varepsilon}^{x_1}\cdots\int_{x-\varepsilon}^{x_{n-j-1}}H_m^{(n-j)}(y;\alpha,0)\mathrm{d}y\cdots\mathrm{d}x_1\\ &\geq g(n,j)\varepsilon^{n-j}\sum_{v\geq 0}x^v\binom{n+v\alpha+m-1}{n+m-1}-g(n,j)\varepsilon^{n-j}\sum_{v=0}^{n-j-1}\binom{n+v\alpha+m-1}{n+m-1}\end{aligned}$$

where we used the monotonicity of the function $\sum_{v\geq 0}x^v\binom{n+v\alpha+m-1}{n+m-1}$ in the last inequality. Also, since

$$\lim_{m\to+\infty}\frac{1}{m}\log\left(g(n,j)\sum_{v=0}^{n-j-1}\binom{n+m+v\alpha-1}{n+m-1}\right)=0$$

it follows

$$\begin{aligned}\liminf_{m\to+\infty}\frac{1}{m}\log G_{K_m^{(n)}}(x;\alpha,0)&=\liminf_{m\to+\infty}\frac{1}{m}\log H_m(x;\alpha,0)\\ &\geq\liminf_{m\to+\infty}\frac{1}{m}\log\sum_{v\geq 0}x^v\binom{n+m+v\alpha-1}{n+m-1}=\Lambda_{\alpha}(\lambda).\end{aligned}$$

Accordingly, by means of Lemma 3.3, the proof is completed for the case of $\alpha\in(0,1)$ and $\theta=0$. Now we consider the case $\theta\neq 0$. By means of arguments similar to (3.7) we can write

$$\frac{\left(j+\frac{\theta}{\alpha}\right)_{v\uparrow 1}}{v!\binom{n+\theta+v\alpha-1}{n+\theta-1}}=C_{\theta}(n,j,\alpha,v)v^{j+\frac{\theta}{\alpha}-n}$$

where $C_\theta(n,j,\alpha,v)$ is uniformly bounded in v from above and below and it has a strict positive lower bound. Accordingly, by choosing an ε small and two positive constants c_1 and c_2 such that $xe^{\varepsilon} < 1$ and such that $c_1 e^{-\varepsilon v} \leq C_\theta(n,j,\alpha,v) v^{j+\frac{\theta}{\alpha}-n} \leq c_2 e^{\varepsilon v}$, it follows that

$$\begin{aligned} & c_2(1-x)^{j+\frac{\theta}{\alpha}} \sum_{v\geq 0} (xe^{-\varepsilon})^v \binom{n+\theta+v\alpha+m-1}{n+\theta+m-1} \\ & \leq G_{K_m^{(n)}}(x;\alpha,\theta) \leq c_3(1-x)^{j+\frac{\theta}{\alpha}} \sum_{v\geq 0} (xe^{\varepsilon})^v \binom{n+\theta+v\alpha+m-1}{n+\theta+m-1}. \end{aligned} \tag{3.9}$$

The proof in completed by letting $m \to +\infty$ and $\varepsilon \to 0$ in (3.9). Indeed by taking these limits we obtain $\lim_{m\to+\infty} \log G_{K_m^{(n)}}(x;\alpha,\theta) = \Lambda_\alpha(\lambda)$, which combined with (3.8), implies $\lim_{m\to+\infty} m^{-1} \log \mathbb{E}[e^{\lambda K_m^{(n)}} \mid K_n = j] = \Lambda_\alpha(\lambda)$. Then, the large deviation principle for $m^{-1}K_m^{(n)} \mid K_n$ follows by Gärtner-Ellis theorem. See Dembo and Zeitouni [4] for details. □

According to Theorem 3.4, K_n and its conditional counterparts $T_m^{(n)} \mid (K_n, \mathbf{N}_n)$ behave the same in terms of large deviations. However, in terms of fluctuation limits, Theorem 3.1 shows that the initial sample $(X_1,\ldots,X_n)$ has a long lasting effect. This is caused by the two different scalings involved, namely m^{-1} for large deviations and $m^{-\alpha}$ for the fluctuations. Since the given initial sample leads to an estimation on the parameters, one would expect that the large deviation results will be dramatically different if the sample size n is allowed to grow and leads to large parameters. This kind of behaviour is discussed in Feng [8] 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.

Note that, 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 $m^{-1}T_m^{(n)} \mid (K_n, \mathbf{N}_n)$ may not be easily derived from that of $m^{-1}K_m^{(n)} \mid K_n$ by a direct comparison argument. Hence, it is helpful to study the moment generating of $T_m^{(n)} \mid (K_n, \mathbf{N}_n)$, namely

$$G_{T_m^{(n)}}(x;\alpha,\theta) = \mathbb{E}\left[\left(\frac{1}{1-x}\right)^{T_m^{(n)}} \mid K_n = j, \mathbf{N}_n = \mathbf{n}\right]. \tag{3.10}$$

We intend to pursue this study further in a subsequent project. Here, we conclude by providing an explicit expression for (3.10). As in Lemma (3.2), this expression follows by applying Proposition 2.1.

Lemma 3.5. *Let $(X_1,\ldots,X_n)$ be a sample from $\tilde{P}_{\alpha,\theta,\nu}$ featuring $K_n = j$ blocks with frequencies $\mathbf{N}_n = \mathbf{n}$. Then*

i) for any $\alpha \in (0,1)$ and $\theta > -\alpha$

$$\begin{aligned} & G_{T_m^{(n)}}(x;\alpha,\theta) \\ & = (1-x)^{\frac{\theta}{\alpha}} \sum_{v=0}^{j} (-x)^v \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \sum_{l\geq 0} \frac{x^l}{l!} \left(j+\frac{\theta}{\alpha}\right)_{l\uparrow 1} \frac{\binom{n-\sum_{i=1}^v n_{c_i}+\theta+v\alpha+l\alpha+m-1}{n+\theta+m-1}}{\binom{n-\sum_{i=1}^v n_{c_i}+\theta+v\alpha+l\alpha-1}{n+\theta-1}}; \end{aligned}$$

ii) for $\alpha = 0$ and $\theta > 0$

$$G_{T_m^{(n)}}(x;0,\theta)$$

$$\frac{1}{(1-x)^j}\sum_{v=0}^{j}(-1)^v \sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}} \frac{\binom{n-\sum_{i=1}^{v} n_{c_i}+\frac{\theta}{1-x}+m-1}{n+\theta+m-1}}{\binom{n-\sum_{i=1}^{v} n_{c_i}+\frac{\theta}{1-x}-1}{n+\theta-1}};$$

where we defined $\mathcal{C}_{j,0}=\emptyset$ *and* $\mathcal{C}_{j,v}=\{(c_1,\ldots,c_v): c_k\in\{1,\ldots,j\}, c_k\neq c_\ell,\ \text{if } k\neq\ell\}$ *for any* $v\geq 1$.

Proof. We expand the right-hand side of (3.10) in terms of falling factorial moments of $T_m^{(n)}\,|\,(T_n,\mathbf{N}_n)$ and we apply Proposition 2.1 in which an expression for these moments is given. Specifically,

$$\begin{aligned} G_{T_m^{(n)}}(x;\alpha,\theta) &= \sum_{i\geq 0} x^i \mathbb{E}\left[\binom{i+T_m^{(n)}-1}{T_m^{(n)-1}} \,|\, K_n=j,\mathbf{N}_n=\mathbf{n}\right] \\ &= \sum_{i\geq 0}\frac{x^i}{i!}\sum_{l=0}^{i}|s(i,l,0)|\sum_{t=0}^{l}S(l,t,0)\mathbb{E}[(T_m^{(n)})_{t\downarrow 1}\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}]. \end{aligned} \tag{3.11}$$

For any $\alpha\in(0,1)$ and $\theta>-\alpha$ the falling factorial moment $\mathbb{E}[(T_m^{(n)})_{t\downarrow 1}\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}]$ is read in (2.4) with $r=t$. Then, by Equation 2.60 and Equation 2.15 in Charalambides [3], we can write

$$\begin{aligned} &\sum_{l=0}^{i}|s(i,l,0)|\sum_{t=0}^{l}S(l,t,0)\mathbb{E}[(T_m^{(n)})_{t\downarrow 1}\,|\,K_n=j,\mathbf{N}_n=\mathbf{n}] \\ &\quad=\frac{1}{(\theta+n)_{m\uparrow 1}}\sum_{l=0}^{i}\sum_{v=0}^{i-l}(-1)^v\left(j+\frac{\theta}{\alpha}\right)_{l\uparrow 1}\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}(\theta+n-\sum_{i=1}^{v}n_{c_i}+v\alpha+l\alpha)_{m\uparrow 1} \\ &\qquad\times\sum_{w=v}^{i-l}w!\binom{i}{i}\binom{i-w}{l}\binom{w+j-v-1}{w-v}\left(-j-\frac{\theta}{\alpha}\right)_{(i-w-l)\uparrow 1} \\ &=\frac{1}{(\theta+n)_{m\uparrow 1}}\sum_{l=0}^{i}\sum_{v=0}^{i-l}(-1)^v\left(j+\frac{\theta}{\alpha}\right)_{l\uparrow 1} \\ &\qquad\times\frac{i!}{l!(i-l-v)!}\left(-\frac{\theta}{\alpha}\right)_{(i-l-v)\uparrow 1}\sum_{(c_1,\ldots,c_v)\in\mathcal{C}_{j,v}}(\theta+n-\sum_{i=1}^{v}n_{c_i}+v\alpha+l\alpha)_{m\uparrow 1}. \end{aligned} \tag{3.12}$$

The proof of $i)$ is completed by combining (3.11) with (3.12) and by standard algebraic manipulations. Finally, for $\alpha=0$ and $\theta>0$ the result in ii) follows by exploiting similar arguments. □

4 Discussion

Our results contribute to the study of conditional properties of exchangeable random partitions induced by the Ewens-Pitman sampling model. While focusing on the number $K_m^{(n)}$ of new blocks generated by the additional sample, Lijoi et al. [19] left open the problem of studying the total number $T_m^{(n)}$ of blocks generated by the additional sample. In this paper we presented a comprehensive analysis of distributional properties of $T_m^{(n)}\,|\,(K_n,\mathbf{N})$ for a finite sample size m and for large m. Hereafter we briefly discuss our results with a view toward Bayesian nonparametric inference for species sampling problems.

As pointed out in the Introduction, the distribution of $T_m^{(n)}\,|\,(K_n,\mathbf{N})$ takes on the interpretation of the posterior distribution of the number of species generated by the additional sample, given an initial observed sample featuring K_n species with frequencies

$\mathbf{N}_n$. Accordingly, the corresponding Bayesian nonparametric estimator, with respect to a squared loss function, are recovered from (2.4) and (2.5) by setting $r = 1$. Then, one obtains

$$\begin{aligned}&\mathbb{E}[T_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &\quad = j - \frac{\sum_{i=1}^n m_i(\theta + n - i + \alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} + \left(j + \frac{\theta}{\alpha}\right)\left(\frac{(\theta+n+\alpha)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} - 1\right)\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}&\mathbb{E}[T_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \\ &\quad = j - \frac{\sum_{i=1}^n m_i(\theta + n - i)_{m\uparrow 1}}{(\theta+n)_{m\uparrow 1}} + \sum_{i=1}^m \frac{\theta}{\theta + n + i - 1}\end{aligned} \tag{4.2}$$

respectively, where m_i denotes the number of distinct species with frequency i in $\mathbf{n}$. In particular, Theorem 3.4 shows that probabilities of "large" deviations away from the point estimations (4.1) and (4.2) decay exponentially with rate functions I^α and I_θ, respectively. Formulae (4.1) and (4.2) generalize the Bayesian nonparametric estimator for the number of new species generated by the additional sample. See Favaro et al. [6] for details.

Besides point estimators (4.1) and (4.2), one would also like to determine highest posterior density (HPD) intervals since they provide a measure of uncertainty on the point estimates. The problem of determining HPD intervals for $\mathbb{E}[T_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$ reduces to the problem of evaluating the distribution of $T_m^{(n)} \,|\, (K_n, \mathbf{N})$. An explicit expression for this distribution has been determined in (2.11) and (2.12). Then a simulation algorithm can be implemented in order to evaluate quantiles for determining HPD intervals of $\mathbb{E}[T_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$. There are, however, situations of practical interest where j, n and m are very large and the computational burden for evaluating the posterior distributions (2.11) and (2.12) becomes overwhelming. This happens, for instance, in several genomic applications where one has to deal with relevant portions of complementary DNA libraries which typically consist of millions of genes. To overcome this drawback we can exploit Theorem 3.1. Indeed, for instance, for $\alpha \in (0,1)$ and $\theta > 0$ one has

$$\mathbb{P}[T_m^{(n)} = x \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}] \approx \mathbb{P}[S_{\alpha,\theta}^{(n,j)} = m^\alpha x]$$

for a large m. Then, resorting the simulation algorithm for $S_{\alpha,\theta}^{(n,j)}$ developed in Favaro et al. [6], we can evaluate appropriate quantiles of the limiting posterior distributions in order to obtain an approximate evaluation of HPD credible sets for $\mathbb{E}[T_m^{(n)} \,|\, K_n = j, \mathbf{N}_n = \mathbf{n}]$.

In this paper we focused on distributional properties of $T_m^{(n)} \,|\, (K_n, \mathbf{N})$ under the Ewens-Pitman sampling model. A natural extension of our results consists in considering more general sampling models. With this regards, a noteworthy generalization of the Ewens-Pitman sampling model is the so-called Gibbs-type sampling model introduced by Gnedin and Pitman [13]. Specifically, let $\alpha \in (-\infty, 1)$ and let $V = (V_{n,j})_{j \le n, n \ge 1}$ be a collection of nonnegative weights satisfying the recursion $V_{n,j} = V_{n+1,j+1} + (n - j\alpha)V_{n+1,j}$, with $V_{1,1} = 1$. Then, a Gibbs-type sampling model with parameter (α, V, ν) is defined as follows

$$\mathbb{P}[X_{i+1} \in \cdot \,|\, X_1, \ldots, X_i] = \frac{V_{n+1,j+1}}{V_{n,j}}\nu(\cdot) + \frac{V_{n+1,j}}{V_{n,j}}\sum_{l=1}^j (n_l - \alpha)\delta_{X_l^*}(\cdot) \tag{4.3}$$

for any $i \ge 1$, with $X_1^*, \ldots, X_j^*$ being the j distinct observations in $(X_1, \ldots, X_i)$ with frequencies $\mathbf{n} = (n_1, \ldots, n_j)$. If $V_{n,j} = \prod_{i=0}^{j-1}(\theta + i\alpha)/(\theta)_{n\uparrow 1}$ then (4.3) reduces to (1.1).

Under the Gibbs-type sampling model with $\alpha \in (0,1)$, we derived an explicit expression for the distribution of $T_m^{(n)} \,|\, (K_n, \mathbf{N})$ and a fluctuation limit as m tends to infinity. The corresponding unconditional results for K_n are known from Gnedin and Pitman [13] and Pitman [25]. Work on unconditional and conditional large deviation principles is ongoing. For any $\alpha \in (0,1)$ our conjecture is that $n^{-1}K_n$ and $m^{-1}T_m^{(n)} \,|\, (K_n, \mathbf{N})$ satisfies a large deviation principle with speed n and m, respectively, and with the same rate function I^α in (1.5). In other words, we conjectured that large deviation principles for $n^{-1}K_n$ and $m^{-1}T_m^{(n)} \,|\, (K_n, \mathbf{N})$ are invariant in the class of Gibbs-type sampling models with $\alpha \in (0,1)$.

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

[2] Bacallado, S., Favaro, S. and Trippa, L. (2013). Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli*, to appear.

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

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

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

[6] 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

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

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

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

[10] 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

[11] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230. MR-0350949

[12] 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

[13] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci* **138**, 5674–5685 MR-2160320

[14] 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

[15] Hoppe, F.M. (1984). Pólya-like urns and the Ewens sampling formula. *J. Math. Biol.*, **20**, 91–94. MR-0758915

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

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

[18] 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

[19] 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

[20] 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

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

[22] Pitman, J. (1995). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli*, **3**, 79–66. MR-1466546

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

[24] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed* (D.R. Goldstein, Ed.) *Lecture Notes Monograph Series* **40** 1-34. IMS, Beachwood, OH. MR-2004330

[25] 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

[26] Tavaré, S. (1987). The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.*, **25**, 161–168. MR-0896431

[27] Zabell, S.L. (1997). The continuum of inductive methods revisited. In *The cosmos of science: essays of exploration*, Earman, J. and Norton, J.D. University of Pittsburgh Press.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)

Economical model of EJP-ECP

- Low cost, based on free software (OJS[1])
- Non profit, sponsored by IMS[2], BS[3], PKP[4]
- Purely electronic and secure (LOCKSS[5])

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]OJS: Open Journal Systems `http://pkp.sfu.ca/ojs/`
[2]IMS: Institute of Mathematical Statistics `http://www.imstat.org/`
[3]BS: Bernoulli Society `http://www.bernoulli-society.org/`
[4]PK: Public Knowledge Project `http://pkp.sfu.ca/`
[5]LOCKSS: Lots of Copies Keep Stuff Safe `http://www.lockss.org/`
[6]IMS Open Access Fund: `http://www.imstat.org/publications/open.htm`