J. Appl. Probab. **59,** 797–824 (2022)
*doi:*10.1017/jpr.2021.87

ASYMPTOTICS OF THE OVERFLOW IN URN MODELS

RAUL GOUET,* *Universidad de Chile*
PAWEŁ HITCZENKO ,** *** *National Science Foundation and Drexel University*
JACEK WESOŁOWSKI,**** *Warsaw University of Technology*

Abstract

Consider a finite or infinite collection of urns, each with capacity r, and balls randomly distributed among them. An overflow is the number of balls that are assigned to urns that already contain r balls. When $r = 1$, this is the number of balls landing in non-empty urns, which has been studied in the past. Our aim here is to use martingale methods to study the asymptotics of the overflow in the general situation, i.e. for arbitrary r. In particular, we provide sufficient conditions for both Poissonian and normal asymptotics.

Keywords: Urn model; occupancy problem; random allocation; weak limit theorem

2020 Mathematics Subject Classification: Primary 60F05; 60K30
Secondary 60K35

1. Introduction

Urn models are one of the fundamental objects in classical probability theory and have been studied for a long time in various degrees of generality. We refer the reader to classical sources [12, 16, 17, 18] for a complete account of the theory and discussions of different models, and, e.g., to [4, 7, 9] for some of the more recent developments. Perhaps the most heavily studied characteristic is the number of occupied urns after n balls have been thrown in. One reason for this is that it is often interpreted as a measure of diversity of a given population. Actually, more refined characteristics, e.g. the number of urns containing the prescribed number of balls (and its asymptotics), have subsequently been studied for various urn models; see, e.g., [13, 16, 20], or [2] and references therein for more recent developments. In diversity analysis, the number K_r of urns with exactly r balls is called the abundance count of order r. In particular, the popular estimator of species richness called the Chao estimator is based on K_1 and K_2 (with a more sophisticated version also using K_3 and K_4); see, e.g., [5]. In [9] the authors used analytical methods based on Poissonization and de-Poissonization to prove that the number of empty urns is asymptotically normal as long as its variance grows to infinity (this is clearly the minimal requirement). As a by-product of their method they established the Poissonian

Received 1 February 2021; revision received 11 October 2021.
* Postal address: Departamento de Ingenieria Matemática and CMM (IRL 2807, CNRS), Universidad de Chile, Beauchef 851, 8370456 Santiago, Chile. Email: rgouet@dim.uchile.cl
** Postal address: Division of Mathematical Sciences, National Science Foundation, 2415 Eisenhower Avenue, Alexandria, VA 22314, USA.
*** Postal address: Department of Mathematics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. Email: ph33@drexel.edu
**** Postal address: Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, Warsaw, Poland. Email: j.wesolowski@mini.pw.edu.pl

asymptotics of the number of balls that fall into non-empty urns when the variance is finite and under additional assumptions on the distribution among boxes. We mention in passing that the number of balls falling into non-empty urns is sometimes called the number of collisions. Under the uniformity assumption for the distribution of balls, it has been used, for example, for testing random number generators (see [14, Section 3.3.2 I] for more details). We also refer to [1] and references therein for another illustration of how this concept is used, e.g. in cryptology.

Our main aim here is to consider the number of balls falling into urns already containing r balls (thus, the number of collisions corresponds to $r = 1$). Relying on martingale-type methods, we provide sufficient conditions for both Poissonian and normal asymptotics for the number of balls falling into such urns.

One way to formulate the problem is as follows. There is a collection (possibly infinite) of distinct containers in which balls are to be inserted. All containers have the same finite capacity r. Each arriving ball is to be placed in one of the containers, randomly and independently of other balls. However, if the container selected for a given ball is already full, the ball lands in the overflow basket. We are interested in the number of balls in that basket when more and more balls appear. The notion of overflow is not entirely new and has appeared, for example, in the context of collision resolution for hashing algorithms; see the discussion under 'External searching' in [15, Section 6.4]. We also refer to subsequent work [21, 23] for the computation of the probability that there is no overflow (under the uniformity assumption), and to [6], which, in part, concerns the estimation of the probability of unusually large overflow. As far as we are aware, however, the asymptotic behavior of the overflow has not been systematically investigated.

More precisely, we consider the following model. For any $n \geq 1$, let $X_{n,1}, \ldots, X_{n,n}$ be independent and identically distributed (i.i.d.) random variables with values in $M_n \subset \mathbb{N} := \{1, 2, \ldots\}$, and let $p_{n,m} = \mathbb{P}(X_{n,1} = m)$, $m \in M_n$, be the common distribution among the boxes for each of the n balls in the nth experiment. Here, $X_{n,j}$ is interpreted as a random label of the urn selected for the jth ball in the nth experiment. Also let

$$N_{n,k}(m) = \sum_{j=1}^{k-1} \mathbf{1}_{\{X_{n,j}=m\}} \tag{1}$$

for any $n \in \mathbb{N}$, $k \in \{1, \ldots, n, n+1\}$, and $m \in M_n$, where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator of the event within brackets. That is, $N_{n,k}(m)$ is the number of balls among the first $k-1$ balls for which the mth box was selected.

Let r be a given positive integer that denotes the (same) capacity of every container. Then

$$Y_{n,k} = \sum_{m \in M_n} \mathbf{1}_{\{X_{n,k}=m\}} \mathbf{1}_{\{N_{n,k}(m) \geq r\}} \tag{2}$$

is 1 if the kth ball lands in the overflow, and is 0 otherwise. Naturally, $Y_{n,k} = 0$ for $k = 1, \ldots, r$. Consequently, the size of the overflow, denoted $V_{n,r}$, can be written as

$$V_{n,r} = \sum_{k=1}^{n} Y_{n,k}. \tag{3}$$

We are interested in the asymptotic distribution of $V_{n,r}$ as $n \to \infty$. We show that there are regimes related to $p_{n,m}$ under which the limiting distribution of $V_{n,r}$ (possibly standardized)

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

is either Poisson or normal. These regimes are defined through the limiting behavior of the sequences np_n^* and $n^{r+1}\sum_{m\in M_n} p_{n,m}^{r+1}$, where $p_n^* = \sup_{m\in M_n} p_{n,m}$.

We find it convenient to introduce auxiliary sequences of random variables X_n, Y_n, $n \geq 1$, such that, for any $n \in \mathbb{N}$, the random variables X_n, Y_n, $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. This allows us to simplify expressions in general because sums over $m \in M_n$ can be represented as expectations, and the computations are compactly carried out by means of conditional expectations. For example, $\sum_{m\in M_n} p_{n,m}^{r+1} = \mathbb{E}p_{X_n}^r$, where p_{X_n} stands for the random variable p_{n,X_n}. Convergence in probability and in distribution (as $n \to \infty$) are denoted by $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathrm{d}}$, respectively.

We present our main results in Section 2; these give conditions for Poissonian and Gaussian asymptotics of the overflow. We also describe the limiting behavior for the number of full containers. Intermediate technical results are found in Section 3, and the proofs of the main theorems are presented in Section 4. Section 5 includes some remarks concerning the asymptotic behavior of the mean of $V_{n,r}$.

2. Main results

2.1. Poissonian asymptotics

Theorem 1. *Let* Pois(μ) *denote the Poisson distribution with parameter* $\mu \in (0, \infty)$. *If*

$$n^{r+1}\mathbb{E}\, p_{X_n}^r \to (r+1)!\mu, \tag{4}$$

$$n\, p_n^* \to 0, \tag{5}$$

then $V_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$.

Proof. See Section 4.1. □

Example 1. Consider the uniform case, that is, $p_{n,j} = 1/m_n$, for $j \in M_n = \{1, \ldots, m_n\}$. Then

$$np_n^* = \frac{n}{m_n}, \qquad n^{r+1}\mathbb{E}\, p_{X_n}^r = \frac{n^{r+1}}{m_n^r}. \tag{6}$$

Take $m_n = \left\lfloor an^{\frac{r+1}{r}} \right\rfloor$, $a > 0$. Then, by (6), $np_n^* \to 0$ and $n^{r+1}\mathbb{E}\, p_{X_n}^r \to 1/a^r$. Consequently, Theorem 1 yields $V_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$, with $\mu = 1/(a^r(r+1)!)$. Illustrative simulations are shown in Fig. 1.

Example 2. Consider the geometric case, with $p_{n,j} = p_n(1-p_n)^j, j \in M_n = \{0, 1, 2, \ldots\}$. Then

$$np_n^* = np_n, \qquad n^{r+1}\mathbb{E}\, p_{X_n}^r = \frac{(np_n)^{r+1}}{1-(1-p_n)^{r+1}}. \tag{7}$$

Take $p_n = a/n^{(r+1)/r}$, $a > 0$. Then, by (7), $np_n^* = a/n^{1/r} \to 0$ and

$$n^{r+1}\mathbb{E}\, p_{X_n}^r = \frac{a^r p_n}{(r+1)p_n + o(p_n)} \to \frac{a^r}{r+1}.$$

Consequently, Theorem 1 yields $V_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$, with $\mu = a^r/[(r+1)!(r+1)]$. Illustrative simulations are shown in Fig. 2.

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

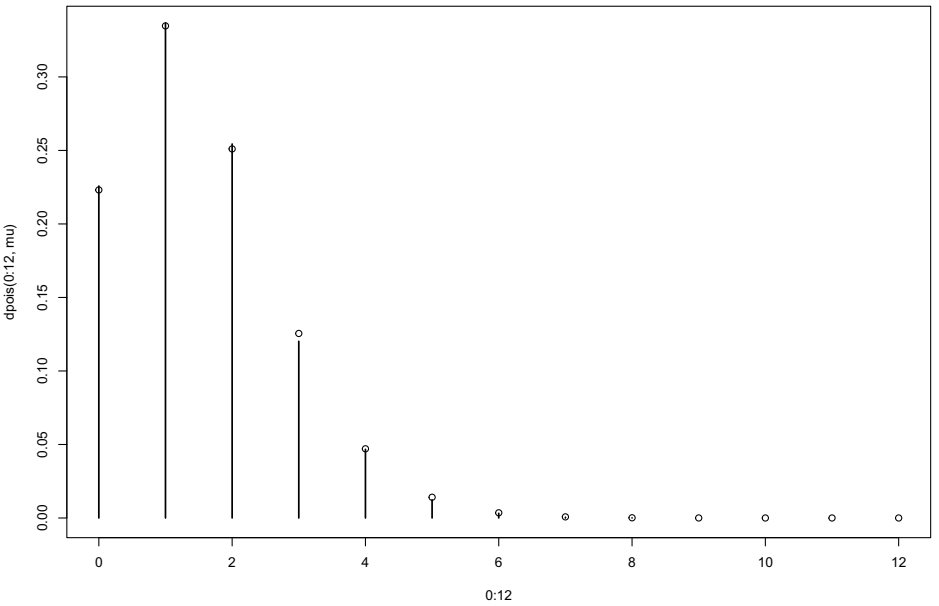

FIGURE 1. Simulations of the overflow in the uniform case with $r = 2$, $n = 10^5$, $m_n = \left\lfloor an^{\frac{r+1}{r}} \right\rfloor$, and $a = 1/3$ (i.e. $m_{10^5} = 10\,540\,925$ and $\mu = 1.5$) are shown as vertical lines (10^4 repetitions), while Poisson probabilities for $k = 0, \ldots, 12$, dpois$(0:12, \mu)$, are depicted by circles.

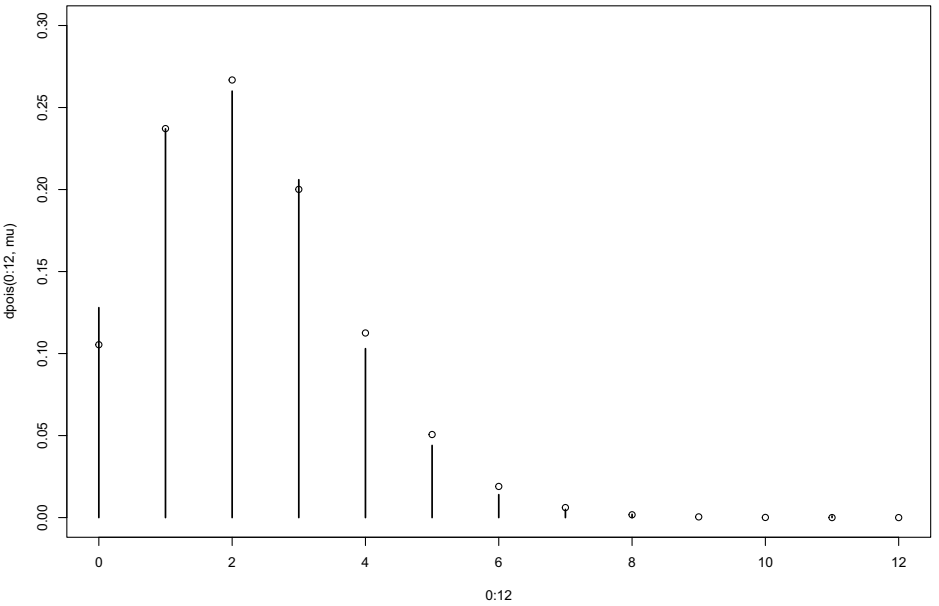

FIGURE 2. Simulations of the overflow in the geometric case with $r = 3$, $n = 10^5$, $p_n = an^{-(r+1)/r}$ with $a = 6$ (i.e. $p_{10^5} \approx 1.29 \times 10^{-6}$ and $\mu = 2.25$) are shown as vertical lines (10^3 repetitions), while Poisson probabilities for $k = 0, \ldots, 12$, dpois$(0:12, \mu)$, are depicted by circles.

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Hwang and Janson [9] used the method of Poissonization and de-Poissonization to establish the asymptotic normality for the number of occupied boxes under the weakest possible assumption that its variance tends to infinity. As a by-product of their approach they derived Theorem 1 for $r = 1$ (see [9, Theorem 8.2]). The proof we present in Section 4.1 is entirely different and relies on a martingale-type convergence result from [3].

2.2. Normal asymptotics

Theorem 2. *Assume that np_n^* is bounded, and that $n^{r+1}\mathbb{E}\, p_{X_n}^r \to \infty$. Then*

$$\frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\operatorname{Var} V_{n,r}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,\, 1).$$

Proof. See Section 4.2. □

The boundedness of np_n^* may be interpreted as the asymptotic negligibility condition, which is a natural requirement for central limit theorem (CLT) type results. The assumption $n^{r+1}\mathbb{E}\, p_{X_n}^r \to \infty$ is to ensure that the variance $\operatorname{Var} V_{n,r}$ grows to infinity with n, a necessary condition for the CLT. In Proposition 1 we show that $\mathbb{E}\, V_{n,r}$ and $\operatorname{Var} V_{n,r}$ are of order $n^{r+1}\mathbb{E}\, p_{X_n}^r$.

Proposition 1. *Assume that np_n^* is bounded and let $\lambda = \limsup np_n^* \geq 0$. Then*

$$\frac{\Gamma_\lambda(r+1)}{r!} \leq \liminf \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \leq \limsup \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \leq \frac{1}{(r+1)!}, \tag{8}$$

where, for $p > 0$ and $x \geq 0$ we have set $\Gamma_x(p) := \int_0^1 t^{p-1}\mathrm{e}^{-xt}\,\mathrm{d}t$. If, in addition, $n^{r+1}\mathbb{E}\, p_{X_n}^r \to \infty$, then

$$\frac{\mathrm{e}^{-2\lambda}}{(r+1)!} \leq \liminf \frac{\operatorname{Var} V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \leq \limsup \frac{\operatorname{Var} V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \leq \frac{1}{r!}. \tag{9}$$

Proof. See Section 4.2.2. □

Remark 1. Since $\Gamma_0(p) = 1/p$, it follows immediately that if $\lambda = 0$ then the limit of $\mathbb{E}\, V_{n,r}/\left[n^{r+1}\mathbb{E}\, p_{X_n}^r\right]$ in (8) exists and equals $1/(r+1)!$, regardless of the values of $(p_{n,k})$. As we will see in Section 5, the situation is more complex when $\lambda > 0$.

Example 3. Consider the uniform case, with $p_{n,j} = 1/m_n$, $j \in M_n = \{1, \ldots, m_n\}$. Then $n^{r+1}\mathbb{E}\, p_{X_n}^r = n^{r+1}/m_n^r \to \infty$ and $np_n^* = n/m_n \to \lambda \geq 0$. Thus, by Theorem 2,

$$\frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\operatorname{Var} V_{n,r}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,\, 1).$$

In particular, $m_n = \lfloor \kappa n^a \rfloor$ with $a \in [1,\ 1 + r^{-1})$ and $\kappa > 0$ yields normal asymptotics. Illustrative simulations are shown in Fig. 3.

Example 4. Consider the geometric case, with $p_{n,j} = p_n(1-p_n)^j$, $j \in M_n = \{0, 1, 2, \ldots\}$, and $p_n = 1/n^a$, $a \in [1,\ 1 + r^{-1})$. Then (7) yields

$$n^{r+1}\mathbb{E}\, p_{X_n}^r = \frac{n^{r+1-ra}}{r + 1 + o(1)} \to \infty.$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

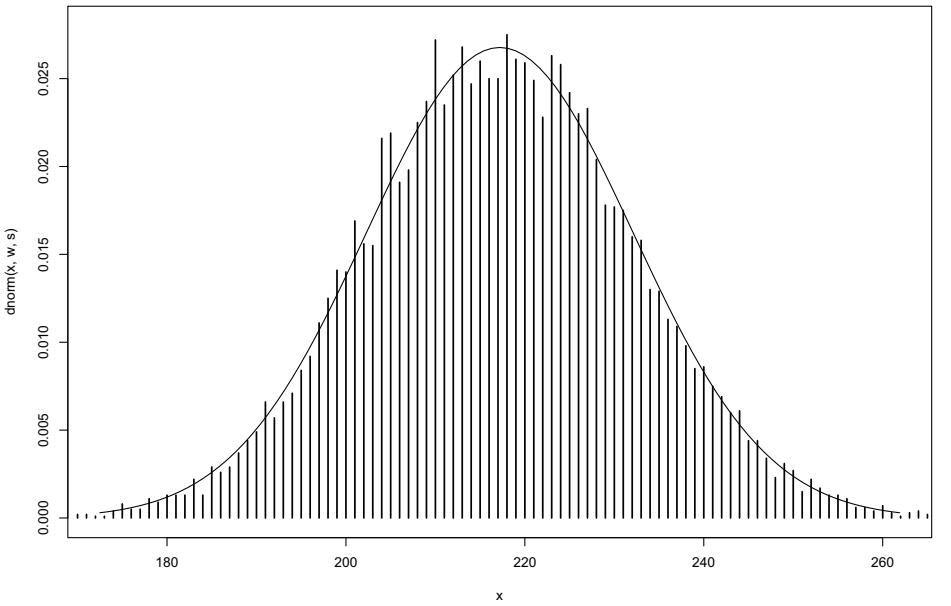

FIGURE 3. Simulations of the overflow in the uniform case with $r=2$, $n=10^4$, $m_n=\lfloor n^a \rfloor$ with $a=1.1$ (i.e. $m_n=25\,118$) are shown as vertical lines (10^4 repetitions) vs. the graph of the normal density dnorm(x, w, s), where $w=217.2$ and $s=14.9$ are the empirical mean and standard deviation, respectively.

Moreover,

$$np_n^* = np_n = n^{1-a} \to \begin{cases} 1, & a=1, \\ 0, & 1<a<1+r^{-1}. \end{cases}$$

Thus, the asymptotic normality of $V_{n,r}$ follows from the above theorem. Illustrative simulations are shown in Fig. 4.

2.3. Phase transition

In this short subsection we combine the results on the Poisson and normal asymptotics given in Theorems 1 and 2 in order to identify a critical capacity r that separates two asymptotic phases, normal and degenerate, the phase transition being through the Poissonian asymptotic regime.

Proposition 2. *Let $np_n^* \to 0$. Assume there exists an $r \in \{1, 2, \ldots\}$ such that (4) holds. Then*

(i) $(V_{n,s} - \mathbb{E}\, V_{n,s})/\sqrt{\operatorname{Var} V_{n,s}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$ *for* $s \in \{1, \ldots, r-1\}$,

(ii) $V_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$,

(iii) $V_{n,s} \xrightarrow{\mathbb{P}} 0$ *for* $s \in \{r+1, r+2, \ldots\}$.

Proof. See Section 4.3. □

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

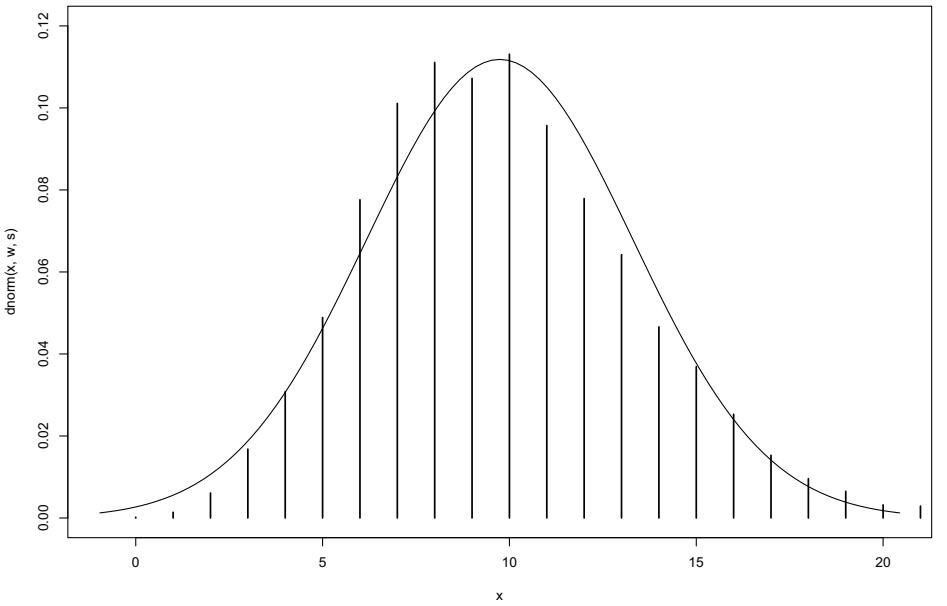

FIGURE 4. Simulations of the overflow in the geometric case with $r = 4$, $n = 10^4$, $a = 1$ are shown as vertical lines (10^4 repetitions) vs. the graph of the normal density dnorm(x, w, s), where $w = 9.74$ and $s = 3.57$ are the empirical mean and standard deviation, respectively.

2.4. Asymptotics for the number of full containers

Let $L_{n,r}$ denote the number of full containers and $K_{n,r}$ the number of full containers without overflow. The main idea is to represent $L_{n,r}$ and $K_{n,r}$ in terms of the size of the overflow $V_{n,r}$.

Note from (1) that $N_{n,n+1}(m)$ is the total number of balls in the sample for which the mth box was selected. Thus, $K_{n,r} = \sum_{j\in M_n} \mathbf{1}_{\{N_{n,n+1}(j)=r\}} = L_{n,r} - L_{n,r+1}$ and $L_{n,r} = \sum_{j\in M_n} \mathbf{1}_{\{N_{n,n+1}(j)\geq r\}}$. We note that

$$\begin{aligned} L_{n,r} &= \sum_{j\in M_n} \sum_{k=1}^{n} \mathbf{1}_{\{X_{n,k}=j\}} \mathbf{1}_{\{N_{n,k}(j)=r-1\}} \\ &= \sum_{j\in M_n} \sum_{k=1}^{n} \mathbf{1}_{\{X_{n,k}=j\}} \mathbf{1}_{\{N_{n,k}(j)\geq r-1\}} - \sum_{j\in M_n} \sum_{k=1}^{n} \mathbf{1}_{\{X_{n,k}=j\}} \mathbf{1}_{\{N_{n,k}(j)\geq r\}}. \end{aligned}$$

That is,

$$L_{n,r} = V_{n,r-1} - V_{n,r}, \qquad K_{n,r} = V_{n,r-1} - 2V_{n,r} + V_{n,r+1}. \tag{10}$$

Note that in the case $r = 1$ we have $V_{n,0} = n$ and thus $L_{n,1}$, which is the number of non-empty boxes, is

$$L_{n,1} = n - V_{n,1}, \tag{11}$$

and $K_{n,1}$, which is the number of singleton boxes, is

$$K_{n,1} = n - 2V_{n,1} + V_{n,2}. \tag{12}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

These representations of $K_{n,r}$ and $L_{n,r}$ in terms of $V_{n,r-1}$, $V_{n,r}$, and $V_{n,r+1}$ allow us to read the Poissonian asymptotics of these two sequences from Theorem 1. For $K_{n,r}$ the forthcoming statement was proved in [16, Theorem III.3.1].

Theorem 3. *Assume that* $np_n^* \to 0$.

(i) If $r > 1$ *and* $n^r \mathbb{E}\, p_{X_n}^{r-1} \to r!\mu$ *then* $L_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$ *and* $K_{n,r} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$.

(ii) If $r = 1$ *and* $n^2 \mathbb{E}\, p_{X_n} \to 2\mu$ *then* $n - L_{n,1} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$ *and* $\frac{1}{2}(n - K_{n,1}) \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$.

Proof. See Section 4.4 □

Note that, under the assumptions of Theorem 3, we have $L_{n,r} - K_{n,r} \xrightarrow{\mathbb{P}} 0$ in case (i) and $L_{n,1} - K_{n,1} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$ in case (ii). To see these two facts, note that $L_{n,r} - K_{n,r} = V_{n,r} - V_{n,r+1}$ for all $r = 1, 2, \ldots$ and, under the assumptions of Theorem 3, both $V_{n,r}$, $V_{n,r+1} \xrightarrow{\mathbb{P}} 0$ in case (i), while $V_{n,2} \xrightarrow{\mathbb{P}} 0$ and $V_{n,1} \xrightarrow{\mathrm{d}} \mathrm{Pois}(\mu)$ in case (ii).

The representations in (10) are also useful for getting Gaussian asymptotics of $L_{n,r}$ and $K_{n,r}$ from Theorem 2 in the case $\lambda = 0$.

Theorem 4. *Assume that* $np_n^* \to 0$ *and* $r \geq 1$.

(i) If $n^{r+1} \mathbb{E}\, p_{X_n}^r \to \infty$ *then*

$$\frac{L_{n,r} - \mathbb{E}\, L_{n,r}}{\sqrt{\mathrm{Var}\, L_{n,r}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1).$$

(ii) If $n^{r+2} \mathbb{E}\, p_{X_n}^{r+1} \to \infty$ *then*

$$\frac{K_{n,r} - \mathbb{E}\, K_{n,r}}{\sqrt{\mathrm{Var}\, K_{n,r}}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1).$$

Proof. See Section 4.4. □

3. Intermediate technical results

In this section we provide the notation, definitions, and technical results that are needed in the proofs of the theorems.

3.1. Multinomial distribution and negative association

Note that, for distinct $m_1, \ldots, m_s \in M_n$ and any $k = 1, \ldots, n$, $(N_{n,k}(m_1), \ldots, N_{n,k}(m_s))$ has a multinomial distribution, denoted $\mathrm{Mn}_s(k-1; p_{n,m_1}, \ldots, p_{n,m_s})$, i.e.

$$\mathbb{P}(N_{n,k}(m_1) = i_1, \ldots, N_{n,k}(m_s) = i_s) = \binom{k-1}{i_1, \ldots, i_s} \left(1 - \sum_{v=1}^{s} p_{n,v}\right)^{k-1-\sum_{v=1}^{s} i_v} \prod_{v=1}^{s} p_{n,m_v}^{i_v}$$

for $i_v \geq 0$, $v = 1, \ldots, s$, and $\sum_{v=1}^{s} i_v \leq k-1$. In particular, $N_{n,k}(m)$ has the binomial distribution $\mathrm{Bin}(k-1, p_{n,m})$, i.e. $\mathbb{P}(N_{n,k}(m) = i) = \binom{k-1}{i} p_{n,m}^i q_{n,m}^{k-1-i}$, $i = 0, \ldots, k-1$, where $q_{n,m} = 1 - p_{n,m}$. Also, let $N_{n,k}^{\ell}(m) = N_{n,\ell}(m) - N_{n,k}(m) = \sum_{j=k}^{\ell-1} \mathbf{1}_{\{X_{n,j}=m\}}$ for $k < \ell$, and $N_{n,k}^{\ell}(m) = 0$ for $k \geq \ell$. Then, for distinct $j_1, \ldots, j_t \in M_n$ and $k < \ell$, $(N_{n,k}^{\ell}(j_1), \ldots, N_{n,k}^{\ell}(j_t))$ has distribution $\mathrm{Mn}_t(\ell - k; p_{n,j_1}, \ldots, p_{n,j_t})$. Moreover, the random vectors $(N_{n,k}(m_1), \ldots, N_{n,k}(m_s))$ and

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

$(N^{\ell}_{n,k}(j_1), \ldots, N^{\ell}_{n,k}(j_t))$ are independent since the former is a function of $X_{n,1}, \ldots, X_{n,k-1}$ and the latter of $X_{n,k}, \ldots, X_{n,\ell-1}$.

Further, it is well known that multinomial random variables are negatively upper-orthant dependent (NUOD; see, e.g., [19], where this observation seems to have appeared for the first time, or [11]). That is, for $m_1 \neq m_2$,

$$\mathbb{P}(N_{n,k}(m_1) \geq x_1,\ N_{n,k}(m_2) \geq x_2) \leq \mathbb{P}(N_{n,k}(m_1) \geq x_1)\, \mathbb{P}(N_{n,k}(m_2) \geq x_2). \tag{13}$$

As such, they are also negatively associated (NA); see [10] for the definition and basic properties P1–P7. We recall three of these properties that we will use:

P4: A subset of two or more NA random variables is NA.

P6: Increasing functions defined on disjoint subsets of a set of NA random variables are NA.

P7: The union of independent sets of NA random variables is NA.

Since $\{N_{n,k}(m_1), \ldots, N_{n,k}(m_t)\}$ and $\{N^{\ell}_{n,k}(j_1), \ldots, N^{\ell}_{n,k}(j_t)\}$ are independent sets of NA random variables, by property P7, $\{N_{n,k}(m_1), \ldots, N_{n,k}(m_t), N^{\ell}_{n,k}(j_1), \ldots, N^{\ell}_{n,k}(j_t)\}$ is also NA. In particular, by P4, for distinct m_1, m_2, n_1, and n_2, the subset $\{N_{n,k}(m_1), N_{n,k}(n_1), N_{n,k}(m_2), N_{n,k}(n_2), N^{\ell}_{n,k}(m_2), N^{\ell}_{n,k}(n_2)\}$ is NA as well. Finally, noting that $N_{n,\ell}(m) = N_{n,k}(m) + N^{\ell}_{n,k}(m)$, we conclude by P6 that $N_{n,k}(m_1), N_{n,k}(n_1), N_{n,\ell}(m_2)$, and $N_{n,\ell}(n_2)$ are NA.

Consequently, the following extended versions of the NUOD property (13) hold:

$$\begin{aligned}\mathbb{P}(N_{n,k}(m_1) \geq x_1, N_{n,k}(n_1) \geq y_1, N_{n,\ell}(m_2) \geq x_2, N_{n,\ell}(n_2) \geq y_2)\\ \leq \mathbb{P}(N_{n,k}(m_1) \geq x_1)\mathbb{P}(N_{n,k}(n_1) \geq y_1)\mathbb{P}(N_{n,\ell}(m_2) \geq x_2)\mathbb{P}(N_{n,\ell}(n_2) \geq y_2),\end{aligned} \tag{14}$$

and, taking $y_1 = y_2 = 0$ in (14),

$$\mathbb{P}(N_{n,k}(m_1) \geq x_1,\ N_{n,\ell}(m_2) \geq x_2) \leq \mathbb{P}(N_{n,k}(m_1) \geq x_1)\, \mathbb{P}(N_{n,\ell}(m_2) \geq x_2). \tag{15}$$

3.2. Conditional expectations

Let $\mathcal{F}_{n,k} = \sigma(X_{n,1}, \ldots, X_{n,k})$ be the σ-algebra generated by $X_{n,1}, \ldots, X_{n,k}$ for $k = 1, \ldots, n$, and note that $N_{n,j}(m)$ is $\mathcal{F}_{n,k}$-measurable, for any $m \in M_n$, $k \geq j-1$. Note also that, for any n, k, X_n is independent of $\mathcal{F}_{n,k}$. Then $Y_{n,j}$ can be written as

$$Y_{n,j} = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,n} \right).$$

So, for $j \geq k$,

$$\begin{aligned}\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1}) &= \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k-1} \right)\\ &= \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \,\middle|\, X_n, \mathcal{F}_{n,j-1} \right) \,\middle|\, \mathcal{F}_{n,k-1} \right)\\ &= \mathbb{E}\left(\mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \,\middle|\, X_n, \mathcal{F}_{n,j-1} \right) \mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k-1} \right)\\ &= \mathbb{E}\left(\mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \mid \mathcal{F}_{n,k-1} \right).\end{aligned} \tag{16}$$

Hence, $\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) = \mathbb{E}(\mathbf{1}_{\{N_{n,j}(X_n)\geq r\}} \mid \mathcal{F}_{n,k})$ for $j > k$, and $\mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k}) = Y_{n,k}$.

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Note that the representation in (16) implies

$$\mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k-1}) = \mathbb{P}(N_{n,k}(X_n) \geq r \mid \mathcal{F}_{n,k-1}) = \mathbb{P}(N_{n,k}(X_n) \geq r \mid \mathcal{F}_{n,n}). \tag{17}$$

Taking expectations of both extremes of (16), we get

$$\mathbb{E}\, Y_{n,j} = \mathbb{P}(N_{n,j}(X_n) \geq r) = \mathbb{E}\, \mathbb{P}(N_{n,j}(X_n) \geq r \mid X_n) = \mathbb{E} \sum_{i=r}^{j-1} \binom{j-1}{i} p_{X_n}^i q_{X_n}^{j-1-i}, \tag{18}$$

where $q_{X_n} = 1 - p_{X_n}$. Furthermore, for $k, \ell = 1 \ldots, n$, (17) yields

$$\mathbb{E}(\mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k-1})\mathbb{E}(Y_{n,\ell} \mid \mathcal{F}_{n,\ell-1})) = \mathbb{E}(\mathbb{P}(N_{n,k}(X_n) \geq r \mid \mathcal{F}_{n,n})\mathbb{P}(N_{n,\ell}(Y_n) \geq r \mid \mathcal{F}_{n,n})),$$

and, because $N_{n,k}(X_n)$ and $N_{n,\ell}(Y_n)$ are conditionally independent given $\mathcal{F}_{n,n}$, it follows that $\mathbb{E}(\mathbb{E}\,(Y_{n,k} \mid \mathcal{F}_{n,k-1})\mathbb{E}(Y_{n,\ell} \mid \mathcal{F}_{n,\ell-1})) = \mathbb{P}(N_{n,k}(X_n) \geq r, N_{n,\ell}(Y_n) \geq r)$. Consequently, for any k, ℓ,

$$\mathrm{Cov}\big(\mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k-1}), \mathbb{E}(Y_{n,\ell} \mid \mathcal{F}_{n,\ell-1})\big) = \mathrm{Cov}(\mathbf{1}_{\{N_{n,k}(X_n)\geq r\}}, \mathbf{1}_{\{N_{n,\ell}(Y_n)\geq r\}}). \tag{19}$$

3.3. Two useful lemmas

In the proof of Theorem 1 we use the following results, obtained from (4) and (5).

Lemma 1. *(i) Let s be a positive integer. If (4) and (5) hold, then*

$$n^s \mathbb{E}\, p_{X_n}^s \to 0, \tag{20}$$

$$n^{s+1} \mathbb{E}\, p_{X_n}^s \to 0, \quad s > r. \tag{21}$$

(ii) If np_n^*, $n \geq 1$, is bounded and $n^{r+1}\mathbb{E}\, p_{X_n}^r \to \infty$ then

$$n^{s+1} \mathbb{E}\, p_{X_n}^s \to \infty, \quad 0 < s < r. \tag{22}$$

Proof. (i) Since $n^s \mathbb{E}\, p_{X_n}^s \leq (np_n^*)^s$, (20) follows from (5). Also, (21) follows from (4) and (5) since $n^{s+1}\mathbb{E}\, p_{X_n}^s \leq (np_n^*)^{s-r} n^{r+1} \mathbb{E}\, p_{X_n}^r$.

(ii) For $s < r$, by (4) and (5) we get $n^{s+1}\mathbb{E}\, p_{X_n}^s \geq n^{r+1}\mathbb{E}\, p_{X_n}^r (1/(np_n^*)^{r-s}) \to \infty$. □

We also need the following simple estimate of the tail of a binomial sum.

Lemma 2. *Let m,n be positive integers such that $m \leq n$, and let $p \in (0, 1)$. Then*

$$\sum_{i=m}^{n} \binom{n}{i} p^i (1-p)^{n-i} \leq \frac{(np)^m}{m!}. \tag{23}$$

Proof. The left-hand side of (23) is $\mathbb{P}(B_n \geq m)$, where B_n has distribution $\mathrm{Bin}(n, p)$. We argue by induction on $n \geq m$. Note that for $n = m$ the left-hand side of (23) is p^m and the right-hand side is $(m^m/m!)p^m$. Since $m^m \geq m!$ the result follows. For any $n \geq 1$, by induction we have

$$\begin{aligned} \mathbb{P}(B_{n+1} \geq m) &= \mathbb{P}(B_n \geq m-1)p + \mathbb{P}(B_n \geq m)(1-p) \\ &\leq \frac{(np)^{m-1}}{(m-1)!}p + \frac{(np)^m}{m!}(1-p) \leq \frac{((n+1)p)^m}{m!}, \end{aligned}$$

where the last inequality follows from $mn^{m-1} + n^m(1-p) \leq (n+1)^m$. □

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

4. Proof of Theorems

4.1. Poisson convergence

The proof of Theorem 1 is based on the following theorem due to [3]; see Corollary 5 therein.

Theorem 5. *Let* $\{Y_{n,k},\ k=1,\dots,n; n\geq 1\}$ *be a double sequence of non-negative random variables adapted to a row-wise increasing double sequence of* σ*-fields* $\{\mathcal{F}_{n,k},\ k=1,\dots,n; n\geq 1\}$*, and let* $\eta>0$*. If*

$$\max_{1\leq k\leq n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})\xrightarrow{\mathbb{P}}0, \tag{24}$$

$$\sum_{k=1}^{n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})\xrightarrow{\mathbb{P}}\eta, \tag{25}$$

$$\sum_{k=1}^{n}\mathbb{E}(Y_{n,k}\mathbf{1}_{\{|Y_{n,k}-1|>\varepsilon\}}\mid\mathcal{F}_{n,k-1})\xrightarrow{\mathbb{P}}0\quad\text{for any }\varepsilon>0, \tag{26}$$

then $\sum_{k=1}^{n}Y_{n,k}\xrightarrow{\mathrm{d}}\mathrm{Pois}(\eta)$.

Proof of Theorem 1. We show that conditions (24), (25) (with $\eta=\mu$), and (26) of Theorem 5 are satisfied for $Y_{n,k}$, defined in (2). First, we note that (26) is trivial because, for $\varepsilon<1$, $Y_{n,k}=0$ if and only if $\mathbf{1}_{\{|Y_{n,k}-1|>\varepsilon\}}=1$.

The rest of the proof is divided into three steps. In Step I we check (24). Then we prove that (25) holds in quadratic mean, i.e. $\mathbb{E}\big(\sum_{k=1}^{n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})-\mu\big)^2\to 0$. To that end we show that $\sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}\to\mu$ and $\mathrm{Var}\sum_{k=1}^{n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})\to 0$ in Steps II and III, respectively.

Step I: We prove (24) using (17). Clearly, $\mathbf{1}_{\{N_{n,k}(m)\geq r\}}\leq\mathbf{1}_{\{N_{n,l}(m)\geq r\}}$ for $k\leq l$, so $\max_{1\leq k\leq n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})=\mathbb{E}(Y_{n,n}\mid\mathcal{F}_{n,n-1})$. Note also that, due to (18), (23), and (20), $\mathbb{E}\,Y_{n,n}=\mathbb{E}\sum_{i=r}^{n-1}\binom{n-1}{i}p_{X_n}^{i}q_{X_n}^{n-1-i}\leq n^r\mathbb{E}\,p_{X_n}^{r}\to 0$. Consequently, Markov's inequality implies that $\mathbb{E}(Y_{n,n}\mid\mathcal{F}_{n,n-1})\xrightarrow{\mathbb{P}}0$, and (24) follows.

Step II: To prove that $\lim\sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}=\mu$ we show that lim sup and lim inf are respectively bounded above and below by μ. From (18), (23), and (4),

$$\begin{aligned}\sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}=\mathbb{E}\sum_{k=1}^{n}\sum_{i=r}^{k-1}\binom{k-1}{i}p_{X_n}^{i}q_{X_n}^{k-1-i}&\leq\mathbb{E}\sum_{k=1}^{n}\frac{(k-1)^r p_{X_n}^{r}}{r!}\\&\leq\frac{n^{r+1}}{(r+1)!}\mathbb{E}\,p_{X_n}^{r}\to\mu,\end{aligned} \tag{27}$$

so $\limsup\sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}\leq\mu$.

Additionally, since by (18) we have $\mathbb{E}\,Y_{n,k}=\mathbb{P}(N_{n,k}(X_n)\geq r)\geq\mathbb{P}(N_{n,k}(X_n)=r)$ and $(1-p)^k\geq 1-kp$, $p\in(0,1)$, then

$$\begin{aligned}\sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}&\geq\sum_{k=1}^{n}\binom{k-1}{r}\mathbb{E}\,p_{X_n}^{r}q_{X_n}^{k-1-r}\\&\geq\mathbb{E}\,p_{X_n}^{r}\sum_{k=1}^{n}\binom{k-1}{r}-\mathbb{E}\,p_{X_n}^{r+1}\sum_{k=1}^{n}\binom{k-1}{r}(k-1-r).\end{aligned} \tag{28}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Further, observe that

$$\frac{(r+1)!}{n^{r+1}}\sum_{k=1}^{n}\binom{k-1}{r}\to 1,\qquad \frac{r!(r+2)}{n^{r+2}}\sum_{k=1}^{n}\binom{k-1}{r}(k-1-r)\to 1.$$

Thus, by (4) and (21), the right-hand side of (28) converges to μ, and so we have $\liminf \sum_{k=1}^{n}\mathbb{E}\,Y_{n,k}\geq\mu$.

Step III: We prove that $W_n := \operatorname{Var}\sum_{k=1}^{n}\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1})\to 0$ by relying on the NUOD property of $N_{n,k}(m_1)$ and $N_{n,\ell}(m_2)$, for distinct $m_1, m_2\in M_n$. In what follows we compute and bound some expectations that add up to W_n. First, note from (19) that

$$\begin{aligned}W_n &= \sum_{k,\ell=1}^{n}\operatorname{Cov}(\mathbb{E}(Y_{n,k}\mid\mathcal{F}_{n,k-1}),\mathbb{E}(Y_{n,\ell}\mid\mathcal{F}_{n,l-1}))\\ &= \sum_{k,\ell=1}^{n}\operatorname{Cov}(\mathbf{1}_{\{N_{n,k}(X_n)\geq r\}},\mathbf{1}_{\{N_{n,\ell}(Y_n)\geq r\}}).\end{aligned}$$

For U, V square-integrable random variables and $\mathcal{G}$ a σ-algebra, let the conditional covariance be defined as $\operatorname{Cov}(U,V\mid\mathcal{G})=\mathbb{E}(UV\mid\mathcal{G})-\mathbb{E}(U\mid\mathcal{G})\mathbb{E}(V\mid\mathcal{G})$. Also, let $\mathbf{1}_k(m)=\mathbf{1}_{\{N_{n,k}(m)\geq r\}}$ (for simplicity) and $k\wedge\ell=\min\{k,l\}$. Then, by the i.i.d. assumption on $X_{n,1},\ldots,X_{n,n}$, X_n, Y_n, we have

$$\begin{aligned}\operatorname{Cov}(\mathbf{1}_k(X_n),\mathbf{1}_\ell(Y_n)\mid X_n,Y_n) &= \mathbb{E}(\mathbf{1}_k(X_n)\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\\ &\quad-\mathbb{E}(\mathbf{1}_k(X_n)\mid X_n)\mathbb{E}(\mathbf{1}_\ell(Y_n)\mid Y_n).\end{aligned}\tag{29}$$

Furthermore,

$$\begin{aligned}\mathbb{E}(\mathbf{1}_k(X_n)\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\mathbf{1}_{\{X_n=Y_n\}} &= \mathbb{E}(\mathbf{1}_{\{X_n=Y_n\}}\mathbf{1}_k(X_n)\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\\ &= \mathbb{E}(\mathbf{1}_{\{X_n=Y_n\}}\mathbf{1}_k(X_n)\mathbf{1}_\ell(X_n)|X_n,Y_n)\\ &= \mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n)\mid X_n)\mathbf{1}_{\{X_n=Y_n\}},\end{aligned}\tag{30}$$

where the last equality follows from $\mathbf{1}_k(m)\leq\mathbf{1}_\ell(m)$ for $k\leq\ell$, because $N_{n,k}(m)\geq r$ implies $N_{n,\ell}(m)\geq r$. So, from (29) and (30), we get

$$\operatorname{Cov}(\mathbf{1}_k(X_n),\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\mathbf{1}_{\{X_n=Y_n\}}\leq\mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n)\mid X_n)\mathbf{1}_{\{X_n=Y_n\}}.\tag{31}$$

Furthermore, by the NUOD property (15),

$$\begin{aligned}\mathbb{E}(\mathbf{1}_k(X_n)\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\mathbf{1}_{\{X_n\neq Y_n\}} &= \mathbb{E}(\mathbf{1}_{\{X_n\neq Y_n\}}\mathbf{1}_k(X_n)\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\\ &\leq \mathbb{E}(\mathbf{1}_{\{X_n\neq Y_n\}}\mathbf{1}_k(X_n)\mid X_n,Y_n)\mathbb{E}(\mathbf{1}_{\{X_n\neq Y_n\}}\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\\ &= \mathbb{E}(\mathbf{1}_k(X_n)\mid X_n)\mathbb{E}(\mathbf{1}_\ell(Y_n)\mid Y_n)\mathbf{1}_{\{X_n\neq Y_n\}}.\end{aligned}\tag{32}$$

Hence, from (29) and (32), we have

$$\operatorname{Cov}(\mathbf{1}_k(X_n),\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\mathbf{1}_{\{X_n\neq Y_n\}}\leq 0.\tag{33}$$

And, finally, from (31) and (33),

$$\operatorname{Cov}(\mathbf{1}_k(X_n),\mathbf{1}_\ell(Y_n)\mid X_n,Y_n)\leq\mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n)\mid X_n)\mathbf{1}_{\{X_n=Y_n\}}.\tag{34}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Let us write

$$\begin{aligned}\operatorname{Cov}(\mathbf{1}_k(X_n), \mathbf{1}_\ell(Y_n)) &= \mathbb{E}\operatorname{Cov}(\mathbf{1}_k(X_n), \mathbf{1}_\ell(Y_n) \mid X_n, Y_n)\\ &\quad + \operatorname{Cov}(\mathbb{E}(\mathbf{1}_k(X_n) \mid X_n, Y_n), \mathbb{E}(\mathbf{1}_\ell(Y_n) \mid X_n, Y_n)).\end{aligned}$$

By the independence of X_n and Y_n we can write $\mathbb{E}(\mathbf{1}_k(X_n) \mid X_n, Y_n) = \mathbb{E}(\mathbf{1}_k(X_n) \mid X_n)$ and $\mathbb{E}(\mathbf{1}_\ell(Y_n) \mid X_n, Y_n) = \mathbb{E}(\mathbf{1}_\ell(Y_n) \mid Y_n)$. Thus, again referring to the independence of X_n and Y_n, we conclude that the second term above vanishes. Applying (34), we thus get

$$\begin{aligned}\operatorname{Cov}(\mathbf{1}_k(X_n), \mathbf{1}_\ell(Y_n)) &\le \mathbb{E}\,\mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n) \mid X_n)\mathbf{1}_{\{X_n=Y_n\}}\\ &= \mathbb{E}\,\mathbf{1}_{k\wedge\ell}(X_n)\mathbf{1}_{\{X_n=Y_n\}} = \mathbb{E}\,\mathbf{1}_{k\wedge\ell}(X_n)p_{X_n}. \qquad (35)\end{aligned}$$

Note that in the first equality above we used the identity $\mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n) \mid X_n) = \mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n) \mid X_n, Y_n)$, i.e. we referred again to the independence of X_n and Y_n.

Also, by (23),

$$\mathbb{E}(\mathbf{1}_{k\wedge\ell}(X_n)p_{X_n} \mid X_n) = \sum_{i=r}^{(k\wedge\ell)-1} \binom{(k\wedge\ell)-1}{i} p_{X_n}^i q_{X_n}^{(k\wedge\ell)-1-i} p_{X_n} \le (k-1)^r p_{X_n}^{r+1}.$$

Finally, taking expectation above and adding over k and ℓ, from (35) we obtain

$$W_n \le \sum_{k,\ell=1}^{n} (k-1)^r \mathbb{E}\, p_{X_n}^{r+1} \le n^{r+2}\mathbb{E}\, p_{X_n}^{r+1} \to 0,$$

where convergence to 0 follows from (21). Since $W_n \ge 0$, it holds that $W_n \to 0$. □

4.2. Gaussian convergence

The proof of Theorem 2 is split into several steps, which are presented in four subsections below. In Section 4.2.1 we decompose $V_{n,r} - \mathbb{E}\, V_{n,r}$ as the sum of martingale differences $\sum_{k=1}^n d_{n,k}$, with suitably defined (uniformly bounded) $d_{n,k}$. In Section 4.2.2 we prove Proposition 1. In Section 4.2.3 we show that $\operatorname{Var}\sum_{k=1}^n \operatorname{Var}(d_{n,k} \mid \mathcal{F}_{n,k-1})$ is $o((n^{r+1}\mathbb{E}\, p_{X_n}^r)^2)$. In the final part of the proof in Section 4.2.4 we use this bound and the growth rate for $\operatorname{Var} V_{n,r}$ to conclude the proof by the martingale central limit theorem.

4.2.1. *Martingale differences decomposition.*

Lemma 3. *The centered size of the overflow can be represented as* $V_{n,r} - \mathbb{E}\, V_{n,r} = \sum_{k=1}^n d_{n,k}$, *where the* $d_{n,k}$ *are martingale differences defined by*

$$d_{n,k} = \sum_{j=k}^{n} \big(\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) - \mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1})\big). \qquad (36)$$

Proof. Clearly, $\mathbb{E}(d_{n,k} \mid \mathcal{F}_{n,k-1}) = 0$. Further, noting that $\mathcal{F}_{n,0}$ is the trivial σ-algebra,

$$\begin{aligned}\sum_{k=1}^{n} d_{n,k} &= \sum_{j=1}^{n}\sum_{k=1}^{j} (\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) - \mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1}))\\ &= \sum_{j=1}^{n} (\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,j}) - \mathbb{E}\, Y_{n,j}) = V_{n,r} - \mathbb{E}\, V_{n,r}.\end{aligned}$$

□

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Lemma 4. *The martingale differences $d_{n,k}$ of (36) are uniformly bounded and can be represented as*

$$d_{n,k} = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \mathbf{1}_{\{N_{n,k}(X_n)+N^{n+1}_{n,k+1}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k} \right).$$

Proof. Let $n, r \in \mathbb{N}, j > k$ and note that $N_{n,j}(X_n) = N_{n,k}(X_n) + \mathbf{1}_{\{X_{n,k}=X_n\}} + N^j_{n,k+1}(X_n) = U_j + J$, where, for simplicity, we let $V = N_{n,k}(X_n)$, $U_j = V + N^j_{n,k+1}(X_n) \geq V$, and $J = \mathbf{1}_{\{X_{n,k}=X_n\}}$. Then $\{U_j + J \geq r\} = \{U_j \geq r\} \cup \{U_j = r-1, J = 1\}$. Clearly, $\{V \geq r\} \subseteq \{U_j \geq r\}$, and thus we have $\{U_j + J \geq r\} = \{V \geq r\} \cup \{U_j \geq r, V < r\} \cup \{U_j = r-1, J = 1\}$. Consequently, by (16) we can write

$$\begin{aligned}\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) &= \mathbb{E}(\mathbf{1}_{\{V\geq r\}} \mid \mathcal{F}_{n,k}) + \mathbb{E}(\mathbf{1}_{\{U_j\geq r, V<r\}} \mid \mathcal{F}_{n,k}) + \mathbb{E}(\mathbf{1}_{\{U_j=r-1, J=1\}} \mid \mathcal{F}_{n,k}) \\ &= \mathbb{E}(\mathbf{1}_{\{V\geq r\}} \mid \mathcal{F}_{n,k-1}) + \mathbb{E}(\mathbf{1}_{\{U_j\geq r, V<r\}} \mid \mathcal{F}_{n,k-1}) + \mathbb{E}(J\mathbf{1}_{\{U_j=r-1\}} \mid \mathcal{F}_{n,k}),\end{aligned}$$

where $\mathcal{F}_{n,k}$ is changed to $\mathcal{F}_{n,k-1}$ in the first two conditional expectations due to the independence of $X_{n,k}$ and $(X_{n,j},\, j \in \{1, \ldots, n\} \setminus \{k\})$ (note that V and U_j depend only on the latter set of variables and do not depend on $X_{n,k}$). Similarly, $\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1}) = \mathbb{E}(\mathbf{1}_{\{V\geq r\}} \mid \mathcal{F}_{n,k-1}) + \mathbb{E}(\mathbf{1}_{\{U_j\geq r, V<r\}} \mid \mathcal{F}_{n,k-1}) + \mathbb{E}(J\mathbf{1}_{\{U_j=r-1\}} \mid \mathcal{F}_{n,k-1})$. Also, note that $\mathbb{E}\big(J\mathbf{1}_{\{U_j=r-1\}} \mid \mathcal{F}_{n,k-1}\big) = \mathbb{E}\Big(\mathbf{1}_{\{U_j=r-1\}} p_{X_n} \mid \mathcal{F}_{n,k}\Big)$. Therefore, for $j > k$, $\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) - \mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1}) = \mathbb{E}\Big(\mathbf{1}_{\{U_j=r-1\}}(J - p_{X_n}) \mid \mathcal{F}_{n,k}\Big)$. Thus,

$$e_{n,k} := \sum_{j=k+1}^{n} \left(\mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k}) - \mathbb{E}(Y_{n,j} \mid \mathcal{F}_{n,k-1})\right) = \mathbb{E}\left((J - p_{X_n}) \sum_{j=k+1}^{n} \mathbf{1}_{\{U_j=r-1\}} \,\middle|\, \mathcal{F}_{n,k} \right).$$

Observe that, for $j > k$,

$$\mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \,\middle|\, X_n, \mathcal{F}_{n,j-1} \right) = 1.$$

Then

$$\begin{aligned} e_{n,k} &= \mathbb{E}\left((J - p_{X_n}) \sum_{j=k+1}^{n} \mathbf{1}_{\{U_j=r-1\}} \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,j}=X_n\}}}{p_{X_n}} \,\middle|\, X_n, \mathcal{F}_{n,j-1} \right) \,\middle|\, \mathcal{F}_{n,k} \right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\frac{J - p_{X_n}}{p_{X_n}} \sum_{j=k+1}^{n} \mathbf{1}_{\{U_j=r-1\}} \mathbf{1}_{\{X_{n,j}=X_n\}} \,\middle|\, X_n, \mathcal{F}_{n,j-1} \right) \,\middle|\, \mathcal{F}_{n,k} \right) \\ &= \mathbb{E}\left(\frac{J - p_{X_n}}{p_{X_n}} \mathbf{1}_{\{V<r\}} \sum_{j=k+1}^{n} \mathbf{1}_{\{U_j=r-1\}} \mathbf{1}_{\{X_{n,j}=X_n\}} \,\middle|\, \mathcal{F}_{n,k} \right), \end{aligned}$$

where in the last expression we used the fact that $\{V < r\} \supset \{U_j = r-1\}$ for $j > k$. Observe also that $\sum_{j=k+1}^{n} \mathbf{1}_{\{U_j=r-1\}} \mathbf{1}_{\{X_{n,j}=X_n\}} = \mathbf{1}_{\{U_j=r-1,\, U_{j+1}=r,\ \text{for some } j\in\{k+1,\ldots,n\}\}}$ is equal to $\mathbf{1}_{\{U_{n+1}\geq r\}}$ on the event $\{V < r\}$ (note that $U_{k+1} = V$). That is, using the original notation,

$$\sum_{j=k+1}^{n} \mathbf{1}_{\{N_{n,k}(X_n)+N^j_{n,k+1}(X_n)=r-1\}} \mathbf{1}_{\{X_{n,j}=X_n\}} = \mathbf{1}_{\{N_{n,k}(X_n)+N^{n+1}_{n,k+1}(X_n)\geq r\}}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

on the event $\{N_{n,k}(X_n) < r\}$, and so

$$e_{n,k} = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \mathbf{1}_{\{N_{n,k}(X_n)<r\}} \mathbf{1}_{\{N_{n,k}(X_n)+N^{n+1}_{n,k+1}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k} \right).$$

Finally, since

$$Y_{n,k} - \mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k-1}) = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \mathbf{1}_{\{N_{n,k}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k} \right),$$

we conclude that

$$\begin{aligned} d_{n,k} &= Y_{n,k} - \mathbb{E}(Y_{n,k} \mid \mathcal{F}_{n,k-1}) + e_{n,k} \\ &= \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \left(\mathbf{1}_{\{N_{n,k}(X_n)\geq r\}} + \mathbf{1}_{\{N_{n,k}(X_n)<r\}} \mathbf{1}_{\{N_{n,k}(X_n)+N^{n+1}_{n,k+1}(X_n)\geq r\}} \right) \,\middle|\, \mathcal{F}_{n,k} \right) \\ &= \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \mathbf{1}_{\{N_{n,k}(X_n)+N^{n+1}_{n,k+1}(X_n)\geq r\}} \,\middle|\, \mathcal{F}_{n,k} \right). \end{aligned} \tag{37}$$

For the boundedness of $d_{n,k}$, note that

$$|d_{n,k}| \leq \mathbb{E}\left(\left| \frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}} \right| \,\middle|\, \mathcal{F}_{n,k} \right) \leq \sum_{m\in M_n} \left| \mathbf{1}_{\{X_{n,k}=m\}} - p_{n,m} \right| \leq 2. \qquad \square$$

4.2.2. *Growth rate of the expected value and variance.*

Proof of Proposition 1. We first prove (8). For the upper bound in (8) note that, by the representation in (3) and the estimates in (27), it follows that

$$\frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p^r_{X_n}} \leq \frac{1}{(r+1)!}.$$

Similarly, for the lower bound, by the first part of (28) we have

$$\mathbb{E}\, V_{n,r} \geq \mathbb{E} \sum_{k=1}^{n} \binom{k-1}{r} p^r_{X_n} (1-p_{X_n})^{k-r-1} \geq \mathbb{E}\, p^r_{X_n} \sum_{k=1}^{n} \binom{k-1}{r} (1-p^*_n)^{k-r-1}.$$

Since, for any m and any odd j, we have $(1-x)^m \geq \sum_{i=0}^{j} \binom{m}{i} (-x)^i$, we get, for any odd j,

$$\begin{aligned} \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p^r_{X_n}} &\geq \frac{1}{n^{r+1}} \sum_{k=1}^{n} \binom{k-1}{r} \sum_{i=0}^{j} \binom{k-r-1}{i} (-p^*_n)^i \\ &= \sum_{i=0}^{j} (-1)^i (np^*_n)^i \frac{\sum_{k=1}^{n} \binom{k-1}{r}\binom{k-r-1}{i}}{n^{r+i+1}}. \end{aligned}$$

Note that

$$\begin{aligned} \frac{1}{n^{r+i+1}} \sum_{k=1}^{n} \binom{k-1}{r}\binom{k-r-1}{i} &= \frac{1}{n^{r+i+1} r!\, i!} \sum_{k=1}^{n} \frac{(k-1)!}{(k-i-r-1)!} \\ &= \frac{\binom{r+i}{r}\binom{n}{r+i+1}}{n^{r+i+1}} \to \frac{1}{r!\, i!\, (r+i+1)}. \end{aligned}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Consequently,

$$\liminf \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \geq \frac{1}{r!}\sum_{i=0}^{\infty}\frac{(-\lambda)^i}{i!(i+r+1)}$$
$$= \frac{1}{r!}\sum_{i=0}^{\infty}\frac{(-\lambda)^i}{i!}\int_0^1 x^{r+i}\,\mathrm{d}x = \frac{1}{r!}\int_0^1 x^r \mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{\Gamma_\lambda(r+1)}{r!},$$

which proves the lower bound in (8).

To prove (9), let $p_x = p_{n,x}$, $q_x = 1 - p_x$, and

$$T_{n,k}(x) = \sum_{i=r-N_{n,k}(x)}^{n-k}\binom{n-k}{i}p_x^i q_x^{n-k-i}. \tag{38}$$

Then

$$\mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}\mathbf{1}_{\{N_{n,k}(X_n)+N_{n,k+1}^{n+1}(X_n)\geq r\}}\,\middle|\, X_n, \mathcal{F}_{n,k}\right)$$
$$= \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}T_{n,k}(X_n)\,\middle|\, X_n, \mathcal{F}_{n,k}\right),$$

and so

$$d_{n,k} = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}T_{n,k}(X_n)\,\middle|\, \mathcal{F}_{n,k}\right).$$

Also, recalling that X_n, Y_n, $X_{n,1}, \ldots, X_{n,n}$ are i.i.d.,

$$d_{n,k}^2 = \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}T_{n,k}(X_n)\,\middle|\, \mathcal{F}_{n,k}\right)\mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=Y_n\}} - p_{Y_n}}{p_{Y_n}}T_{n,k}(Y_n)\,\middle|\, \mathcal{F}_{n,k}\right)$$
$$= \mathbb{E}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}T_{n,k}(X_n)\frac{\mathbf{1}_{\{X_{n,k}=Y_n\}} - p_{Y_n}}{p_{Y_n}}T_{n,k}(Y_n)\,\middle|\, \mathcal{F}_{n,k}\right),$$

where the second equality follows from the conditional independence, given $\mathcal{F}_{n,k}$, of

$$\frac{\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n}}{p_{X_n}}T_{n,k}(X_n) \quad \text{and} \quad \frac{\mathbf{1}_{\{X_{n,k}=Y_n\}} - p_{Y_n}}{p_{Y_n}}T_{n,k}(Y_n).$$

In what follows we compute $\mathbb{E}\big(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}\big)$ by considering the cases $X_n = Y_n$ and $X_n \neq Y_n$. We get

$$\mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}) = \mathbb{E}\left(\mathbf{1}_{\{X_n=Y_n\}}\left(\frac{\mathbf{1}_{\{X_{n,k}=X_n\}}-p_{X_n}}{p_{X_n}}\right)^2 T_{n,k}^2(X_n)\,\middle|\, \mathcal{F}_{n,k-1}\right)$$
$$+ \mathbb{E}\left(\mathbf{1}_{\{X_n\neq Y_n\}}\frac{(\mathbf{1}_{\{X_{n,k}=X_n\}} - p_{X_n})(\mathbf{1}_{\{X_{n,k}=Y_n\}} - p_{Y_n})}{p_{X_n}p_{Y_n}}T_{n,k}(X_n)T_{n,k}(Y_n)\,\middle|\, \mathcal{F}_{n,k-1}\right)$$
$$= \mathbb{E}\left(\mathbf{1}_{\{X_n=Y_n\}}\frac{1-p_{X_n}}{p_{X_n}}T_{n,k}^2(X_n)\,\middle|\, \mathcal{F}_{n,k-1}\right)$$
$$- \mathbb{E}\big(\mathbf{1}_{\{X_n\neq Y_n\}}T_{n,k}(X_n)T_{n,k}(Y_n) \mid \mathcal{F}_{n,k-1}\big),$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

where the second equality above is obtained from conditioning inside both expectations with respect to X_n, Y_n, $\mathcal{F}_{n,k-1}$. Finally, integrating out Y_n in the first expectation we get

$$\begin{aligned}\mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}) &= \mathbb{E}(q_{X_n} T_{n,k}^2(X_n) \mid \mathcal{F}_{n,k-1}) \\ &\quad - \mathbb{E}(\mathbf{1}_{\{X_n \neq Y_n\}} T_{n,k}(X_n) T_{n,k}(Y_n) \mid \mathcal{F}_{n,k-1}),\end{aligned} \tag{39}$$

and consequently

$$\operatorname{Var} d_{n,k} = \mathbb{E}\, d_{n,k}^2 = \mathbb{E}\, q_{X_n} T_{n,k}^2(X_n) - \mathbb{E}\, \mathbf{1}_{\{X_n \neq Y_n\}} T_{n,k}(X_n) T_{n,k}(Y_n). \tag{40}$$

For the upper bound of the variance, note that $0 < T_{n,k}(X_n) \leq 1$ and thus (40) implies $\operatorname{Var} d_{n,k} \leq \mathbb{E}\, d_{n,k}^2 \leq \mathbb{E}\, T_{n,k}(X_n)$. Also, $\mathbb{E}(T_{n,k}(X_n) \mid X_n, \mathcal{F}_{n,k}) = \mathbb{P}(N_{n,k}(X_n) + N_{n,k+1}^{n+1}(X_n) \geq r \mid X_n, \mathcal{F}_{n,k})$, and so

$$\begin{aligned}\mathbb{E}(T_{n,k}(X_n) \mid X_n) &= \mathbb{P}(N_{n,k}(X_n) + N_{n,k+1}^{n+1}(X_n) \geq r \mid X_n) \\ &\leq \mathbb{P}(N_{n,n+1}(X_n) \geq r \mid X_n).\end{aligned} \tag{41}$$

Now, recalling that $N_{n,n+1}(m)$ has distribution $\operatorname{Bin}(n, p_{n,m})$ for $m \in M_n$, and using (23), the right-hand side of (41) is bounded by $n^r p_{X_n}^r / r!$. Last, taking expectations, we obtain $\operatorname{Var} d_{n,k} \leq n^r \mathbb{E}\, p_{X_n}^r / r!$, and consequently

$$\operatorname{Var} V_{n,r} = \sum_{k=1}^{n} \operatorname{Var} d_{n,k} \leq \frac{n^{r+1} \mathbb{E}\, p_{X_n}^r}{r!}. \tag{42}$$

In order to bound $\operatorname{Var} d_{n,k}$ from below we first find an upper bound for the last term (with the minus sign) in (40). To that end, note that $T_{n,k}(x)$, defined in (38), can be written as $T_{n,k}(x) = \mathbb{P}(B_{n-k}(x) + N_{n,k}(x) \geq r \mid \mathcal{F}_{n,k})$, where $B_{n-k}(x)$ is $\operatorname{Bin}(n-k, p_x)$, independent of X_n, Y_n, $\mathcal{F}_{n,n}$, so $T_{n,k}(X_n) = \mathbb{P}(B_{n-k}(X_n) + N_{n,k}(X_n) \geq r \mid X_n, \mathcal{F}_{n,k})$ and

$$\mathbb{E}(T_{n,k}(x) \mid X_n) = \mathbb{P}(B_{n-k}(x) + N_{n,k}(x) \geq r \mid X_n). \tag{43}$$

Furthermore, for $y \neq x$ let $B_{n-k}(y)$ be a $\operatorname{Bin}(n-k, p_y)$ random variable, independent of X_n, Y_n, $\mathcal{F}_{n,n}$ and independent of $B_{n-k}(x)$. Then $J_{n,k} := \mathbb{E}(\mathbf{1}_{\{X_n \neq Y_n\}} T_{n,k}(X_n) T_{n,k}(Y_n) \mid X_n, Y_n, \mathcal{F}_{n,k})$ can be written as $J_{n,k} = \mathbb{P}(X_n \neq Y_n, B_{n,k}(X_n) + N_{n,k}(X_n) \geq r, B_{n,k}(Y_n) + N_{n,k}(Y_n) \geq r \mid X_n, Y_n, \mathcal{F}_{n,k})$, and so $\mathbb{E}(J_{n,k} \mid X_n, Y_n) = \mathbb{P}(X_n \neq Y_n, B_{n,k}(X_n) + N_{n,k}(X_n) \geq r, B_{n,k}(Y_n) + N_{n,k}(Y_n) \geq r \mid X_n, Y_n)$. Then, since conditionally on X_n, Y_n, $(N_{n,k}(X_n), N_{n,k}(Y_n))$ is $\operatorname{Mn}_2(k-1, p_{X_n}, p_{Y_n})$, and because of the NUOD property, we have

$$\begin{aligned}\mathbb{E}(J_{n,k} \mid X_n, Y_n) &\leq \mathbf{1}_{\{X_n \neq Y_n\}} \mathbb{P}(B_{n,k}(X_n) + N_{n,k}(X_n) \geq r \mid X_n, Y_n) \\ &\quad \times \mathbb{P}(B_{n,k}(Y_n) + N_{n,k}(Y_n) \geq r \mid X_n, Y_n) \\ &= \mathbf{1}_{\{X_n \neq Y_n\}} \mathbb{P}(B_{n,k}(X_n) + N_{n,k}(X_n) \geq r \mid X_n) \\ &\quad \times \mathbb{P}(B_{n,k}(Y_n) + N_{n,k}(Y_n) \geq r \mid Y_n) \\ &= \mathbf{1}_{\{X_n \neq Y_n\}} \mathbb{E}(T_{n,k}(X_n) \mid X_n) \mathbb{E}(T_{n,k}(Y_n) \mid Y_n) \\ &= \mathbb{E}(T_{n,k}(X_n) \mid X_n) \mathbb{E}(T_{n,k}(Y_n) \mid Y_n) - \mathbf{1}_{\{X_n = Y_n\}} (\mathbb{E}(T_{n,k}(X_n) \mid X_n))^2,\end{aligned}$$

where the second equality follows from the NUOD property and the third from (43). Finally, taking expectations and using the independence of X_n and Y_n, we get $\mathbb{E}\, J_{n,k} \leq (\mathbb{E}\, T_{n,k}(X_n))^2 -$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

$\mathbb{E}\, p_{X_n}(\mathbb{E}(T_{n,k}(X_n) \mid X_n))^2$. Replacing the rightmost expectation in (40) by this bound, we have

$$\begin{aligned}\operatorname{Var} d_{n,k} &\geq \mathbb{E}\, T_{n,k}^2(X_n) - \mathbb{E}\, p_{X_n} T_{n,k}^2(X_n) - (\mathbb{E}\, T_{n,k}(X_n))^2 + \mathbb{E}\, p_{X_n}(\mathbb{E}\, T_{n,k}(X_n) \mid X_n)^2 \\ &= \operatorname{Var} T_{n,k}(X_n) - \mathbb{E}\, p_{X_n} \operatorname{Var}(T_{n,k}(X_n) \mid X_n) \\ &\geq \mathbb{E} \operatorname{Var}(T_{n,k}(X_n) \mid X_n) - \mathbb{E}\, p_{X_n} \operatorname{Var}(T_{n,k}(X_n) \mid X_n).\end{aligned}$$

Note also that

$$\begin{aligned}\mathbb{E}\, p_{X_n} \operatorname{Var}(T_{n,k}(X_n) \mid X_n) &\leq \mathbb{E}\, p_{X_n} \mathbb{E}(T_{n,k}^2(X_n) \mid X_n) \\ &\leq \mathbb{E}\, p_{X_n} T_{n,k}(X_n) \leq \frac{n^r \mathbb{E}\, p_{X_n}^{r+1}}{r!} \leq \frac{np_n^*}{nr!}\, n^r \mathbb{E}\, p_{X_n}^r.\end{aligned}$$

Thus, since np_n^* is bounded,

$$\sum_{k=1}^{n} \operatorname{Var} d_{n,k} \geq \sum_{k=1}^{n} \mathbb{E} \operatorname{Var}(T_{n,k}(X_n) \mid X_n) + o(n^{r+1} \mathbb{E}\, p_{X_n}^r).$$

Finally, observe that $T_{n,k}(x)$ can be written in the form $T_{n,k}(x) = \sum_{j=0}^{\infty} P_j(x) \mathbf{1}_j(x)$, where $P_j(x) = \binom{n-k}{j} p_x^j\, q_x^{n-k-j}$ and $\mathbf{1}_j(x) = \mathbf{1}_{\{N_{n,k}(x) \geq r-j\}}$. Therefore,

$$\operatorname{Var} T_{n,k}(x) = \sum_{j=0}^{\infty} P_j^2(x) \operatorname{Var} \mathbf{1}_j(x) + 2 \sum_{j_1 < j_2} P_{j_1}(x) P_{j_2}(x) \operatorname{Cov}\left(\mathbf{1}_{j_1}(x), \mathbf{1}_{j_2}(x)\right).$$

Since $\mathbf{1}_{j_1}(x) \leq \mathbf{1}_{j_2}(x)$, it follows that the double sum above is non-negative and so

$$\begin{aligned}\operatorname{Var} T_{n,k}(x) &\geq \sum_{j=0}^{\infty} P_j^2(x) \operatorname{Var} \mathbf{1}_j(x) \geq P_0^2(x) \operatorname{Var} \mathbf{1}_0(x) \\ &= (1-p_x)^{2(n-k)} \mathbb{P}(N_{n,k}(x) \geq r) \mathbb{P}(N_{n,k}(x) < r) \\ &\geq (1-p_x)^{2(n-k)} \mathbb{P}(N_{n,k}(x) = r) \mathbb{P}(N_{n,k}(x) = 0) \\ &= \binom{k-1}{r} p_x^r (1-p_x)^{2n-r-2} \geq \binom{k-1}{r} p_x^r (1-p_n^*)^{2n}.\end{aligned}$$

Consequently, $\operatorname{Var}(T_{n,k}(X_n) \mid X_n) \geq \binom{k-1}{r} p_{X_n}^r (1-p_n^*)^{2n}$, and so $\sum_{k=1}^{n} \operatorname{Var} d_{n,k} \geq (1-p_n^*)^{2n} \mathbb{E}\, p_{X_n}^r \sum_{k=1}^{n} \binom{k-1}{r} + o(n^{r+1} \mathbb{E}\, p_{X_n}^r)$. Finally, since $\limsup np_n^* = \lambda$,

$$\liminf \frac{\sum_{k=1}^{n} \operatorname{Var} d_{n,k}}{n^{r+1} \mathbb{E}\, p_{X_n}^r} \geq \liminf \frac{(1-p_n^*)^{2n}}{(r+1)!} = \frac{\mathrm{e}^{-2\lambda}}{(r+1)!}.$$

□

4.2.3. *Variance of the sum of conditional variances.*

Lemma 5. *Under the hypotheses of Theorem 2,*

$$W_n := \operatorname{Var} \sum_{k=1}^{n} \mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}) = o((n^{r+1} \mathbb{E}\, p_{X_n}^r)^2). \tag{44}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Proof. We first rewrite (39) as $\mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}) = \mathbb{E}(A_{n,k}(X_n) - B_{n,k}(X_n, Y_n) \mid \mathcal{F}_{n,k-1}) = \alpha_{n,k} - \beta_{n,k}$, where $A_{n,k}(x) = q_x T_{n,k}^2(x)$, $B_{n,k}(x, y) = \mathbf{1}_{\{x \neq y\}} T_{n,k}(x) T_{n,k}(y)$, $\alpha_{n,k} = \mathbb{E}(A_{n,k}(X_n) \mid \mathcal{F}_{n,k-1})$, and $\beta_{n,k} = \mathbb{E}(B_{n,k}(X_n, Y_n) \mid \mathcal{F}_{n,k-1})$. So, letting $W_n^\alpha = \operatorname{Var} \sum_{k=1}^n \alpha_{n,k}$ and $W_n^\beta = \operatorname{Var} \sum_{k=1}^n \beta_{n,k}$, and noting that $\operatorname{Var}(X + Y) \leq 2(\operatorname{Var} X + \operatorname{Var} Y)$, we have

$$W_n \leq 2W_n^\alpha + 2W_n^\beta. \tag{45}$$

Then,

$$W_n^\alpha = \sum_{k=1}^n \operatorname{Var} \alpha_{n,k} + 2 \sum_{1 \leq k < \ell \leq n} \operatorname{Cov}(\alpha_{n,k}, \alpha_{n,\ell}), \tag{46}$$

and the analogous formula holds for W_n^β. In what follows we express the variances and covariances of $\alpha_{n,k}$ and $\beta_{n,k}$ in terms of $A_{n,k}(X_n)$ and $B_{n,k}(X_n, Y_n)$. For simplicity, let $Z_n = (X_n, Y_n)$, $Z_n' = (X_n', Y_n')$; then $\operatorname{Var} \alpha_{n,k} = \operatorname{Cov}(A_{n,k}(X_n), A_{n,k}(X_n'))$, $\operatorname{Cov}(\alpha_{n,k}, \alpha_{n,\ell}) = \operatorname{Cov}(A_{n,k}(X_n), A_{n,\ell}(X_n'))$, $\operatorname{Var} \beta_{n,k} = \operatorname{Cov}(B_{n,k}(Z_n), B_{n,k}(Z_n'))$, and $\operatorname{Cov}(\beta_{n,k}, \beta_{n,\ell}) = \operatorname{Cov}(B_{n,k}(Z_n), B_{n,\ell}(Z_n'))$, where X_n' and Y_n' are such that $X_n, X_n', Y_n, Y_n', X_{n,1}, \ldots, X_{n,n}$ are i.i.d. for any $n \geq 1$. We only check the first formula; the others are obtained similarly:

$$\begin{aligned}
\mathbb{E}\, \alpha_{n,k}^2 &= \mathbb{E}(\mathbb{E}(A_{n,k}(X_n) \mid \mathcal{F}_{n,k-1}) \mathbb{E}(A_{n,k}(X_n') \mid \mathcal{F}_{n,k-1})) \\
&= \mathbb{E}(\mathbb{E}(A_{n,k}(X_n) A_{n,k}(X_n') \mid \mathcal{F}_{n,k-1})) \\
&= \mathbb{E}(A_{n,k}(X_n) A_{n,k}(X_n')), \\
(\mathbb{E}\, \alpha_{n,k})^2 &= (\mathbb{E}\, A_{n,k}(X_n))^2 = (\mathbb{E}\, A_{n,k}(X_n))(\mathbb{E}\, A_{n,k}(X_n')),
\end{aligned}$$

and the formula for $\operatorname{Var} \alpha_{n,k}$ follows.

We now compute bounds for the covariances defined above. Since $A_{n,k}(x)$ and $B_{n,k}(x, y)$ are bounded above by $T_{n,k}(x) \leq 1$, reasoning as in the paragraph preceding (42) we have

$$\operatorname{Cov}(A_{n,k}(X_n), A_{n,k}(X_n')) \leq \mathbb{E}\, A_{n,k}(X_n) A_{n,k}(X_n') \leq \mathbb{E}\, T_{n,k}(X_n) \leq n^r \mathbb{E}\, p_{X_n}^r \tag{47}$$

$$\operatorname{Cov}(B_{n,k}(Z_n), B_{n,k}(Z_n')) \leq \mathbb{E}\, B_{n,k}(Z_n) B_{n,k}(Z_n') \leq \mathbb{E}\, T_{n,k}(X_n) \leq n^r \mathbb{E}\, p_{X_n}^r. \tag{48}$$

Next, we handle $\operatorname{Cov}(A_{n,k}(X_n), A_{n,\ell}(X_n'))$, which requires somewhat more effort than the previous covariances because the crude bounds do not yield the right order in n. Since $A_{n,k}(x) = (1 - p_x) T_{n,k}^2(x)$,

$$\operatorname{Cov}(A_{n,k}(X_n), A_{n,\ell}(X_n')) = \operatorname{Cov}(T_{n,k}^2(X_n), T_{n,\ell}^2(X_n')) + O(n^r \mathbb{E}\, p_{X_n}^{r+1}) \tag{49}$$

because each of the remaining three covariances is bounded by an expression of the form $\mathbb{E}\, p_{X_n} T_{n,k}(X_n) \leq c n^r \mathbb{E}\, p_{X_n}^{r+1}$. To bound the covariance between $T_{n,k}^2(X_n)$ and $T_{n,\ell}^2(X_n')$ we write

$$\begin{aligned}
\mathbb{E}\, T_{n,k}^2(X_n) T_{n,\ell}^2(X_n') &= \mathbb{E}\, \mathbf{1}_{\{X_n = X_n'\}} T_{n,k}^2(X_n) T_{n,\ell}^2(X_n) \\
&\quad + \mathbb{E}\, \mathbf{1}_{\{X_n \neq X_n'\}} T_{n,k}^2(X_n) T_{n,\ell}^2(X_n')
\end{aligned} \tag{50}$$

and note that the first expectation in (50) is bounded by

$$\mathbb{E}\, \mathbf{1}_{\{X_n = X_n'\}} T_{n,k}(X_n) = \mathbb{E}\, p_{X_n} T_{n,k}(X_n) \leq c n^r \mathbb{E}\, p_{X_n}^{r+1}, \tag{51}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

where c is a positive constant. For the second expectation in (50) we have the following expression, written in terms of (conditionally independent) binomial random variables B_1, B_2, B'_1, B'_2:

$$\mathbb{E}\,\mathbf{1}_{\{X_n\neq X'_n\}}\mathbb{P}(B_1 \geq r - N_{n,k}(X_n),\, B_2 \geq r - N_{n,k}(X_n),\\ B'_1 \geq r - N_{n,\ell}(X'_n),\, B'_2 \geq r - N_{n,\ell}(X'_n) \mid X_n, X'_n). \tag{52}$$

Conditionally on (X_n, X'_n), B_1, B_2, B'_1, and B'_2 are independent, with B_1, B_2 distributed as $\mathrm{Bin}(n-k, p_{X_n})$ and B'_1, B'_2 as $\mathrm{Bin}(n-k, p_{X'_n})$. Further, B_1, B_2, B'_1, and B'_2 are independent of $\mathcal{F}_{n,k}$, $\mathcal{F}_{n,\ell}$, conditionally on (X_n, X'_n).

Observe that (52) can be rewritten as

$$\mathbb{E}\,\mathbf{1}_{\{X_n\neq X'_n\}}\mathbb{P}(N_{n,k}(X_n) \geq r - B_{12},\, N_{n,\ell}(X'_n) \geq r - B'_{12}\} \mid X_n, X'_n), \tag{53}$$

where $B_{12} = \min\{B_1, B_2\}$ and $B'_{12} = \min\{B'_1, B'_2\}$. Note also that, for $x \neq y$, $N_{n,k}(x)$ and $N_{n,l}(y)$ are NUOD; see (15). Thus, conditioning on the values of the binomials, using the NUOD property, then integrating over the Bs and using the independence of X_n and X'_n, we have the following upper bound for (53): $\mathbb{E}\,\mathbf{1}_{\{X_n\neq X'_n\}}\mathbb{P}(N_{n,k}(X_n) \geq r - B_{12} \mid X_n)\mathbb{P}(N_{n,\ell}(X'_n) \geq r - B'_{12}\} \mid X'_n)$, which, after ignoring the indicator and noting that the conditional probabilities (on X_n and X'_n) are independent random variables, can be finally bounded by

$$\mathbb{E}\,\mathbb{P}(N_{n,k}(X_n) \geq r - B_{12} \mid X_n)\mathbb{E}\,\mathbb{P}(N_{n,\ell}(X'_n) \geq r - B'_{12}\} \mid X'_n) = \mathbb{E}\,T^2_{n,k}(X_n)\mathbb{E}\,T^2_{n,\ell}(X'_n). \tag{54}$$

Therefore, from (49), (50), (51), and (54), we have

$$\mathrm{Cov}(T^2_{n,k}(X_n),\, T^2_{n,\ell}(X'_n)) \leq cn^r\mathbb{E}\,p^{r+1}_{X_n}. \tag{55}$$

It remains to bound the covariances $\mathrm{Cov}(B_{n,k}(Z_n),\, B_{n,\ell}(Z'_n))$. To that end we first consider the expected value of the product:

$$\mathbb{E}\,B_{n,k}(Z_n)B_{n,\ell}(Z'_n) \leq \mathbb{E}\,\mathbf{1}_D T_{n,k}(X_n)T_{n,k}(Y_n)T_{n,\ell}(X'_n)T_{n,\ell}(Y'_n)\\ + \mathbb{E}\,\mathbf{1}_{D^c} T_{n,k}(X_n)T_{n,k}(Y_n)T_{n,\ell}(X'_n)T_{n,\ell}(Y'_n), \tag{56}$$

where D is the event that X_n, Y_n, X'_n, and Y'_n are all distinct. Then,

$$\mathbb{E}\,\mathbf{1}_{D^c} T_{n,k}(X_n)T_{n,k}(Y_n)T_{n,\ell}(X'_n)T_{n,\ell}(Y'_n) \leq \binom{4}{2}\,\mathbb{E}\,p_{X_n}T_{n,k}(X_n) \leq cn^r\mathbb{E}\,p^{r+1}_{X_n}. \tag{57}$$

Note that, as in (52), the first term on the right-hand side of (56) can be written as

$$\mathbb{E}\,\mathbf{1}_D\mathbb{P}(B_1 \geq r - N_{n,k}(X_n),\, B_2 \geq r - N_{n,k}(Y_n),\\ B'_1 \geq r - N_{n,\ell}(X'_n),\, B'_2 \geq r - N_{n,\ell}(Y'_n) \mid Z_n, Z'_n). \tag{58}$$

Conditionally on (Z_n, Z'_n), B_1, B_2, B'_1, and B'_2 are independent, where B_1 is $\mathrm{Bin}(n-k, p_{X_n})$, B_2 is $\mathrm{Bin}(n-k, p_{Y_n})$, B'_1 is $\mathrm{Bin}(n-k, p_{X'_n})$, and B'_2 is $\mathrm{Bin}(n-k, p_{Y'_n})$. Also, B_1, B_2, B'_1, and B'_2

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

are independent of $\mathcal{F}_{n,k}$, $\mathcal{F}_{n,\ell}$ conditionally on (Z_n, Z'_n). Now, using the NUOD property (14) and the independence of X_n, Y_n, X'_n, and Y'_n, the expression in (58) is bounded above by

$$\begin{aligned}
&\mathbb{E}\,\mathbb{P}(B_1 \geq r - N_{n,k}(X_n),\, B_2 \geq r - N_{n,k}(Y_n) \mid Z_n) \\
&\quad \times \mathbb{P}(B'_1 \geq r - N_{n,\ell}(X'_n),\, B'_2 \geq r - N_{n,\ell}(Y'_n) \mid Z'_n) \\
&= \mathbb{E}\,\mathbb{P}(B_1 \geq r - N_{n,k}(X_n),\, B_2 \geq r - N_{n,k}(Y_n) \mid Z_n) \\
&\quad \times \mathbb{E}\,\mathbb{P}(B'_1 \geq r - N_{n,\ell}(X'_n),\, B'_2 \geq r - N_{n,\ell}(Y'_n) \mid Z'_n) \\
&= \mathbb{E}\,T_{n,k}(X_n)T_{n,k}(Y_n)\mathbb{E}\,T_{n,\ell}(X'_n)T_{n,l}(Y'_n) \\
&= \mathbb{E}\,B_{n,k}(Z_n)\mathbb{E}\,B_{n,\ell}(Z'_n) + O(n^r\mathbb{E}\,p_{X_n}^{r+1}).
\end{aligned} \tag{59}$$

Therefore, from (56), (57), and (59),

$$\operatorname{Cov}(B_{n,k}(Z_n),\, B_{n,\ell}(Z'_n)) \leq cn^r\mathbb{E}\,p_{X_n}^{r+1}. \tag{60}$$

We complete the proof of (44) by collecting the partial results above to obtain bounds for W_n^α and W_n^β, using formula (46). From (47) and (48), $\sum_{k=1}^n \operatorname{Var}\alpha_{n,k} \leq n^{r+1}\mathbb{E}\,p_{X_n}^r$ and $\sum_{k=1}^n \operatorname{Var}\beta_{n,k} \leq n^{r+1}\mathbb{E}\,p_{X_n}^r$. From (49) and (55),

$$\sum_{1\leq k<\ell\leq n} \operatorname{Cov}(\alpha_{n,k},\, \alpha_{n,\ell}) \leq c\binom{n}{2} n^r\mathbb{E}\,p_{X_n}^{r+1} \leq cnp_n^*\big(n^{r+1}\mathbb{E}\,p_{X_n}^r\big) = o((n^{r+1}\mathbb{E}\,p_{X_n}^r)^2).$$

Finally, from (60),

$$\sum_{1\leq k<\ell\leq n} \operatorname{Cov}(\beta_{n,k},\, \beta_{n,\ell}) \leq c\binom{n}{2} n^r\mathbb{E}\,p_{X_n}^{r+1} \leq cnp_n^*\big(n^{r+1}\mathbb{E}\,p_{X_n}^r\big) = o((n^{r+1}\mathbb{E}\,p_{X_n}^r)^2).$$

The conclusion follows from (45), (46), and the bounds for the sums of variances and covariances above. □

4.2.4. *The final step: the martingale CLT.*

Proof of Theorem 2. We establish asymptotic normality by applying the martingale central limit theorem (see, e.g., [8, Theorem 2.5]) to the martingale differences $(d_{n,k})$. Since the $d_{n,k}$ are uniformly bounded, the conditional Lindeberg condition ([8, Condition (2.5)]) follows from the fact that the variance of the sum grows to infinity. The remaining condition to be checked, [8, Condition (2.7)], is

$$\frac{\sum_{k=1}^n \mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1})}{\sum_{k=1}^n \mathbb{E}\,d_{n,k}^2} \xrightarrow{\mathbb{P}} 1,$$

or, equivalently,

$$\frac{\sum_{k=1}^n (\mathbb{E}(d_{n,k}^2 \mid \mathcal{F}_{n,k-1}) - \mathbb{E}\,d_{n,k}^2)}{\sum_{k=1}^n \mathbb{E}\,d_{n,k}^2} \xrightarrow{\mathbb{P}} 0.$$

But this follows immediately from the second part of Proposition 1, Lemma 5, and Chebyshev's inequality. □

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

4.3. Phase transition

Proof of Proposition 2. Note that the Poissonian asymptotics at the critical capacity r follows immediately from Theorem 1.

When $r \geq 2$ and $s \in \{1, \ldots, r-1\}$, it follows from (22) of Lemma 1 that the assumptions of Theorem 2 are satisfied and thus we have the normal asymptotics.

When $r \geq 1$ and $s \in \{r+1, r+2, \ldots\}$, following (27), we can write, for any $\varepsilon > 0$,

$$\mathbb{P}(V_{n,s} > \varepsilon) \leq \frac{\mathbb{E}\, V_{n,s}}{\varepsilon} \leq \frac{n^{s+1}\mathbb{E}\, p_{X_n}^s}{\varepsilon\,(s+1)!}.$$

Thus, (21) of Lemma 1 yields $V_{n,s} \xrightarrow{\mathbb{P}} 0$. □

4.4. Asymptotics for full containers

Proof of Theorem 3. For the case $r > 1$, due to the representations in (10), to prove both results it suffices to show that $\mathbb{E}\, V_{n,s} \to 0$ for any fixed $s \geq r$. Since (27) yields

$$\mathbb{E}\, V_{n,s} \leq \frac{n^{s+1}}{(s+1)!}\mathbb{E}\, p_{X_n}^s,$$

the result follows from (22) of Lemma 1.

For the case $r = 1$, the first part follows from Theorem 1 since (11) implies $n - L_{n,1} = V_{n,1}$. The second also follows from Theorem 1 since (12) gives $\frac{1}{2}(n - K_{n,1}) = V_{n,1} - \frac{1}{2}V_{n,2}$ and, similarly to the case $r > 1$, we have $\mathbb{E}\, V_{n,2} \to 0$. □

Proof of Theorem 4. By the representation in (10) we can write

$$\frac{\operatorname{Var} L_{n,r}}{\operatorname{Var} V_{n,r-1}} = 1 + \frac{\operatorname{Var} V_{n,r}}{\operatorname{Var} V_{n,r-1}} - 2\frac{\operatorname{Cov}(V_{n,r-1}, V_{n,r})}{\operatorname{Var} V_{n,r-1}}.$$

Since $n^{r+1}\mathbb{E}\, p_{X_n}^r \leq np_n^* n^r \mathbb{E}\, p_{X_n}^{r-1}$, it follows that $n^r \mathbb{E}\, p_{X_n}^{r-1} \to \infty$. Therefore, by Proposition 1, we have

$$\frac{\operatorname{Var} V_{n,r}}{\operatorname{Var} V_{n,r-1}} \leq c\frac{n^{r+1}\mathbb{E}\, p_{X_n}^r}{n^r \mathbb{E}\, p_{X_n}^{r-1}} \leq cnp_n^* \to 0.$$

Thus,

$$\left|\frac{\operatorname{Cov}(V_{n,r-1}, V_{n,r})}{\operatorname{Var} V_{n,r-1}}\right| \leq \sqrt{\frac{\operatorname{Var} V_{n,r}}{\operatorname{Var} V_{n,r-1}}} \to 0,$$

and so

$$\frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\operatorname{Var} L_{n,r}}} \xrightarrow{L^2} 0.$$

Hence, the first result is a consequence of Theorem 2 since, in view of the representation in (10),

$$\frac{L_{n,r} - \mathbb{E}\, L_{n,r}}{\sqrt{\operatorname{Var} L_{n,r}}} = \frac{V_{n,r-1} - \mathbb{E}\, V_{n,r-1}}{\sqrt{\operatorname{Var} V_{n,r-1}}}\sqrt{\frac{\operatorname{Var} V_{n,r-1}}{\operatorname{Var} L_{n,r}}} - \frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\operatorname{Var} L_{n,r}}}.$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

For the second case, by the representation in (10) we can write

$$\frac{\operatorname{Var} K_{n,r}}{\operatorname{Var} V_{n,r-1}} = 1 + 4\frac{\operatorname{Var} V_{n,r}}{\operatorname{Var} V_{n,r-1}} + \frac{\operatorname{Var} V_{n,r+1}}{\operatorname{Var} V_{n,r-1}} - 4\frac{\operatorname{Cov}(V_{n,r-1}, V_{n,r})}{\operatorname{Var} V_{n,r-1}} - 4\frac{\operatorname{Cov}(V_{n,r}, V_{n,r+1})}{\operatorname{Var} V_{n,r-1}} + 2\frac{\operatorname{Cov}(V_{n,r-1}, V_{n,r+1})}{\operatorname{Var} V_{n,r-1}}.$$

Similarly to the previous case, we conclude that $n^s \mathbb{E}\, p_{X_n}^{s-1} \to \infty$ for $s = r, r+1$. Therefore, by the same argument as above, it follows that each of the summands on the right-hand side of the expression above, excluding the first one, converges to 0. Consequently, for $s = r, r+1$,

$$\frac{V_{n,s} - \mathbb{E}\, V_{n,s}}{\sqrt{\operatorname{Var} K_{n,r}}} \xrightarrow{L^2} 0.$$

Thus, the second result is a consequence of Theorem 2 since, in view of (10),

$$\frac{K_{n,r} - \mathbb{E}\, K_{n,r}}{\sqrt{\operatorname{Var} K_{n,r}}} = \frac{V_{n,r-1} - \mathbb{E}\, V_{n,r-1}}{\sqrt{\operatorname{Var} V_{n,r-1}}} \sqrt{\frac{\operatorname{Var} V_{n,r-1}}{\operatorname{Var} K_{n,r}}} - 2\frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\operatorname{Var} K_{n,r}}} + \frac{V_{n,r+1} - \mathbb{E}\, V_{n,r+1}}{\sqrt{\operatorname{Var} K_{n,r}}}.$$

□

5. More on the asymptotics of the mean

As mentioned in Remark 1, when $\lambda = \limsup np_n^* = 0$ the limit of $\mathbb{E}\, V_{n,r}/n^{r+1}\mathbb{E}\, p_{X_n}^r$ exists, but when $\lambda > 0$ we were only able to obtain lower and upper bounds for that ratio. Here, under additional assumptions, we consider the existence of the limit when $\lambda > 0$. We begin by re-writing an expression for $\mathbb{E}\, V_{n,r}$ in a more convenient form.

Lemma 6.

$$\mathbb{E}\, V_{n,r} = \sum_{s=r}^{n-1} \binom{n}{s+1} \binom{s-1}{s-r} (-1)^{s-r} \mathbb{E}\, p_{X_n}^s. \tag{61}$$

Proof. Recall, for example from (18), that

$$\mathbb{E}\, V_{n,r} = \sum_{k=r+1}^{n} \mathbb{E}\, Y_{n,k} = \mathbb{E} \sum_{k=r+1}^{n} \sum_{i=r}^{k-1} \binom{k-1}{i} p_{X_n}^i (1 - p_{X_n})^{k-1-i}.$$

Expanding $(1 - p_{X_n})^{k-1-i}$ by the binomial formula, we see that the double sum on the right-hand side is

$$\sum_{k=r+1}^{n} \sum_{i=r}^{k-1} \binom{k-1}{i} \sum_{j=0}^{k-1-i} \binom{k-1-i}{j} (-1)^j p_{X_n}^{i+j} = \sum_{i=r}^{n-1} \sum_{j=0}^{n-1-i} \binom{i+j}{i} (-1)^j p_{X_n}^{i+j} \sum_{k=i+j+1}^{n} \binom{k-1}{i+j}.$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Since $\sum_{k=i+j+1}^{n}\binom{k-1}{i+j}=\binom{n}{i+j+1}$, we get

$$\begin{aligned}\sum_{k=r+1}^{n}\sum_{i=r}^{k-1}\binom{k-1}{i}p_{X_n}^{i}(1-p_{X_n})^{k-1-i}&=\sum_{i=r}^{n-1}\sum_{j=0}^{n-1-i}\binom{i+j}{i}\binom{n}{i+j+1}(-1)^{j}p_{X_n}^{i+j}\\&=\sum_{i=r}^{n-1}\sum_{s=i}^{n-1}\binom{s}{i}\binom{n}{s+1}(-1)^{s-i}p_{X_n}^{s}\\&=\sum_{s=r}^{n-1}\binom{n}{s+1}p_{X_n}^{s}\sum_{i=r}^{s}\binom{s}{i}(-1)^{s-i}.\end{aligned}$$

The final result follows from the identity $\sum_{i=r}^{s}\binom{s}{i}(-1)^{s-i}=\binom{s-1}{s-r}(-1)^{s-r}$. □

To state our condition we need to introduce one more definition. Let $\mathcal{X}_n$ denote the set of distinct values among $p_{n,k}/p_n^*$, $k\in M_n$. For $x\in\mathcal{X}_n$, denote $K(x)=\{k\in M_n : x=p_{n,k}/p_n^*\}$. Define random variables P_n, $n\geq 1$, as follows:

$$\mathbb{P}(P_n=x)=\frac{1}{\mathbb{E}\,p_{X_n}^{r}}\sum_{k\in K(x)}p_{n,k}^{r+1},\qquad x\in\mathcal{X}_n.$$

Definition 1. We say that the sequence $(X_n)_{n\geq 1}$ is in the class $\mathcal{T}(r)$ if the sequence $(P_n)_{n\geq 1}$ converges in distribution.

Denote by U the limit of $(P_n)_{n\geq 1}$ when it exists, and let ν_r denote the distribution of U. Note that since U is a [0, 1]-valued random variable, $(X_n)_{n\geq 1}\in\mathcal{T}(r)$ if and only if $\mathbb{E}\,P_n^{j}\to\mathbb{E}\,U^{j}$, $j\geq 1$.

We will show that, for $(X_n)_{n\geq 1}\in\mathcal{T}(r)$, if $\lim_{n\to\infty}np_n^*$ exists and is positive then

$$H(r,\lambda):=\lim_{n\to\infty}\frac{\mathbb{E}\,V_{n,r}}{n^{r+1}\mathbb{E}\,p_{X_n}^{r}}$$

also exists.

Recall that the generalized hypergeometric function ${}_pF_q$ is defined by

$${}_pF_q(a_1,\dots,a_p;b_1,\dots,b_q;z)=\sum_{k=0}^{\infty}\frac{(a_1)_k\cdots(a_p)_k}{(b_1)_k\cdots(b_q)_k}\frac{z^k}{k!},\tag{62}$$

where $(a)_0=1$ and $(a)_k=a(a+1)\cdots(a+k-1)$ for $k\geq 1$. We derive the following representation for $H(r,\lambda)$.

Proposition 3. *Let $(X_n)_{n\geq 1}$ be in $\mathcal{T}(r)$ and $np_n^*\to\lambda>0$. Then*

$$H(r,\lambda)=\frac{\mathbb{E}\,\mathrm{e}^{-\lambda BU}}{(r+1)!}=\frac{1}{(r+1)!}\int {}_1F_1(r;r+2;\,-\lambda u)\,\nu_r(\mathrm{d}u),\tag{63}$$

where B is a beta $\mathrm{B}_I(r,2)$ random variable with density $f(b)=r(r+1)b^{r-1}(1-b)\mathbf{1}_{(0,1)}(b)$.

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Proof. Using (61) we get

$$\begin{aligned}\frac{\mathbb{E}\,V_{n,r}}{n^{r+1}\mathbb{E}\,p_{X_n}^r} &= \frac{1}{n^{r+1}}\sum_{s=r}^{n-1}\frac{n(n-1)\dots(n-s)}{s(s+1)}\frac{(-1)^{s-r}}{(s-r)!(r-1)!}\frac{\mathbb{E}\,p_{X_n}^s}{\mathbb{E}\,p_{X_n}^r}\\ &= \frac{1}{(r-1)!}\sum_{s=r}^{n-1}\frac{n(n-1)\dots(n-s)}{n^{s+1}s(s+1)}\frac{(-1)^{s-r}}{(s-r)!}(np_n^*)^{s-r}\mathbb{E}\,P_n^{s-r}\\ &\to \frac{1}{(r-1)!}\sum_{\ell\ge 0}\frac{(-\lambda)^\ell}{\ell!}\frac{\mathbb{E}\,U^\ell}{(r+\ell)(r+1+\ell)} = H(r,\lambda),\end{aligned}$$

where the last line holds by the Lebesgue dominated convergence theorem.

Since

$$\frac{1}{(r+j+1)(r+j)} = \frac{1}{r+j} - \frac{1}{r+j+1} = \int_0^1 \left(x^{r+j-1} - x^{r+j}\right)\mathrm{d}x,$$

we get

$$\begin{aligned}H(r,\lambda) &= \frac{1}{(r-1)!}\sum_{\ell\ge 0}\frac{(-\lambda)^\ell}{\ell!}\int_{[0,1]}u^\ell\nu_r(\mathrm{d}u)\int_0^1 x^{r+\ell-1}(1-x)\mathrm{d}x\\ &= \frac{1}{(r-1)!}\int_{[0,1]}\int_0^1 x^{r-1}(1-x)\sum_{\ell\ge 0}\frac{(-\lambda xu)^\ell}{\ell!}\,\mathrm{d}x\,\nu_r(\mathrm{d}u)\\ &= \frac{1}{(r+1)!}\int_0^1\int_{[0,1]}\mathrm{e}^{-\lambda xu}r(r+1)x^{r-1}(1-x)\nu_r(\mathrm{d}u)\mathrm{d}x = \frac{\mathbb{E}\,\mathrm{e}^{-\lambda BU}}{(r+1)!},\end{aligned}$$

where B is as specified earlier. This proves the first equality in (63). The second follows by recalling that the Laplace transform of the beta $\mathrm{B}_I(\alpha,\beta)$ random variable B is $\mathbb{E}\,\mathrm{e}^{sB} = {}_1F_1(\alpha;\alpha+\beta;s)$. Applying this formula conditionally on U in $\mathbb{E}\,\mathrm{e}^{-\lambda BU}$ and then integrating with respect to U yields the second equality in (63). □

The above representation allows us to give a short proof of the bounds (8) in Proposition 1 for the limit $H(r,\lambda)$ in $\mathcal{T}(r)$ models. While the upper bound remains the same, the lower is tighter and, as we will see in Example 5, is exact.

Proposition 4. *Let $(X_n)_{n\ge 1}$ belong to the class $\mathcal{T}(r)$ and $np_n^* \to \lambda > 0$. Then*

$$\frac{\Gamma_\lambda(r+1)}{r!} < \frac{{}_1F_1(r;r+2;-\lambda)}{(r+1)!} \le H(r,\lambda) < \frac{1}{(r+1)!}. \tag{64}$$

Proof. The upper bound is clear since $\mathbb{E}\,\mathrm{e}^{-\lambda BU} < 1$. The inequality is strict since $\mathbb{P}(U > 0) > 0$.

On the other hand, $\mathbb{P}(B \le 1) = 1$. Thus, by the first part of (63), we get

$$H(r,\lambda) \ge \frac{1}{(r+1)!}\mathbb{E}\,\mathrm{e}^{-\lambda U} = \frac{{}_1F_1(r;r+2;-\lambda)}{(r+1)!},$$

which proves the second inequality in (64). Finally, to prove the first one, note that

$$\frac{{}_1F_1(r;r+2;-\lambda)}{(r+1)!} = \frac{\int_0^1 \mathrm{e}^{-\lambda x}x^{r-1}(1-x)\mathrm{d}x}{(r-1)!}.$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Integrating the numerator by parts yields

$$\frac{1}{r}\int_0^1 (1-x)\mathrm{e}^{-\lambda x}\,\mathrm{d}x^r = \frac{1}{r}\int_0^1 x^r(\mathrm{e}^{-\lambda x}+\lambda(1-x)\mathrm{e}^{-\lambda x})\,\mathrm{d}x$$
$$> \frac{1}{r}\int_0^1 x^r\mathrm{e}^{-\lambda x}\,\mathrm{d}x = \frac{\Gamma_\lambda(r+1)}{r}. \qquad \square$$

We close by considering several special cases, the first of which shows that the middle inequality in (64) is sharp.

Example 5. (*Uniform distribution.*) Let $p_{n,j}=1/m_n$, $j\in M_n=\{1,\dots,m_n\}$, $n\ge 1$. Assume that $n/m_n\to\lambda>0$. Note that in this case $P_n=1$ $\mathbb{P}$-a.s., $n\ge 1$, and thus $\nu_r=\delta_1$. Hence, by (63),

$$H(r,\lambda)=\frac{{}_1F_1(r;r+2;\,-\lambda)}{(r+1)!}.$$

Example 6. (*Geometric distribution.*) Let $p_{n,j}=p_n(1-p_n)^j$, $j\in M_n=\{0,1,\dots\}$, $n\ge 1$. Assume that $np_n\to\lambda>0$ (note that in this case $p_n^*=p_n$). Then, $\nu_r(\mathrm{d}u)=(r+1)u^r\mathbf{1}_{[0,1]}(u)\mathrm{d}u$ and

$$H(r,\lambda)=\frac{{}_2F_2(r,r+1;r+2,r+2;\,-\lambda)}{(r+1)!}. \tag{65}$$

Indeed, with $q_n:=1-p_n$ we have $\mathbb{P}(P_n=q_n^k)=q_n^{(r+1)k}(1-q_n^{r+1})$, $k=0,1,\dots$ Therefore, for any $\ell=1,2,\dots$,

$$\mathbb{E}\,P_n^\ell=\sum_{k\ge 0}q_n^{(\ell+r+1)k}(1-q_n^{r+1})=\frac{1-q_n^{r+1}}{1-q_n^{\ell+r+1}}\to\frac{r+1}{\ell+r+1}=(r+1)\int_0^1 x^{\ell+r}\,\mathrm{d}x,$$

which implies the assertion on the form of ν_r.

Thus, (63) yields

$$H(r,\lambda)=\frac{1}{r!}\int_0^1 {}_1F_1(r,r+2;-\lambda u)u^r\,\mathrm{d}u.$$

Using the Euler integral identity (see, e.g., [22, Eq. 16.5.2]),

$${}_{p+1}F_{q+1}(a_1,\dots,a_p,c;b_1,\dots,b_q,c+d;z)$$
$$=\frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}\int_0^1 t^{c-1}(1-t)^{d-1}{}_pF_q(a_1,\dots,a_p;b_1,\dots,b_q;zt)\mathrm{d}t,$$

with $p=q=1$, $c=r+1$, and $d=1$ yields (65).

Example 7. (*Riemann ζ distribution.*) Let $p_{n,j}=j^{-\alpha_n}/\zeta(\alpha_n)$, $j\in M_n=\{1,2,\dots\}$, and $\alpha_n>1$, $n\ge 1$. Assume that $n(\alpha_n-1)\to\lambda$. Then

$$\nu_r=\sum_{k\ge 1}\frac{k^{-(r+1)}}{\zeta(r+1)}\delta_{1/k}. \tag{66}$$

and

$$H(r,\lambda)=\frac{1}{(r+1)!}\int_{\mathbb{R}}{}_1F_2(r;r+1,r+2;\,-\lambda x)\,\mu_r(\mathrm{d}x), \tag{67}$$

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

where μ_r is the probability measure defined on $(0, \infty)$ by

$$\mu_r(\mathrm{d}x) = \frac{1}{r!\zeta(r+1)} \frac{x^r}{\mathrm{e}^x - 1}\, \mathrm{d}x. \tag{68}$$

We first justify (66). Since $p_n^* = 1/\zeta(\alpha_n)$ we have

$$\mathbb{P}(P_n = k^{-\alpha_n}) = \frac{k^{-\alpha_n(r+1)}}{\zeta(\alpha_n(r+1))}.$$

We have $\alpha_n \to 1$ and thus, by Lebesgue dominated convergence,

$$\mathbb{E}\, P_n^\ell = \sum_{k\geq 1} k^{-\alpha_n \ell} \frac{k^{-\alpha_n(r+1)}}{\zeta(\alpha_n(r+1))} \to \sum_{k\geq 1} k^{-\ell} \frac{k^{-(r+1)}}{\zeta(r+1)}, \qquad \ell = 1, 2, \ldots,$$

which proves the assertion on ν_r.

By (63) we have

$$H(r, \lambda) = \frac{1}{(r+1)!\zeta(r+1)} \sum_{k\geq 1} {}_1F_1\left(r; r+2; -\frac{\lambda}{k}\right) \frac{1}{k^{r+1}}.$$

Expanding ${}_1F_1$ according to (62) and changing the order of the sums, we get

$$H(r, \lambda) = \frac{1}{(r+1)!\zeta(r+1)} \sum_{j\geq 0} \frac{(r)_j}{(r+2)_j} \frac{(-\lambda)^j}{j!} \zeta(j+r+1). \tag{69}$$

Let us recall an integral identity for product of ζ and Γ functions (see, e.g., [22, Eq. 25.5.1]):

$$\zeta(s)\Gamma(s) = \int_0^\infty \frac{x^{s-1}}{\mathrm{e}^x - 1}\, \mathrm{d}x, \qquad \mathrm{Re}(s) > 1. \tag{70}$$

Inserting (70) into the right-hand side of (69) we get

$$\begin{aligned} H(r, \lambda) &= \frac{1}{(r+1)!\zeta(r+1)} \sum_{j\geq 0} \frac{(r)_j}{(r+2)_j} \frac{1}{\Gamma(r+j+1)} \int_0^\infty \frac{x^{r+j}}{\mathrm{e}^x - 1}\, \mathrm{d}x\, \frac{(-\lambda)^j}{j!} \\ &= \frac{1}{r!(r+1)!\zeta(r+1)} \int_0^\infty \left(\sum_{j\geq 0} \frac{(r)_j}{(r+1)_j (r+2)_j} \frac{(-\lambda x)^j}{j!} \right) \frac{x^r}{\mathrm{e}^x - 1}\, \mathrm{d}x, \end{aligned}$$

which proves (67) and (68).

Remark 2. Central limit theorems for various parameters (including the number of occupied urns) for infinite urn models with assumptions on the probabilities (p_j), $j \geq 1$, similar to those on $(p_{n,j})_{j\geq 1}$, $n \geq 1$, in Example 7 have been investigated in, e.g., [13].

Acknowledgements

We wish to thank the Editor for his consideration and careful handling of the manuscript.

https://doi.org/10.1017/jpr.2021.87 Published online by Cambridge University Press

Funding information

The first author acknowledges financial support from grants PIA AFB-170001 and Fondecyt 1161319. The second author wishes to state that this material is based upon work supported by and while serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The third author acknowledges partial support through project 2016/21/B/ST1/00005 of the National Science Center (NCN), Poland.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Arratia, R., Garibaldi, S. and Kilian, J. (2016). Asymptotic distribution for the birthday problem with multiple coincidences, via an embedding of the collision process. *Random Structures Algorithms* **48**, 480–502.
[2] Barbour, A. Gnedin, A. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Prob.* **14**, 13, 365–384.
[3] Beśka, M., Kłopotowski, A. and Słomiński, L. (1982). Limit theorems for random sums of dependent *d*-dimensional random vectors. *Z. Wahrscheinlichkeitsth.* **61**, 43–57.
[4] Bobecka, K., Hitczenko, P., López-Blázquez, F., Rempała, G. and Wesołowski, J. (2013). Asymptotic normality through factorial cumulants and partition identities. *Combinatorics Prob. Comput.* **22**, 213–240.
[5] Chao, A. and Chiu, C.-H. (2016). Species richness: Estimation and comparison. In *Wiley StatsRef: Statistics Reference Online* (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels), John Wiley, London.
[6] Dupuis, P., Nuzman, C. and Whiting, P. (2004). Large deviation asymptotics for occupancy problems. *Ann. Prob.* **32**, 2765–2818.
[7] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Prob. Surv.* **4**, 146–171.
[8] Helland, I. S. (1982). Central limit theorems for martingales with discrete or continuous time. *Scand. J. Statist.* **9**, 79–94.
[9] Hwang, H. K. and Janson, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Prob.* **36**, 992–1022.
[10] Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *Ann. Statist.* **11**, 286–295.
[11] Jogdeo, K and Patil G. P. (1975). Probability inequalities for certain multivariate discrete distribution. *Sankhyā B* **37**, 158–164.
[12] Johnson N. L. and Kotz, S. (1977). *Urn Models and Their Application*. *Wiley Series in Probability and Mathematical Statistics*. John Wiley, New York.
[13] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17**, 373–401.
[14] Knuth, D. E. (1998). *The Art of Computer Programming*, Vol. **2**, 3rd edn. Addison-Wesley, Reading, MA.
[15] Knuth, D. E. (1998). *The Art of Computer Programming*, Vol. **3**, 2nd edn. Addison-Wesley, Reading, MA.
[16] Kolchin, V. F., Sevastyanov, B. A. and Chistyakov, V. P. (1978). *Random Allocations*. Winston & Sons, Washington.
[17] Kotz, S. and Balakrishnan, N. (1977). Advances in urn models during the past two decades. In *Advances in Combinatorial Methods and Applications to Probability and Statistics*, ed. N. Balkrishnan. Birkhäuser, Boston, MA, pp. 203–257.
[18] Mahmoud, H. M. (2009). *Pólya Urn Models*. CRC Press, Boca Raton, FL.
[19] Mallows, C. L. (1968). An inequality involving multinomial probabilities. *Biometrika* **55**, 422–424.
[20] Mikhailov, V. G. (1981). The central limit theorem for the scheme of independent placements of particles among cells. *Trudy Steklov Math. Inst. (MIAN)* **157**, 138–152.
[21] Monahan, J. F. (1987). An alternative method for computing the overflow probabilities. *Commun. Statist. Theory Meth.* **16**, 3355–3357.
[22] National Institute of Standards and Technology (2020). *NIST Digital Library of Mathematical Functions*, Release 1.0.28. (2020). Available at: https://dlmf.nist.gov.
[23] Ramakrishna, M. V. (1987). Computing the probability of hash table/urn overflow. *Commun. Statist. Theory Meth.* **16**, 3343–3353.