Some Limit Theorems on Distributional Patterns of Balls in Urns

Samuel Karlin; Ming-Ying Leung

The Annals of Applied Probability, Vol. 1, No. 4 (Nov., 1991), 513-538.

Stable URL:
http://links.jstor.org/sici?sici=1050-5164%28199111%291%3A4%3C513%3ASLTODP%3E2.0.CO%3B2-H

The Annals of Applied Probability is currently published by Institute of Mathematical Statistics.

The Annals of Applied Probability
1991, Vol. 1, No. 4, 513–538

SOME LIMIT THEOREMS ON DISTRIBUTIONAL PATTERNS OF BALLS IN URNS

BY SAMUEL KARLIN[1] AND MING-YING LEUNG[2]

Stanford University and University of Texas at San Antonio

In an independent, equiprobable allocation urn model, there are various Poisson and normal limit laws for the occupancy of single urns. Applying the Chen–Stein method, we obtain Poisson, compound Poisson and multivariate Poisson limit laws, together with estimates of their rates of convergence, for the number of chunks of κ (fixed) adjacent urns occupied by certain numbers of balls distributed in some specified patterns. Several related results on occupancy, waiting time and spacings at certain random times are also presented.

1. Introduction. The original motivation of the results reported in this paper stems from comparative studies on molecular sequences seeking to characterize repetitive structures in DNA and protein sequences [e.g., see Karlin (1986)]. Some problems concerning long repeats in a random letter sequence are idealized into a ball-in-urn model (urns correspond to all DNA words of a given size and balls refer to the observed words in a given sequence). Limit theorems for several generalized occupancy problems are presented. More discussion of the background problems is given at the close of the introduction and some molecular data applications are set forth in Section 8.

A sequence of n indistinguishable balls are allocated independently into an array of m urns following a uniform distribution. Two prominent Poisson limit laws refer to the variable N_r, the number of urns containing r balls: With $n, m \to \infty$, N_r has a Poisson limit law with parameter $c/r!$ if $n/m^{(r-1)/r} \to c > 0$; and if $n = m \ln m + rm \ln\ln m + mx + o(m)$, then N_r has a Poisson limit law with parameter $e^{-x/r!}$. Limiting distributions for the waiting times T_r^- (T_r^+) until some urn first acquires r balls (all urns acquire greater than r balls) ensue through their duality relations with the occupancy problems. A compendium of results of this kind with references can be found in Johnson and Kotz (1977). Kolchin, Sevast'yanov and Chistyakov (1978) contains detailed information on these limit theorems and further generalizations.

Received June 1990; revised December 1990.

[1]Supported in part by NIH Grants GM-39907-02 and GM-10452-26 and NSF Grant DMS-86-06244.

[2]Supported in part by Texas Higher Education Coordinating Board Advanced Research Grant 010115-012.

AMS 1980 *subject classification*. 60F05.

Key words and phrases. Ball-in-urn models, occupancy distributions, Poisson approximations, Chen–Stein method.

In this paper, we consider the number of κ-chunks (each chunk consists of κ adjacent urns) containing a prescribed configuration of balls under the independence uniform allocation scheme. A κ-chunk is said to have configuration $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_\kappa)$ if the urn components of the chunk, respectively, contain $\sigma_1, \sigma_2, \ldots, \sigma_\kappa$ balls in that order. Theorem 1 indicates that the counts of κ-chunks with the different configurations totalling exactly r balls are distributed approximately multivariate Poisson provided $n, m \to \infty$ in a proper relationship. Invoking the Chen–Stein method of establishing Poisson limit laws, a rate of convergence is ascertained.

Many ball-in-urn distributional problems can be handled more expeditiously via an embedding into an appropriate system of independent Poisson processes or equivalently to distributing a Poisson distributed number of balls in the wins and it is easy to obtain the corresponding results for a fixed number of balls from this. This technique, quite old, is explained in Johnson and Kotz (1977) [see also Karlin (1967)]. It is well known that the embedding into Poisson processes is equivalent to distributing a Poisson number of balls into the urns, and it is easy to obtain the corresponding results for a fixed number of balls from this fact. Thus, our results are presented in the context of a system of m independent Poisson processes, where the acquisition of balls in each urn occurs with the events of these Poisson processes. Exceptions are Theorems 6 and 7, where the embedding does not seem to simplify the proof, and Theorem 13, where compound Poisson processes are used to reflect a Markov allocation of balls. We let $Y_1(t), Y_2(t), \ldots, Y_m(t)$ denote m independent Poisson processes, each with parameter $1/m$. κ-chunks correspond to κ contiguous lines in the embedded processes. For simplicity of exposition, we arrange the urns in a circle with Y_1 adjacent to Y_m so that there will be exactly m distinct κ-chunks. The circular arrangement eases the counting of chunks. All the results below hold with the Poisson processes arranged in a linear array.

The following notation is useful. Let

$$Q(n,\kappa) = \left\{\boldsymbol{\sigma}: \boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_\kappa),\ \sigma_i \text{ nonnegative integers and } \sum_{i=0}^{\kappa} \sigma_i = n\right\}; \tag{1.1}$$

$$Q^*(n,\kappa) = \{\boldsymbol{\sigma}: \boldsymbol{\sigma} \in Q(n,\kappa) \text{ and } \sigma_1 \geq 1\}. \tag{1.2}$$

For each $1 \leq i \leq m$ and each $\boldsymbol{\sigma} \in Q(n,\kappa)$, define

$$X_{i,\boldsymbol{\sigma}}(t) = \begin{cases} 1, & \text{if } (Y_i(t), Y_{i+1}(t), \ldots, Y_{i+\kappa-1}(t)) = \boldsymbol{\sigma}, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$N_{\boldsymbol{\sigma}}(m,t) = \sum_{i=1}^{m} X_{i,\boldsymbol{\sigma}}(t)$$

so that $N_{\sigma}(m,t)$ is the number of κ-chunks at time t exhibiting the configuration of events described by σ.

For any two discrete (univariate or vector) random variables W and Z, let

$$\|\mathscr{L}(W) - \mathscr{L}(Z)\| = \sum_{\sigma} |\mathrm{pr}\{W = \sigma\} - \mathrm{pr}\{Z = \sigma\}| \tag{1.3}$$

[$\mathscr{L}(\cdot)$ denotes the probability law], where σ ranges over all possible values for W and Z. This is the familiar total variation distance except for a factor of $\frac{1}{2}$.

THEOREM 1. (a) *If $t, m \to \infty$ in the manner that $t^r/m^{r-1} \to c$ for some constant $0 < c < \infty$ and*

$$\left|\frac{t^r}{m^{r-1}} - c\right| = O\left(\frac{1}{m^p}\right) \quad \text{for some } p > 0, \tag{1.4}$$

then the total variation distance for the joint probability law of $N_{\sigma}(m,t)$, σ restricted to $Q^*(r,\kappa)$, satisfies

$$\left\|\mathscr{L}\left(\{N_{\sigma}(m,t)\}_{\sigma \in Q^*(r,\kappa)}\right) - \mathscr{L}\left(\{Z_{\lambda_{\sigma}}\}_{\sigma \in Q^*(r,\kappa)}\right)\right\| = O(1/m^q),$$

where $q = \min(p, 1/r)$ and $\{Z_{\lambda_{\sigma}}\}_{\sigma \in Q^*(r,\kappa)}$ is a family of independent Poisson random variables with parameter

$$\lambda_{\sigma} = \frac{c}{\sigma_1!\sigma_2! \cdots \sigma_{\kappa}!}.$$

(b) *For $m, t \to \infty$ obeying*

$$t = \frac{1}{\kappa} m \ln m + \frac{r}{\kappa} m \ln\ln m + mc(m), \qquad c(m) \to c, 0 < c < \infty, \tag{1.5}$$

then

$$\left\|\mathscr{L}\left(\{N_{\sigma}(m,t)\}_{\sigma \in Q(r,\kappa)}\right) - \mathscr{L}\left(\{Z_{\lambda_{\sigma}}\}_{\sigma \in Q(r,\kappa)}\right)\right\| = O(c(m) - c) + O\left(\frac{\ln\ln m}{\ln m}\right),$$

where $Z_{\lambda_{\sigma}}$ are independent Poisson random variables with parameter

$$\lambda = \frac{e^{-\kappa c}}{\sigma_1!\sigma_2! \cdots \sigma_{\kappa}!\kappa^r}.$$

Without the Poisson process embedding, when $t \approx m^{(r-1)/r}$ balls are distributed successively into the urns, the error from the Poisson process model is an additional $O(m^{-1/r})$.

For the κ-chunks with configurations $(\sigma_1, \ldots, \sigma_{\kappa})$ fulfilling $\sigma_1 + \cdots + \sigma_{\kappa} \geq r$ $(\leq r)$ the corresponding count variables also tend in distribution to a Poisson limit law as set forth in the following two theorems. These results reduce to the classical occupancy and waiting time theorems if $\kappa = 1$.

THEOREM 2. *For fixed positive integers* κ *and* r *where* $r \geq 2$, *define*

$$X_i(t) = \begin{cases} 1, & \text{if } Y_i(t) \geq 1 \text{ and } Y_i(t) + \cdots + Y_{i+\kappa-1}(t) \geq r, \\ 0, & \text{otherwise.} \end{cases}$$

(*Note the restriction that the first line in the chunk is nonempty*; *otherwise*, *see Theorem* 5.) *Set*

$$N(m,t) = \sum_{i=1}^{m} X_i(t).$$

Let $t, m \to \infty$ *satisfying* $t^r/m^{r-1} \to c$, $0 < c < \infty$, *and*

$$\left| \frac{t^r}{m^{r-1}} - c \right| = O\left(\frac{1}{m^p}\right) \quad \text{for some } p > 0. \tag{1.6}$$

Then

$$\| \mathscr{L}(N(m,t)) - \mathscr{L}(Z_\lambda) \| = O(1/m^q),$$

where $q = \min(p, 1/r)$ *and* Z_λ *is a Poisson random variable with parameter*

$$\lambda = c\left(\kappa^r - (\kappa - 1)^r\right)/r!.$$

THEOREM 3. *For any fixed positive integer* κ *and nonnegative integer* r, *define*

$$X_i(t) = \begin{cases} 1, & \text{if } Y_i(t) + Y_{i+1}(t) + \cdots + Y_{i+\kappa-1}(t) \leq r, \\ 0, & \text{otherwise.} \end{cases}$$

Set

$$L(m,t) = \sum_{i=1}^{m} X_i(t).$$

If $m, t \to \infty$ *satisfying*

$$t = \frac{1}{\kappa} m \ln m + \frac{r}{\kappa} m \ln\ln m + mc(m), \qquad c(m) \to c,\ 0 < c < \infty, \tag{1.7}$$

then

$$\| \mathscr{L}(L(m,t)) - \mathscr{L}(Z_\lambda) \| = O(c(m) - c) + O\left(\frac{\ln\ln m}{\ln m}\right),$$

where Z_λ *is a Poisson random variable with parameter* $\lambda = e^{-\kappa c}/r!$.

In Theorem 3, the error of the model of the Poisson process from the straight multinomial procedure is of the order $O(1/m)$.

Let $T^-_{r,\kappa}$ be the waiting time until some κ-chunk first accumulates r events (balls) and let $T^+_{r,\kappa}$ be the waiting time until all κ-chunks accrue more than r balls. The limiting distributional behavior of these waiting times follows from the previous two theorems by virtue of the duality relations of the events

$\{T^-_{r,\kappa} > t\} \equiv \{N(m,t) = 0\}$ and $\{T^+_{r,\kappa} \le t\} \equiv \{L(m,t) = 0\}$. Thus we have:

COROLLARY 4.

$$\lim_{m\to\infty} \operatorname{pr}\left\{ \frac{T^-_{r,\kappa}}{m^{(r-1)/r}} > x \right\} = \exp\left\{ -\frac{x^r\left(\kappa^r - (\kappa-1)^r\right)}{r!} \right\}, \tag{1.8}$$

$$\lim_{m\to\infty} \operatorname{pr}\left\{ \frac{T^+_{r,\kappa}}{m} - \left(\frac{1}{\kappa} \ln m + \frac{r}{\kappa} \ln\ln m \right) \le x \right\} = \exp\{-e^{-\kappa x/r!}\}. \tag{1.9}$$

In Theorem 2, the condition that the κ-chunk starts with a nonempty urn is necessary for convergence to a Poisson limit law. The following result indicates that without any restrictions, the limiting distribution is intrinsically compound Poisson.

THEOREM 5. *For any fixed positive integers κ and r where $r \ge 2$, define*

$$\tilde{X}_i(t) = \begin{cases} 1, & \text{if } Y_i(t) + \cdots + Y_{i+\kappa-1}(t) \ge r, \\ 0, & \text{otherwise.} \end{cases}$$

Set

$$\tilde{N}(m,t) = \sum_{i=1}^{m} \tilde{X}_i(t).$$

Let $t, m \to \infty$ with $t^r/m^{r-1} \to c$, $0 < c < \infty$. Then

$$\tilde{N}(m,t) \to C_{\lambda_1,\ldots,\lambda_\kappa} \quad \textit{in distribution},$$

where $C_{\lambda_1,\ldots,\lambda_\kappa}$ is a compound Poisson random variable with probability generating function

$$\exp\left\{ \sum_{k=1}^{\kappa} \lambda_k (s^{\kappa-k+1} - 1) \right\},$$

where

$$\lambda_1 = c/r!, \qquad \lambda_k = c\left(k^r - 2(k-1)^r + (k-2)^r\right)/r! \quad \text{for } k = 2,\ldots,\kappa.$$

Limit results involving compound Poisson processes also occur in recent works of Arratia, Goldstein and Gordon (1990).

We shall derive several other results of interest related to the above urn problems. Corollary 4 indicates that $T^-_{r,1}/m^{(r-1)/r}$ has a limiting distribution function $1 - \exp(-x^r/r!)$. To simplify notation, we write T_r instead of $T^-_{r,1}$. Consider the exact times in acquiring the previous $r-1$ balls by *that* urn which first accumulates r balls. Theorem 6 describes these times as asymptotically uniformly distributed relative to the time T_r. It is well known and trivial that for an unencumbered Poisson process, given the rth event at time T_r, the successive earlier events relative to T_r are distributed as the order statistics of $r-1$ uniform random variables. However, for the times among all the m urns

when some urn first acquires r balls (this involves conditions on the total composition of all the urns), only in the limit as $m \to \infty$ do the corresponding uniform order statistics hold as stated in Theorem 6. In Theorem 7, we investigate the number $N_r^{(i)}$ of urns occupied by exactly i balls, $1 \le i \le r-1$ at time T_r.

THEOREM 6. *Let $T_r^{(1)}, T_r^{(2)}, \ldots, T_r^{(r-1)}$ be the times at which the urn that is the first to accumulate r balls receives its first, second, $\ldots, r-1$st ball. As $m \to \infty$, the limiting joint distribution of*

$$\left(\frac{T_r^{(1)}}{T_r}, \frac{T_r^{(2)}}{T_r}, \ldots, \frac{T_r^{(r-1)}}{T_r} \right)$$

is the same as that of the order statistics of $r-1$ independent, uniformly distributed random variables on $[0,1]$.

THEOREM 7. *For each positive integer l and $1 \le i \le r-1$,*

$$\lim_{m\to\infty} E\left[\left(\frac{N_r^{(i)}}{m^{(r-i)/r}} \right)^l \right] = \lim_{m\to\infty} \frac{1}{(i!)^l} E\left[\left(\frac{T_r}{m^{(r-1)/r}} \right)^{il} \right]. \tag{1.10}$$

Hence, $N_r^{(i)}/m^{(r-i)/r}$ has a limiting distribution with density given by

$$g(y) = \frac{(i-1)!}{(r-1)!} (i!y)^{(r-i)/i} \exp\left\{ -\frac{(i!y)^{r/i}}{r!} \right\} \tag{1.11}$$

as $m \to \infty$.

In Section 7, we discuss the limiting behavior of T_r if each allocation assigns a random number of balls into the selected urn (i.e., each line generates events according to a compound Poisson process).

It is a common technique in the solution of classical occupancy problems to consider the sequence of Bernoulli random variables X_i, which is defined to be 1 if urn i contains the required number of balls and 0 otherwise. Then $N_r = X_1 + \cdots + X_m$ and hence the limiting distribution of N_r can be found by calculating moments or generating functions directly [see Kolchin, Sevast'yanov and Chistyakov (1978)]. However, these calculations become arduous in our problems when $\kappa > 1$ because the chunks of overlapping urns entail additional dependence relations among the Bernoulli random variables.

In 1975, Chen adapted Stein's differential method for obtaining normal limit laws [see Stein (1972)] and provided error estimates in establishing Poisson limit laws for a sequence of dependent Bernoulli random variables by effectively computing only first and second moments. The method provides an upper bound on the total variation distance to the difference from the Poisson distribution. Since then, Chen's result has been further refined and applied in myriad contexts [e.g., see Arratia, Goldstein and Gordon (1989) and Barbour and Holst (1989) and references therein]. The proofs of Theorems 2 and 3,

which will be given in Section 4, are based upon the Chen-Stein method and Theorem 1 will be proved in Section 3 by applying the multivariate version as extended in Arratia, Goldstein and Gordon (1989). The proof of Theorem 10 exploits some further decompositions of the random variable and the multivariate Poisson limit law. Theorems 6, 7 and 13 are analyzed by more direct means.

We describe three practical biomolecular sequence problems to which the results of this paper may give some *qualitative* insights.

1. Consider a long random letter (DNA or protein) sequence sampled from an alphabet of size a (e.g., $a = 20$ for proteins). We examine s-words (a contiguous set of s letters in the sequence). There are potentially $a^s = m$ urns reflecting all s words and m would be quite large for $a = 20$, $s = 3$. For a given sequence of length N, we can view each position $i = 1, \ldots, N - s + 1$ in the sequence to determine an s-word engendering a ball in an urn. Of course, overlapping words are certainly dependent but as an approximation we assume that the balls (the collection of s-words in the sequence) are independently randomly generated. Allowing for few errors, we may coalesce words and thereby obtain an idealization of balls falling into neighboring urns or a related set of urns.

   The concept and identification of "rare" and "frequent" words in a letter sequence is a useful assessment. For a given sequence of length N, we determine s (the size of the word) to satisfy

$$s - 1 < \frac{\log N}{\log a} \leq s \tag{1.12}$$

   and then choose r such that

$$\frac{r-1}{r} < \frac{\log N}{\log a^s} \leq \frac{r}{r+1}. \tag{1.13}$$

   Here, s words occurring greater than or equal to r times are considered *frequent words*. On the basis of Theorem 2, we propose that the number of frequent words, defined by the clump size κ, is Poisson distributed with parameter $c(\kappa^r - (\kappa - 1)^r)/r!$ where $c = N^r/m^{r-1}$. By exploiting the concepts of Theorem 3, we characterize *rare words* as follows. Determine the word size s to satisfy

$$a^s \ln a^s < N \leq a^{s+1} \ln a^{s+1} \tag{1.14}$$

   and then determine r such that (for $m = a^s$)

$$m(\ln m + r \ln\ln m) < N \leq m(\ln m + (r+1)\ln\ln m). \tag{1.15}$$

   By this prescription, s-words occurring less than or equal to r times are considered *rare words*. Real data applications of these notions of frequent and rare words with some interpretations are given in Section 8.
2. Given a sequence of N independent random letters from a finite alphabet with equal probability of sampling each letter, let L_r denote the length of

the longest r-fold repeat (words occurring at least r times in the sequence). It is of interest to determine the expected spacings of the L_r maximal length r-fold repeat. The result of Theorem 6 suggests that asymptotically as $N \to \infty$, the locations of these copies are distributed like r observations from a uniform distribution over the sequence length.

3. The following problems are motivated by the human genome initiative having the objective of sequencing the totality of human DNA. The ideas of this paper can be used to give a rough estimate of the following problem. Given a sequence of DNA bases (say 1600 long and assuming the 4 bases occur equally likely and independently), it is of interest to estimate the number of repeated 8-words that occur in a 1600 genomic stretch. There are $4^8 \approx 65{,}000$ possible 8-words (urns) and about 1600 balls corresponding to the sequence to be distributed to the urns. Note $65{,}000^{2/3} \approx 1600$. Therefore we would expect a triply repeated ($r = 3$) 8-word to be very rare. Guided by Theorem 7 on the variable $N_r^{(i)}$, we would expect an order $65{,}000^{1-2/3} \approx 40$ double repeats. This kind of information is useful in developing apparatus for matrix sequencing as described in Drmanac, Labat, Brukner and Crkvenjakov (1989).

2. The Chen–Stein method.

THEOREM 8 [As formulated in Arratia, Goldstein and Gordon (1989)]. *Let I be an index set. For any $\alpha \in I$, X_α is a Bernoulli random variable with parameter p_α and $B(\alpha)$ is a subset of I containing α, called the neighborhood of dependence of α. Let*

$$W = \sum_{\alpha \in I} X_\alpha$$

and let Z_λ be a Poisson random variable with parameter $\lambda = \sum_{\alpha \in I} p_\alpha$. Define

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B(\alpha)} p_\alpha p_\beta,$$

$$b_2 = \sum_{\alpha \in I} \sum_{\beta \in B(\alpha) \setminus \{\alpha\}} p_{\alpha\beta} \quad \text{where } p_{\alpha\beta} = E\left[X_\alpha X_\beta\right]$$

$$b_3 = \sum_{\alpha \in I} s_\alpha, \quad \text{where } s_\alpha = E\left|E\left[X_\alpha - p_\alpha | X_\gamma, \gamma \notin B(\alpha)\right]\right|.$$

Then

$$\left\|\mathscr{L}(W) - \mathscr{L}(Z_\lambda)\right\| \le 2\left[(b_1 + b_2)\frac{1 - e^{-\lambda}}{\lambda} + \min\left(1, \sqrt{\frac{2}{\lambda}}\right) b_3\right].$$

For many levels of applications, see Aldous (1989), Barbour (1982), Barbour and Eagleson (1984), Barbour and Hall (1984), Barbour and Holst (1989), Godbole (1991) and Janson (1987), among others. The formulation above and the multivariate version given below have been elegantly set forth in Arratia, Goldstein and Gordon (1989).

THEOREM 9. *Retaining the notations in Theorem* 8, *suppose the index set* I *can be partitioned into a finite disjoint union of subsets* $I_1, I_2, \ldots, I_n$ *and*

$$W_i = \sum_{\alpha \in I_i} X_\alpha,$$

$$\lambda_i = \sum_{\alpha \in I_i} E[X_\alpha].$$

Then,

$$\| \mathscr{L}(W_1, W_2, \ldots, W_n) - \mathscr{L}(Z_1, Z_2, \ldots, Z_n) \| \leq 8(b_1 + b_2 + b_3),$$

where $Z_1, Z_2, \ldots, Z_n$ *are independent Poisson random variables with parameters* $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Generally, b_3 serves as a measure of far away dependence. In all the applications discussed below, X_β is always independent of X_α whenever $X_\beta \notin B(\alpha)$. Hence, $b_3 \equiv 0$. In these cases, to derive a Poisson limit law, it suffices to show that b_1 and b_2 tend to 0.

3. Poisson limit laws for occupancy problems of κ-chunks.

PROOF OF THEOREM 2. Take the index set I to be $\{1, 2, \ldots, m\}$ and for any $i \in I$, take the neighborhood of dependence to be $\{k \in I: |k - i| < \kappa\}$. Each X_i is then a Bernoulli random variable with mean

$$p_i = E[X_i] = \sum_{n=r}^{\infty} \sum_{\sigma \in Q^*(n,\kappa)} \prod_{k=1}^{\kappa} \frac{e^{-t/m}(t/m)^{\sigma_k}}{\sigma_k!}$$

$$= (e^{-t/m})^{\kappa} \sum_{n=r}^{\infty} \left(\frac{t}{m}\right)^n \sum_{\sigma \in Q^*(n,\kappa)} \frac{1}{\sigma_1! \cdots \sigma_\kappa!},$$

where $Q^*(n, \kappa)$ is defined in (1.2). Note that

$$\sum_{\sigma \in Q^*(n,\kappa)} \frac{1}{\sigma_1! \cdots \sigma_\kappa!} = \sum_{\sigma_1=1}^{n} \frac{1}{\sigma_1!} \sum_{\sigma_2 + \cdots + \sigma_n = n - \sigma_1} \frac{1}{\sigma_2! \cdots \sigma_\kappa!}$$

$$= \frac{1}{n!}\left(\kappa^n - (\kappa - 1)^n\right), \quad \text{abbreviated } f(n, \kappa).$$

Set $\lambda_m = \sum_{i=1}^{m} p_i = E[N(m, t)]$. Then,

$$\lambda_m = me^{-t\kappa/m} \sum_{n=r}^{\infty} \left(\frac{t}{m}\right)^n f(n, \kappa)$$

$$= \frac{t^r}{m^{r-1}} e^{-t\kappa/m} \left[f(r, \kappa) + \sum_{n=1}^{\infty} \left(\frac{t}{m}\right)^n f(n + r, \kappa) \right].$$

We then have

$$|\lambda_m - \lambda| \le f(r,\kappa)\left|\frac{t^r}{m^{r-1}}e^{-t\kappa/m} - ce^{-t\kappa/m}\right| + f(r,\kappa)|ce^{-t\kappa/m} - c| + \frac{t^r}{m^{r-1}}\sum_{n=1}^{\infty}\left(\frac{t}{m}\right)^n f(n+r,\kappa).$$

The first term is $O(1/m^p)$ by condition (1.6); the second and the third terms are both $O(t/m)$. Since $t^r \approx cm^{r-1}$, $O(t/m) = O(1/m^{1/r})$. These estimates combined yield

$$|\lambda_m - \lambda| = O(1/m^q), \qquad q = \min(p, 1/r). \tag{3.1}$$

Now we compute the relevant b_1 and b_2 of Theorem 8 (note that $b_3 = 0$):

$$\begin{aligned} b_1 &= \sum_{i=1}^{m}\sum_{k\in B(i)} p_i p_k \\ &= (2\kappa - 1)me^{-2\kappa t/m}\left(\sum_{n=r}^{\infty}(t/m)^n f(n,\kappa)\right)^2 \\ &= O(t^{2r}/m^{2r-1}) = O(1/m). \end{aligned}$$

To get b_2, we examine $E[X_iX_k]$ for $0 < |k-i| < \kappa$. We may obviously assume $i < k$ and set $d = k - i \ge 1$. Then, denoting the configurations of the ith and kth κ-chunks by $\boldsymbol{\eta}$ and $\boldsymbol{\sigma}$, respectively, we have

$$\begin{aligned} &E[X_iX_k] \\ &\quad = \mathrm{pr}\{X_i = X_k = 1\} \\ &\quad = \sum_{n=r}^{\infty}\sum_{\boldsymbol{\sigma}\in Q^*(n,\kappa)}\frac{e^{-t\kappa/m}(t/m)^n}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!}\sum_{u=r}^{\infty}\sum_{\boldsymbol{\eta}\in Q^*_{\boldsymbol{\sigma}}(u,\kappa)}\frac{e^{-td/m}(t/m)^{\eta_1+\eta_2+\cdots+\eta_d}}{\eta_1!\eta_2!\cdots\eta_d!} \\ &\quad = e^{-(\kappa+d)t/m}O\left(\left(\frac{t}{m}\right)^{r+1}\right) \quad \text{since } \eta_1 \ge 1, \end{aligned}$$

where $Q^*_{\boldsymbol{\sigma}}(u,\kappa) = \{(\eta_1,\eta_2,\ldots,\eta_\kappa) \in Q^*(u,\kappa)\colon\ \eta_{d+l} = \sigma_l,\ l = 1,2,\ldots,\kappa - d\}$ is the subset of κ-tuples of $Q^*(u,\kappa)$ whose last $\kappa - d$ components are in agreement with the first $\kappa - d$ components of σ.

Thus

$$b_2 = \sum_{i=1}^{m}\sum_{k\in B(i)\setminus\{i\}} E[X_iX_k] \le m(2\kappa-2)e^{-(\kappa+1)t/m}O\left(\left(\frac{t}{m}\right)^{r+1}\right) = O\left(\frac{1}{m^{1/r}}\right).$$

Now by Theorem 8,

$$\left\|\mathscr{L}(N(m,t)) - \mathscr{L}(Z_{\lambda_m})\right\| \le O(b_1 + b_2) = O\left(\frac{1}{m^{1/r}}\right). \tag{3.2}$$

Trivially, in view of (3.1),

$$\left\| \mathscr{L}(Z_{\lambda_m}) - \mathscr{L}(Z_\lambda) \right\| = O(1/m^q). \tag{3.3}$$

Combining (3.2) and (3.3), the theorem is confirmed. □

The condition that the first urn is nonempty is essential in the counts of Theorem 2, as is to be shown in Theorem 5. This condition, however, is not necessary in Theorem 3.

PROOF OF THEOREM 3. Let λ_m be the mean of $L(m,t)$. Then

$$\lambda_m = mE[X_1(t)] = me^{-\kappa t/m} \sum_{n=0}^{r} \left(\frac{t}{m}\right)^n \sum_{\sigma \in Q(n,\kappa)} \frac{1}{\sigma_1!\sigma_2! \cdots \sigma_\kappa!},$$

with $Q(n,\kappa)$ defined in (1.1). Since

$$\sum_{\sigma \in Q(n,\kappa)} \frac{1}{\sigma_1!\sigma_2! \cdots \sigma_\kappa!} = \frac{\kappa^n}{n!},$$

we have

$$\lambda_m = me^{-\kappa t/m} \sum_{n=0}^{r} \frac{(t/m)^n \kappa^n}{n!}.$$

Inserting $t = (1/\kappa)m \ln m + (r/\kappa)m \ln\ln m + mc(m)$ yields

$$\begin{aligned}\lambda_m &= e^{-\kappa c(m)}(\ln m)^{-r}\left[\left(\frac{\ln m}{\kappa}\right)^r\left(\frac{\kappa^r}{r!}\right) + O\big((\ln m)^{r-1}\ln\ln m\big)\right] \\ &= e^{-\kappa c(m)}\left[\frac{1}{r!} + O\left(\frac{\ln\ln m}{\ln m}\right)\right]\end{aligned}$$

and therefore

$$\lambda_m - \lambda = O(c(m) - c) + O(\ln\ln m/\ln m). \tag{3.4}$$

Taking the neighborhood of dependence $B(i)$ of $i \in I = \{1, 2, \ldots, m\}$ to be the set of all k's satisfying $|k - i| < \kappa$,

$$\begin{aligned}b_1 &= m(2\kappa - 1)\left[e^{-\kappa t/m} \sum_{n=0}^{r} \left(\frac{t}{m}\right)^n \sum_{\sigma_1 + \cdots + \sigma_\kappa = n} \frac{1}{\sigma_1! \cdots \sigma_\kappa!}\right]^2 \\ &= m(2\kappa - 1)e^{-2\kappa c(m)}m^{-2}(\ln m)^{-2r}\left[\frac{(\ln m)^r}{r!} + o\big((\ln m)^r\big)\right]^2 \\ &= O\left(\frac{1}{m}\right).\end{aligned}$$

To obtain b_2, note that for any $i = 1, 2, \ldots, m$ and for any k such that $\kappa > |k - i| = d > 0$,

$$
\begin{aligned}
&E[X_i X_k] \\
&= \sum_{n=0}^{r} e^{-\kappa t/m} \left(\frac{t}{m}\right)^n \sum_{\sigma \in Q(n,\kappa)} \frac{1}{\sigma_1! \cdots \sigma_\kappa!} \sum_{u=0}^{r-\sigma_1-\cdots-\sigma_{\kappa-d}} e^{-dt/m} \left(\frac{t}{m}\right)^u \\
&\quad \times \sum_{\eta \in Q(u,d)} \frac{1}{\eta_1! \cdots \eta_d!} \\
&= e^{-(\kappa+d)t/m} \sum_{n=0}^{r} \sum_{\sigma \in Q(n,\kappa)} \sum_{u=0}^{r-\sigma_1-\cdots-\sigma_{\kappa-d}} \sum_{\eta \in Q(u,d)} \frac{(t/m)^{n+u}}{\sigma_1! \cdots \sigma_\kappa! \eta_1! \cdots \eta_d!}.
\end{aligned}
$$

It is easy to see that the highest power of t/m in the above sum is $2r$ and it is attained whenever $\eta_1 + \cdots + \eta_d = r$, $\eta_{d+1} = \cdots = \eta_\kappa = \sigma_1 = \cdots = \sigma_{\kappa-d} = 0$ and $\sigma_{\kappa-d+1} + \cdots + \sigma_\kappa = r$. This gives

$$
E[X_i X_k] = e^{-(\kappa+d)t/m} O\big((t/m)^{2r}\big)
$$

and hence

$$
\begin{aligned}
b_2 &= \sum_{i=1}^{m} \sum_{0<|k-i|<\kappa} E[X_i X_k] \\
&= m(2\kappa - 2) e^{-(\kappa+1)t/m} O\big((t/m)^{2r}\big) \\
&= O\big(m^{-1/\kappa} (\ln m)^{r(1-1/\kappa)}\big).
\end{aligned}
$$

Now, appealing to Theorem 8, we have

$$
\big\| \mathscr{L}(L(m,t)) - \mathscr{L}(Z_{\lambda_m}) \big\| = O\big(m^{-1/\kappa} (\ln m)^{r(1-1/\kappa)}\big). \tag{3.5}
$$

From (3.4), we have

$$
\big\| \mathscr{L}(Z_{\lambda_m}) - \mathscr{L}(Z_\lambda) \big\| = O(c(m) - c) + O(\ln \ln m / \ln m),
$$

and consequently the conclusion of Theorem 3. □

PROOF OF COROLLARY 4. Observe that

$$
\mathrm{pr}\{T^-_{r,\kappa} > m^{(r-1)/r} x\} = \mathrm{pr}\{N(m, m^{(r-1)/r} x) = 0\}.
$$

Since $(m^{(r-1)/r} x)^r / m^{r-1} = x^r$, Theorem 2 implies

$$
\lim_{m \to \infty} \mathrm{pr}\{N(m, m^{(r-1)/r} x) = 0\} = \mathrm{pr}\{Z_\lambda = 0\} = \exp\{-x^r f(r,\kappa)\},
$$

proving the first statement (1.8) of Corollary 4. By similar arguments, the statement (1.9) follows from Theorem 3. □

4. Multivariate Poisson limit laws. The κ-chunks with a total of r or more balls may be classified into κ distinct groups according to which line in the chunk is the first having a partial sum of greater than or equal to r events. For $k = 1, 2, \ldots, \kappa$, let $P^{(k)}(r, \kappa)$ denote the set of κ-tuples of nonnegative integers whose first element is nonzero and whose kth partial sum is the first to reach the level r or more, that is,

$$P^{(1)}(r, \kappa) = \{(\sigma_1, \sigma_2, \ldots, \sigma_\kappa) : \sigma_1 \geq r\},$$

$$P^{(k)}(r, \kappa) = \{(\sigma_1, \sigma_2, \ldots, \sigma_\kappa) : \sigma_1 \geq 1, \sigma_1 + \cdots + \sigma_{k-1} < r \text{ and } \sigma_1 + \cdots + \sigma_k \geq r\}, \tag{4.1}$$

for $k = 2, 3, \ldots, \kappa$. Define

$$X_i^{(k)}(t) = \begin{cases} 1, & \text{if } (Y_i(t), \ldots, Y_{i+\kappa-1}(t)) \in P^{(k)}(r, \kappa), \\ 0, & \text{otherwise,} \end{cases}$$

then

$$N^{(k)}(m, t) = \sum_{i=1}^{m} X_i^{(k)}(t)$$

counts the number of κ-chunks given by some configuration $(\sigma_1, \ldots, \sigma_\kappa) \in P^{(k)}(r, \kappa)$. Theorem 10 shows that the joint distribution of the numbers of chunks in these groups tends to a multivariate Poisson limit when $t^r \approx cm^{r-1}$. This result will be used later to prove Theorem 5.

THEOREM 10. *Suppose $t, m \to \infty$ in such a way that $t^r/m^{r-1} \to c$, $0 < c < \infty$, and*

$$\left| \frac{t^r}{m^{r-1}} - c \right| = O\left(\frac{1}{m^p}\right) \quad \text{for some } p > 0. \tag{4.2}$$

Then

$$\left\| \mathscr{L}\left(\{N^{(k)}(m, t)\}_{k=1}^{\kappa}\right) - \mathscr{L}\left(\{Z_{\lambda_k}\}_{k=1}^{\kappa}\right) \right\| = O(1/m^q), \tag{4.3}$$

where $q = \min(p, 1/r)$ and $\{Z_{\lambda_k}\}_{k=1}^{\kappa}$ is a family of Poisson random variables with parameters

$$\lambda_1 = c/r!, \qquad \lambda_k = c\left(k^r - 2(k-1)^r + (k-2)^r\right)/r! \quad \text{for } k = 2, \ldots, \kappa.$$

PROOF. Let the index set I be the ordered pairs $\{(i, k) : 1 \leq i \leq m, 1 \leq k \leq \kappa\}$. Then I is the disjoint union of the subsets $I(k) = \{(i, k) : 1 \leq i \leq m\}$, $k = 1, 2, \ldots, \kappa$. For any $(i, k) \in I$, define its neighborhood of dependence $B(i, k)$ to be the set of all ordered pairs (i', k') such that $|i' - i| < \kappa$ with no restriction on k'. Let $P_n^{(k)}(r, \kappa) = \{\boldsymbol{\sigma} \in P^{(k)}(r, \kappa) : \sigma_1 + \cdots + \sigma_\kappa = n\}$. Note that

the sets $P_n^{(k)}(r,\kappa)$, $n = r, r+1, \ldots,$ are mutually disjoint and $P^{(k)}(r,\kappa) = \bigcup_{n=r}^{\infty} P_n^{(k)}(r,\kappa)$. Hence,

$$E\big[X_i^{(k)}(t)\big] = \sum_{n=r}^{\infty} \operatorname{pr}\big\{\big(Y_i(t), \ldots, Y_{i+\kappa-1}(t)\big) \in P_n^{(k)}(r,\kappa)\big\}$$

$$= \sum_{n=r}^{\infty} e^{-t\kappa/m}\left(\frac{t}{m}\right)^n \sum_{\sigma \in P_n^{(k)}(r,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!}.$$

Let $\lambda_m^{(k)} = E[N^{(k)}(m,t)]$. Then

$$\lambda_m^{(k)} = \sum_{i=1}^{m} E\big[X_i^{(k)}(t)\big] = me^{-t\kappa/m}\left(\frac{t}{m}\right)^r \sum_{n=r}^{\infty}\left(\frac{t}{m}\right)^{n-r} \sum_{\sigma \in P_n^{(k)}(r,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!}.$$

Since $P_n^{(k)}(r,\kappa) \subset Q^*(n,\kappa)$ [defined in (1.2)],

$$\sum_{\sigma \in P_n^{(k)}(r,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!} \le \sum_{\sigma \in Q^*(n,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!} = \frac{\kappa^n - (\kappa-1)^n}{n!},$$

which is bounded above for all values of n. So, we have

$$\lambda_m^{(k)} = e^{-t\kappa/m}\frac{t^r}{m^{r-1}}\left(\sum_{\sigma \in P_r^{(k)}(r,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!} + O\left(\frac{t}{m}\right)\right).$$

Now observe that $\sigma \in P_r^{(k)}(r,\kappa)$ if and only if $\sigma_1 \ge 1$, $\sigma_k \ge 1$, $\sigma_1 + \cdots + \sigma_k = r$ and $\sigma_{k+1} = \cdots = \sigma_\kappa = 0$. This implies that

$$\sum_{\sigma \in P_n^{(k)}(r,\kappa)} \frac{1}{\sigma_1!\sigma_2!\cdots\sigma_\kappa!} = \sum_{\sigma_1=1}^{r} \frac{1}{\sigma_1!} \sum_{\sigma_k=1}^{r-\sigma_1} \frac{1}{\sigma_k!} \sum_{\sigma_2+\cdots+\sigma_{k-1}=r-\sigma_1-\sigma_k} \frac{1}{\sigma_2!\cdots\sigma_{k-1}!}$$

$$= \begin{cases} \big(k^r - 2(k-1)^r + (k-2)^r\big)/r!, & \text{for } k \ge 2, \\ 1/r!, & \text{for } k = 1. \end{cases}$$

When $m, t \to \infty$ as given by (4.2), we have $\lambda_m^{(k)} \to \lambda_k$. Furthermore $|\lambda_m^{(k)} - \lambda_k| = O(1/m^q)$ with $q = \min(p, 1/r)$.

Next, we estimate b_1 and b_2. Note first that $\sum_{k=1}^{\kappa} X_i^{(k)}(t) = X_i(t)$ with $X_i(t)$ defined in Theorem 2. So,

$$b_1 = \sum_{i=1}^{m} \sum_{|i'-i|<\kappa} E\big[X_i(t)\big]E\big[X_{i'}(t)\big]$$

and as in the proof of Theorem 2, this quantity is $O(1/m)$. Similarly, b_2 is identical to that in Theorem 2 and therefore $O(1/m^{1/r})$. Hence by Theorem 9, the desired result is established. □

PROOF OF THEOREM 1 (Stated in Section 1). We shall only prove part (a). To apply Theorem 9, let the index set be $I = \bigcup_{\sigma \in Q^*(r,\kappa)} I(\sigma)$, where $I(\sigma) = \{(i,\sigma)$:

$1 \le i \le m\}$. For any $(i, \boldsymbol{\sigma}) \in I$, define its neighborhood of dependence $B(i, \boldsymbol{\sigma})$ to be the set of all ordered pairs $(k, \boldsymbol{\omega})$ such that $|k - i| < \kappa$ and $\boldsymbol{\omega} \in Q^*(r, \kappa)$.

Denoting $E[N_{\boldsymbol{\sigma}}(m, t)]$ by $\lambda_{\boldsymbol{\sigma}}(m)$, we have

$$\begin{aligned}\lambda_{\boldsymbol{\sigma}}(m) &= \sum_{i=1}^{m} E\left[X_{i,\boldsymbol{\sigma}}(t)\right] \\ &= \frac{t^r}{m^{r-1}} e^{-t\kappa/m} \frac{1}{\sigma_1!\sigma_2! \cdots \sigma_\kappa!} \\ &\to \frac{c}{\sigma_1!\sigma_2! \cdots \sigma_\kappa!} = \lambda_{\boldsymbol{\sigma}}.\end{aligned}$$

Furthermore, by (1.4), $|\lambda_{\boldsymbol{\sigma}}(m) - \lambda_{\boldsymbol{\sigma}}| = O(1/m^p)$.

Now we estimate b_1 and b_2:

$$\begin{aligned}b_1 &= \sum_{i=1}^{m} \sum_{\boldsymbol{\sigma} \in Q^*(r,\kappa)} \frac{e^{-t\kappa/m}(t/m)^r}{\sigma_1!\sigma_2! \cdots \sigma_\kappa!} \sum_{|k-i|<\kappa} \sum_{\boldsymbol{\omega} \in Q^*(r,\kappa)} \frac{e^{-t\kappa/m}(t/m)^r}{\omega_1!\omega_2! \cdots \omega_\kappa!} \\ &= me^{-2t\kappa/m}\left(\frac{t}{m}\right)^{2r} (2\kappa - 1)\left[\kappa^r - (\kappa - 1)^r\right]^2 = O\left(\frac{1}{m}\right), \\ b_2 &= \sum_{i=1}^{m} \sum_{\boldsymbol{\sigma} \in Q^*(r,\kappa)} \sum_{0<|k-i|<\kappa} \sum_{\boldsymbol{\omega} \in Q^*(r,\kappa)} E[X_{i,\boldsymbol{\sigma}} X_{k,\boldsymbol{\omega}}].\end{aligned}$$

For any i, k such that $0 < k - i = d < \kappa$,

$$E[X_{i,\boldsymbol{\sigma}} X_{k,\boldsymbol{\omega}}] = \begin{cases} \dfrac{e^{-(\kappa+d)t/m}(t/m)^{r+\sigma_1+\cdots+\sigma_d}}{\sigma_1!\sigma_2! \cdots \sigma_d!\omega_1! \cdots \omega_\kappa!}, & \text{if } \omega_1 = \sigma_{d+1}, \ldots, \omega_{\kappa-d} = \sigma_\kappa, \\ 0, & \text{otherwise.} \end{cases}$$

As $t^r \approx cm^{r-1}$,

$$b_2 \le m(t/m)^{r+1}(2\kappa - 2)e^{-(\kappa+1)t/m}\left[\left(\kappa^r - (\kappa - 1)^r\right)/r!\right] = O(1/m^{1/r}).$$

By Theorem 9 and the foregoing estimates, we get

$$\begin{aligned}&\left\| \mathscr{L}\left(\{N_{\boldsymbol{\sigma}}(m, t)\}_{\boldsymbol{\sigma} \in Q^*(r,\kappa)}\right) - \mathscr{L}\left(\{Z_{\lambda_{\boldsymbol{\sigma}}}\}_{\boldsymbol{\sigma} \in Q^*(r,\kappa)}\right) \right\| \\ &\quad = O(b_1 + b_2) + O(1/m^p) \\ &\quad = O(1/m^q), \qquad q = \min(p, 1/r). \end{aligned}$$ □

5. Compound Poisson limit law. In Theorem 2, we showed that the number of chunks which begin with a nonempty urn and have acquired a total of r or more balls has a Poisson limiting distribution. The condition that the chunk starts with a nonempty urn is essential for convergence to a Poisson

limit law. Theorem 5 shows that without this condition, the limiting distribution is genuinely compound Poisson. The following notation is useful. Let

$$U = \{\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_\kappa) : \sigma_i \text{ nonnegative integers and } \sigma_1 + \cdots + \sigma_\kappa \geq r\},$$

$$U_j = \{\boldsymbol{\sigma} \in U : \sigma_i = 0 \text{ for all } i < j,\ \sigma_j \geq 1\},$$

$$U_j^{(1)} = \{\boldsymbol{\sigma} \in U_j,\ \sigma_j \geq r\},$$

$$U_j^{(k)} = \{\boldsymbol{\sigma} \in U_j : \sigma_j + \cdots + \sigma_{j+k-2} < r,\ \sigma_j + \cdots + \sigma_{j+k-1} \geq r\},$$

$$2 \leq k \leq \kappa - j + 1.$$

Thus U_j consists of all $\boldsymbol{\sigma} \in U$ whose first nonzero component is the jth, and $U_j^{(k)}$ consists of all $\boldsymbol{\sigma} \in U_j$ that require k components thereafter to attain a level of r or more. Clearly $U_j^{(k)}$ is analogous to $P^{(k)}(r, \kappa)$ defined in Theorem 10 except for beginning at the jth component of $\boldsymbol{\sigma}$. Denote by $N_j(m,t)$ $[N_j^{(k)}(m,t)]$ the number of κ-chunks with configurations in U_j $(U_j^{(k)})$. We have the following lemma.

LEMMA 11. *For any $j = 1, \ldots, \kappa$ and $k = 1, \ldots, \kappa - j + 1$, as $m, t \to \infty$ such that $t^r/m^{r-1} \to c$, $0 < c < \infty$,*

$$N^{(k)}(m,t) - N_j^{(k)}(m,t) \to 0 \quad \textit{in distribution}, \tag{5.1}$$

where $N^{(k)}(m,t)$ is the number of chunks of κ contiguous lines whose configuration is in $P^{(k)}(r,\kappa)$ as defined in (4.1).

PROOF. We claim that $N^{(k)}(m,t) - N_j^{(k)}(m,t) \geq 0$. In fact, whenever $(Y_i, \ldots, Y_{i+\kappa-1})$ is in $U_j^{(k)}$, $(Y_{i+j}, \ldots, Y_{i+j+\kappa-1})$ is necessarily in $P^{(k)}(r,\kappa)$. Furthermore,

$$\begin{aligned}
E\left[N_j^{(k)}(m,t)\right] &= m \sum_{\boldsymbol{\sigma} \in U_j^{(k)}} e^{-\kappa t/m} \frac{(t/m)^{\sigma_1 + \cdots + \sigma_\kappa}}{\sigma_1! \cdots \sigma_\kappa!} \\
&= m e^{-\kappa t/m} \sum_{n=r}^{\infty} \left(\frac{t}{m}\right)^n \sum_{\substack{\boldsymbol{\sigma} \in U_j^{(k)} \\ \sigma_j + \cdots + \sigma_\kappa = n}} \frac{1}{\sigma_j! \cdots \sigma_\kappa!} \\
&= m e^{-\kappa t/m} \left(\frac{t}{m}\right)^r \left[\sum_{\substack{\boldsymbol{\sigma} \in U_j^{(k)} \\ \sigma_j + \cdots + \sigma_\kappa = r}} \frac{1}{\sigma_j! \cdots \sigma_\kappa!} + O\left(\frac{t}{m}\right) \right] \\
&\to c \sum_{\substack{\boldsymbol{\sigma} \in U_j^{(k)} \\ \sigma_1 + \cdots + \sigma_\kappa = r}} \frac{1}{\sigma_j! \cdots \sigma_\kappa!}
\end{aligned}$$

as $m, t \to \infty$ such that $t^r/m^{r-1} \to c$. For any $\boldsymbol{\sigma} \in U_j^{(k)}$, $\sigma_1 + \cdots + \sigma_\kappa = r$ if

and only if $\sigma_1 = \cdots = \sigma_{j-1} = \sigma_{j+k} = \cdots = \sigma_\kappa = 0$, $\sigma_j \geq 1$, $\sigma_{j+k-1} \geq 1$ and $\sigma_j + \cdots + \sigma_{j+\kappa-1} = r$. We therefore have

$$\sum_{\substack{\boldsymbol{\sigma} \in U_j^{(k)} \\ \sigma_j + \cdots + \sigma_\kappa = r}} \frac{1}{\sigma_j! \cdots \sigma_\kappa!}$$

$$= \sum_{\sigma_j=1}^{r} \frac{1}{\sigma_j!} \sum_{\sigma_{j+k-1}=1}^{r-\sigma_1} \frac{1}{\sigma_{j+k-1}!} \sum_{\sigma_{j+1}+\cdots+\sigma_{j+k-2}=r-\sigma_j-\sigma_{j+k-1}} \frac{1}{\sigma_{j+1}! \cdots \sigma_{j+k-2}!}$$

$$= \begin{cases} \left(k^r - 2(k-1)^r + (k-2)^r\right)/r!, & \text{for } k \geq 2, \\ 1/r!, & \text{for } k = 1. \end{cases}$$

Comparing with the limit of $E[N^{(k)}(m,t)]$ in the proof of Theorem 10, we see that as $m, t \to \infty$, $E[N^{(k)}(m,t) - N_j^{(k)}(m,t)] \to 0$, which implies that $N^{(k)}(m,t) - N_j^{(k)}(m,t) \to 0$ in distribution. □

PROOF OF THEOREM 5. First observe the disjoint decomposition

$$U = \bigcup_{j=1}^{\kappa} U_j = \bigcup_{j=1}^{\kappa} \bigcup_{k=1}^{\kappa-j+1} U_j^{(k)}.$$

Hence,

$$\tilde{N}(m,t) = \sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa-j+1} N_j^{(k)}(m,t)$$

$$= \sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa-j+1} N^{(k)}(m,t) + \sum_{j=1}^{\kappa} \sum_{k=1}^{\kappa-j+1} \left[N_j^{(k)}(m,t) - N^{(k)}(m,t)\right].$$

By Lemma 11, the second term tends to 0 in distribution. Interchanging the order of summation, the first term reduces to $\sum_{k=1}^{\kappa}(\kappa - k + 1)N^{(k)}(m,t)$. By Theorem 10, the joint distribution of $\{N^{(k)}(m,t)\}_{k=1,\ldots,\kappa}$ converges to that of $\{Z_{\lambda_k}\}_{k=1,\ldots,\kappa}$, where $Z_{\lambda_1}, \ldots, Z_{\lambda_\kappa}$ are independent Poisson random variables. $\tilde{N}(m,t)$ therefore converges in distribution to the compound Poisson random variable $C_{\lambda_1,\ldots,\lambda_\kappa}$ with probability generating function

$$E[s^{\kappa Z_{\lambda_1} + (\kappa-1)Z_{\lambda_2} + \cdots + Z_{\lambda_\kappa}}] = \exp\left\{\sum_{k=1}^{\kappa} \lambda_k (s^{\kappa-k+1} - 1)\right\}. \qquad \square$$

6. Spacings and counts of balls in urns at certain random times. In Theorem 6 (see the Introduction), we consider the times of acquisition of the previous $r-1$ balls by the urn which first acquires r balls and determine the limiting joint distribution of these variables. Next, we consider the number $N_r^{(i)}$ of urns occupied by exactly i balls, $1 \leq i \leq r-1$, at time T_r. The limiting distribution of $N_r^{(i)}/m^{(r-i)/r}$ is described in Theorem 7.

PROOF OF THEOREM 6 (Stated in Section 1; consult the notation there). For any $0 < x_1 < x_2 < \cdots < x_{r-1} < 1$, conditioning on the value of T_r, consider

$$\operatorname{pr}\left\{\frac{T_r^{(1)}}{T_r} \le x_1, \frac{T_r^{(2)}}{T_r} \le x_2, \ldots, \frac{T_r^{(r-1)}}{T_r} \le x_{r-1}\right\}$$

$$= \sum_{b=0}^{\infty} \operatorname{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1} | T_r = b\}\operatorname{pr}\{T_r = b\}.$$

The infinite sum is conveniently split into three parts, the first summing on b up to $um^{1-1/r}$, the second from $um^{1-1/r}$ to $vm^{1-1/r}$ and the third from $vm^{1-1/r}$ to ∞, where u (small) and v (large) are fixed. Since

$$\lim_{m\to\infty} \operatorname{pr}\left\{\frac{T_r}{m^{(r-1)/r}} \le x\right\} = 1 - e^{-x^r/r!},$$

we have, when m is sufficiently large,

$$\sum_{b \le um^{(r-1)/r}} \operatorname{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1} | T_r = b\}\operatorname{pr}\{T_r = b\}$$

$$\le \operatorname{pr}\{T_r \le um^{(r-1)/r}\} \le 1 - \exp\{-(u+\varepsilon)^r/r!\} \quad \text{for any } \varepsilon > 0$$

and

$$\sum_{b > vm^{(r-1)/r}} \operatorname{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1} | T_r = b\}\operatorname{pr}\{T_r = b\}$$

$$\le \operatorname{pr}\{T_r \ge vm^{(r-1)/r}\} \le 2e^{-v^r/r!}.$$

These estimates are arbitrarily small for u, ε small and v large.

To evaluate the remaining sum, observe that the event $\{T_r = b\}$ occurs if and only if exactly $r-1$ of the first $b-1$ balls go into the urn receiving the bth ball, and the rest of the balls distribute among the other $m-1$ urns such that each contains fewer than r balls. So

$$\operatorname{pr}\{T_r = b\} = \frac{1}{m^{b-1}}\binom{b-1}{r-1} D_{b-r,m-1,r}, \tag{6.1}$$

where $D_{b,g,r}$ is the number of distinct assignments of b distinguishable balls into g distinguishable urns such that each urn contains fewer than r balls. The factor $1/m^{b-1}$ rather than $1/m^b$ occurs because by symmetry any urn can be specified to receive the r balls. Also

$$\operatorname{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1}, T_r = b\}$$

$$= \frac{1}{m^{b-1}}\left(\sum_{i_{r-2} < i_{r-1} \le bx_{r-1}} \cdots \sum_{i_1 < i_2 \le bx_2} \sum_{0 < i_1 \le bx_1} 1\right) D_{b-r,m-1,r}. \tag{6.2}$$

The ratio of (6.2) to (6.1) gives

$$\begin{aligned}&\mathrm{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1} | T_r = b\}\\ &\quad = \frac{(r-1)!}{(b-1)(b-2)\cdots(b-r+1)}\Gamma,\end{aligned}$$

where

$$\Gamma = \sum_{i_{r-2} < i_{r-1} \le bx_{r-1}} \cdots \sum_{i_1 < i_2 \le bx_2} \sum_{0 < i_1 \le bx_1} 1.$$

So, we can write

$$\begin{aligned}&\sum_{um^{1-1/r} \le b \le vm^{1-1/r}} \mathrm{pr}\{T_r^{(1)} \le bx_1, \ldots, T_r^{(r-1)} \le bx_{r-1} | T_r = b\}\mathrm{pr}\{T_r = b\}\\ &= (r-1)! \sum_{um^{1-1/r} \le b \le vm^{1-1/r}} \left[\frac{b^{r-1}}{(b-1)(b-2)\cdots(b-r+1)} \times \frac{\Gamma}{b^{r-1}} \mathrm{pr}\{T_r = b\}\right]. \qquad (6.3)\end{aligned}$$

As $m \to \infty$, $b \to \infty$, clearly

$$(r-1)!\frac{\Gamma}{b^{r-1}} \to F_{(X_1^*, X_2^*, \ldots, X_{r-1}^*)}(x_1, x_2, \ldots, x_{r-1}),$$

where $F_{(X_1^*, X_2^*, \ldots, X_{r-1}^*)}$ is the distribution function of the order statistics of $r-1$ independent uniformly distributed random variables $X_1, \ldots, X_{r-1}$ on $[0, 1]$. Thus the limit as $m \to \infty$ of the sum in (6.3) is

$$F_{(X_1^*, X_2^*, \ldots, X_{r-1}^*)}(x_1, x_2, \ldots, x_{r-1}) \int_u^v f(x)\, dx,$$

$f(x)$ being the limiting density function of $T_r/m^{(r-1)/r}$. Since u and v are arbitrary, letting $u \to 0$ and $v \to \infty$ yields the desired result. □

PROOF OF THEOREM 7. Take i fixed and $1 \le i \le r-1$. Let I_α be the indicator random variable which takes the value 1 if urn α contains exactly i balls at time T_r and 0 otherwise so that $N_r^{(i)} = \sum_{\alpha=1}^m I_\alpha$ is the count of urns containing i balls at time T_r. Expanding the lth power of $N_r^{(i)}$, we have

$$\left(N_r^{(i)}\right)^l = \left(\sum_{\alpha=1}^m I_\alpha\right)^l = \sum_{u=1}^l A_{lu} \sum_{1 \le \alpha_1 < \cdots < \alpha_u \le m} I_{\alpha_1} I_{\alpha_2} \cdots I_{\alpha_u},$$

where A_{lu} is the number of different assignments of l distinguishable balls to u distinguishable urns such that each urn contains at least one ball. Then, by symmetry, we have

$$E\left[\left(\frac{N_r}{m^{(r-i)/r}}\right)^l\right] = \frac{1}{m^{l(r-i)/r}} \sum_{u=1}^l A_{lu} \binom{m}{u} E[I_1 I_2 \cdots I_u].$$

We need to estimate $E[I_1 I_2 \cdots I_u]$ conditioned on T_r. In the rest of the proof, we shall use $T_{r,g}$ to denote the waiting time until some urn among an array of g urns first acquires r balls. Obviously by the multinomial theorem and the meaning of T_r $(= T_{r,m})$, we have

$$E[I_1 I_2 \cdots I_u | T_r = b] = \binom{b-1}{iu} \frac{(iu)!}{(i!)^u m^{iu}} \frac{\text{pr}\{T_{r,m-u} = b - iu\}}{\text{pr}\{T_r = b\}}$$

$$= \left[\frac{b^{iu}}{(i!)^u m^{iu}} + o\left(\frac{b^{iu}}{m^{iu}} \right) \right] \frac{\text{pr}\{T_{r,m-u} = b - iu\}}{\text{pr}\{T_r = b\}}$$

with i and u fixed and b large. Observe that for $b = ym^{(r-1)/r}$, $0 < \varepsilon \le y \le K < \infty$, we have

$$\frac{\text{pr}\{T_{r,m-u} = ym^{(r-1)/r} - iu\}}{\text{pr}\{T_r = ym^{(r-1)/r}\}} \to 1 \quad \text{uniformly in } y.$$

Observe next that

$$\binom{m}{u} E[I_1 I_2 \cdots I_u | T_r = ym^{(r-1)/r}] = O\left(\frac{m^u y^{iu} m^{iu(r-1)/r}}{m^{iu}} \right) \tag{6.4}$$

and therefore dividing (6.3) by $1/m^{l(r-i)/r}$ yields

$$O\left(\frac{m^{u(r-i)/r} y^i u}{m^{l(r-i)/r}} \right) \to 0 \quad \text{for } u < l.$$

When $u = l$, we have $A_{l,l} = l!$ So,

$$E\left[\left(\frac{N_r^{(i)}}{m^{(r-i)/r}} \right)^l \right] = \frac{\binom{m}{l} l!}{m^{l(r-i)/r}} \int_0^\infty E[I_1 \cdots I_l | T_r = ym^{(r-1)/r}]$$

$$\times \text{pr}\{T_r = ym^{(r-1)/r}\}\, dy$$

and hence

$$\lim_{m\to\infty} E\left[\left(\frac{N_r^{(i)}}{m^{(r-1)/r}} \right)^l \right] = \lim_{m\to\infty} \frac{1}{(i!)^l} E\left[\left(\frac{T_r}{m^{(r-1)/r}} \right)^{il} \right]. \tag{6.5}$$

Denote by $g(y)$ the limiting density of $N_r^{(i)}/m^{(r-i)/r}$. The next equation expresses (6.5):

$$\int_0^\infty y^l g(y)\, dy = \frac{1}{(i!)^l} \int_0^\infty (x^i)^l f(x)\, dx$$

$$= \frac{1}{(i!)^l} \int_0^\infty \xi^l f(\xi^{1/i}) \frac{1}{i\xi^{(i-1)/i}}\, d\xi,$$

and incorporating the geometric factor $1/(i!)^l$ yields

$$g(y) = \frac{(i-1)!\, f\big((i!y)^{1/i}\big)}{(i!y)^{(i-1)/i}},$$

where $f(x) = (x^{r-1}/(r-1)!)e^{-x^r/r!}$ is the limiting density of $T_r/m^{1-1/r}$. Thus we have the formula of (1.11):

$$g(y) = \frac{(i-1)!}{(r-1)!}(i!y)^{(r-i)/i}\exp\left\{-\frac{(i!y)^{r/i}}{r!}\right\}. \qquad \square$$

7. Limiting waiting times for lines of compound Poisson processes. The following simple Markov structure for distributing balls in urns postulates that the next ball is added to the previously selected urn with probability p and otherwise (with probability $1-p$) an urn is selected equally likely from the totality of urns to receive the ball. The corresponding continuous time embedding (the geometric case 7.1 following) considers m independent compound Poisson process lines such that at each Poisson event, a random number N of balls is added to the urn represented by the line. We shall consider two cases:

$$\text{pr}\{N = k\} = (1-p)p^{k-1}, \qquad k = 1,2,3,\ldots \tag{7.1}$$

and

$$\text{pr}\{N = k\} = \begin{cases} 1-p, & k = 1, \\ pc_k, & k = 2,3,\ldots, \end{cases} \tag{7.2}$$

where $c_k \geq 0$ are fixed and independent of p and $\sum_{k=2}^{\infty} c_k = 1$.

Our objective is to obtain the limit law of T_r, the waiting time until at least r balls have accumulated in some line in the context of compound Poisson processes.

LEMMA 12. *Let S_r be the waiting time until at least r balls have accumulated for a single compound Poisson process with rate parameter $1/m$ where a random number N (≥ 1) of balls are created at each Poisson event. Let $f(\xi)$ be the density function of S_r.*

(a) *For N geometrically distributed with parameter p, as in* (7.1),

$$f(\xi) = e^{-\xi/m}\sum_{k=0}^{r-1}\binom{r-1}{k}\frac{\xi^k}{k!m^{k+1}}p^{r-1-k}(1-p)^k. \tag{7.3}$$

(b) *For N distributed as in* (7.2),

$$f(\xi) = e^{-\xi/m}\sum_{k=0}^{r-1}A_k\frac{\xi^k}{k!m^{k+1}}, \tag{7.4}$$

where

$$A_0 = p\sum_{i=r}^{\infty} c_i, \qquad A_{r-1} = (1-p)^{r-1}$$

and for $k = 1, 2, \ldots, r-2$, A_k *is of the order* $O(p)$.

PROOF. (a) In order to calculate $f(\xi)\,\Delta\xi$, we condition on k events of the Poisson process in time $(0, \xi)$, producing $(n_1, n_2, \ldots, n_k)$ contributions of balls, $\sum_{i=1}^{k} n_i \le r-1$, and a Poisson event during the infinitesimal time interval $(\xi, \xi + \Delta\xi)$ adding at least $r - \sum_{i=0}^{k} n_i$ further balls. This combined probability is

$$e^{-\xi/m}\left(\frac{\xi}{m}\right)^k \frac{1}{k!}(1-p)^k \prod_{i=1}^{k} p^{n_i - 1}\,\frac{\Delta\xi}{m}\,p^{r-1-n_1-\cdots-n_k}$$

$$= e^{-\xi/m}\,\frac{\xi^k}{k!\,m^{k+1}}(1-p)^k p^{r-k-1}\,\Delta\xi.$$

Summing over all configurations $(n_1, \ldots, n_k)$ such that $n_i \ge 1$, $i = 1, \ldots, k$, and $\sum_{i=1}^{k} n_i \le r-1$,

$$\sum_{(n_1,\ldots,n_k)} 1 = \binom{r-1}{k},$$

yielding the formula (7.3). The derivation of (7.4) is similar. □

THEOREM 13. *Consider a system of* m *independent Poisson process lines each with parameter* $1/m$. *At the occurrence of each event,* N *balls are contributed, where* N *is a random integer greater than or equal to* 1. *Let* T_r *be the waiting time until at least* r *balls have accumulated in some line.*

(a) *Suppose* N *is geometrically distributed with parameter* p, *where* $p = a/m^{\beta}$ *for some constant* a *independent of* m. *Then the limiting behavior of* T_r, *as* $m \to \infty$, *is given as*:

$$\text{if} \quad \begin{Bmatrix} \beta < \dfrac{1}{r} \\ \beta = \dfrac{1}{r} \\ \beta > \dfrac{1}{r} \end{Bmatrix}$$

(7.5)

$$\text{then} \quad \begin{Bmatrix} \operatorname{pr}\left\{\dfrac{T_r}{m^{(r-1)\beta}} > x\right\} \to \exp\{-a^{r-1}x\} \\ \operatorname{pr}\left\{\dfrac{T_r}{m^{(r-1)/r}} > x\right\} \to \exp\left\{-\displaystyle\sum_{k=0}^{r-1}\binom{r-1}{k}\frac{a^{r-1-k}x^{k+1}}{(k+1)!}\right\} \\ \operatorname{pr}\left\{\dfrac{T_r}{m^{(r-1)/r}} > x\right\} \to \exp\left\{\dfrac{-x^r}{r!}\right\} \end{Bmatrix}.$$

(b) *Suppose N is distributed such that*

$$\operatorname{pr}\{N = 1\} = 1 - p,$$
$$\operatorname{pr}\{N = k\} = pc_k, \qquad k = 2, 3, \ldots,$$

where $p = a/m^{\beta}$ for some constant a independent of m and some $\{c_k\}_{k=2}^{\infty}$ independent of p with $\sum_{k=2}^{\infty} c_k = 1$. Then the limiting behavior of T_r is given as

$$\text{if} \quad \left\{ \begin{array}{l} \beta < 1 - \dfrac{1}{r} \\ \beta = 1 - \dfrac{1}{r} \\ \beta > 1 - \dfrac{1}{r} \end{array} \right\}$$

(7.6)

$$\text{then} \quad \left\{ \begin{array}{l} \operatorname{pr}\left\{ \dfrac{T_r}{m^{\beta}} > x \right\} \to \exp\left\{ -ax \sum_{i=r}^{\infty} c_i \right\} \\ \operatorname{pr}\left\{ \dfrac{T_r}{m^{(r-1)/r}} > x \right\} \to \exp\left\{ -\dfrac{x^r}{r!} - ax \sum_{i=r}^{\infty} c_i \right\} \\ \operatorname{pr}\left\{ \dfrac{T_r}{m^{(r-1)/r}} > x \right\} \to \exp\left\{ \dfrac{-x^r}{r!} \right\} \end{array} \right\}.$$

PROOF. By symmetry and independence of the Poisson lines,

$$\operatorname{pr}\{T_r/m^{\alpha} > x\} = (1 - F(xm^{\alpha}))^m,$$

where F denotes the distribution function of S_r. For T_r/m^{α} to have a nontrivial limiting distribution, $F(xm^{\alpha})$ must be of the order $1/m$.

(a) Integrating the density function of S_r with a change of variable $\eta = \xi/m$ gives

$$F(xm^{\alpha}) = \sum_{i=0}^{r-1} \binom{r-1}{i} \frac{p^{r-1-i}(1-p)^i}{i!} \int_0^{x/m^{1-\alpha}} \eta^i e^{-\eta}\, d\eta.$$

Substituting $p = a/m^{\beta}$ gives the estimates

$$\sum_{i=0}^{r-1} \binom{r-1}{i} \frac{a^{r-1-i}}{(i+1)!} \left[\frac{1}{m^{(r-1-i)\beta}} \frac{x^{i+1}}{m^{(1-\alpha)(i+1)}} + o\left(\frac{1}{m^{\beta(r-1-i)+(1-\alpha)(i+1)}} \right) \right].$$

(7.7)

For $\beta < 1/r$, set $\alpha = (r-1)\beta$. The above expression reduces to (only the first term remains involved)

$$\frac{a^{r-1}x}{m} + o\left(\frac{1}{m}\right)$$

(7.8)

and no other choice of α can produce a term of order $1/m$. In fact, for $\alpha > (r-1)\beta$, $[1 - F(xm^{\alpha})]^m \to 0$, and for $\alpha < (r-1)\beta$, $[1 - F(xm^{\alpha})]^m \to 1$.

For $\beta > 1/r$, the determination $\alpha = (r-1)/r$ gives

$$F(xm^{(r-1)/r}) = \frac{x^r}{r!m} + o\left(\frac{1}{m}\right) \tag{7.9}$$

coming from the last summand of (7.7). Finally for $\beta = 1/r$, with $\alpha = (r-1)/r$, all terms of (7.7) contribute to produce

$$F(xm^{(r-1)/r}) = \frac{1}{m}\sum_{i=0}^{r-1}\binom{r-1}{i}\frac{a^{r-1-i}x^{i+1}}{(i+1)!} + o\left(\frac{1}{m}\right). \tag{7.10}$$

The results of (7.5) follow immediately on the basis of (7.8)–(7.10).

(b) Following the same procedure as above, by (7.4) of Lemma 12, we have

$$\begin{aligned} F(xm^{\alpha}) &= \sum_{k=0}^{r-1} A_k \int_0^{xm^{\alpha}} \frac{\xi^k}{k!m^{k+1}} e^{-\xi/m}\, d\xi \\ &= A_0\left[\frac{x}{m^{1-\alpha}} + O\left(\frac{1}{m^{2(1-\alpha)}}\right)\right] + A_1\left[\frac{1}{2!}\left(\frac{x}{m^{1-\alpha}}\right)^2 + O\left(\frac{1}{m^{3(1-\alpha)}}\right)\right] \\ &\quad + \cdots + A_{r-1}\left[\frac{1}{r!}\left(\frac{x}{m^{1-\alpha}}\right)^r + O\left(\frac{1}{m^{(r+1)(1-\alpha)}}\right)\right]. \end{aligned}$$

With $p = a/m^{\beta}$ and the A_k's as given in Lemma 12,

$$F(xm^{\alpha}) = \frac{ax\sum_{n=r}^{\infty} c_n}{m^{\beta+(1-\alpha)}} + \frac{(1 - a/m^{\beta})^{r-1}x^r}{r!m^{1-\alpha}} + o\left(\frac{x}{m^{\beta+(1-\alpha)}}\right).$$

If $\beta < 1 - 1/r$, the first term dominates. Setting $\alpha = \beta$ gives

$$(1 - F(xm^{\alpha}))^m \to \exp\left\{-ax\sum_{n=r}^{\infty} c_n\right\}.$$

If $\beta > 1 - 1/r$, the second term dominates. Setting $\alpha = 1 - 1/r$ gives

$$(1 - F(xm^{\alpha}))^m \to \exp\left\{-\frac{x^r}{r!}\right\}.$$

If $\beta = 1 - 1/r$, setting $\alpha = 1 - 1/r$ gives

$$(1 - F(xm^{\alpha}))^m \to \exp\left\{-\frac{x^r}{r!} - ax\sum_{n=r}^{\infty} c_n\right\}. \qquad \square$$

8. Some data examples and interpretations.

A protein example. All *E. coli* (bacterium) proteins of the current data base (culled for repetitions) were concatenated, producing a sequence of $N = 160{,}247$ amino acids ($a = 20$) and similarly the human protein collection totalled $N = 170{,}737$. Following the prescription of (1.12) and (1.13), *frequent*

words for these data sets correspond to word size $s = 5$ (so $m = 20^5$) with at least $r = 5$ occurrences. [Data results not shown. See Karlin and Burge (1991).] The following interesting contrasts emerge.

1. High frequency words in the *E. coli* collection (the highest word count is 8) occur mostly in distinct protein sequences, whereas in the human collection, there are many copies (often tandem repeats) in relatively few sequences.
2. Many of the frequent words of the human set are homopeptides (i.e., iterations on one or two amino acid types, especially proline and glycine). This may reflect the abundance of collagen type proteins in humans.

A DNA example. Consider the genome (the totality of DNA) of all the human herpes viruses (herpes simplex virus 1, varicello zoster virus, cytomegalo virus and Epstein–Barr virus). The DNA totals 678,780 base pairs; alphabet {A, T, C, G}. The criterion for a rare word is size $s = 7$ and at most $r = 12$ occurrences. For these characteristics, 728 rare words qualified from a totality of $4^7 \approx 16{,}000$. All 7-words occurred at least once and the bottom 11 of least occurrences are TCTAGTA (1 occurrence), ACTAGGC (3), CTAACTC (3), TCTAGTC (3), AAGTTAG (4), ACTTAGG (4), ATCACTC (4), CTTAGCT (4), GACCTAA (4), GGACTAG (4) and TACTAAG (4). It is interesting that all but one of these words center on the stop codons (the DNA triplets that signal the termination of a gene sequence) TAG, TAA or TGA; see Karlin and Burge (1991).

REFERENCES

ALDOUS, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic. Applied Mathematical Sciences* **77**. Springer, New York.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25.

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5** 403–424.

BARBOUR, A. D. (1982). Poisson convergence and random graphs. *Math. Proc. Cambridge Philos. Soc.* **92** 349–359.

BARBOUR, A. D. and EAGLESON, G. K. (1984). Poisson convergence for dissociated statistics. *J. Roy. Statist. Soc. Ser. B* **46** 397–402.

BARBOUR, A. D. and HALL, P. (1984). On the rate of Poisson convergence. *Math. Proc. Cambridge Philos. Soc.* **95** 473–480.

BARBOUR, A. D. and HOLST, L. (1989). Some applications of the Stein–Chen method for proving Poisson convergence. *Adv. Appl. Probab.* **21** 74–90.

CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.

DRMANAC, R., LABAT, I., BRUKNER, I. and CRKVENJAKOV, R. (1989). Sequencing of megabase plus DNA by hybridization: Theory of the method. *Genomics* **4** 114–128.

GODBOLE, A. P. (1991). Poisson approximations for runs and patterns of rare events. *Adv. Appl. Probab.* To appear.

JANSON, S. (1987). Poisson convergence and Poisson processes with applications to random graphs. *Stochastic Processes Appl.* **26** 1–30.

JOHNSON, N. L. and KOTZ, S. (1977). *Urn Models and Their Applications*. Wiley, New York.

KARLIN, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17** 373–401.

KARLIN, S. (1986). Comparative analysis of structural relationships in DNA and protein sequences. In *Evolutionary Processes and Theory* (S. Karlin and E. Nevo, eds.) 329. Academic, Orlando, Fla.

KARLIN, S. and BURGE, C. (1991). Studies of DNA inhomogeneities in prokaryote species and phages. *Proc. Nat. Acad. Sci.* To appear.

KOLCHIN, V. F., SEVAST'YANOV, B. A. and CHISTYAKOV, V. P. (1978). *Random Allocations* (A. V. Balakrishnan, trans. and ed.) Wiley, New York.

STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2** 583–602. Univ. California Press, Berkeley.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DIVISION OF MATHEMATICS, COMPUTER SCIENCE AND STATISTICS
UNIVERSITY OF TEXAS AT SAN ANTONIO
SAN ANTONIO, TEXAS 78285