Adv. Appl. Prob. **46,** 168–185 (2014)
Printed in Northern Ireland

THE NUMBER OF COLLISIONS FOR THE OCCUPANCY PROBLEM WITH UNEQUAL PROBABILITIES

TOSHIO NAKATA,* *Fukuoka University of Education*

Abstract

In this article we study a number of collisions concerning a simple occupancy problem with unequal probabilities. Using combinatorial arguments and negative associations of random variables, we have several limit theorems, namely, a weak law of large numbers and a Poisson law of small numbers including the Chen–Stein estimate.

Keywords: Urn model; collision; negative association; Poisson approximation

2010 Mathematics Subject Classification: Primary 60C05
Secondary 91A60

1. Introduction

1.1. Problem and background

Throughout this article, we use the notation $\mathbb{R} := (-\infty, \infty)$ and $\mathbb{N} := \{1, 2, \ldots\}$. For $t \in \mathbb{N}$, we also use the convenient notation $[t] := \{1, 2, \ldots, t\}$.

The classical *occupancy problem* or the classical *ball-and-bin problem* has been studied extensively. See [7] and [8] for a survey and [12] for a detailed study of the asymptotic results. There also exist some variants of the classical occupancy problem. One was studied by Wendl [20], who considered the following problem.

> Letting $m, n, t \in \mathbb{N}$, we throw m white balls, denoted by A, and n black balls, denoted by B, into urns $[t]$ with probabilities $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$, respectively, where $p_i \geq 0$ and $q_i \geq 0$ for $i \in [t]$ and $\sum_{i\in[t]} p_i = \sum_{i\in[t]} q_i = 1$. For $k \in [t] \cup \{0\}$ what is the probability that the number of urns containing both colors, that is, the number of *collisions* between A and B, is k?

Prior to Wendl's [20] work, Popova [17] investigated a joint distribution of a number of collision urns and a number of non-collision urns and Selivanov [19] studied the first collision time for infinite numbers of colors. Moreover, there also exists some applied research concerning this problem. From the beginning Wendl [20] pointed out that this problem contained many applications to practical problems, for example, a clone mapping problem, a collision problem of airborne planes, etc. In [21] he focused on a DNA fingerprint mapping problem, and gave some numerical discussions. Motivated by computational learning theory, Boucheron and Gardy [3] studied a variant of the collision problem. More recently, Bodini *et al.* [2] investigated a *Boltzmann sampling* algorithm for an Hadamard product of two combinatorial classes. To evaluate the algorithm, they efficiently used some results of the collision problem. Nishimura

Received 29 November 2011; revision received 12 December 2012.
* Postal address: Department of Mathematics, Fukuoka University of Education, Akama-Bunkyomachi, Munakata, Fukuoka, 811-4192, Japan. Email address: nakata@fukuoka-edu.ac.jp
Dedicated to Professor Masafumi Yamashita on the occasion of his 60th birthday.

and Sibuya [15] noted that this problem was applicable to cryptography. They studied several variants of this problem under the condition that the throwing probabilities $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$ are uniform, namely, $p_i = q_i = 1/t$ for $i \in [t]$. In virtue of the uniform property the method of *enumerating surjections* using *Stirling numbers of the second kind* (see [9, Section II.3.1]) is applicable to the problem. Under the uniform condition, Nakata [13] directly studied this problem, and gave both the exact probability distribution of the number of collisions (see (5)) and its factorial moment (see (12)). These results were also essentially given in [15]. In [13] he gave several limit theorems, namely, a weak law of large numbers and Poisson approximations with the Chen–Stein method under some conditions.

1.2. Our contributions

In this article we present a further extension of [13] when both $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$ are not necessarily uniform. We denote by $X_t = X_t(m, n)$ the number of collisions between A, namely, m white balls, and B, namely, n black balls, into t urns. First, we indicate the exact probability distribution of X_t using $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$ instead of the Stirling numbers of the second kind (see Proposition 1). We also indicate the factorial moment (see Proposition 2). In the proof of [13, Theorem 2] the factorial moment was directly calculated using combinatorial arguments. In this article we utilize indicator random variables concerning collisions to obtain the factorial moment. In many previous works generating functions were often analyzed. However, in this article we give only the explicit form of the generating function, and we do not directly investigate it. Instead of analyzing generating functions, we examine indicator random variables concerning collisions. We show that these indicator random variables are *negatively associated*. The negative association of random variables is one qualitative version of the negative dependence of random variables, which was investigated intensively in [4], [5, Section 3.1], [11], and [16]. This implies that the variance of X_t does not exceed the expectation of X_t (see Proposition 3). In Proposition 5 we provide an upper and a lower bound on the second moment of X_t using the smart arguments given in [10, Section 4]. As a corollary, we give a simple upper bound of the probability without collisions (see Corollary 2 and compare with [20, Section 3]). Proposition 5 plays a role in giving concrete bounds with respect to Poisson approximations in Section 4.2.

Moreover, using these results, we argue a weak law of large numbers. In [13, Proposition 1] the weak law of large numbers was proved under some strong conditions including the uniform condition. However, we can prove it under a weak condition (see Theorem 1). We also give some examples satisfying the conditions of Theorem 1 (see Examples 1 and 2).

We investigate Poisson approximations of X_t using the moment method (see Theorem 2). Specifically, we check convergences of all factorial moments using the same method as that used in [13, Theorem 3]. Although the proof is more complicated, we can overcome this by estimating terms of a multinomial expansion (see Lemma 3). Using Theorem 2, we give some examples of Poisson convergences. An application of Theorem 2 is an approximation of the probability without collisions (see Corollary 3, and compare with [20, Section 3] and [6, Section 2.6, Example 6.6]). In Example 3.1 we give a numerical verification with respect to Corollary 3. It turns out that no collision probabilities (see (6)) under suitable conditions are well approximated by the Gumbel distribution. Finally, we give an error bound for the Poisson approximation using the Chen–Stein method (see Section 4.2). The estimates are easily given in virtue of the negative association of the random variables. In Example 4 we concretely show the Chen–Stein estimate corresponding to Example 3.

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

1.3. Plan of the article

In Section 2 we give combinatorial arguments. In Sections 2.1, 2.2, and 2.5 we respectively give the exact probability distribution, the factorial moment, and the generating function. In Section 2.3 we discuss negatively associated random variables and in Section 2.4 we give bounds concerning the second moment of X_t. In Section 3 we state the weak law of large numbers. Finally, in Section 4 we establish the law of small numbers and Poisson approximations, including the Chen–Stein estimate.

2. Combinatorial arguments

2.1. Exact probability distribution

In this subsection we show the exact probability distribution of X_t which is the number of collisions between m white balls, denoted by A, and n black balls, denoted by B. First, we need to introduce some notation. For A, let $\alpha_i \in [t]$ be a target urn of the ith white ball for $i \in [m]$. The multiset $A_m := \{\alpha_1, \dots, \alpha_m\}$ is called *a path* from A. Similarly, let $\beta_j \in [t]$ be a target urn of the jth black ball for $j \in [n]$. The multiset $B_n := \{\beta_1, \dots, \beta_n\}$ is also called *a path* from B. Fix $k \in [t] \cup \{0\}$. Considering k collisions on $[t]$, we introduce a family of pairs (I, J) such that both I and J are subsets of $[t]$ and $|I \cap J| = k$, where $|K|$ denotes the cardinality of the set K. Namely, we set

$$\mathcal{I}_k^{(i,j)}([t]) := \{(I, J) \colon I, J \subset [t],\ |I| = i,\ |J| = j,\ |I \cap J| = k\} \quad \text{for } i, j \in [t]. \tag{1}$$

For a subset $I \subset [t]$ and a path A_m, the notation $A_m \sqsupseteq I$ means that each element of I is contained in the path A_m, and each element of the path A_m is contained in I. Namely, the multiset A_m is equivalent to the ordinary set I, if $A_m \sqsupseteq I$. We similarly define $B_n \sqsupseteq J$ for a subset $J \subset [t]$ and B_n.

Using this notation, we state the following proposition.

Proposition 1. *Throw m white balls and n black balls into urns $[t]$ with probabilities $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$, respectively. Then the probability of k collisions is*

$$\mathbb{P}(X_t = k) = \sum_{i\in[m]} \sum_{j\in[n]} \sum_{(I,J)\in\mathcal{I}_k^{(i,j)}([t])} \Bigg(\sum_{A_m \sqsupseteq I} \prod_{l\in A_m} p_l\Bigg)\Bigg(\sum_{B_n \sqsupseteq J} \prod_{l\in B_n} q_l\Bigg) \quad \text{for } k \in [t] \cup \{0\}, \tag{2}$$

where '$\sum_{A_m \sqsupseteq I}$' and '$\sum_{B_n \sqsupseteq J}$' denote the summations for all paths satisfying $A_m \sqsupseteq I$ and $B_n \sqsupseteq J$, respectively.

Proof. We fix two subsets of urns $I \subset [t]$ and $J \subset [t]$. Let $\{A \Rightarrow I\}$ be the event that the target urns from A is I. The event $\{B \Rightarrow J\}$ is defined similarly. Then we have

$$\mathbb{P}(A \Rightarrow I) = \mathbb{P}\Bigg(\bigcup_{A_m \sqsupseteq I} \{A \text{ has a path } A_m\}\Bigg) = \sum_{A_m \sqsupseteq I} \prod_{l\in A_m} p_l. \tag{3}$$

Similarly, we have

$$\mathbb{P}(B \Rightarrow J) = \sum_{B_n \sqsupseteq J} \prod_{l\in B_n} q_l. \tag{4}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

If the number of collisions is k then $|I \cap J| = k$. Hence, we have

$$\begin{aligned}\mathbb{P}(X_t = k) &= \mathbb{P}\Bigg(\bigcup_{\{I,J\,:\,|I\cap J|=k\}} \{A \Rightarrow I\} \cap \{B \Rightarrow J\}\Bigg) \\ &= \sum_{\{I,J\,:\,|I\cap J|=k\}} \mathbb{P}(\{A \Rightarrow I\} \cap \{B \Rightarrow J\}) \\ &= \sum_{\{I,J\,:\,|I\cap J|=k\}} \mathbb{P}(A \Rightarrow I)\mathbb{P}(B \Rightarrow J) \\ &= \sum_{i\in[m]} \sum_{j\in[n]} \sum_{(I,J)\in\mathcal{I}_k^{(i,j)}([t])} \Bigg(\sum_{A_m \supseteq I} \prod_{l\in A_m} p_l\Bigg)\Bigg(\sum_{B_n \supseteq J} \prod_{l\in B_n} q_l\Bigg).\end{aligned}$$

The third equality holds because the events $\{A \Rightarrow I\}$ and $\{B \Rightarrow J\}$ are independent. This completes the proof.

If we assume the uniform condition, namely, $p_i = q_i = 1/t$ for $i \in [t]$, then (2) has a somewhat simpler form. Indeed, probability (3) is

$$\mathbb{P}(A \Rightarrow I) = |I|! \left\{ {m \atop |I|} \right\} t^{-m},$$

where $\left\{ {n \atop k} \right\}$ denotes Stirling numbers of the second kind (see [9, Section II.3.1]). Similarly, probability (4) is

$$\mathbb{P}(B \Rightarrow J) = |J|! \left\{ {n \atop |J|} \right\} t^{-n}.$$

By (1) and the uniform condition, probability (2) is simplified as

$$\begin{aligned}\mathbb{P}(X_t = k) &= \sum_{i\in[m]} \sum_{j\in[m]} \binom{t}{i+j-k} \binom{i+j-k}{i-k, j-k, k} \left[i! \left\{ {m \atop i} \right\} t^{-m} \right] \left[j! \left\{ {n \atop j} \right\} t^{-n} \right] \\ &= \frac{1}{t^{m+n}} \sum_{i=k}^{m} \sum_{j=k}^{n} \left\{ {m \atop i} \right\} \left\{ {n \atop j} \right\} \frac{(i)_k (j)_k (t)_{i+j-k}}{k!} \quad \text{for } k = 0, \ldots, \min\{m, n, t\}, \quad (5)\end{aligned}$$

where $(t)_k$ denotes $\binom{t}{k} k! = t(t-1)\cdots(t-k+1)$ for nonnegative integers $t \geq 0$ and $k \geq 0$. Note that (5) was explicitly given as [14, Equation (4)]. In particular, the probability without collisions,

$$\mathbb{P}(X_t = 0) = \frac{1}{t^{m+n}} \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ {m \atop i} \right\} \left\{ {n \atop j} \right\} (t)_{i+j}, \quad (6)$$

was previously given in [20, Theorem 1] and [15, Equation (3.4)].

2.2. Factorial moment

In this subsection we indicate the factorial moment of X_t. We can obtain the expectation of X_t and the variance of X_t using the first and second factorial moments.

Proposition 2. *Assume that the conditions of Proposition 1 hold. Then, for $l \geq 1$, the lth factorial moment is*

$$\mathbb{E}((X_t)_l) = l! \sum_{L\subset[t],\,|L|=l} \left\{ \sum_{k=0}^{l} \sum_{I\subset L,\,|I|=k} (-1)^k (1-p_I)^m \right\} \left\{ \sum_{k=0}^{l} \sum_{J\subset L,\,|J|=k} (-1)^k (1-q_J)^n \right\}, \quad (7)$$

where $p_I := \sum_{i\in I} p_i$ and $q_J := \sum_{j\in J} q_j$.

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

Proof. We fix an urn $i \in [t]$. Let $A(i)$ be the event that there exists a white ball from A in the ith urn when throwing m white balls. Similarly, let $B(i)$ be the event that there exists a black ball from B in the ith urn when throwing n black balls. Note that the events $A(i)$ and $B(i)$ are independent because all balls are independently thrown.

Given $L \subset [t]$, we investigate the joint probability of the events $\{A(i)\}_{i\in L}$. Using E^{c} to denote the complement of an event E, we have

$$\begin{aligned}
\mathbb{P}\Big(\bigcap_{i\in L} A(i)\Big) &= 1 - \mathbb{P}\Big(\bigcup_{i\in L} A(i)^{\mathrm{c}}\Big) \\
&= 1 - \sum_{i=1}^{|L|} \sum_{I\subset L,\,|I|=i} (-1)^{i+1} \mathbb{P}\Big(\bigcap_{j\in I} A(j)^{\mathrm{c}}\Big) \\
&= 1 - \sum_{i=1}^{|L|} \sum_{I\subset L,\,|I|=i} (-1)^{i+1} (1-p_I)^m \\
&= \sum_{i=0}^{|L|} \sum_{I\subset L,\,|I|=i} (-1)^{i} (1-p_I)^m. \qquad (8)
\end{aligned}$$

The second equality holds due to the inclusion–exclusion principle. Namely, for events $\{C_i\}$,

$$\mathbb{P}\Big(\bigcup_{i=1}^{n} C_i\Big) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{I\subset[n],|I|=k} \mathbb{P}\Big(\bigcap_{i\in I} C_i\Big).$$

Similarly, for the given $L \subset [t]$, we have

$$\mathbb{P}\Big(\bigcap_{i\in L} B(i)\Big) = \sum_{i=0}^{|L|} \sum_{J\subset L,\,|J|=i} (-1)^{i} (1-q_J)^n. \qquad (9)$$

We now define an indicator function as

$$\xi_i := \begin{cases} 1 & \text{if a collision occurs in the } i\text{th urn,} \\ 0 & \text{if a collision does not occur in the } i\text{th urn.} \end{cases} \qquad (10)$$

By definition, $X_t = \sum_{i=1}^{t} \xi_i$ holds. Note that the $\{\xi_i\}_{i\in[t]}$ are *not* independent. More precisely, they are *negatively associated*. Indeed, we will show the negative association in Proposition 4 below. If the throwing probabilities are uniform, they are *exchangeable* (see [7, Section 2.9]).

Since $\xi_i^2 = \xi_i$, we obtain, for $l \geq 1$,

$$\begin{aligned}
\mathbb{E}((X_t)_l) &= \mathbb{E}\Big(\prod_{j=0}^{l-1}\Big(\sum_{i=1}^{t}\xi_i - j\Big)\Big) \\
&= \mathbb{E}\Big(1\cdot 2 \sum_{i_1<i_2} \xi_{i_1}\xi_{i_2} \prod_{j=2}^{l-1}\Big(\sum_{i=1}^{t}\xi_i - j\Big)\Big) \\
&= \mathbb{E}\Big(1\cdot 2\cdot 3 \sum_{i_1<i_2<i_3} \xi_{i_1}\xi_{i_2}\xi_{i_3} \prod_{j=3}^{l-1}\Big(\sum_{i=1}^{t}\xi_i - j\Big)\Big) \\
&= \cdots \\
&= l! \sum_{L\subset[t],\,|L|=l} \mathbb{E}\Big(\prod_{i\in L}\xi_i\Big). \qquad (11)
\end{aligned}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

By (10) and the definitions of $A(i)$ and $B(i)$, we have

$$\mathbb{E}\Big(\prod_{i\in L}\xi_i\Big)=\mathbb{P}\Big(\bigcap_{i\in L}\{\xi_i=1\}\Big)=\mathbb{P}\Big(\bigcap_{i\in L}\{A(i)\cap B(i)\}\Big)=\mathbb{P}\Big(\bigcap_{i\in L}A(i)\Big)\mathbb{P}\Big(\bigcap_{i\in L}B(i)\Big).$$

The last equality holds because the events $A(i)$ and $B(i)$ are independent. Hence, by (8), (9), and (11), we obtain the desired result.

Under the uniform condition, the factorial moment of X_t was given in [13, Equation (1)] and [15, Equation (3.2)]. Indeed, for a fixed $l\geq 1$,

$$\mathbb{E}((X_t)_l)=(t)_l\Big(\sum_{i=0}^{l}\binom{l}{i}(-1)^i\Big(1-\frac{i}{t}\Big)^m\Big)\Big(\sum_{j=0}^{l}\binom{l}{j}(-1)^j\Big(1-\frac{j}{t}\Big)^n\Big). \tag{12}$$

Using (7) with respect to $l=1$ and 2, we have the expectation of X_t and the variance of X_t.

Corollary 1. *Assume that the conditions of Proposition 1 hold. Then we have*

$$\mathbb{E}(X_t)=\sum_{i=1}^{t}(1-a_i)(1-b_i), \tag{13}$$

$$\begin{aligned}\mathrm{var}(X_t)=&\sum_{i,j}\{(1-a_i-a_j)(b_{ij}-b_ib_j)+(1-b_i-b_j)(a_{ij}-a_ia_j)+a_{ij}b_{ij}-a_ia_jb_ib_j\}\\&+\sum_{i}\{(1-a_i)(b_i-b_{ii})+(1-b_i)(a_i-a_{ii})-(a_i-a_{ii})(b_i-b_{ii})\},\end{aligned} \tag{14}$$

where

$$\begin{aligned}a_i=a_i(t):=(1-p_i)^m,\qquad b_i=b_i(t):=(1-q_i)^n,\\ a_{ij}=a_{ij}(t):=(1-p_i-p_j)^m,\qquad b_{ij}=b_{ij}(t):=(1-q_i-q_j)^n.\end{aligned} \tag{15}$$

Proof. By (7) with respect to $l=1$, we have

$$\mathbb{E}(X_t)=\sum_{i=1}^{t}\{1-(1-p_i)^m\}\{1-(1-q_i)^n\}=\sum_{i=1}^{t}(1-a_i)(1-b_i).$$

Hence, we obtain (13). Moreover, by (7) with respect to $l=2$, we have

$$\begin{aligned}\mathrm{var}(X_t)&=\mathbb{E}((X_t)_2)+\mathbb{E}(X_t)-\mathbb{E}^2(X_t)\\&=\sum_{i,j}(1-a_i-a_j+a_{ij})(1-b_i-b_j+b_{ij})-\sum_{i=1}^{t}(1-2a_i+a_{ii})(1-2b_i+b_{ii})\\&\quad+\sum_{i=1}^{t}(1-a_i)(1-b_i)-\sum_{i,j}\{(1-a_i)(1-b_i)(1-a_j)(1-b_j)\}\\&=\sum_{i,j}(1-a_i-a_j+a_{ij})(1-b_i-b_j+b_{ij})-(1-a_i)(1-b_i)(1-a_j)(1-b_j)\\&\quad+\sum_{i=1}^{t}(1-a_i)(1-b_i)-(1-2a_i+a_{ii})(1-2b_i+b_{ii})\end{aligned}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

$$= \sum_{i,j}(1-a_i-a_j)(b_{ij}-b_ib_j) + (1-b_i-b_j)(a_{ij}-a_ia_j) + a_{ij}b_{ij} - a_ia_jb_ib_j$$
$$+ \sum_{i=1}^{t}(1-a_i)(b_i-b_{ii}) + (1-b_i)(a_i-a_{ii}) - (a_i-a_{ii})(b_i-b_{ii}).$$

Hence, we obtain (14).

2.3. The expectation and the variance via the negative association

To investigate the concentration of X_t around $\mathbb{E}(X_t)$ we would like to compare the expectation of X_t and the variance of X_t. The main result of this subsection is the following.

Proposition 3. *Given two probability distributions $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$, we have*

$$\mathbb{E}(X_t) \geq \mathrm{var}(X_t). \tag{16}$$

To obtain (16), we usually calculate both (13) and (14) directly. However, we would like to avoid the calculations because they are complicated. Instead of direct calculations we investigate the dependence among $\{\xi_i\}_{i\in[t]}$ defined by (10).

First, we give the definition of negatively associated random variables (see [5, Section 3.1]). Random variables $\{Y_i\}_{i\in[t]}$ are *negatively associated* if, for all disjoint subsets $I, J \subset [t]$ and all nondecreasing functions $f: \mathbb{R}^I \to \mathbb{R}$ and $g: \mathbb{R}^J \to \mathbb{R}$, the following inequality holds:

$$\mathbb{E}(f(Y_i, i\in I)g(Y_j, j\in J)) \leq \mathbb{E}(f(Y_i, i\in I))\mathbb{E}(g(Y_j, j\in J)).$$

Then we have the following result.

Proposition 4. *The random variables $\{\xi_i\}_{i\in[t]}$ defined by (10) are negatively associated. In particular,*

$$\mathbb{E}(\xi_i\xi_j) \leq \mathbb{E}(\xi_i)\mathbb{E}(\xi_j) \quad \textit{for } i \neq j \in [t]. \tag{17}$$

We will show this later. By virtue of Proposition 4, it is not difficult to prove Proposition 3.

Proof of Proposition 3. If we obtain

$$\mathbb{E}((X_t)_2) \leq \mathbb{E}^2(X_t), \tag{18}$$

then we obtain (16) because

$$\mathrm{var}(X_t) = \mathbb{E}(X_t^2) - \mathbb{E}^2(X_t) = \mathbb{E}((X_t)_2) + \mathbb{E}(X_t) - \mathbb{E}^2(X_t) \leq \mathbb{E}(X_t),$$

using (18). Therefore, it is sufficient to show (18):

$$\mathbb{E}((X_t)_2) = \sum_{i\neq j}\mathbb{E}(\xi_i\xi_j) \leq \sum_{i\neq j}\mathbb{E}(\xi_i)\mathbb{E}(\xi_j) \leq \sum_{i\in[t]}\sum_{j\in[t]}\mathbb{E}(\xi_i)\mathbb{E}(\xi_j) = \mathbb{E}^2(X_t).$$

The first inequality holds by (17). This completes the proof.

To prove Proposition 4, we introduce the following subdividing indicator random variables with respect to (10). For $w \in [m]$, $b \in [n]$, and $i \in [t]$, we define

$$\xi_i^{w,b} := \begin{cases} 1 & \text{if a collision occurs in the } i\text{th urn between the white ball } w \\ & \text{and the black ball } b, \\ 0 & \text{otherwise.} \end{cases}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

By definition, we have

$$\xi_i = \begin{cases} 1 & \text{if } \sum_{w\in[m],b\in[n]} \xi_i^{w,b} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We now show that the $\{\xi_i^{w,b}\}_{i\in[t]}$ are negatively associated for each $w \in [m]$ and $b \in [n]$.

Lemma 1. *For each $w \in [m]$ and $b \in [n]$, the random variables $\{\xi_i^{w,b}\}_{i\in[t]}$ are negatively associated.*

Proof. For each $I, J \subset [t]$ satisfying $I \cap J = \varnothing$, we consider nondecreasing functions $f\colon \mathbb{R}^I \to \mathbb{R}$ and $g\colon \mathbb{R}^J \to \mathbb{R}$. We can assume that $f\colon \mathbb{R}^I \to [0,\infty)$, $g\colon \mathbb{R}^J \to [0,\infty)$, and $f(0,\dots,0) = g(0,\dots,0) = 0$, because we deal with $f(a_i, i \in I) - f(0,\dots,0)$ for all $(a_i, i \in I) \in \mathbb{R}^I$ and $g(b_j, j \in J) - g(0,\dots,0)$ for all $(b_j, j \in J) \in \mathbb{R}^J$ if necessary. Then we have

$$\mathbb{E}(f(\xi_i^{w,b},\ i \in I)g(\xi_j^{w,b},\ j \in J)) = 0 \le \mathbb{E}(f(\xi_i^{w,b},\ i \in I))\mathbb{E}(g(\xi_j^{w,b},\ j \in J)).$$

The equality holds since either $\{\xi_i^{w,b}\}_{i\in I}$ or $\{\xi_j^{w,b}\}_{j\in J}$ must be all zero for the fixed $w \in [m]$ and $b \in [n]$. This completes the proof.

To show Proposition 4, we quote the following property from [5].

Closure Under Products. *([5, p. 35].)* Suppose that $\{X_i\}_{i\in[n]}$ are negatively associated, and $\{Y_j\}_{j\in[m]}$ are also negatively associated. If $\{X_i\}_{i\in[n]}$ and $\{Y_j\}_{j\in[m]}$ are independent, then $\{X_i, Y_j\}_{i\in[n],\, j\in[m]}$ are also negatively associated.

Using this property, we show the following lemma.

Lemma 2. *All random variables $\{\xi_i^{w,b}\}_{i\in[t],\ w\in[m],\ b\in[n]}$ are negatively associated.*

Proof. We make use of the same method at that used in the last part of [5, Example 3.1, p. 36]. By Lemma 1, for each $w \in [m]$ and $b \in [n]$, the random variables $\{\xi_i^{w,b}\}_{i\in[t]}$ are negatively associated. For fixed $i \in [t]$ and $w \in [m]$ the random variables $\{\xi_i^{w,b}\}_{b\in[n]}$ are independent. Therefore, for a fixed $w \in [m]$, applying the closure under products property to

$$\{\xi_i^{w,1}\}_{i\in[t]}, \{\xi_i^{w,2}\}_{i\in[t]}, \dots, \{\xi_i^{w,n}\}_{i\in[t]},$$

it follows that $\{\xi_i^{w,b}\}_{i\in[t],b\in[n]}$ are negatively associated. Applying the closure under products property again to

$$\{\xi_i^{1,b}\}_{i\in[t],\ b\in[n]}, \{\xi_i^{2,b}\}_{i\in[t],\ b\in[n]}, \dots, \{\xi_i^{m,b}\}_{i\in[t],\ b\in[n]},$$

all members of $\{\xi_i^{w,b}\}_{i\in[t],\ w\in[m],\ b\in[n]}$ are also negatively associated. This completes the proof.

Using Lemma 2, we can now prove Proposition 4.

Proof of Proposition 4. For any disjoint subsets $I, J \subset [t]$ and any nondecreasing functions $f\colon \mathbb{R}^I \to \mathbb{R}$ and $g\colon \mathbb{R}^J \to \mathbb{R}$, set $f_1\colon \mathbb{R}^I \times \mathbb{R}^{[m]} \times \mathbb{R}^{[n]} \to \mathbb{R}$ and $g_1\colon \mathbb{R}^J \times \mathbb{R}^{[m]} \times \mathbb{R}^{[n]} \to \mathbb{R}$ as

$$f_1(\xi_i^{w,b},\ (i,w,b) \in R) := f(\xi_i,\ i \in I), \qquad g_1(\xi_j^{w,b},\ (j,w,b) \in S) \quad := g(\xi_j,\ j \in J),$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

where $R := I \times [m] \times [n]$ and $S := J \times [m] \times [n]$. Then $R \cap S = \varnothing$ since $I \cap J = \varnothing$, and, by definition, both f_1 and g_1 are nondecreasing. Therefore, by virtue of Lemma 2, we have

$$\begin{aligned}\mathbb{E}(f_1(\xi_i^{w,b},\ (i,w,b) \in R) g_1(\xi_j^{w,b},\ (j,w,b) \in S)) \\ \leq \mathbb{E}(f_1(\xi_i^{w,b},\ (i,w,b) \in R)) \mathbb{E}(g_1(\xi_j^{w,b},\ (j,w,b) \in S)).\end{aligned}$$

As a result,

$$\mathbb{E}(f(\xi_i,\ i \in I) g(\xi_j,\ j \in J)) \leq \mathbb{E}(f(\xi_i,\ i \in I)) \mathbb{E}(g(\xi_j,\ j \in J)).$$

This yields the desired result.

By virtue of negatively associated random variables, we can easily give an upper bound of $\mathbb{P}(X_t = 0)$.

Corollary 2. *We have*

$$\mathbb{P}(X_t = 0) \leq \prod_{i \in [t]} (a_i + b_i - a_i b_i).$$

Proof. Using (17), we have $\mathbb{P}(\xi_i = 1,\ \xi_j = 1) \leq \mathbb{P}(\xi_i = 1)\mathbb{P}(\xi_j = 1)$ for $i \neq j \in [t]$. This yields

$$\begin{aligned}\mathbb{P}(\xi_i = 0,\ \xi_j = 0) &= 1 - \mathbb{P}(\{\xi_i = 1\} \cup \{\xi_j = 1\}) \\ &= 1 - \mathbb{P}(\xi_i = 1) - \mathbb{P}(\xi_j = 1) + \mathbb{P}(\xi_i = 1, \xi_j = 1) \\ &\leq 1 - \mathbb{P}(\xi_i = 1) - \mathbb{P}(\xi_j = 1) + \mathbb{P}(\xi_i = 1)\mathbb{P}(\xi_j = 1) \\ &= \mathbb{P}(\xi_i = 0)\mathbb{P}(\xi_j = 0)\end{aligned}$$

for $i \neq j \in [t]$. Hence,

$$\mathbb{P}(X_t = 0) = \mathbb{P}\left(\bigcap_{i \in [t]} \{\xi_i = 0\}\right) \leq \prod_{i \in [t]} \mathbb{P}(\xi_i = 0) = \prod_{i \in [t]} (a_i + b_i - a_i b_i).$$

The last equality holds because

$$\mathbb{P}(\xi_i = 0) = 1 - \mathbb{P}(\xi_i = 1) = 1 - (1 - a_i)(1 - b_i) = a_i + b_i - a_i b_i.$$

This completes the proof.

2.4. A bound on the second moment

In this subsection we give a bound on the second moment of X_t using the properties of a_i, b_i, a_{ij}, and b_{ij} defined in (15). Indeed, we estimate the gap between $\mathbb{E}^2(X_t)$ and $\mathbb{E}((X_t)_2)$.

Proposition 5. *Assume that the conditions of Proposition 1 hold. Then we have*

$$0 \leq \mathbb{E}^2(X_t) - \mathbb{E}((X_t)_2) \leq mn\left\{(m+n)\left(\sum_{i \in [t]} p_i q_i\right)^2 + mn \sum_{i \in [t]} p_i^2 q_i^2\right\}. \tag{19}$$

Proof. The first inequality of (19) holds because of (18). However, we also give a direct calculation of $\mathbb{E}^2(X_t) - \mathbb{E}((X_t)_2)$ because some facts are needed later. First, we investigate some properties of a_i, b_i, a_{ij}, and b_{ij} defined in (15). Namely, we have

$$\begin{aligned}0 \leq a_i a_j - a_{ij} \leq m p_i p_j, \\ 0 \leq b_i b_j - b_{ij} \leq n q_i q_j.\end{aligned} \tag{20}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

By symmetry, we only show (20). The first inequality of (20) follows from

$$a_{ij} = (1 - p_i - p_j)^m \le (1 - p_i - p_j + p_i p_j)^m = (1 - p_i)^m (1 - p_j)^m = a_i a_j.$$

The second inequality of (20) follows from

$$a_i a_j - a_{ij} = (1 - p_i - p_j + p_i p_j)^m - (1 - p_i - p_j)^m \le 1^m - (1 - p_i p_j)^m \le m p_i p_j. \tag{21}$$

The first inequality of (21) follows from the convexity of $x \mapsto x^m$ for $m \ge 1$ and $0 \le x \le 1$, an idea used in [10, Section 4]. Hence, (20) holds. Using Proposition 2, we have

$$\mathbb{E}((X_t)_2) = \sum_{i \ne j} (1 - a_i - a_j + a_{ij})(1 - b_i - b_j + b_{ij}). \tag{22}$$

Since $(1 - 0)^m - (1 - p_i)^m \ge \{(1 - p_j) - 0\}^m - \{(1 - p_j) - p_i\}^m$, we obtain

$$1 - (1 - p_i)^m - (1 - p_j)^m + (1 - p_i - p_j)^m \ge 0.$$

Namely, we have $1 - a_i - a_j + a_{ij} \ge 0$. Combining (20), we have

$$0 \le 1 - a_i - a_j + a_{ij} \le 1 - a_i - a_j + a_i a_j = (1 - a_i)(1 - a_j). \tag{23}$$

Similarly, we have

$$0 \le 1 - b_i - b_j + b_{ij} \le 1 - b_i - b_j + b_i b_j = (1 - b_i)(1 - b_j). \tag{24}$$

By (22), (23), and (24), we obtain

$$\begin{aligned}
\mathbb{E}((X_t)_2) &\le \sum_{i \ne j} (1 - a_i)(1 - a_j)(1 - b_i)(1 - b_j) \\
&\le \sum_{i \ne j} (1 - a_i)(1 - a_j)(1 - b_i)(1 - b_j) + \sum_{i=1}^{t} (1 - a_i)^2 (1 - b_i)^2 \\
&= \sum_{i,j} (1 - a_i)(1 - a_j)(1 - b_i)(1 - b_j) \\
&= \sum_{i=1}^{t} (1 - a_i)(1 - b_i) \sum_{j=1}^{t} (1 - a_j)(1 - b_j) \\
&= \mathbb{E}^2(X_t).
\end{aligned}$$

Hence, we have the first inequality of (19).

On the other hand,

$$\begin{aligned}
&\mathbb{E}^2(X_t) - \mathbb{E}((X_t)_2) \\
&\quad = \sum_{i,j} (1 - a_i)(1 - a_j)(1 - b_i)(1 - b_j) - \sum_{i \ne j} (1 - a_i - a_j + a_{ij})(1 - b_i - b_j + b_{ij}) \\
&\quad = \sum_{i \in [t]} (1 - a_i)^2 (1 - b_i)^2 \\
&\qquad + \sum_{i \ne j} \{(1 - a_i - a_j + a_i a_j)(1 - b_i - b_j + b_i b_j) \\
&\qquad\qquad - (1 - a_i - a_j + a_{ij})(1 - b_i - b_j + b_{ij})\}
\end{aligned}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

$$
\begin{aligned}
&= \sum_{i\in[t]}(1-a_i)^2(1-b_i)^2 \\
&\quad + \sum_{i\neq j}\{(1-a_i)(1-a_j)(b_ib_j-b_{ij}) + (1-b_i)(1-b_j)(a_ia_j-a_{ij}) \\
&\qquad\qquad - (a_ia_j-a_{ij})(b_ib_j-b_{ij})\} \\
&\le \sum_{i\in[t]}(1-a_i)^2(1-b_i)^2 + \sum_{i\neq j}\{(1-a_i)(1-a_j)nq_iq_j + (1-b_i)(1-b_j)mp_ip_j\}.
\end{aligned}
$$

The last inequality holds because of (20). Since $(1-x)^m \ge 1-mx$ holds for $0<x<1$ and $m\ge 1$, we obtain

$$0 \le 1-a_i = 1-(1-p_i)^m \le mp_i.$$

This yields

$$
\begin{aligned}
\mathbb{E}^2(X_t) - \mathbb{E}((X_t)_2) &\le \sum_{i\in[t]}(mp_i)^2(nq_i)^2 + mn(m+n)\sum_{i\neq j}p_ip_jq_iq_j \\
&= mn\left\{(m+n)\left(\sum_{i\in[t]}p_iq_i\right)^2 + (mn-m-n)\sum_{i\in[t]}p_i^2q_i^2\right\} \\
&\le mn\left\{(m+n)\left(\sum_{i\in[t]}p_iq_i\right)^2 + mn\sum_{i\in[t]}p_i^2q_i^2\right\}.
\end{aligned}
$$

Hence, we have the second inequality of (19), completing the proof.

2.5. Generating function

In this subsection we give a generating function of X_t. Indeed, the method of calculation for the generating function is the same as that given in [15], [17], and [19]. Considering the probability generating function $\mathbb{E}(s^{X_t(m,n)}) = \sum_{k\ge 0}s^k\mathbb{P}(X_t(m,n)=k)$, we set

$$\Phi^{(t)}(s,x,y) := \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathbb{E}(s^{X_t(m,n)})\frac{x^m}{m!}\frac{y^n}{n!}t^{m+n}.$$

Then we have

$$\Phi^{(t)}(s,x,y) = \prod_{k=1}^{t}\{s(\mathrm{e}^{p_ktx}-1)(\mathrm{e}^{p_kty}-1) + \mathrm{e}^{p_ktx} + \mathrm{e}^{p_kty} - 1\}.$$

In particular, if the probabilities are uniform, then

$$\Phi^{(t)}(s,x,y) = \{s(\mathrm{e}^x-1)(\mathrm{e}^y-1) + \mathrm{e}^x + \mathrm{e}^y - 1\}^t.$$

By the Cauchy coefficient formula, we obtain (2). Moreover, (7) is also given by differentiating $\Phi^{(t)}(s,x,y)$ several times.

3. Weak law of large numbers

In [13, Proposition 1] it was shown that $X_t/\mathbb{E}(X_t)$ converges to 1 in probability as $t\to\infty$ under some additional conditions when the thrown probabilities are uniform, namely, $p_i = q_i = 1/t$ for $i\in[t]$. Here, we give some generalizations. Namely, we show a weak law of large numbers under a general condition even if the thrown probabilities are not uniform.

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

Note that when we consider limits concerning the number of urns t, the numbers of balls m, n and thrown probabilities $\{p_i\}_{i\in[t]}$, $\{q_i\}_{i\in[t]}$ depend on t. Moreover, we introduce some notation concerning the asymptotics along t. Namely, $g(t) = o(f(t))$ means that $\lim_{t\to\infty} g(t)/f(t) = 0$, and $g(t) = O(f(t))$ means that $\limsup_{t\to\infty} g(t)/f(t) < \infty$ for positive $f(t) > 0$ and $g(t) > 0$.

By Proposition 3, we see that the variance is not so large. Hence, X_t might concentrate around $\mathbb{E}(X_t)$. For convenience we use the following notation as the expectation of the number of collisions.

$$E(t) := \mathbb{E}(X_t) = \sum_{i=1}^{t} \{1 - (1 - p_i)^m\}\{1 - (1 - q_i)^n\}.$$

We have the following statement under the condition of infinite collisions as t goes to ∞.

Theorem 1. *If the expectation diverges, namely,*

$$\lim_{t\to\infty} E(t) = \infty, \tag{25}$$

then we have

$$\lim_{t\to\infty} \frac{X_t}{E(t)} = 1 \quad \textit{in probability.}$$

Proof. By (16) and (25), we have

$$0 \le \frac{\mathrm{var}(X_t)}{(E(t))^2} \le \frac{1}{E(t)} \to 0 \quad \text{as } t \to \infty.$$

Hence, by [6, Theorem I.5.4], we have the desired result.

We give the following concrete claim for (25) to be satisfied.

Claim 1. *Assume that there exist $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ satisfying*

$$\lim_{t\to\infty} |I_t(\varepsilon_1, \varepsilon_2)| = \infty, \tag{26}$$

where $I_t(\varepsilon_1, \varepsilon_2) := \{1 \le i \le t : mp_i \ge \varepsilon_1,\ nq_i \ge \varepsilon_2\}$. Then (25) holds.

Proof. For $\varepsilon > 0$, $m \ge 1$, and $0 < p < 1$, if $mp \ge \varepsilon$ then

$$1 - (1 - p)^m \ge 1 - \left(1 - \frac{\varepsilon}{m}\right)^m.$$

Hence, we have

$$\begin{aligned}
E(t) &= \sum_{i=1}^{t} \{1 - (1 - p_i)^m\}\{1 - (1 - q_i)^n\} \\
&\ge \sum_{i\in I_t(\varepsilon_1,\varepsilon_2)} \{1 - (1 - p_i)^m\}\{1 - (1 - q_i)^n\} \\
&\ge \sum_{i\in I_t(\varepsilon_1,\varepsilon_2)} \left\{1 - \left(1 - \frac{\varepsilon_1}{m}\right)^m\right\}\left\{1 - \left(1 - \frac{\varepsilon_2}{n}\right)^n\right\} \\
&= \left\{1 - \left(1 - \frac{\varepsilon_1}{m}\right)^m\right\}\left\{1 - \left(1 - \frac{\varepsilon_2}{n}\right)^n\right\} |I_t(\varepsilon_1, \varepsilon_2)| \\
&\ge (1 - \mathrm{e}^{-\varepsilon_1})(1 - \mathrm{e}^{-\varepsilon_2})|I_t(\varepsilon_1, \varepsilon_2)| \\
&\to \infty \quad \text{as } t \to \infty.
\end{aligned}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

The last inequality holds because $1-(1-\varepsilon/m)^m$ is monotone decreasing with respect to m, and the limit is $1-\mathrm{e}^{-\varepsilon}$. This completes the proof.

Example 1. We give examples which satisfy the condition of Claim 1.

1. Put $m=n:=t$ and $p_i=q_i:=1/t$ for $i\in[t]$. Then $I_t(1,1)=[t]$ holds. Hence, we have (26).
2. Put $m=n:=t$, $p_i:=i/\binom{t+1}{2}$, and $q_i:=(t-i+1)/\binom{t+1}{2}$ for $i\in[t]$. Since $mp_i=2i/(t+1)$ and $nq_i=2(t-i+1)/(t+1)$, we have
$$\left\{i\in[t]\colon mp_i\geq\frac{2}{3}\right\}\supset\left\{\left\lfloor\frac{t}{3}\right\rfloor+1,\dots,t\right\},\quad \left\{i\in[t]\colon nq_i\geq\frac{2}{3}\right\}\supset\left\{1,\dots,\left\lfloor\frac{2t}{3}\right\rfloor\right\}.$$
Therefore, we obtain
$$I_t\left(\frac{2}{3},\frac{2}{3}\right)\supset\left\{\left\lfloor\frac{t}{3}\right\rfloor+1,\dots,\left\lfloor\frac{2t}{3}\right\rfloor\right\}.$$
Hence, we have (26).

Claim 1 means that if the number of urns satisfying the condition that the collision probability is strictly positive is large then (25) holds. The condition of Claim 1 above is natural but strong. In Claim 2 we state that even if the collision probabilities are not so large, (25) may be obtained.

Claim 2. *Assume that*
$$\max_{1\leq i\leq t}\{p_i\}=o\left(\frac{1}{m}\right),\qquad \max_{1\leq i\leq t}\{q_i\}=o\left(\frac{1}{n}\right),\tag{27}$$
and
$$\lim_{t\to\infty}mn\sum_{i=1}^{t}p_iq_i=\infty.\tag{28}$$
Then (25) holds.

Proof. We use
$$(1-x)^n=1-nx+O(n^2x^2)\quad\text{for } 0<x<\frac{1}{n}.$$
Since $p_i=o(1/m)$ and $q_i=o(1/n)$ for $i\in[t]$, we have
$$1-(1-p_i)^m=mp_i+O(m^2p_i^2),\qquad 1-(1-q_i)^n=nq_i+O(n^2q_i^2).$$
Therefore, using (27), we have
$$\begin{aligned}E(t)&=\sum_{i=1}^{t}\{mp_i+O(m^2p_i^2)\}\{nq_i+O(n^2q_i^2)\}\\&=mn\sum_{i=1}^{t}\{(1+o(1))p_i\}\{(1+o(1))q_i\}\\&=(1+o(1))mn\sum_{i=1}^{t}p_iq_i\\&\to\infty\quad\text{as } t\to\infty.\end{aligned}$$
This completes the proof.

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

Example 2. We give an example which satisfies the conditions of Claim 2. Put $m = n := \lfloor t^{3/4} \rfloor$ and $p_i = q_i := 1/t$ for $i \in [t]$. Then we have $\lim_{t\to\infty} |I_t(\varepsilon_1, \varepsilon_2)| = 0$ for arbitrary $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, since $\lim_{t\to\infty} mp_i = \lim_{t\to\infty} nq_i = 0$. Hence, (26) does not hold. However, (27) holds by definition. Moreover, (28) also holds since $mn \sum_{i=1}^{t} p_i q_i \geq \lfloor t^{1/2} \rfloor \to \infty$ as $t \to \infty$.

4. Poisson approximations and the Chen–Stein method

4.1. Poisson law of small numbers

We investigate Poisson approximations of the number of collisions. Before stating a theorem, we prove the following lemma needed later.

Lemma 3. *Let $\{c_i(t)\}_{i\in[t]}$ be a positive sequence for $t \in \mathbb{N}$. If*

$$\lim_{t\to\infty} \max_{1\leq i\leq t} c_i(t) = 0 \tag{29}$$

and $\{\sum_{i=1}^{t} c_i(t)\}$ is positively bounded, namely, there exists $M > 0$ which satisfies

$$\sum_{i=1}^{t} c_i(t) \geq M > 0 \quad \textit{for all } t > 0, \tag{30}$$

then, for all integer $l \geq 1$, we have

$$\lim_{t\to\infty} \frac{\sum_{L\subset[t],\,|L|=l} l! \prod_{i\in L} c_i(t)}{(\sum_{i=1}^{t} c_i(t))^l} = 1.$$

Proof. By the multinomial expansion, we have

$$\left(\sum_{i=1}^{t} c_i(t)\right)^l = \sum_{L\subset[t],\,|L|=l} l! \prod_{i\in L} c_i(t) + \sum_{(i_1,\ldots,i_l)}^{*} \prod_{k=1}^{l} c_{i_k}(t), \tag{31}$$

where '$\sum^{*}_{(i_1,\ldots,i_l)}$' denotes the summation for $i_1, \ldots, i_l \in [t]$ and $1 \leq |\{i_1, \ldots, i_l\}| < l$. Therefore, we have

$$\begin{aligned} 0 &< \frac{\sum^{*}_{(i_1,\ldots,i_l)} \prod_{k=1}^{l} c_{i_k}(t)}{(\sum_{i=1}^{t} c_i(t))^l} \\ &\leq \frac{l!\,(\sum_{i=1}^{t} c_i(t))^{l-1} \max_{1\leq i\leq t} c_i(t)}{(\sum_{i=1}^{t} c_i(t))^l} \\ &\leq \frac{l!\,\max_{1\leq i\leq t} c_i(t)}{M} \\ &\to 0 \quad \text{as } t \to \infty. \end{aligned} \tag{32}$$

The first and second inequalities hold because of the positivity of $\{c_i(t)\}$. The last inequality holds because of (30). The convergence holds because of (29). Hence, by (31) and (32), we have the desired result.

Theorem 2. *Suppose that*

$$\lim_{t\to\infty} \max_{1\leq i\leq t} mp_i = 0, \qquad \lim_{t\to\infty} \max_{1\leq i\leq t} nq_i = 0, \tag{33}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

and that there exists $\lambda > 0$ *satisfying*

$$\lim_{t\to\infty} mn \sum_{i=1}^{t} p_i q_i = \lambda. \tag{34}$$

Then we have $\lim_{t\to\infty} \mathbb{E}(X_t) = \lambda$, *and*

$$\mathcal{L}(X_t) \to \mathrm{Poi}(\lambda) \quad \textit{in distribution},$$

where $\mathcal{L}(X)$ *is the distribution of the random variable* X *and* $\mathrm{Poi}(\gamma)$ *denotes a Poisson distribution with parameter* $\gamma > 0$.

Remark 1. Popova [17, Theorem 2] showed a similar statement using a generating function. However, it needs the condition that there exist $0 < C, C' < \infty$ satisfying $C \le m^2/t \le C'$ and $C \le n^2/t \le C'$, which our Theorem 2 does not require.

Proof of Theorem 2. We show that the lth factorial moment of X_t converges to λ^l for each fixed $l \ge 1$. In fact, it is known that the lth factorial moment of a Poisson random variable with parameter λ is

$$\mathbb{E}((Y)_l) = \lambda^l$$

(see [7, Equation (2.13)]).

Letting $L \subset [t]$ be a subset of urns whose cardinality is l, we consider the following term presented in (7):

$$\begin{aligned}\sum_{k=0}^{l} \sum_{I\subset L,\,|I|=k} (-1)^k (1-p_I)^m &= \sum_{k=0}^{l} \sum_{I\subset L,\,|I|=k} (-1)^k \mathrm{e}^{-mp_I}\left(1 - \frac{mp_I^2}{2} + O(mp_I^3)\right) \\ &= (1+o(1)) \prod_{i\in L} (1-\mathrm{e}^{-mp_i}).\end{aligned}$$

The first equality holds because

$$(1-p)^m = \mathrm{e}^{-mp}\left(1 - \frac{mp^2}{2} + O(mp^3)\right) \quad \text{for } 0 < p < 1.$$

The second equality holds because, for all $I \subset L \subset [t]$,

$$0 \le mp_I^2 \le m\Big(|I| \max_{1\le i\le t} p_i\Big)^2 \le \Big(l^2 \max_{1\le i\le t} p_i\Big)\Big(m \max_{1\le i\le t} p_i\Big) \to 0 \quad \text{as } t\to\infty,$$

which follows from (33). Therefore, we have

$$\begin{aligned}\mathbb{E}((X_t)_l) &= (1+o(1))l! \sum_{L\subset[t],\,|L|=l} \prod_{i\in L} \{(1-\mathrm{e}^{-mp_i})(1-\mathrm{e}^{-nq_i})\} \\ &= (1+o(1))l! \sum_{L\subset[t],\,|L|=l} \prod_{i\in L} \{(mp_i + O(m^2p_i^2))(nq_i + O(n^2q_i^2))\} \\ &= (1+o(1))l! \sum_{L\subset[t],\,|L|=l} \prod_{i\in L} \{(1+o(1))mnp_iq_i\} \\ &= (1+o(1))\Big(\sum_{i=1}^{t} mnp_iq_i\Big)^l \\ &\to \lambda^l \quad \text{as } t\to\infty.\end{aligned}$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

The third equality holds because of (33). The last equality holds using Lemma 3. In fact, we put $c_i = mnp_iq_i$ for $i \in [t]$ in Lemma 3. Then (29) and (30) follow from (33) and (34), respectively. The last convergence holds because of (34). Therefore, we have the convergence of all the lth moments. We complete the proof using the same argument as that given in [13, Equation (14), Theorem 3].

Corollary 3. *Let $F_t(x) := \mathbb{P}(X_t = 0)$ be the probability without collisions when $m = n := \lfloor\sqrt{\mathrm{e}^{-x}t}\rfloor$ for $x \in \mathbb{R}$ and $p_i = q_i = 1/t$ for $i \in [t]$. Then we have*

$$\lim_{t\to\infty} F_t(x) = \mathrm{e}^{-\mathrm{e}^{-x}} \quad \textit{for } x \in \mathbb{R}, \tag{35}$$

which is the Gumbel distribution (see [18, Proposition 0.3]).

Proof. Fix $x \in \mathbb{R}$. Since $mp_i = nq_i = \lfloor\sqrt{\mathrm{e}^{-x}t}\rfloor/t = O(t^{-1/2})$ for $i \in [t]$, we have (33). Since

$$mn\sum_{i=1}^{t} p_iq_i = \frac{\lfloor\sqrt{\mathrm{e}^{-x}t}\rfloor^2}{t} \to \mathrm{e}^{-x} \quad \text{as } t \to \infty,$$

we have $\mathcal{L}(X_t) \to \mathrm{Poi}(\mathrm{e}^{-x})$ in distribution as $t \to \infty$ from Theorem 2. Namely, we have

$$\lim_{t\to\infty} \mathbb{P}(X_t = k) = \mathrm{e}^{-\mathrm{e}^{-x}}\frac{(\mathrm{e}^{-x})^k}{k!} \quad \text{for } k = 0, 1, 2, \ldots,$$

which yields (35).

Example 3. We give examples which satisfy the conditions of Theorem 2.

1. Put $m = n := \lfloor\sqrt{\mathrm{e}^{-x}t}\rfloor$ for $x \in \mathbb{R}$ and $p_i = q_i = 1/t$ for $i \in [t]$. Then, by the proof of Corollary 3, we have $\mathcal{L}(X_t) \to \mathrm{Poi}(\mathrm{e}^{-x})$ in distribution as $t \to \infty$. In this setup we perform a numerical verification. Assuming that $m = n = 10$ and $t = \lfloor 100\mathrm{e}^x\rfloor$, we present values for both $F_t(x)$ defined in Corollary 3 and $\mathrm{e}^{-\mathrm{e}^{-x}}$ for $x \in [-3, 3]$ in Table 1. It turns out that the probabilities are well approximated even if m, n, and t are not so large.

TABLE 1: $F_t(x)$ and $\mathrm{e}^{-\mathrm{e}^{-x}}$ if $m = n = 10$ and $t = \lfloor 100\mathrm{e}^x\rfloor$ for $x \in [-3, 3]$.

x	$F_t(x)$	$\mathrm{e}^{-\mathrm{e}^{-x}}$
−3.0	$1.243\,665 \times 10^{-5}$	$1.892\,178\,69 \times 10^{-9}$
−2.4	$1.693\,3537 \times 10^{-4}$	$1.631\,9066 \times 10^{-5}$
−1.8	0.004 006 545	0.002 358 693
−1.2	0.037 01089	0.036 148 60
−0.6	0.161 8725	0.161 6828
0.0	0.366 8358	0.367 8794
0.6	0.576 5954	0.577 6358
1.2	0.739 6319	0.739 9340
1.8	0.847 5404	0.847 6403
2.4	0.913 2151	0.913 2752
3.0	0.951 4309	0.951 4319

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

2. Put $m = n := \lfloor \sqrt{t} \rfloor$, $p_i := i/\binom{t+1}{2}$, and $q_i := (t-i+1)/\binom{t+1}{2}$ for $i \in [t]$. Since

$$mn \sum_{i=1}^{t} p_i q_i = (1 + o(1)) \frac{2(t+2)}{3(t+1)} \to \frac{2}{3} \quad \text{as } t \to \infty,$$

we have $\lim_{t\to\infty} \mathcal{L}(X_t) = \mathrm{Poi}(\frac{2}{3})$ in distribution.

3. Put $m := \lfloor \sqrt{\log t} \rfloor$, $n := \lfloor t/\sqrt{\log t} \rfloor$, $p_i := (i \sum_{j=1}^{t} 1/j)^{-1}$, and $q_i := 1/t$ for $i \in [t]$. Since $mn \sum_{i=1}^{t} p_i q_i \to 1$ as $t \to \infty$, we have $\lim_{t\to\infty} \mathcal{L}(X_t) = \mathrm{Poi}(1)$ in distribution.

4.2. The Chen–Stein estimate

Finally, we give a simple estimate by the Chen–Stein method. We can give the same argument as in [13, Theorem 4]. However, we simply utilize [1, Corollary 2.C.2] in virtue of the negative association among $\{\xi_i\}$.

Theorem 3. *Throw m white balls and n black balls into urns $[t]$ with probabilities $\{p_i\}_{i\in[t]}$ and $\{q_i\}_{i\in[t]}$, respectively. Then we have*

$$\mathrm{d_{TV}}(\mathcal{L}(X_t), \mathrm{Poi}(\lambda)) \le \frac{1-\mathrm{e}^{-\lambda}}{\lambda} mn \left\{ (m+n) \left(\sum_{i\in[t]} p_i q_i \right)^2 + mn \sum_{i\in[t]} p_i^2 q_i^2 \right\}, \tag{36}$$

where $\lambda := \mathbb{E}(X_t)$ and $\mathrm{d_{TV}}$ denotes the total variation distance between two distributions, namely,

$$\mathrm{d_{TV}}(\mathcal{L}(X), \mathcal{L}(Y)) = \frac{1}{2} \sum_{k\ge 0} |\mathbb{P}(X=k) - \mathbb{P}(Y=k)|.$$

Proof. By the negative association among $\{\xi_i\}$, we make use of [1, Corollary 2.C.2]. Namely, we have

$$\mathrm{d_{TV}}(\mathcal{L}(X_t), \mathrm{Poi}(\lambda)) \le \frac{1-\mathrm{e}^{-\lambda}}{\lambda} (\mathbb{E}^2(X_t) - \mathbb{E}((X_t)_2)).$$

By (19) we have the desired result.

Example 4. To see (36), we illustrate concrete bounds appearing in Example 3.

1. Put $m = n := \lfloor \sqrt{\mathrm{e}^{-x} t} \rfloor$ for $x \in \mathbb{R}$, and $p_i = q_i := 1/t$ for $i \in [t]$. Then we have

$$\mathrm{d_{TV}}(\mathcal{L}(X_t), \mathrm{Poi}(\mathrm{e}^{-x})) \le \frac{2\mathrm{e}^{-x/2}(1-\mathrm{e}^{-\mathrm{e}^{-x}})}{\sqrt{t}} + O\left(\frac{1}{t}\right).$$

2. Put $m = n := \lfloor \sqrt{t} \rfloor$, $p_i := i/\binom{t+1}{2}$, and $q_i := (t-i+1)/\binom{t+1}{2}$ for $i \in [t]$. Then we have

$$\mathrm{d_{TV}}\left(\mathcal{L}(X_t), \mathrm{Poi}\left(\frac{2}{3}\right)\right) \le \frac{4(1-\mathrm{e}^{-2/3})}{3\sqrt{t}} + O\left(\frac{1}{t}\right).$$

3. Put $m := \lfloor \sqrt{\log t} \rfloor$, $n := \lfloor t/\sqrt{\log t} \rfloor$, $p_i := (i \sum_{j=1}^{t} 1/j)^{-1}$, and $q_i := 1/t$ for $i \in [t]$. Then we have

$$\mathrm{d_{TV}}(\mathcal{L}(X_t), \mathrm{Poi}(1)) \le \frac{1-\mathrm{e}^{-1}}{\sqrt{\log t}} + O(t^{-1}(\log t)^{1/2}).$$

https://doi.org/10.1239/aap/1396360108 Published online by Cambridge University Press

Acknowledgements

The author would like to thank the anonymous referee for helpful comments and suggestions, and Professor Danièle Gardy for interesting discussions during the Lattice Path Conference in Siena 2010. This research was supported by the Japan Society for the Promotion of Science, under grant KAKENHI 21540133.

References

[1] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation* (Oxford Studies Prob. **2**). Oxford University Press.
[2] Bodini, O., Gardy, D. and Roussel, O. (2012). Boys-and-girls birthdays and Hadamard products. *Fund. Inform.* **117,** 85–104.
[3] Boucheron, S. and Gardy, D. (1997). An urn model from learning theory. *Random Structures Algorithms* **10,** 43–67.
[4] Borcea, J., Brändén, P. and Liggett, T. M. (2009). Negative dependence and the geometry of polynomials. *J. Amer. Math. Soc.* **22,** 521–567.
[5] Dubhashi, D. and Panconesi, A. (2009). *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press.
[6] Durrett, R. (1996). *Probability: Theory and Examples*, 2nd edn. Duxbury Press, Belmont, CA.
[7] Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Application*. John Wiley, New York.
[8] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I, 3rd edn. John Wiley, New York.
[9] Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press.
[10] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Prob. Surveys* **4,** 146–171.
[11] Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables, with applications. *Ann. Statist.* **11,** 286–295.
[12] Kolchin, V. F., Sevast′yanov, B. A. and Chistyakov, V. P. (1978). *Random Allocations*. V. H. Winston, Washington, DC.
[13] Nakata, T. (2008). A Poisson approximation for an occupancy problem with collisions. *J. Appl. Prob.* **45,** 430–439.
[14] Nakata, T. (2008). Collision probability for an occupancy problem. *Statist. Prob. Lett.* **78,** 1929–1932.
[15] Nishimura, K. and Sibuya, M. (1988). Occupancy with two types of balls. *Ann. Inst. Statist. Math.* **40,** 77–91.
[16] Pemantle, R. (2000). Towards a theory of negative dependence. *J. Math. Phys.* **41,** 1371–1390.
[17] Popova, T. Yu. (1968). Limit theorems in a model of distribution of particles of two types. *Theory Prob. Appl.* **13,** 511–516.
[18] Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer, New York.
[19] Selivanov, B. I. (1995). On the waiting time in a scheme for the random allocation of colored particles. *Discrete Math. Appl.* **5,** 73–82.
[20] Wendl, M. C. (2003). Collision probability between sets of random variables. *Statist. Prob. Lett.* **64,** 249–254.
[21] Wendl, M. C. (2005). Probabilistic assessment of clone overlaps in DNA fingerprint mapping via a priori models. *J. Comput. Biol.* **12,** 283–297.