Теорія ймовір. та матем. статист.
Вип. 87, 2012, стор. 28–37

Teor. Ĭmovir. ta Matem. Statyst.
No. 87, 2012, pp. 28–37

ON THE ASYMPTOTIC BEHAVIOR OF A SEQUENCE OF RANDOM VARIABLES OF INTEREST IN THE CLASSICAL OCCUPANCY PROBLEM

UDC 519.21

RITA GIULIANO AND CLAUDIO MACCI

Abstract. In the classical occupancy problem one puts balls in n boxes, and each ball is independently assigned to any fixed box with probability $\frac{1}{n}$. It is well known that, if we consider the random number T_n of balls required to have all the n boxes filled with at least one ball, the sequence $\{T_n/(n\log n)\colon n \geq 2\}$ converges to 1 in probability. Here we present the large deviation principle associated to this convergence. We also discuss the use of the Gärtner Ellis Theorem for the proof of some parts of this large deviation principle.

Анотація. У класичній задачі про розміщення кулі розміщують в n урн, і кожна з куль незалежним чином потрапляє у фіксовану урну з ймовірністю $\frac{1}{n}$. Розглянемо випадкову величину T_n — кількість куль, яка потрібна для того, щоб заповнити кожну з n урн хоча б однією кулею. Добре відомо, що послідовність $\{T_n/(n\log n)\colon n \geq 2\}$ збігається до 1 за ймовірністю. У цій статті представлено принцип великих відхилень, пов'язаний з цією збіжністю, а також обговорюється корисність теореми Гертнер–Елліса для доведення деяких частин принципу великих відхилень у цьому випадку.

Аннотация. В классической задаче о размещении шары размещаются в n урн, и каждый из шаров независимым образом назначается в фиксированную урну с вероятностью $\frac{1}{n}$. Рассмотрим случайную величину T_n — количество шаров, которое требуется для того, чтобы заполнить каждую из n урн хотя бы одним шаром. Хорошо известно, что последовательность $\{T_n/(n\log n)\colon n \geq 2\}$ сходится к 1 по вероятности. В этой статье мы представляем принцип больших уклонений, связанный с этой сходимостью, а также обсуждаем полезность теоремы Гертнер–Эллиса для доказательства некоторых частей принципа больших уклонений в этом случае.

1. Introduction

There is a wide literature on urn models (see e.g. [8] and [9] for a wide source of results): we have closed formulas based on combinatorial methods, and asymptotic methods which often give a good approximation with a modest effort. Some asymptotic methods are based on Poisson approximation (see e.g. [1] as a general reference on this topic).

In the classical occupancy problem one puts balls in n boxes, and each ball is independently assigned to any fixed box with probability $\frac{1}{n}$; then, if we consider the random number T_n of balls required to have all the n boxes filled with at least one ball, it is known that the sequence $\{T_n/(n\log n)\colon n \geq 2\}$ converges to 1 in probability. We remark that a different formulation of the same problem in the literature leads to the well known coupon collector's problem: a coupon collector chooses at random among n coupon types, and let T_n be the number of coupons required to collect all the n coupon types.

2000 *Mathematics Subject Classification.* Primary 60F10, 60C05.

Key words and phrases. Large deviation principle, coupon collector's problem, triangular array, Poisson approximation.

The financial support of the Research Grant PRIN 2008 *Probability and Finance* is gratefully acknowledged.

We thank Francesco Pasquale for useful comments on the proof of (3) for $F \in \mathcal{C}_2$.

The theory of large deviations gives an asymptotic computation of small probabilities on exponential scale (see e.g. [2, 3, 12] as references on this topic). The basic concept of large deviation principle (see e.g. [2, pages 4–5]) consists of an upper bound for all closed sets and a lower bound for all open sets. In this paper we present the large deviation principle (LDP from now on) for the sequence $\{T_n/(n\log n)\colon n\geq 2\}$; in particular the proof of the lower bound is more interesting because the upper bound is an easy consequence of some results in the literature.

The interest of our LDP relies on the two following facts: (i) the speed function is $v_n = \log n$ instead of $v_n = n$ as it happens in other results on large deviations for sequences of interest in some occupancy problems (see e.g. [4, 5, 6, 13]); (ii) we cannot derive our LDP by using the Gärtner Ellis Theorem (see e.g. Theorem 2.3.6 in [2]), which in this case provides only a trivial non-sharp lower bound for open sets in terms of the exposed points of the rate function.

The outline of the paper is the following: in section 2 we present some preliminaries and the main result (Proposition 2.1); in section 3 we discuss the use of the Gärtner Ellis Theorem for the proof of some parts of the LDP of $\{T_n/(n\log n)\colon n\geq 2\}$. For the sake of completeness, the statement of Gärtner Ellis Theorem is recalled in the final Appendix A. Throughout the paper we write $[x] := \max\{k\in\mathbb{Z}\colon k\leq x\}$ for any $x\in\mathbb{R}$, and $x_n\sim y_n$ (as $n\to\infty$) to mean $\lim_{n\to\infty} x_n/y_n = 1$.

2. Preliminaries and main result

In view of Propositions 2.1–3.1 below, we recall some preliminaries. Firstly (see e.g. [7], Examples 6.5–6.6 and Theorem 6.6 in Chapter 2, pages 143–144) we have

$$P(T_n\leq m) = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\left(1-\frac{k}{n}\right)^m \quad \text{for each integer } m\geq 1. \tag{1}$$

Furthermore (see e.g. [7], Example 5.3 in Chapter 1, page 38) we have

$$T_n = \sum_{k=1}^{n} X_{n,k} \tag{2}$$

where $\{X_{n,k}\colon k\in\{1,\dots,n\}\}$ are independent random variables, and $X_{n,k}$ is geometric distributed with parameter $p_{n,k} = 1-(k-1)/n$; moreover $\{T_n/(n\log n)\colon n\geq 2\}$ converges to 1 in probability as $n\to\infty$.

Proposition 2.1. *The sequence $\{T_n/(n\log n)\colon n\geq 2\}$ satisfies the LDP with speed $v_n = \log n$ and good rate function I defined by*

$$I(x) = \begin{cases} x-1, & \text{if } x\geq 1\\ \infty, & \text{if } x<1.\end{cases}$$

This means that

$$\limsup_{n\to\infty}\frac{1}{\log n}\log \mathsf{P}\left(\frac{T_n}{n\log n}\in F\right)\leq -\inf_{x\in F} I(x) \quad \text{for all closed sets } F, \tag{3}$$

$$\liminf_{n\to\infty}\frac{1}{\log n}\log \mathsf{P}\left(\frac{T_n}{n\log n}\in G\right)\geq -\inf_{x\in G} I(x) \quad \text{for all open sets } G, \tag{4}$$

and the level sets $\{\{x\in\mathbb{R}\colon I(x)\leq\eta\}\colon \eta\geq 0\}$ are compact.

Proof. The proof is divided in two parts: the proof of (3) and the proof of (4). The compactness of the level sets $\{\{x\in\mathbb{R}\colon I(x)\leq\eta\}\colon\eta\geq 0\}$ is immediate and we omit the details.

Proof of (3). Firstly we remark that (3) trivially holds if $1\in F$ and, from now on, we assume that $1\notin F$. We also assume that both $F\cap(-\infty,1)$ and $F\cap(1,\infty)$ are nonempty

(at least one of them is nonempty and, if one of them is empty, the proof can readily adapted). Then we can define $x_1 := \max(F \cap (-\infty, 1))$ and $x_2 := \min(F \cap (1, \infty))$, and, since $F \subset (-\infty, x_1] \cup [x_2, \infty)$, we have

$$\limsup_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(\frac{T_n}{n\log n} \in F\right) \leq \max\left\{\limsup_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(\frac{T_n}{n\log n} \leq x_1\right), \limsup_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(\frac{T_n}{n\log n} \geq x_2\right)\right\}$$

by Lemma 1.2.15 in [2]. Thus we only have to check the upper bound (3) for $F \in \mathcal{C}_1 \cup \mathcal{C}_2 \cup \mathcal{C}_3$, where

$$\mathcal{C}_1 := \{[x, \infty) \colon x > 1\}, \qquad \mathcal{C}_2 := \{(-\infty, x] \colon x \in (0, 1)\}, \qquad \mathcal{C}_3 := \{(-\infty, x] \colon x \leq 0\},$$

and this will be a straightforward consequence of the following estimates.

- The case $F \in \mathcal{C}_3$ is trivial because $P(T_n/(n\log n) \leq 0) = 0$ (for all $n \geq 2$).
- For $F \in \mathcal{C}_1$, we consider $x > 1$ and $\varepsilon > 0$ small enough to have $x - \varepsilon > 1$; then, by a well known estimate (see e.g Exercise 3.10 in [11], page 58), we get (for all $n \geq 2$)
$$\mathsf{P}\left(\frac{T_n}{n\log n} \geq x\right) \leq \mathsf{P}(T_n > (x-\varepsilon) n \log n) \leq n^{1-x+\varepsilon},$$
and we let ε go to zero.
- For $F \in \mathcal{C}_2$, we consider $x \in (0, 1)$ and, by a well known estimate on Poisson approximation (see e.g. Theorem 5.10 and Corollary 5.11 in [10]), we get (for all $n \geq 2$)

$$\mathsf{P}\left(\frac{T_n}{n\log n} \leq x\right) = \mathsf{P}(T_n \leq xn\log n) \leq 2\left(1 - \exp\left\{-\frac{xn\log n}{n}\right\}\right)^n = 2(1 - n^{-x})^n.$$

Proof of (4). It is known (see e.g. condition (b) with eq. (1.2.8) in [2]) that it is enough to prove that

$$\liminf_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(\frac{T_n}{n\log n} \in G\right) \geq 1 - x$$

for all $x \geq 1$ and for all open set G such that $x \in G$; thus, if we find $\varepsilon > 0$ small enough to have $(x - \varepsilon, x + \varepsilon] \subset G$, we get the above inequality if we prove that

$$\liminf_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(x - \varepsilon < \frac{T_n}{n\log n} \leq x + \varepsilon\right) \geq 1 - x. \tag{5}$$

The latter condition holds trivially if $x = 1$ because of the convergence of

$$\{T_n/(n\log n) \colon n \geq 2\}$$

to 1 in probability; thus, in what follows, we prove (5) for $x > 1$ and $\varepsilon > 0$ small enough to have

$$(x - \varepsilon, x + \varepsilon] \subset G \cap (1, \infty).$$

We also assume $n \geq 2$ sufficiently large. We start noting that

$$\begin{aligned}\mathsf{P}\left(x - \varepsilon < \frac{T_n}{n\log n} \leq x + \varepsilon\right) &= \mathsf{P}\big((x-\varepsilon)n\log n < T_n \leq (x+\varepsilon)n\log n\big) \\ &\geq F_{T_n}([(x+\varepsilon)n\log n]) - F_{T_n}([(x-\varepsilon)n\log n] + 1),\end{aligned}$$

where, by (1),

$$\begin{cases} F_{T_n}([(x+\varepsilon)n\log n]) = \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^{[(x+\varepsilon)n\log n]}, \\ F_{T_n}([(x-\varepsilon)n\log n] + 1) = \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^{[(x-\varepsilon)n\log n]+1}. \end{cases}$$

We recall that, for every fixed $\gamma \geq 0$, for all $n \geq \gamma$ we have $(1-\gamma/n)^n \leq e^{-\gamma} \leq (1-\gamma/n)^{n-\gamma}$. Then we obtain

$$\begin{aligned}
F_{T_n}([(x+\varepsilon)n\log n]) &= \sum_{\text{even } k} \binom{n}{k} \left(1-\frac{k}{n}\right)^{[(x+\varepsilon)n\log n]} - \sum_{\text{odd } k} \binom{n}{k} \left(1-\frac{k}{n}\right)^{[(x+\varepsilon)n\log n]} \\
&= \sum_{\text{even } k} \binom{n}{k} \left(\left(1-\frac{k}{n}\right)^{n-k}\left(1-\frac{k}{n}\right)^{k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} \\
&\quad - \sum_{\text{odd } k} \binom{n}{k} \left(\left(1-\frac{k}{n}\right)^{n}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} \\
&\geq \sum_{\text{even } k} \binom{n}{k} \left(e^{-k}\left(1-\frac{k}{n}\right)^{k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} - \sum_{\text{odd } k} \binom{n}{k} \left(e^{-k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} \\
&= \left(\sum_{\text{even } k} \binom{n}{k} \left(e^{-k}\left(1-\frac{k}{n}\right)^{k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} - \sum_{\text{even } k} \binom{n}{k} \left(e^{-k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}}\right) \\
&\quad + \sum_{k=0}^{n} (-1)^k \binom{n}{k} \left(e^{-k}\right)^{\frac{[(x+\varepsilon)n\log n]}{n}} \\
&= -A_n^{(+)} + \left(1-e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^n
\end{aligned}$$

where

$$A_n^{(+)} := \sum_{\text{even } k} \binom{n}{k} e^{-k\frac{[(x+\varepsilon)n\log n]}{n}} \left(1-\left(1-\frac{k}{n}\right)^{k\frac{[(x+\varepsilon)n\log n]}{n}}\right).$$

Similarly, we also obtain

$$F_{T_n}([(x-\varepsilon)n\log n]+1) \leq A_n^{(-)} + \left(1-e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n$$

where

$$A_n^{(-)} := \sum_{\text{odd } k} \binom{n}{k} e^{-k\frac{[(x-\varepsilon)n\log n]+1}{n}} \left(1-\left(1-\frac{k}{n}\right)^{k\frac{[(x-\varepsilon)n\log n]+1}{n}}\right).$$

Then, if we consider

$$A_n := \left(1-e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^n - \left(1-e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n,$$

we have

$$\begin{aligned}
&\mathsf{P}\left(x-\varepsilon < \frac{T_n}{n\log n} \leq x+\varepsilon\right) \\
&\quad \geq -A_n^{(+)} + \left(1-e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^n - \left(A_n^{(-)} + \left(1-e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n\right) \\
&\quad = A_n - \left(A_n^{(+)} + A_n^{(-)}\right),
\end{aligned}$$

where $A_n, A_n^{(+)}, A_n^{(-)} \geq 0$. Thus we obtain

$$\begin{aligned}
\frac{1}{\log n}\log \mathsf{P}\left(x-\varepsilon < \frac{T_n}{n\log n} \leq x+\varepsilon\right) &\geq \frac{\log\left(A_n - \left(A_n^{(+)} + A_n^{(-)}\right)\right)}{\log n} \\
&= \frac{\log A_n}{\log n} + \frac{\log\left(1-\frac{A_n^{(+)}+A_n^{(-)}}{A_n}\right)}{\log n},
\end{aligned}$$

and we complete the proof of (4) by proving the following relations (actually we only need the lower bound in the first one):

(i): $1-x-\varepsilon \le \liminf_{n\to\infty} \frac{\log A_n}{\log n}$, $\limsup_{n\to\infty} \frac{\log A_n}{\log n} \le 1-x+\varepsilon$;
(ii): $\lim_{n\to\infty} \frac{A_n^{(+)}}{A_n} = 0$;
(iii): $\lim_{n\to\infty} \frac{A_n^{(-)}}{A_n} = 0$.

Indeed, since $\varepsilon > 0$ is arbitrary, for all $K > 1$ we have

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(x-\varepsilon < \frac{T_n}{n\log n} \le x+\varepsilon\right) &\ge \liminf_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(x-\frac{\varepsilon}{K} < \frac{T_n}{n\log n} \le x+\frac{\varepsilon}{K}\right)\\ &\ge 1-x-\frac{\varepsilon}{K}\end{aligned}$$

by the above conditions with ε/K in place of ε, and then we conclude letting K go to infinity.

Proof of (i). Put $f(y) = (1-e^{-y})^n$. By Lagrange Theorem there exists

$$\xi_n \in \left[\frac{[(x-\varepsilon)n\log n]+1}{n}, \frac{[(x+\varepsilon)n\log n]}{n}\right]$$

such that

$$\begin{aligned}A_n &= f\left(\frac{[(x+\varepsilon)n\log n]}{n}\right) - f\left(\frac{[(x-\varepsilon)n\log n]+1}{n}\right)\\ &= f'(\xi_n)\left(\frac{[(x+\varepsilon)n\log n]}{n} - \left(\frac{[(x-\varepsilon)n\log n]+1}{n}\right)\right)\\ &= \big([(x+\varepsilon)n\log n] - [(x-\varepsilon)n\log n] - 1\big)\left(1-e^{-\xi_n}\right)^{n-1} e^{-\xi_n},\end{aligned}$$

whence we obtain

$$\frac{\log A_n}{\log n} = \frac{\log\big([(x+\varepsilon)n\log n] - [(x-\varepsilon)n\log n] - 1\big)}{\log n} + \frac{(n-1)\log\left(1-e^{-\xi_n}\right)}{\log n} - \frac{\xi_n}{\log n}.$$

We complete the proof of (i) by considering the following relations (as $n\to\infty$):

$$\frac{\log\left([(x+\varepsilon)n\log n] - [(x-\varepsilon)n\log n] - 1\right)}{\log n} \sim \frac{\log\left(2\varepsilon n\log n\right)}{\log n} \to 1;$$

$$\begin{cases}(n-1)\log\left(1-e^{-\xi_n}\right)\\ \quad \le (n-1)\log\left(1-e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right) \sim (n-1)\log\left(1-n^{-(x+\varepsilon)}\right) \sim -\frac{n-1}{n^{x+\varepsilon}} \to 0,\\ (n-1)\log\left(1-e^{-\xi_n}\right)\\ \quad \ge (n-1)\log\left(1-e^{-\frac{[(x-\varepsilon)n\log n]-1}{n}}\right) \sim (n-1)\log\left(1-n^{-(x-\varepsilon)}\right) \sim -\frac{n-1}{n^{x-\varepsilon}} \to 0,\end{cases}$$

$$\begin{cases}-\frac{\xi_n}{\log n} \le -\frac{[(x-\varepsilon)n\log n]+1}{n\log n} \to -x+\varepsilon,\\ -\frac{\xi_n}{\log n} \ge -\frac{[(x+\varepsilon)n\log n]}{n\log n} \to -x-\varepsilon.\end{cases}$$

Proof of (ii). Put $f(y) = y^{[(x+\varepsilon)n\log n]/n}$. Then, for all $k \in \{0,\dots,n\}$, we have $1-(1-k/n)^k \le k^2/n$ and, by Lagrange Theorem, there exists $\xi_{n,k} \in \big[(1-k/n)^k, 1\big]$

such that

$$\begin{aligned}
0 \le A_n^{(+)} &= \sum_{\text{even } k} \binom{n}{k} e^{-k\frac{[(x+\varepsilon)n\log n]}{n}} \left(f(1) - f\left(\left(1-\frac{k}{n}\right)^k \right) \right) \\
&= \sum_{\text{even } k} \binom{n}{k} e^{-k\frac{[(x+\varepsilon)n\log n]}{n}} f'(\xi_{n,k}) \left(1 - \left(1-\frac{k}{n}\right)^k \right) \\
&\le \sum_{\text{even } k} \binom{n}{k} e^{-k\frac{[(x+\varepsilon)n\log n]}{n}} \left(\frac{[(x+\varepsilon)n\log n]}{n} \xi_{n,k}^{\frac{[(x+\varepsilon)n\log n]}{n}-1} \right) \frac{k^2}{n} \\
&\le \frac{[(x+\varepsilon)n\log n]}{n^2} \sum_{k=0}^{n} k^2 \binom{n}{k} e^{-k\frac{[(x+\varepsilon)n\log n]}{n}}.
\end{aligned}$$

Now recall the known formula $\sum_{k=0}^n k^2\binom{n}{k}y^k = ny(1+ny)(1+y)^{n-2}$ (for all $y \in \mathbb{R}$); since there exists $C > 0$ such that $0 < 1 + n\exp\{-[(x+\varepsilon)n\log n]/n\} < C$ for all $n \ge 1$, we get

$$0 \le A_n^{(+)} \le C\frac{[(x+\varepsilon)n\log n]}{n} e^{-\frac{[(x+\varepsilon)n\log n]}{n}} \left(1 + e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^{n-2}.$$

As far as A_n is concerned, we have

$$\begin{aligned}
A_n &= \left(1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n \left(\frac{\left(1 - e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^n}{\left(1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n} - 1 \right) \\
&= \left(1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n \left(\exp\left\{ n\log\left(\frac{1 - e^{-\frac{[(x+\varepsilon)n\log n]}{n}}}{1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}} \right) \right\} - 1 \right)
\end{aligned}$$

and, noting that

$$\lim_{n\to\infty} \left(1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n = 1,$$

we obtain the following estimate with some tedious computations:

$$\begin{aligned}
A_n &\sim \exp\left\{ n\log\left(1 + \frac{e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}} - e^{-\frac{[(x+\varepsilon)n\log n]}{n}}}{1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}} \right) \right\} - 1 \\
&\sim n\log(1 + n^{-(x-\varepsilon)}) \sim n^{1-x+\varepsilon}.
\end{aligned} \tag{6}$$

Then, noting that $\lim_{n\to\infty}\left(1+\exp\{-[(x+\varepsilon)n\log n]/n\}\right)^{n-2} = 1$, we complete the proof of (ii) as follows:

$$\begin{aligned}
0 \le \frac{A_n^{(+)}}{A_n} &\le \frac{C\frac{[(x+\varepsilon)n\log n]}{n} e^{-\frac{[(x+\varepsilon)n\log n]}{n}} \left(1 + e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^{n-2}}{\left(1 - e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^n - \left(1 - e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^n} \\
&\sim \frac{C(x+\varepsilon)\log n \cdot n^{-(x+\varepsilon)}}{n^{1-x+\varepsilon}} = \frac{C(x+\varepsilon)\log n}{n^{1+2\varepsilon}} \to 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Proof of (iii). We follow the lines of the proof of (ii). Firstly, if we set $f(y) = y^{\frac{[(x-\varepsilon)n\log n]+1}{n}}$, we get

$$0 \le A_n^{(-)} \le D\frac{[(x-\varepsilon)n\log n]+1}{n} e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}} \left(1 + e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^{n-2}$$

for a constant $D > 0$. Then, noting that $\lim_{n\to\infty}\big(1+\exp\big\{-\frac{[(x-\varepsilon)n\log n]+1}{n}\big\}\big)^{n-2} = 1$, by (6) we complete the proof of (iii) as follows:

$$0 \le \frac{A_n^{(-)}}{A_n} \le \frac{D\frac{[(x-\varepsilon)n\log n]+1}{n} e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}} \left(1+e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^{n-2}}{\left(1-e^{-\frac{[(x+\varepsilon)n\log n]}{n}}\right)^{n} - \left(1-e^{-\frac{[(x-\varepsilon)n\log n]+1}{n}}\right)^{n}}$$
$$\sim \frac{D(x-\varepsilon)\log n \cdot n^{-(x-\varepsilon)}}{n^{1-x+\varepsilon}} = \frac{D(x-\varepsilon)\log n}{n} \to 0 \quad \text{as } n\to\infty. \qquad \square$$

3. A discussion on the use of the Gärtner Ellis Theorem

In this section we discuss the use of the Gärtner Ellis Theorem for the proof of some parts of the LDP in Proposition 2.1. The application of this theorem consists in checking the existence of the function $\Lambda\colon \mathbb{R} \to (-\infty,\infty]$ defined by

$$\Lambda(\theta) := \lim_{n\to\infty} \frac{1}{\log n} \log \mathsf{E}\left[e^{\theta T_n/n}\right]; \tag{7}$$

then, if $0 \in (\{\theta\in\mathbb{R} : \Lambda(\theta) < \infty\})^\circ$ and if we consider the function I defined by

$$I(x) := \sup_{\theta\in\mathbb{R}}\{\theta x - \Lambda(\theta)\}, \tag{8}$$

we have the following three results: the upper bound (3); the lower bound

$$\liminf_{n\to\infty} \frac{1}{\log n} \log \mathsf{P}\left(\frac{T_n}{n\log n} \in G\right) \ge -\inf_{x\in G\cap\mathcal{F}} I(x) \quad \text{for all open sets } G, \tag{9}$$

where $\mathcal{F}$ is the set of exposed points (see e.g. Definition 2.3.3 in [2]); if Λ is essentially smooth (see e.g. Definition 2.3.5 in [2]) and lower semi-continuous, the LDP holds with the good rate function I. See also Appendix A.

In the next Proposition 3.1 we prove the existence of the limit (7) showing that

$$\Lambda(\theta) = \begin{cases} \theta, & \text{if } \theta \le 1, \\ \infty, & \text{if } \theta > 1; \end{cases} \tag{10}$$

therefore the function I in (8) meets the rate function in Proposition 2.1 because we have

$$I(x) = \sup_{\theta\le 1}\{\theta x - \theta\} = \sup_{\theta\le 1}\{\theta(x-1)\} = \begin{cases} x-1, & \text{if } x-1 \ge 0, \\ \infty, & \text{if } x-1 < 0. \end{cases}$$

Thus Gärtner Ellis Theorem provides an alternative proof of the upper bound (3) based on the sums in (2) expressed in terms of the random variables of a triangular array, and we do not need to consider the Poisson approximation. However we cannot derive the LDP from a complete application of Gärtner Ellis Theorem because the non-sharp lower bound (9) with $\mathcal{F} = \{1\}$ coincides with the sharp lower bound (4) if and only if $1 \in G$. Thus the LDP in Proposition 2.1 provides an example in which we can improve the consequences of Gärtner Ellis Theorem because we can prove the sharp lower bound (4) in place of the lower bound (9) in terms of the exposed points. Other two examples of the same situation can be found in Remark (d) after the statement of Theorem 2.3.6 in [2] and in Exercise 2.3.24 in [2]; in the first case the rate function J (say) is similar to the rate function I in Proposition 2.1 in this paper because we have $J(x) = I(x-1)$ for all $x \in \mathbb{R}$.

Proposition 3.1. *For all $\theta \in \mathbb{R}$, the limit* (7) *exists and Λ is given by* (10).

Proof. Firstly, by (2) and the hypotheses on the random variables $\{X_{n,k}\colon k \in \{1,\dots,n\}\}$, we have

$$\frac{1}{\log n}\log \mathsf{E}\left[e^{\theta T_n/n}\right] = \frac{\log \mathsf{E}\left[\exp\{\frac{\theta}{n}\sum_{k=1}^n X_{n,k}\}\right]}{\log n} = \frac{\sum_{k=1}^n \log \mathsf{E}\left[\exp\{\frac{\theta}{n}X_{n,k}\}\right]}{\log n}$$

and

$$\begin{aligned}\log \mathsf{E}\left[e^{\frac{\theta}{n}X_{n,k}}\right] &= \log\left(\sum_{j=1}^{\infty} e^{\frac{\theta}{n}j}\left(\frac{k-1}{n}\right)^{j-1}\left(1-\frac{k-1}{n}\right)\right)\\ &= \begin{cases}\log\left(\frac{\left(1-\frac{k-1}{n}\right)e^{\theta/n}}{1-\frac{k-1}{n}e^{\theta/n}}\right), & \text{if } \frac{k-1}{n}e^{\theta/n}<1,\\ \infty, & \text{if } \frac{k-1}{n}e^{\theta/n}\geq 1.\end{cases}\end{aligned}$$

Therefore (we recall that $n \geq 2$)

$$\begin{aligned}\log \mathsf{E}\left[e^{\theta T_n/n}\right] &= \begin{cases}\sum_{k=1}^n \log\left(\frac{\left(1-\frac{k-1}{n}\right)e^{\theta/n}}{1-\frac{k-1}{n}e^{\theta/n}}\right), & \text{if } \frac{k-1}{n}e^{\theta/n}<1 \text{ for all } k\in\{1,\dots,n\},\\ \infty, & \text{otherwise}\end{cases}\\ &= \begin{cases}\sum_{k=1}^n \log\left(\frac{\left(1-\frac{k-1}{n}\right)e^{\theta/n}}{1-\frac{k-1}{n}e^{\theta/n}}\right), & \text{if } \frac{n-1}{n}e^{\theta/n}<1,\\ \infty, & \text{otherwise}\end{cases}\\ &= \begin{cases}\sum_{k=1}^n \log\left(\frac{\left(1-\frac{k-1}{n}\right)e^{\theta/n}}{1-\frac{k-1}{n}e^{\theta/n}}\right), & \text{if } \theta < n\log\frac{n}{n-1},\\ \infty, & \text{if } \theta \geq n\log\frac{n}{n-1}.\end{cases}\end{aligned}$$

Then, since $n\log\frac{n}{n-1}\downarrow 1$ as $n\uparrow\infty$, the proof for $\theta>1$ is completed because $\theta \geq n\log\frac{n}{n-1}$ holds eventually, and therefore $\log \mathsf{E}\left[e^{\theta T_n/n}\right]=\infty$ eventually.

Hence, from now on, we restrict our attention to the case $\theta \leq 1$ and we can neglect the case $\theta = 0$ because the equality $\Lambda(0) = 0$ is trivial. Let us consider the function $h_{n,\theta}\colon [0,n)\to\mathbb{R}$ defined by $h_{n,\theta}(x) := (n - xe^{\theta/n})/(n-x)$; $h_{n,\theta}$ is increasing if $\theta<0$ and is decreasing if $\theta\in(0,1]$. Then we have

$$\begin{aligned}\log \mathsf{E}\left[e^{\theta T_n/n}\right] &= \sum_{k=1}^n \log\left(\frac{\left(1-\frac{k-1}{n}\right)e^{\theta/n}}{1-\frac{k-1}{n}e^{\theta/n}}\right) = \sum_{k=1}^n\left\{\log\left(\frac{1-\frac{k-1}{n}}{1-\frac{k-1}{n}e^{\theta/n}}\right)+\frac{\theta}{n}\right\}\\ &= \theta - \sum_{k=1}^n \log\left(\frac{1-\frac{k-1}{n}e^{\theta/n}}{1-\frac{k-1}{n}}\right) = \theta - \sum_{k=1}^n \log\left(\frac{n-(k-1)e^{\theta/n}}{n-(k-1)}\right)\\ &= \theta - \sum_{k=0}^{n-1}\log\left(\frac{n-ke^{\theta/n}}{n-k}\right) = \theta - \sum_{k=0}^{n-1}\log h_{n,\theta}(k),\end{aligned}$$

whence we obtain

$$\frac{1}{\log n}\log \mathsf{E}\left[e^{\log n\cdot\theta\frac{T_n}{n\log n}}\right] = \frac{\log \mathsf{E}\left[e^{\theta T_n/n}\right]}{\log n} = \frac{\theta}{\log n} - \frac{\sum_{k=0}^{n-1}\log h_{n,\theta}(k)}{\log n}.$$

Moreover we have the following bounds: for $\theta<0$,

$$\int_0^{n-1}\log h_{n,\theta}(x)\,dx + \log h_{n,\theta}(0) \leq \sum_{k=0}^{n-1}\log h_{n,\theta}(k) \leq \int_0^{n-1}\log h_{n,\theta}(x)\,dx + \log h_{n,\theta}(n-1);$$

for $\theta\in(0,1]$,

$$\int_0^{n-1}\log h_{n,\theta}(x)\,dx + \log h_{n,\theta}(n-1) \leq \sum_{k=0}^{n-1}\log h_{n,\theta}(k) \leq \int_0^{n-1}\log h_{n,\theta}(x)\,dx + \log h_{n,\theta}(0).$$

Hence, noting that $\log h_{n,\theta}(0) = 0$ and

$$\begin{aligned}\log h_{n,\theta}(n-1) &= \log\left(n-(n-1)e^{\theta/n}\right)\\ &= \log\left(e^{\theta/n}+n(1-e^{\theta/n})\right) \to \log(1-\theta) \quad \text{as } n\to\infty,\end{aligned}$$

we complete the proof for $\theta \le 1$ by checking that

$$\lim_{n\to\infty} \frac{\int_0^{n-1} \log h_{n,\theta}(x)\,dx}{\log n} = -\theta. \tag{11}$$

To this aim we note that

$$\begin{aligned}\frac{\int_0^{n-1} \log h_{n,\theta}(x)dx}{\log n} &= \frac{1}{\log n}\left[\frac{e^{\theta/n}x-n}{e^{\theta/n}}\log\left(n-e^{\theta/n}x\right)+(n-x)\log(n-x)\right]_{x=0}^{x=n-1}\\ &= \frac{1}{\log n}\left(\left(n-1-ne^{-\theta/n}\right)\log\left(n-(n-1)e^{\theta/n}\right)-\left\{-ne^{-\theta/n}\log n+n\log n\right\}\right)\\ &= \frac{1}{\log n}\left(\left(n\left(1-e^{-\theta/n}\right)-1\right)\log\left(n\left(1-e^{\theta/n}\right)+e^{\theta/n}\right)+n\left(e^{-\theta/n}-1\right)\log n\right);\end{aligned}$$

thus (11) can be checked by observing that

$$\lim_{n\to\infty}\left(n\left(1-e^{-\theta/n}\right)-1\right)\log\left(n\left(1-e^{\theta/n}\right)+e^{\theta/n}\right) = (\theta-1)\log(1-\theta)$$

and $\lim_{n\to\infty} n(e^{-\theta/n}-1) = -\theta$. □

Appendix A. Statement of Gärtner Ellis Theorem

In this Appendix we recall the statement of Gärtner Ellis Theorem. We refer to Theorem 2.3.6 in [2] with some changes: we prefer to give a presentation for a sequence of real valued random variables $\{Z_n : n \ge 1\}$ (instead of $\mathbb{R}^d$-valued random variables for some $d \ge 1$) and a general speed function $\{v_n : n \ge 1\}$ (it is a sequence that tends to infinity); in our results we consider the case $Z_n = T_n/(n\log n)$ and $v_n = \log n$.

We start with Assumption 2.3.2 in [2]: for each $\theta \in \mathbb{R}$, there exists

$$\Lambda(\theta) := \lim_{n\to\infty} \frac{1}{v_n}\log \mathsf{E}\left[e^{\theta v_n Z_n}\right]$$

as an extended real number; further 0 belongs to the interior of

$$\mathcal{D}_\Lambda := \{\theta \in \mathbb{R} : \Lambda(\theta) < \infty\}.$$

In this setting, we consider the Legendre transform I of Λ, i.e. the function I is defined by (8) above. Moreover we recall the definitions of exposed point (of I) and essentially smooth function.

Definition A.1. *We say that $y \in \mathbb{R}$ is an exposed point of I if, for some $\theta \in \mathbb{R}$, we have $\theta y - I(y) > \theta x - I(x)$ for all $x \ne y$.*

Definition A.2. *A convex function $\Lambda : \mathbb{R} \to (-\infty, \infty]$ is essentially smooth if $\mathcal{D}_\Lambda^\circ$ is non-empty, Λ is essentially smooth throughout $\mathcal{D}_\Lambda^\circ$ and Λ is steep (namely*

$$\lim_{n\to\infty} |\Lambda'(\theta_n)| \to \infty$$

whenever $\{\theta_n : n \ge 1\}$ is a sequence in $\mathcal{D}_\Lambda^\circ$ converging to a boundary point of $\mathcal{D}_\Lambda^\circ$).

Now we are ready to give the statement of Gärtner Ellis Theorem.

Theorem A.3. *Let Assumption 2.3.2 in* [2] *hold. Then:*

$$\limsup_{n\to\infty} \frac{1}{v_n} \log \mathsf{P}\left(Z_n \in F\right) \leq -\inf_{x\in F} I(x) \quad \textit{for all closed sets } F;$$

$$\liminf_{n\to\infty} \frac{1}{v_n} \log \mathsf{P}\left(Z_n \in G\right) \geq -\inf_{x\in G\cap\mathcal{F}} I(x) \quad \textit{for all open sets } G,$$

where $\mathcal{F}$ is the set of exposed points; if Λ is essentially smooth and lower semi-continuous, the LDP holds with the good rate function I.

References

1. A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*, The Clarendon Press–Oxford University Press, New York, 1992.
2. A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Second Edition, Springer-Verlag, New York, 1998.
3. P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, New York, 1997.
4. P. Dupuis, C. Nuzman, and P. Whiting, *Large deviation asymptotics for occupancy problems*, Ann. Probab. **32** (2004), 2765–2818.
5. P. Dupuis, C. Nuzman, and P. Whiting, *Large deviations principle for occupancy problems with colored balls*, J. Appl. Probab. **44** (2007), 115–141.
6. P. Dupuis, J. Zhang, and P. Whiting, *Refined large deviation asymptotics for the classical occupancy problem*, Methodol. Comput. Appl. Probab. **8** (2006), 467–496.
7. R. Durrett, *Probability: Theory and Examples*, Second Edition, Duxbury Press, Belmont CA, 1996.
8. N. L. Johnson and S. Kotz, *Urn Models and Their Applications*, John Wiley & Sons, New York, 1977.
9. H. M. Mahmoud, *Polya Urns Models*, CRC Press, Boca Raton, 2009.
10. M. Mitzenmacher and E. Upfal, *Probability and Computing*, Cambridge University Press, Cambridge, 2005.
11. R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, Cambridge, 1995.
12. A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*, Chapman & Hall, London, 1995.
13. J. X. Zhang and P. Dupuis, *Large-deviation approximations for general occupancy models*, Combin. Probab. Comput. **17** (2008), 437–470.

Dipartimento di Matematica "L. Tonelli", Università di Pisa, Largo Bruno Pontecorvo 5, I-56127 Pisa, Italy

E-mail address: giuliano@dm.unipi.it

Dipartimento di Matematica, Università di Roma Tor Vergata, Via della Ricerca Scientifica, I-00133 Rome, Italy

E-mail address: macci@mat.uniroma2.it

Received 13/12/2011