Statistics and Probability Letters 92 (2014) 232–240

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Rate of convergence and large deviation for the infinite color Pólya urn schemes

Antar Bandyopadhyay *, Debleena Thacker

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Delhi Centre, 7 S. J. S. Sansanwal Marg, New Delhi 110016, India

ARTICLE INFO

Article history:
Received 17 November 2013
Received in revised form 22 May 2014
Accepted 27 May 2014
Available online 11 June 2014

MSC:
primary 60F05
60F10
secondary 60G50

Keywords:
Berry–Esseen bound
Infinite color urn
Large deviation principle
Rate of convergence
Urn models

ABSTRACT

In this work we consider the *infinite color urn model* associated with a bounded increment random walk on $\mathbb{Z}^d$. This model was first introduced in Bandyopadhyay and Thacker (2013). We prove that the rate of convergence of the expected configuration of the urn at time n with appropriate centering and scaling is of the order $\mathcal{O}\left((\log n)^{-1/2}\right)$. Moreover we derive bounds similar to the classical Berry–Esseen bound. Further we show that for the expected configuration a *large deviation principle (LDP)* holds with a good rate function and speed $\log n$.

1. Introduction

Pólya urn scheme is one of the most well studied stochastic process which has plenty of applications in different fields. Since the time of its introduction by Pólya (1930) there has been a vast number of different variants and generalizations studied in the literature (see Pemantle (2007) for an extensive survey). In general one considers the model with finitely many colors and starts with an urn containing finitely many balls of different colors. At any time $n \geq 1$, a ball is selected uniformly at random from the urn, and its color is noted. The selected ball is then returned to the urn along with a set of balls of various colors which may depend on the color of the selected ball.

Recently Bandyopadhyay and Thacker (2013) has introduced a new generalization of the classical model with infinite but countably many colors with replacement mechanism corresponding to random walks in d-dimension. This generalization is essentially different than that of the classical Pólya urn scheme, as well as, the model introduced by Blackwell and MacQueen (1973), where the replacement mechanism is diagonal. The generalization by Bandyopadhyay and Thacker (2013) considers replacement mechanism with non-zero off diagonal entries and provides a novel connection between the two classical models, namely, Pólya urn scheme and random walks on d-dimensional Euclidean space. In the current work we exploit this connection to derive the *rate of convergence* and the *large deviation principle* for the $(n+1)$th selected color. In the following subsection we describe the specific model which we study.

* Corresponding author. Tel.: +91 11 4149 3932; fax: +91 11 4149 3981.
E-mail addresses: antar@isid.ac.in (A. Bandyopadhyay), thackerdebleena@gmail.com (D. Thacker).
URL: http://www.isid.ac.in/~antar/ (A. Bandyopadhyay).

http://dx.doi.org/10.1016/j.spl.2014.05.016

1.1. Infinite color urn model associated with random walks

Let $\left(X_j\right)_{j\geq 1}$ be i.i.d. random vectors taking values in $\mathbb{Z}^d$ with probability mass function $p\left(\mathbf{u}\right) := \mathbf{P}\left(X_1 = \mathbf{u}\right), \mathbf{u} \in \mathbb{Z}^d$. We assume that the distribution of X_1 is bounded, that is there exists a non-empty finite subset $B \subseteq \mathbb{Z}^d$ such that $p\left(u\right) = 0$ for all $u \notin B$. Throughout this paper we take the convention of writing all vectors as row vectors. Thus for a vector $\mathbf{x} \in \mathbb{R}^d$ we will write $\mathbf{x}^T$ to denote it as a column vector. The notations $\langle \cdot, \cdot \rangle$ will denote the usual Euclidean inner product on $\mathbb{R}^d$ and $\| \cdot \|$ the Euclidean norm. We will always write $\boldsymbol{\mu} := \mathbf{E}\left[X_1\right], \Sigma := \mathbf{E}\left[X_1^T X_1\right]$ and $e\left(\boldsymbol{\lambda}\right) := \mathbf{E}\left[e^{\langle \boldsymbol{\lambda}, X_1 \rangle}\right], \boldsymbol{\lambda} \in \mathbb{Z}^d$. When the dimension $d = 1$ we will denote the mean and variance simply by μ and σ^2 respectively.

Let $S_n := X_0 + X_1 + \cdots + X_n, n \geq 0$, be the random walk on $\mathbb{Z}^d$ starting at X_0 and with increments $\left(X_j\right)_{j\geq 1}$ which are independent. Needless to say that $\left(S_n\right)_{n\geq 0}$ is Markov chain with state-space $\mathbb{Z}^d$, initial distribution given by the distribution of X_0 and the transition matrix $R := \left(\left(p\left(\mathbf{u} - \mathbf{v}\right)\right)\right)_{u,v\in\mathbb{Z}^d}$.

In Bandyopadhyay and Thacker (2013) the following infinite color generalization of Pólya urn scheme was introduced where the colors were indexed by $\mathbb{Z}^d$. Let $U_n := \left(U_{n,\mathbf{v}}\right)_{\mathbf{v}\in\mathbb{Z}^d} \in [0, \infty)^{\mathbb{Z}^d}$ denote the configuration of the urn at time n, that is,

$$\mathbf{P}\left((n+1)\text{th selected ball has color } \mathbf{v} \,|\, U_n, U_{n-1}, \ldots, U_0\right) \propto U_{n,\mathbf{v}}, \quad \mathbf{v} \in \mathbb{Z}^d.$$

Starting with U_0 which is a probability distribution we define $\left(U_n\right)_{n\geq 0}$ recursively as follows:

$$U_{n+1} = U_n + \chi_{n+1} R \tag{1}$$

where $\chi_{n+1} = \left(\chi_{n+1,\mathbf{v}}\right)_{\mathbf{v}\in\mathbb{Z}^d}$ is such that $\chi_{n+1,V} = 1$ and $\chi_{n+1,\mathbf{u}} = 0$ if $\mathbf{u} \neq V$ where V is a random color chosen from the configuration U_n. Following Bandyopadhyay and Thacker (2013) we define the process $\left(U_n\right)_{n\geq 0}$ as the *infinite color urn model* with initial configuration U_0 and the replacement matrix R. We will also refer to it as the *infinite color urn model associated with the random walk* $\left(S_n\right)_{n\geq 0}$ *on* $\mathbb{Z}^d$. Throughout this paper we will assume that $U_0 = \left(U_{0,\mathbf{v}}\right)_{\mathbf{v}\in\mathbb{Z}^d}$ is such that $U_{0,\mathbf{v}} = 0$ for all but finitely many $\mathbf{v} \in \mathbb{Z}^d$. It is worth noting that $\sum_{\mathbf{u}\in\mathbb{Z}^d} U_{n,\mathbf{u}} = n + 1$ for all $n \geq 0$. So if Z_n denotes the $(n+1)$th selected color, then

$$\mathbf{P}\left(Z_n = \mathbf{v} | U_n, U_{n-1}, \ldots, U_0\right) = \frac{U_{n,\mathbf{v}}}{n+1} \Rightarrow \mathbf{P}\left(Z_n = \mathbf{v}\right) = \frac{\mathbf{E}\left[U_{n,\mathbf{v}}\right]}{n+1}. \tag{2}$$

In other words the expected configuration of the urn at time n is given by the distribution of Z_n.

1.2. Outline of the main contribution of the paper

In Bandyopadhyay and Thacker (2013) the authors studied the asymptotic distribution of Z_n, in particular, it has been proved (see Theorem 2.1 of Bandyopadhyay and Thacker (2013)) that as $n \to \infty$,

$$\frac{Z_n - \boldsymbol{\mu} \log n}{\sqrt{\log n}} \xrightarrow{d} N_d\left(\mathbf{0}, \Sigma\right). \tag{3}$$

In Section 2 we find the rate of convergence for the above asymptotic and show that the classical Berry–Esseen type bound holds at any dimension $d \geq 1$, which is of the order $\mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right)$.

It is easy to see that (3) implies

$$\frac{Z_n}{\log n} \xrightarrow{d} \boldsymbol{\mu} \quad \text{as } n \to \infty \Rightarrow \frac{Z_n}{\log n} \xrightarrow{p} \boldsymbol{\mu} \quad \text{as } n \to \infty. \tag{4}$$

In Section 3 we show that the above sequence of measures satisfy an LDP with a good rate function and speed $\log n$. A characterization of the rate function is also provided.

1.3. Fundamental representation

We end the introduction with the following very important observation made by Bandyopadhyay and Thacker (2013) (see Theorem 3.1 by Bandyopadhyay and Thacker (2013)):

Proposition 1. *Suppose Z_n be as defined before, then*

$$Z_n \stackrel{d}{=} Z_0 + \sum_{j=1}^{n} I_j X_j \tag{5}$$

where $\left(X_j\right)_{j\geq 1}$ are as above and $\left(I_j\right)_{j\geq 1}$ are independent Bernoulli variables such that $I_j \sim$ Bernoulli $\left(\frac{1}{j+1}\right)$ and are independent of $\left(X_j\right)_{j\geq 1}$. $Z_0 \sim U_0$ and is independent of $\left(\left(X_j\right)_{j\geq 1}, \left(I_j\right)_{j\geq 1}\right)$.

We will use this representation to derive the Berry–Esseen type bounds and also the LDP. The proof of Proposition 1 is given in detail in Bandyopadhyay and Thacker (2013), but for the sake of completeness we are presenting it here as well.

Proof of Proposition 1. From (2) we know that

$$\mathbf{P}\left(Z_n = \mathbf{v}\right) = \frac{\mathbf{E}\left[U_{n,\mathbf{v}}\right]}{n+1}.$$

So for $\boldsymbol{\lambda} \in \mathbb{R}^d$, the moment generating function of Z_n is given by

$$m_n\left(\boldsymbol{\lambda}\right) = \frac{1}{n+1} \sum_{v \in \mathbb{Z}^d} e^{\langle \boldsymbol{\lambda}, v\rangle} \mathbb{E}\left[U_{n,v}\right].$$

Note that $e\left(\boldsymbol{\lambda}\right)$ is an eigenvalue of R corresponding to the right eigenvector $x\left(\boldsymbol{\lambda}\right) = \left(e^{\langle \boldsymbol{\lambda}, v\rangle}\right)^T_{v \in \mathbb{Z}^d}$. Thus $m_n\left(\boldsymbol{\lambda}\right) = \frac{1}{n+1}\mathbb{E}\left[U_n\right] x\left(\boldsymbol{\lambda}\right)$.

Let $\mathcal{F}_n = \sigma\left(U_j : 0 \le j \le n\right), n \ge 0$ be the natural filtration. From (1) we obtain,

$$\mathbb{E}\left[U_{n+1} x\left(\boldsymbol{\lambda}\right) \middle| \mathcal{F}_n\right] = U_n x\left(\boldsymbol{\lambda}\right) + e\left(\boldsymbol{\lambda}\right) \mathbb{E}\left[\mathcal{X}_{n+1} x\left(\boldsymbol{\lambda}\right) \middle| \mathcal{F}_n\right] = \left(1 + \frac{e\left(\boldsymbol{\lambda}\right)}{n+1}\right) U_n x\left(\boldsymbol{\lambda}\right).$$

Therefore,

$$\begin{aligned} m_n\left(\boldsymbol{\lambda}\right) &= \frac{\mathbf{E}\left[U_0 x\left(\boldsymbol{\lambda}\right)\right]}{n+1} \prod_{j=1}^{n} \left(1 + \frac{e\left(\boldsymbol{\lambda}\right)}{j}\right) \\ &= \mathbf{E}\left[U_0 x\left(\boldsymbol{\lambda}\right)\right] \prod_{j=1}^{n} \left(\frac{j}{j+1} + \frac{e(\boldsymbol{\lambda})}{j+1}\right) \\ &= \mathbf{E}\left[U_0 x\left(\boldsymbol{\lambda}\right)\right] \prod_{j=1}^{n} \left(1 - \frac{1}{j+1} + \frac{e(\boldsymbol{\lambda})}{j+1}\right). \end{aligned} \tag{6}$$

This completes the proof. □

2. Berry–Esseen bounds for the expected configuration

2.1. Berry–Esseen bound for $d = 1$

We first consider the case when the associated random walk is a one dimensional walk and the set of colors is indexed by the set of integers $\mathbb{Z}$.

Theorem 1. *Suppose $U_0 = \delta_0$ then*

$$\sup_{x \in \mathbb{R}} \left| \mathbf{P}\left(\frac{Z_n - \mu h_n}{\sqrt{n \rho_2}} \le x\right) - \Phi\left(x\right) \right| \le 2.75 \times \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right), \tag{7}$$

where $h_n := \sum_{j=1}^{n} \frac{1}{j+1}$, Φ is the standard normal distribution function and

$$\rho_2 := \frac{1}{n}\left(\sigma^2 h_n - \mu^2 \sum_{j=1}^{n} \frac{1}{(j+1)^2}\right) \tag{8}$$

and

$$\rho_3 := \frac{1}{n}\left(\sum_{j=1}^{n} \frac{1}{j+1} \mathbf{E}\left[\left|X_1 - \frac{\mu}{j+1}\right|^3\right] + |\mu|^3 \sum_{j=1}^{n} \frac{j}{(j+1)^4}\right). \tag{9}$$

Proof. We first note that when $U_0 = \delta_0$ then (5) can be written as

$$Z_n \stackrel{d}{=} \sum_{j=1}^{n} I_j X_j \tag{10}$$

where $\left(X_j\right)_{j \ge 1}$ are i.i.d. increments of the random walk $(S_n)_{n \ge 0}$, $\left(I_j\right)_{j \ge 1}$ are independent Bernoulli variables such that $I_j \sim \text{Bernoulli}\left(\frac{1}{j+1}\right)$ and are independent of $\left(X_j\right)_{j \ge 1}$.

Now observe that

$$n\rho_2 = \sum_{j=1}^{n} \mathbf{E}\left[\left(I_jX_j - \mathbf{E}\left[I_jX_j\right]\right)^2\right] \quad \text{and} \quad n\rho_3 = \sum_{j=1}^{n} \mathbf{E}\left[\left|I_jX_j - \mathbf{E}\left[I_jX_j\right]\right|^3\right].$$

Thus from the *Berry–Esseen Theorem* for the independent but non-identical increments (see Theorem 12.4 of Bhattacharya and Ranga (1976)) we get

$$\sup_{x\in\mathbb{R}} \left| \mathbf{P}\left(\frac{\sum_{j=1}^{n} I_jX_j - \mu h_n}{\sqrt{n\rho_2}} \le x \right) - \Phi(x) \right| \le 2.75 \times \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}}. \tag{11}$$

Eqs. (10) and (11) imply the inequality in (7).

Finally to prove the last part of Eq. (7) we note that from the definition of $n\rho_2 \sim C_1 \log n$ and $n\rho_3 \sim C_2 \log n$ where $0 < C_1, C_2 < \infty$ are some constants. Thus

$$\frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right).$$

This completes the proof of the theorem. □

The following result follows easily from the above theorem by observing the facts $h_n \sim \log n$ and $n\rho_2 \sim C_1 \log n$ where $C_1 > 0$ is a constant.

Theorem 2. *Suppose $U_{0,k} = 0$ for all but finitely many $k \in \mathbb{Z}$ then there exists a constant $C > 0$ such that*

$$\sup_{x\in\mathbb{R}} \left| \mathbf{P}\left(\frac{Z_n - \mu \log n}{\sigma\sqrt{\log n}} \le x \right) - \Phi(x) \right| \le C \times \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right), \tag{12}$$

Φ is the standard normal distribution function and ρ_2 and ρ_3 are as defined in (8) and (9) respectively.

It is worth noting that unlike in Theorem 1 the constant C which appears in (12), is not a universal constant, it may depend on the increment distribution, as well as on U_0.

Proof. Observe that

$$\sup_{x\in\mathbb{R}} \left| \mathbf{P}\left(\frac{Z_n - \mu \log n}{\sigma\sqrt{\log n}} \le x \right) - \Phi(x) \right| \le \sup_{x\in\mathbb{R}} J_n(x) + \sup_{x\in\mathbb{R}} K_n(x)$$

where

$$J_n(x) = \left| \mathbf{P}\left(\frac{Z_n - \mu h_n}{\sqrt{n\rho_2}} \le x_n \right) - \Phi(x_n) \right|$$

and $x_n = \mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} + x \frac{\sigma\sqrt{\log n}}{\sqrt{n\rho_2}}$ and

$$K_n(x) = \left| \Phi\left(\mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} + x \frac{\sigma\sqrt{\log n}}{\sqrt{n\rho_2}} \right) - \Phi(x) \right|.$$

From Theorem 1 we observe that

$$\sup_{x\in\mathbb{R}} J_n(x) \le 2.75 \times \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right). \tag{13}$$

For a suitable choice of $C_1 > 0$, we have

$$\begin{aligned} K_n(x) &= \left| \frac{1}{\sqrt{2\pi}} \int_x^{\mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} + x \frac{\sigma\sqrt{\log n}}{\sqrt{n\rho_2}}} e^{\frac{-t^2}{2}}\, dt \right| \\ &\le C_1 e^{-\frac{x^2}{2}} \left| \mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} + x \frac{\sigma\sqrt{\log n}}{\sqrt{n\rho_2}} - x \right| \\ &\le C_1 \left| \mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} \right| + C_1 e^{-\frac{x^2}{2}} |x| \left| \sigma \frac{\sqrt{\log n}}{\sqrt{n\rho_2}} - 1 \right|. \end{aligned}$$

Observe that $h_n = \log n + \gamma + \epsilon_n$ where $\epsilon_n \longrightarrow 0$ as $n \to \infty$ and γ is the Euler constant. Also $\sqrt{n\rho_2} \sim \sqrt{\log n}$. Therefore, there exists a constant $C_2 > 0$ such that for all $n \in \mathbb{N}$

$$C_1 \left| \mu \frac{(\log n - h_n)}{\sqrt{n\rho_2}} \right| \le C_2 \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}} \right). \tag{14}$$

Note that the function $e^{-\frac{x^2}{2}} |x|$ attains its maximum at $x = 1$. Therefore,

$$C_1 e^{-\frac{x^2}{2}} |x| \left| \sigma \frac{\sqrt{\log n}}{\sqrt{n\rho_2}} - 1 \right| \le C_1 e^{-\frac{1}{2}} \left| \sigma \frac{\sqrt{\log n}}{\sqrt{n\rho_2}} - 1 \right|.$$

Since $\left|\sqrt{x} - 1\right| \le \sqrt{|x-1|}$ for all $x \in \mathbb{R}$, we obtain

$$C_1 e^{-\frac{1}{2}} \left| \sigma \frac{\sqrt{\log n}}{\sqrt{n\rho_2}} - 1 \right| \le C_3 \sqrt{\left| \frac{\sigma^2 \log n - n\rho_2}{n\rho_2} \right|} \tag{15}$$

for an appropriate constant $C_3 > 0$. Observe that for some constant $C_4 > 0$,

$$\begin{aligned} \frac{n\rho_2 - \sigma^2 \log n}{\sqrt{n\rho_2}} &= \frac{\sigma^2 \left(\sum_{j=1}^{n} \frac{1}{j+1} - \log n \right) - \mu^2 \sum_{j=1}^{n} \frac{1}{(j+1)^2}}{\sqrt{n\rho_2}} \\ &\le C_4 \frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}} \right). \end{aligned} \tag{16}$$

Therefore, combining (14)–(16) we can choose an appropriate constant $C > 0$ such that (12) holds. □

2.2. Berry–Esseen bound for $d \ge 2$

We now consider the case when the associated random walk is $d \ge 2$ dimensional and the colors are indexed by $\mathbb{Z}^d$. Before we present our main result we introduce few notations.

Notations. For a vector $\mathbf{x} \in \mathbb{R}^d$ is written as $\left(x^{(1)}, x^{(2)}, \ldots, x^{(d)}\right)$. For example the vector $\boldsymbol{\mu}$ will be written as $\left(\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(d)}\right)$. For a matrix $A = \left((a_{ij})\right)_{1 \le i,j \le d}$ we denote by $A(i,j)$ the $(d-1) \times (d-1)$ sub-matrix of A, obtained by deleting the ith row and jth column. Let

$$\rho_2^{(d)} := \frac{1}{n} \sum_{j=1}^{n} \frac{1}{(j+1)} \frac{\det\left(\Sigma - \frac{1}{j+1} M\right)}{\det\left(\Sigma(1,1) - \frac{1}{j+1} M(1,1)\right)}, \tag{17}$$

where $M := \left(\left(\mu^{(i)}\mu^{(j)}\right)\right)_{1 \le i,j \le d}$ and

$$\rho_3^{(d)} := \frac{1}{nd} \sum_{i=1}^{d} \gamma_n^3(i) \left(\sum_{j=1}^{n} \beta_j(i) \right), \tag{18}$$

where

$$\gamma_n^2(i) := \max_{1 \le j \le n} \frac{\det\left(\Sigma(i,i) - \frac{1}{(j+1)} M(i,i)\right)}{\det\left(\Sigma(1,1) - \frac{1}{j+1} M(1,1)\right)}$$

and

$$\beta_j(i) = \frac{1}{j+1} \mathbf{E}\left[\left| X_1^{(i)} - \frac{\mu^{(i)}}{j+1} \right|^3 \right] + \frac{j}{(j+1)^4} \left| \mu^{(i)} \right|^3.$$

For any two vectors $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^d$ we will write $\mathbf{x} \le \mathbf{y}$, if the inequality holds coordinate wise. Finally for a positive definite matrix B, we write $B^{1/2}$ for the unique positive definite square root of it.

Theorem 3. *Suppose $U_0 = \delta_{\mathbf{0}}$ then there exists an universal constant $C(d) > 0$ which may depend on the dimension d such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left|\mathbf{P}\left((Z_n - \boldsymbol{\mu} h_n)\, \Sigma_n^{-1/2} \le \mathbf{x}\right) - \Phi_d(\mathbf{x})\right| \le C(d) \frac{\sqrt{n}\rho_3^{(d)}}{\left(\rho_2^{(d)}\right)^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right), \tag{19}$$

where $\Sigma_n := \sum_{j=1}^n \frac{1}{j+1}\left(\Sigma - \frac{1}{j+1}M\right)$ and Φ_d is the distribution function of a standard d-dimensional normal random vector.

Proof. Like in the one dimensional case, we start by observing that when $U_0 = \delta_0$ then (5) can be written as

$$Z_n \stackrel{d}{=} \sum_{j=1}^n I_j X_j \tag{20}$$

where $(X_j)_{j\ge1}$ are i.i.d. increments of the random walk $(S_n)_{n\ge0}$, $(I_j)_{j\ge1}$ are independent Bernoulli variables such that $I_j \sim$ Bernoulli $\left(\frac{1}{j+1}\right)$ and are independent of $(X_j)_{j\ge1}$.

Now the proof of the inequality in (19) follows from equation (D) of Bergström (1949) which deals with d-dimensional version of the classical Berry–Esseen inequality for independent but non-identical summands, which in our case are the random variables $(I_jX_j)_{j\ge1}$. It is enough to notice that

$$\beta_j(i) = \mathbf{E}\left[\left|I_jX_1^{(i)} - \mathbf{E}\left[I_jX_j^{(i)}\right]\right|^3\right]$$

and

$$\Sigma_n = \sum_{j=1}^n \mathbf{E}\left[\left(I_jX_j - \mathbf{E}\left[I_jX_j\right]\right)^T\left(I_jX_j - \mathbf{E}\left[I_jX_j\right]\right)\right].$$

Finally to prove the last part of Eq. (19) just like in the one dimensional case we note that from the definition of $n\rho_2^{(d)} \sim C_1' \log n$ and $n\rho_3^{(d)} \sim C_2' \log n$ where $0 < C_1', C_2' < \infty$ are some constants. Thus

$$\frac{\sqrt{n}\rho_3}{\rho_2^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right).$$

This completes the proof of the theorem. □

Remark 2.1. If we define that $\Sigma(1, 1) = 1$ and $M(1, 1) = 0$ when $d = 1$ then Theorem 1 follows from the above theorem except in Theorem 1 the constant is more explicit.

Just like in the one dimensional case the following result follows easily from the above theorem by observing $h_n \sim \log n$.

Theorem 4. *Suppose $U_0 = (U_{0,\mathbf{v}})_{\mathbf{v}\in\mathbb{Z}^d}$ is such that $U_{0,\mathbf{v}} = 0$ for all but finitely many $\mathbf{v} \in \mathbb{Z}^d$ then there exists a constant $C > 0$ which may depend on the increment distribution, such that*

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left|\mathbf{P}\left(\left(\frac{Z_n - \boldsymbol{\mu}\log n}{\sqrt{\log n}}\right)\Sigma^{-1/2} \le \mathbf{x}\right) - \Phi_d(\mathbf{x})\right| \le C \times \frac{\sqrt{n}\rho_3^{(d)}}{\left(\rho_2^{(d)}\right)^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right), \tag{21}$$

where Φ_d is the distribution function of a standard d-dimensional normal random vector.

Proof. Observe that

$$\sup_{\mathbf{x}\in\mathbb{R}^d}\left|\mathbf{P}\left(\left(\frac{Z_n - \boldsymbol{\mu}\log n}{\sqrt{\log n}}\right)\Sigma^{-1/2} \le \mathbf{x}\right) - \Phi_d(\mathbf{x})\right| \le \sup_{\mathbf{x}\in\mathbb{R}^d} J_n(\mathbf{x}) + \sup_{\mathbf{x}\in\mathbb{R}^d} K_n(\mathbf{x})$$

where

$$J_n(\mathbf{x}) = \left|\mathbf{P}\left((Z_n - \boldsymbol{\mu}h_n)\,\Sigma_n^{-1/2} \le \mathbf{x}_n\right) - \Phi_d(\mathbf{x}_n)\right| \tag{22}$$

where $\mathbf{x}_n = \boldsymbol{\mu}(\log n - h_n)\,\Sigma_n^{-1/2} + \mathbf{x}\sqrt{\log n}\,\Sigma^{1/2}\Sigma_n^{-1/2}$ and

$$K_n(\mathbf{x}) = \left|\Phi_d(\mathbf{x}_n) - \Phi_d(\mathbf{x})\right|. \tag{23}$$

It follows from Theorem 3 that

$$\sup_{\mathbf{x}\in\mathbb{R}^d} J_n(\mathbf{x}) \leq C\,(d)\,\frac{\sqrt{n}\rho_3^{(d)}}{\left(\rho_2^{(d)}\right)^{3/2}} = \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right).$$

Further writing $\mathbf{x}_n := \left(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(d)}\right)$ we get

$$K_n(x) \leq \sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \left| \int_{x^{(i)}}^{x_n^{(i)}} e^{\frac{-t^2}{2}}\, dt \right|.$$

Note that $\Sigma_n = h_n\Sigma - \left(\sum_{j=1}^{n} \frac{1}{(j+1)^2}\right) M$, so $h_n^{-1}\Sigma_n \longrightarrow \Sigma$. The rest of the argument is exactly similar to that of the one dimensional case. This completes the proof. □

3. Large deviations for the expected configuration

In this section we discuss the asymptotic behavior of the tail probabilities of $\frac{Z_n}{\log n}$. Following standard notations are used in rest of the paper. For any subset $A \subseteq \mathbb{R}^d$ we write A° to denote the *interior* of A and $\bar{A}$ to denote the *closer* of A under the usual Euclidean metric.

Theorem 5. *The sequence of measures* $\mathbf{P}\left(\frac{Z_n}{\log n} \in \cdot\right)_{n\geq 2}$ *satisfy a LDP with rate function* $I(\cdot)$ *and speed* $\log n$*, that is,*

$$-\inf_{\mathbf{x}\in A^\circ} I(\mathbf{x}) \leq \varliminf_{n\to\infty} \frac{\log \mathbf{P}\left(\frac{Z_n}{\log n} \in A\right)}{\log n} \leq \varlimsup_{n\to\infty} \frac{\log \mathbf{P}\left(\frac{Z_n}{\log n} \in A\right)}{\log n} \leq -\inf_{\mathbf{x}\in\bar{A}} I(\mathbf{x}) \tag{24}$$

where $I(\cdot)$ *is the Fenchel–Legendre dual of* $e(\cdot) - 1$*, that is for* $x \in \mathbb{R}^d$*,*

$$I(x) = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^d} \{\langle \mathbf{x}, \boldsymbol{\lambda}\rangle - e(\boldsymbol{\lambda}) + 1\}. \tag{25}$$

Moreover $I(\cdot)$ *is convex and a good rate function.*

Proof. Using representation (5) without loss we may assume that $Z_0 = \mathbf{0}$ with probability one, that is, $U_0 = \delta_{\mathbf{0}}$. Now we define

$$\Lambda_n(\boldsymbol{\lambda}) := \frac{1}{\log n} \log \mathbb{E}\left[e^{\langle \boldsymbol{\lambda}, Z_n\rangle}\right]. \tag{26}$$

From (5) it follows that

$$\mathbb{E}\left[e^{\langle \boldsymbol{\lambda}, Z_n\rangle}\right] = \frac{1}{n+1} \Pi_n\left(e\left(\boldsymbol{\lambda}\right)\right)$$

where $\Pi_n(z) = \prod_{j=1}^{n}\left(1 + \frac{z}{j}\right)$, $z \in \mathbb{C}$. Using Gauss's formula (see page 178 of Conway (1978)) we have

$$\lim_{n\to\infty} \frac{\Pi_n(z)}{n^z} \Gamma(z+1) = 1 \tag{27}$$

and the convergence happens uniformly on compact subsets of $\mathbb{C} \setminus \{-1, -2, \ldots\}$. Therefore we get

$$\Lambda_n(\boldsymbol{\lambda}) \longrightarrow e(\boldsymbol{\lambda}) - 1 < \infty \quad \forall\, \boldsymbol{\lambda} \in \mathbb{R}^d. \tag{28}$$

Thus the LDP as stated in (24) follows from the Gärtner–Ellis Theorem (see Remark (a) on page 45 of Dembo and Zeitouni (1993) or page 66 of Chakrabarty (2010)).

We next note that $I(\cdot)$ is a convex function because it is the Fenchel–Legendre dual of $e(\boldsymbol{\lambda}) - 1$ which is finite for all $\boldsymbol{\lambda} \in \mathbb{R}^d$.

Finally, we will show that $I(\cdot)$ is good rate function, that is, the level sets $A(\alpha) = \{\mathbf{x}: I(\mathbf{x}) \leq \alpha\}$ are compact for all $\alpha > 0$. Since I is a rate function so by definition it is lower semicontinuous. So it is enough to prove that $A(\alpha)$ is bounded for all $\alpha \in \mathbb{R}$. Observe that for all $\mathbf{x} \in \mathbb{R}^d$,

$$I(\mathbf{x}) \geq \sup_{\|\boldsymbol{\lambda}\|=1} \{\langle \mathbf{x}, \boldsymbol{\lambda}\rangle - e(\boldsymbol{\lambda}) + 1\}.$$

Now the function $\boldsymbol{\lambda} \mapsto e(\boldsymbol{\lambda})$ is continuous and $\{\boldsymbol{\lambda}: \|\boldsymbol{\lambda}\| = 1\}$ is a compact set. So $\exists\, \boldsymbol{\lambda}_0 \in \{\boldsymbol{\lambda}: \|\boldsymbol{\lambda}\| = 1\}$ such that $\sup_{|\boldsymbol{\lambda}|=1} e(\boldsymbol{\lambda}) = e(\boldsymbol{\lambda}_0)$. Therefore for $\|\mathbf{x}\| \neq 0$ choosing $\boldsymbol{\lambda} = \frac{\mathbf{x}}{\|\mathbf{x}\|}$, we have $I(\mathbf{x}) \geq \|\mathbf{x}\| - e(\boldsymbol{\lambda}_0) + 1$. So if $\mathbf{x} \in A(\alpha)$ then

$$\|\mathbf{x}\| \leq (\alpha + e(\boldsymbol{\lambda}_0) - 1).$$

This proves that the level sets are bounded, which completes the proof. □

Our next result is an easy consequence of (25) which can be used to compute the explicit formula for the rate function I.

Theorem 6. *The rate function I is same as the rate function for the large deviation of the empirical means of i.i.d. random vectors with distribution corresponding to the distribution of the following random vector:*

$$W = \sum_{i=1}^{N} X_i, \tag{29}$$

where $N \sim$ Poisson (1) and is independent of $(X_j)_{j\geq 1}$ which are the i.i.d. increments of the associated random walk.

Proof. We first observe that $\log \mathbf{E}\left[e^{\langle \boldsymbol{\lambda}, W\rangle}\right] = e(\boldsymbol{\lambda}) - 1$. The rest then follows from (25) and Cramér's Theorem (see Theorem 2.2.30 of Dembo and Zeitouni (1993)). □

For $d = 1$, one can get more information about the rate function I, in particular the following result is a consequence of Theorem 6 and Lemma 2.2.5 of Dembo and Zeitouni (1993).

Proposition 2. *Suppose $d = 1$ then $I(x)$ is non-decreasing when $x \geq \mu$ and non-increasing when $x \leq \mu$. Moreover,*

$$I(x) = \begin{cases} \sup\limits_{\lambda\geq 0}\{x\lambda - e(\lambda) + 1\} & \text{if } x \geq \mu \\ \sup\limits_{\lambda\leq 0}\{x\lambda - e(\lambda) + 1\} & \text{if } x \leq \mu. \end{cases} \tag{30}$$

In particular, $I(\mu) = \inf_{x\in\mathbb{R}} I(x)$.

Following is an immediate corollary of the above result and Theorem 5.

Corollary 7. *Let $d = 1$ then for any $\epsilon > 0$*

$$\lim_{n\to\infty} \frac{1}{\log n} \log \mathbf{P}\left(\frac{Z_n}{\log n} \geq \mu + \epsilon\right) = -I(\mu + \epsilon) \tag{31}$$

and

$$\lim_{n\to\infty} \frac{1}{\log n} \log \mathbf{P}\left(\frac{Z_n}{\log n} \leq \mu - \epsilon\right) = -I(\mu - \epsilon). \tag{32}$$

We end the section with explicit computations of the rate functions for two examples of infinite color urn models associated with random walks on a one dimensional integer lattice.

Example 3.1. Our first example is the case when the random walk is trivial, which moves deterministically one step at a time. In other words $X_1 = 1$ with probability one. In this case $\mu = 1$ and $\sigma^2 = 1$. Also the moment generating function of X_1 is given by $e(\lambda) := e^{\lambda}$, $\lambda \in \mathbb{R}$. By Theorem 6 the rate function for the associated infinite color urn model is same as the rate function for a Poisson random variable with mean 1, that is, $I(x) = x\log x - x + 1$, if $x > 0$, $I(0) = 1$ and $I(x) = \infty$ when $x < 0$.

Example 3.2. Our next example is the case when the random walk is the *simple symmetric random walk* on the one dimensional integer lattice. For this case we note that $\mu = 0$, $\sigma^2 = 1$ and the moment generating function X_1 is $e(\lambda) = \cosh \lambda$, $\lambda \in \mathbb{R}$. The rate function for the associated infinite color urn model turns out to be

$$I(x) = x \sinh^{-1} x - \sqrt{1 + x^2} + 1. \tag{33}$$

References

Bandyopadhyay, A., Thacker, D., 2013 Pólya urn schemes with infinitely many colors. http://arxiv.org/pdf/1303.7374v2.pdf.

Bergström, H., 1949. On the central limit theorem in the case of not equally distributed random variables. Skand. Aktuarietidskr. 32, 37–62.

Bhattacharya, R.N., Ranga Rao, R., 1976. Normal Approximation and Asymptotic Expansions. John Wiley & Sons, New York, London, Sydney, (Wiley Series in Probability and Mathematical Statistics).

Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Pólya urn schemes. Ann. Statist. 1, 353–355.
Chakrabarty, A., 2010. When is a truncated heavy tail heavy? (Ph.D. thesis). Cornell University.
Conway, J.B., 1978. Functions of One Complex Variable, second ed. In: Graduate Texts in Mathematics, vol. 11. Springer-Verlag, New York.
Dembo, A., Zeitouni, O., 1993. Large Deviations Techniques and Applications. Jones and Bartlett Publishers, Boston, MA.
Pemantle, R., 2007. A survey of random processes with reinforcement. Probab. Surv. 4, 1–79.
Pólya, G., 1930. Sur quelques points de la théorie des probabilités. Ann. Inst. H. Poincaré 1 (2), 117–161.