



# An identification problem in an urn and ball model with heavy tailed distributions

Christine Fricker, Fabrice Guillemin, Philippe Robert

## ► To cite this version:

Christine Fricker, Fabrice Guillemin, Philippe Robert. An identification problem in an urn and ball model with heavy tailed distributions. 2009. inria-00347012v2

**HAL Id: inria-00347012**

**<https://inria.hal.science/inria-00347012v2>**

Preprint submitted on 20 Jun 2009

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# AN IDENTIFICATION PROBLEM IN AN URN AND BALL MODEL WITH HEAVY TAILED DISTRIBUTIONS

CHRISTINE FRICKER, FABRICE GUILLEMIN, AND PHILIPPE ROBERT

**ABSTRACT.** We consider in this paper an urn and ball problem with replacement, where balls are with different colors and are drawn uniformly from a unique urn. The numbers of balls with a given color are i.i.d. random variables with a heavy tailed probability distribution, for instance a Pareto or a Weibull distribution. We draw a small fraction  $p \ll 1$  of the total number of balls. The basic problem addressed in this paper is to know to which extent we can infer the total number of colors and the distribution of the number of balls with a given color. By means of Le Cam's inequality and Chen-Stein method, bounds for the total variation norm between the distribution of the number of balls drawn with a given color and the Poisson distribution with the same mean are obtained. We then show that the distribution of the number of balls drawn with a given color has the same tail as that of the original number of balls. We finally establish explicit bounds between the two distributions when each ball is drawn with fixed probability  $p$ .

## 1. INTRODUCTION

We consider in this paper the following urn and ball scheme *with replacement*: An urn contains a random number of balls with different colors. We draw a small fraction  $p \ll 1$  of the total number of balls. A ball which has been drawn is replaced into the urn. The problem considered in this paper consists of estimating the number of colors together with the distribution of the number of balls with a given color by using information from sampled balls. This problem is motivated by the analysis of packet sampling in the Internet (see Chabchoub *et al.* [5] for details).

To address the above problem, we analyze the non-normalized distribution of the number of balls drawn with a given color. More specifically, let  $W_j$  (respectively,  $W_j^+$ ) denote the number of colors with a number of sampled balls equal to (respectively, equal to or greater than)  $j$ . Denoting by  $\tilde{K}$  the number of colors seen when drawing balls, the quantities  $W_j/\tilde{K}$  and  $W_j^+/\tilde{K}$  are equal to the proportions of colors, which at the end of the trial comprise exactly or at least  $j$  balls, respectively.

The numbers of balls with various colors are assumed to be i.i.d. random variables and the number  $K$  of colors is large. In addition, the distribution of the number of balls with a given color has a heavy tailed probability distribution of Pareto or Weibull type. Finally, balls are drawn uniformly. This means that for each  $i = 1, \dots, K$ , if there are  $v_i$  balls with color  $i$ , the probability of drawing a ball with this color is  $v_i/V$ , where  $V = v_1 + \dots + v_K$  is the total number of balls in the urn.

---

*Key words and phrases.* Chen-Stein method, Pareto distribution, Weibull distribution.

The above model is defined as the “uniform model”. It will be compared against the case when balls are drawn independently one of each other with probability  $p$ . This model will be referred to as probabilistic model. We show that the results obtained in both models are close one to each other when  $p$  is very small. But there are some subtle differences between the two models, notably with regard to the achievable accuracy in the inference of original statistics. It turns out that the probabilistic model is simpler to analyze than the uniform model but yields less accurate results. This is due to the fact that we cannot exploit the fact that the number of colors is very large.

One of the main results of the paper concerns the analysis of the validity of the following simple scaling rule: The distribution of the original number  $v_i$  of balls with color  $i$  could be estimated by that of the random variable  $\tilde{v}_i/p$ , where  $\tilde{v}_i$  is the number of sampled balls with color  $i$ . When each ball is drawn with a fixed probability, it is known that this rule is valid for tails of the distributions as soon as they are heavy tailed. See Asmussen *et al* [3] and Foss and Korshunov [7] where this asymptotic equivalence is proved in a quite general framework. Our main goal here is to get, for  $j \geq 2$ , an *explicit bound* on the quantity

$$\left| \frac{\mathbb{P}(\tilde{v} \geq j)}{\mathbb{P}(v \geq j/p)} - 1 \right|.$$

In the context of packet sampling in the Internet, explicit expressions are especially important for the estimation of the sizes of flows in Internet traffic. In this setting the variable  $j$  is taken to be large but cannot be too large so that the event  $\{\tilde{v} = j\}$  occurs sufficiently often to obtain reliable statistics. Henceforth, the dependence on  $j$  should be made explicit. See Chabchoub *et al.* [5] for a discussion.

The organization of this paper is as follows: The notation and the basic results used in this paper (Le Cam’s inequality and Chen-Stein method) are presented in Section 2. The mean values of the random variables  $W_j$  and  $W_j^+$  are computed in Section 3. The approximation of the distribution of  $W_j^+$  by a Poisson distribution and the validity of the scaling rule are investigated in Section 4. We compare in Section 5 the original distribution of the number of balls with a given color against the rescaled distribution of the number of drawn balls with the same color. Some concluding remarks with regard to sampling are presented in Section 6.

## 2. NOTATION AND BASIC RESULTS

**2.1. Definitions and assumptions.** We consider an urn containing  $v_i$  balls with color  $i$  for  $i = 1, \dots, K$ . The quantities  $v_i$  are independent random variables with a common heavy tailed distribution. In the following we shall consider two families of heavy tailed distributions for the number  $v$  of balls with a given color:

**Pareto distributions:** The distribution of  $v$  is given by

$$(1) \quad \mathbb{P}(v > x) = (b/x)^a, \quad x \geq b,$$

with the shape parameter  $a > 1$  and the location parameter  $b > 0$ . The mean of  $v$  is  $ab/(a-1)$ .

**Weibull distributions:** The distribution of  $v_i$  is given by

$$(2) \quad \mathbb{P}(v > x) = \exp(-(x/\eta)^\beta), \quad x \geq 0,$$

with the skew parameter  $\beta \in (0, 1)$  and the scale parameter  $\eta > 0$ . The mean of  $v$  is  $\frac{\eta}{\beta} \Gamma(1/\beta)$ , where  $\Gamma$  is the classical Euler’s Gamma function.

The total number of balls in the urn is  $V = \sum_{i=1}^K v_i$ . We draw only a fraction  $p$  of this total number of balls. Each ball is drawn at random: A ball with color  $i$  is drawn with probability  $v_i/V$ . After drawing the  $pV$  balls, we have  $\tilde{v}_i$  balls with color  $i$ . Of course, only those colors with  $\tilde{v}_i > 0$  can be seen. The quantity  $\tilde{K} = \sum_{i=1}^K \mathbb{1}_{\{\tilde{v}_i > 0\}}$  is the number of colors seen at the end of a trial.

In the following, we shall be interested in the asymptotic regime when the number of colors  $K \rightarrow \infty$  while the fraction  $p \rightarrow 0$ . Note that by the law of large numbers,  $V \rightarrow \infty$  a.s. (the total number of balls in the urn is very large).

The random variables we consider in this paper to infer the original statistics of the number of balls and colors are the variables  $W_j$  and  $W_j^+$ ,  $j \geq 0$ , defined as follows.

**Definition 1** (Definition of  $W_j$ ). *The random variable  $W_j$  is the number of colors with  $j$  balls at the end of a trial and is given by*

$$j \geq 0, \quad W_j = \mathbb{1}_{\{\tilde{v}_1=j\}} + \mathbb{1}_{\{\tilde{v}_2=j\}} + \cdots + \mathbb{1}_{\{\tilde{v}_K=j\}},$$

where  $\tilde{v}_i \geq 0$  is the number of balls drawn with color  $i$  (which can be equal to 0).

**Definition 2** (Definition of  $W_j^+$ ). *The random variable  $W_j^+$  is the number of colors with at least  $j$  balls at the end of a trial. The random variables  $W_j^+$  are formally defined by*

$$j \geq 0, \quad W_j^+ = \mathbb{1}_{\{\tilde{v}_1 \geq j\}} + \mathbb{1}_{\{\tilde{v}_2 \geq j\}} + \cdots + \mathbb{1}_{\{\tilde{v}_K \geq j\}}.$$

Note that we have

$$\forall j \geq 0, \quad W_j^+ = \sum_{\ell \geq j} W_\ell.$$

The averages of the random variables  $W_j$  are in fact the key quantities we shall use in the following to infer the original numbers of balls per color.

**2.2. Le Cam's inequality and Chen-Stein method.** Le Cam's inequality gives the distance in total variation between the distribution of a sum of independent and identically distributed (i.i.d.) Bernoulli random variables and the Poisson distribution with the same mean (see Barbour *et al.* [4]). Note that if  $V$  and  $W$  are two random variables taking integer values, the distance in total variation between their distributions is defined by

$$\begin{aligned} \|\mathbb{P}(W \in \cdot) - \mathbb{P}(V \in \cdot)\|_{tv} &\stackrel{\text{def.}}{=} \sup_{A \subset \mathbb{N}} |\mathbb{P}(W \in A) - \mathbb{P}(V \in A)| \\ &= \frac{1}{2} \sum_{n \geq 0} |\mathbb{P}(W = n) - \mathbb{P}(V = n)|. \end{aligned}$$

**Theorem 1** (Le Cam's Inequality). *If the random variable  $W = \sum_i I_i$ , where the random variables  $I_i$  are i.i.d. Bernoulli random variables, then*

$$(3) \quad \|\mathbb{P}(W \in \cdot) - \mathbb{P}(Q_{\mathbb{E}(W)} \in \cdot)\|_{tv} \leq \sum_i \mathbb{P}(I_i = 1)^2,$$

where for  $\lambda > 0$ ,  $Q_\lambda$  is a Poisson random variable with mean  $\lambda$ , that is, for all  $n \geq 0$ ,

$$\mathbb{P}(Q_\lambda = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

When the random variables  $I_i$  appearing in the above theorem are not independent but satisfy a specific condition, referred to as monotonic coupling, it is still possible to obtain a bound on the distance between the distribution of the sum  $W = \sum_i I_i$  and the Poisson distribution with mean  $\mathbb{E}(W)$ .

**Definition 3** (Monotonic Coupling). *The variables  $I_i$  are said to be negatively related, when there exist some random variables  $U_i$  and  $V_i$  such that*

- (1)  $U_i \stackrel{\text{dist.}}{=} W$  and  $1 + V_i \stackrel{\text{dist.}}{=} (W \mid I_i = 1)$ ;
- (2)  $V_i \leq U_i$ .

The main result of the Chen-Stein method is given by the following theorem (see Barbour *et al.* [4]).

**Theorem 2.** *If the monotonic coupling condition is satisfied, then the following inequality holds*

$$(4) \quad \|\mathbb{P}(W \in \cdot) - \mathbb{P}(Q_{\mathbb{E}(W)} \in \cdot)\|_{tv} \leq 1 - \frac{\text{Var}(W)}{\mathbb{E}(W)}.$$

When the monotonic coupling condition is satisfied, in order to prove the Poisson approximation, it is sufficient to show that the ratio of the variance to the mean value of  $W$  is close to 1; this is a very weak condition to prove in practice.

It should be noted (see [8]) that Relation (4) can be used not only when  $\mathbb{E}(W)$  takes bounded values so that  $W$  is approximately a Poisson random variable, but also when  $\mathbb{E}(W)$  is large. In this case Chen-Stein Method yields a central limit theorem: If  $\mathcal{N}$  is a standard normal distribution,

$$\begin{aligned} \left\| \mathbb{P} \left( \frac{W - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) - \mathbb{P}(\mathcal{N} \in \cdot) \right\|_{tv} &\leq \\ &\left\| \mathbb{P} \left( \frac{W - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) - \mathbb{P} \left( \frac{Q_{\mathbb{E}(W)} - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) \right\|_{tv} \\ &\quad + \left\| \mathbb{P} \left( \frac{Q_{\mathbb{E}(W)} - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) - \mathbb{P}(\mathcal{N} \in \cdot) \right\|_{tv} \end{aligned}$$

where  $\text{Var}(W)$  is the variance of the random variable  $W$ .

By using Relation (4), we have

$$\begin{aligned} \left\| \mathbb{P} \left( \frac{W - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) - \mathbb{P}(\mathcal{N} \in \cdot) \right\|_{tv} &\leq 1 - \frac{\text{Var}(W)}{\mathbb{E}(W)} \\ &\quad + \left\| \mathbb{P} \left( \frac{Q_{\mathbb{E}(W)} - \mathbb{E}(W)}{\sqrt{\text{Var}(W)}} \in \cdot \right) - \mathbb{P}(\mathcal{N} \in \cdot) \right\|_{tv}. \end{aligned}$$

If the ratio  $\mathbb{E}(W)/\text{Var}(W)$  is close to 1, then the first term in the right hand side of the above relation is negligible. In addition, the classical central limit theorem for Poisson distributions implies that when  $\mathbb{E}(W)$  is large, the second term is negligible too. Therefore, we have  $W \sim \mathbb{E}(W) + \sqrt{\text{Var}(W)}\mathcal{N}$  with a bound on the error.

## 3. COMPUTATION OF MEAN VALUES

**3.1. Bounds for mean values.** By using Le Cam's inequality, we can establish the following result for the mean value of the random variables  $W_j$ .

**Proposition 1** (Mean Value of  $W_j$ ). *If there are  $V$  balls and  $K$  colors in the urn, for  $j \geq 0$ , the mean number  $\mathbb{E}(W_j)$  of colors with  $j$  balls at the end of a trial satisfies the relation*

$$(5) \quad \left| \frac{\mathbb{E}(W_j)}{K} - \mathbb{Q}_j \right| \leq \mathbb{E} \left( \min(pv, 1) \frac{v}{V} \right),$$

where  $\mathbb{Q}$  is the probability distribution defined for  $j \geq 0$  by

$$\mathbb{Q}_j = \mathbb{E} \left( \frac{(pv)^j}{j!} e^{-pv} \right),$$

$p$  is the sampling rate, and  $v$  is distributed as the number of balls with a given color.

*Proof.* We have

$$\tilde{v}_i = B_1^i + B_2^i + \cdots + B_{pV}^i,$$

where  $B_\ell^i$  is equal to one if the  $\ell$ th ball drawn from the urn has color  $i$ , which event occurs with probability  $v_i/V$ , the quantity  $V$  being the total number of balls in the urn.

Conditionally on the values of the set  $\mathcal{F} = \{v_1, \dots, v_K\}$ , the variables  $(B_\ell^i, \ell \geq 1)$  are independent Bernoulli variables. For  $1 \leq i \leq K$ , Le Cam's Inequality (3) therefore gives the relation

$$\|\mathbb{P}(\tilde{v}_i \in \cdot \mid \mathcal{F}) - \mathbb{P}(Q_{pv_i} \in \cdot)\|_{tv} \leq p \frac{v_i^2}{V},$$

and Relation (4) which can also be used in this case yields

$$\|\mathbb{P}(\tilde{v}_i \in \cdot \mid \mathcal{F}) - \mathbb{P}(Q_{pv_i} \in \cdot)\|_{tv} \leq \frac{v_i}{V},$$

By integrating with respect to the variables  $v_1, \dots, v_K$ , these two inequalities give the relation

$$(6) \quad \|\mathbb{P}(\tilde{v}_i \in \cdot) - \mathbb{Q}\|_{tv} \leq \mathbb{E} \left( \min(pv, 1) \frac{v}{V} \right).$$

Since  $\mathbb{E}(W_j) = \sum_{i=1}^K \mathbb{P}(\tilde{v}_i = j)$ , by summing on  $i = 1, \dots, K$ , we obtain

$$|\mathbb{E}(W_j) - K\mathbb{Q}_j| \leq K\mathbb{E} \left( \min(pv, 1) \frac{v}{V} \right).$$

and the result follows.  $\square$

By using the fact that  $\mathbb{E}(W_j^+) = \sum_{i=1}^K \mathbb{P}(\tilde{v}_i \geq j)$ , we can deduce from Equation (6) the following result.

**Proposition 2** (Mean Value of  $W_j^+$ ). *If there are  $V$  balls and  $K$  colors in the urn, the mean number  $\mathbb{E}(W_j^+)$  of colors with at least  $j \geq 0$  balls at the end of an arbitrary trial satisfies the relation*

$$(7) \quad \left| \frac{\mathbb{E}(W_j^+)}{K} - \sum_{\ell \geq j} \mathbb{Q}_\ell \right| \leq \mathbb{E} \left( \min(pv, 1) \frac{v}{V} \right),$$

where the probability distribution  $\mathbb{Q}$  is defined in Proposition 1.

We immediately deduce from Propositions 1 and 2 the following corollary by using the fact that  $V \geq K$ .

**Corollary 1** (Asymptotic Mean Values). *The relations*

$$\lim_{K \rightarrow \infty} \frac{1}{K} \mathbb{E}(W_j) = \mathbb{Q}_j \quad \text{and} \quad \lim_{K \rightarrow \infty} \frac{1}{K} \mathbb{E}(W_j^+) = \sum_{\ell \geq j} \mathbb{Q}_\ell.$$

*hold.*

Note that if balls are drawn with probability  $p$  independently one of each other (probabilistic model), we have  $\tilde{v}_i = \sum_{\ell=1}^{v_i} \tilde{B}_\ell^i$ , where the random variables  $\tilde{B}_\ell^i$  are Bernoulli with mean  $p$ . By adapting the above proofs, we find

$$(8) \quad \left| \frac{\mathbb{E}(W_j)}{K} - \mathbb{Q}_j \right| \leq p.$$

### 3.2. Asymptotic results for specific probability distributions.

**3.2.1. Pareto distributions.** Let us first assume that the number of balls of a given color follows a Pareto distribution given by Equation (1). Then, we have the following result when the number of colors goes to infinity.

**Proposition 3.** *If  $v$  has a Pareto distribution as in Equation (1), then for all  $j > a$ , the relations*

$$(9) \quad \lim_{K \rightarrow +\infty} \frac{\mathbb{E}(W_{j+1})}{\mathbb{E}(W_j)} = 1 - \frac{a+1}{j+1} + O((pb)^{j-a}),$$

$$(10) \quad \lim_{K \rightarrow +\infty} \frac{\mathbb{E}(W_j)}{K} = a(pb)^a \frac{\Gamma(j-a)}{j!} + O((pb)^j),$$

$$(11) \quad \lim_{K \rightarrow +\infty} \frac{\mathbb{E}(W_j^+)}{K} = (pb)^a \frac{\Gamma(j-a)}{(j-1)!} + O\left(\frac{(pb)^j}{1-pb}\right)$$

*hold.*

*Proof.* For  $j > a$ ,

$$(12) \quad \begin{aligned} \mathbb{Q}_j = \mathbb{E} \left( \frac{(pv)^j}{j!} e^{-pv} \right) &= ab^a \frac{p^a}{j!} \int_{pb}^{+\infty} u^{j-a-1} e^{-u} du \\ &= a(pb)^a \frac{\Gamma(j-a)}{j!} - a \frac{(pb)^j}{j!} \int_0^1 u^{j-a-1} e^{-pbu} du. \end{aligned}$$

Therefore, by using the relation  $\Gamma(x+1) = x\Gamma(x)$ , we get the equivalence

$$\frac{\mathbb{Q}_{j+1}}{\mathbb{Q}_j} = \frac{j-a}{j+1} + O((pb)^{j-a}),$$

which gives Equations (9) and (10) by using Corollary 1. For the mean value of  $W_j^+$ , Equation (12) gives the relation

$$\begin{aligned} \lim_{K \rightarrow +\infty} \frac{\mathbb{E}(W_j^+)}{K} &= a(pb)^a \sum_{n \geq j} \frac{\Gamma(n-a)}{n!} + O\left(\frac{(pb)^j}{1-pb}\right) \\ &= a(pb)^a \sum_{n \geq 0} \frac{\Gamma(n+j-a)\Gamma(n+1)}{\Gamma(j+n+1)} \frac{1^n}{n!} + O\left(\frac{(pb)^j}{1-pb}\right) \\ &= a(pb)^a \frac{\Gamma(j-a)}{j!} F(j-a, 1; j+1; 1) + O\left(\frac{(pb)^j}{1-pb}\right), \end{aligned}$$

where  $F(a, b; c; z)$  is the hypergeometric function satisfying

$$F(a, b; c; 1) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)}$$

(see Abramowitz and Stegun [1]), and Equation (11) follows.  $\square$

The shape parameter  $a$  can be estimated via Relation (11) by

$$(13) \quad a = \lim_{K \rightarrow \infty} j \left( 1 - \frac{\mathbb{E}(W_{j+1}^+)}{\mathbb{E}(W_j^+)} \right) + O\left(\frac{(pb)^j}{1-pb}\right)$$

for all  $j > a$ . This gives a means of estimating the shape parameter  $a$ . When observing drawn balls, we have in fact only access to the quantity  $\mathbb{E}(\tilde{K})$  of the number of sampled colors. While this has no impact for the estimation of  $a$ , this correcting term is important when estimating  $b$  from Equation (11). It is straightforward that

$$\tilde{K} = \sum_{i=1}^K \mathbb{1}_{\{\tilde{v}_i > 0\}} = K - W_0$$

and then when  $K \rightarrow \infty$

$$\mathbb{E}(\tilde{K}) \sim K(1 - \mathbb{Q}_0) = K(1 - \mathbb{E}(e^{-pv})).$$

Since

$$(14) \quad 1 - \mathbb{E}(e^{-pv}) = p \int_0^\infty e^{-px} \mathbb{P}(v > x) dx = bp + (bp)^a \Gamma(1-a, bp),$$

where  $\Gamma(a, x)$  is the incomplete Gamma function defined by  $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$ , we can use the above equations together with Equation (11) in order to estimate  $b$  and then  $K$ . It is also worth noting that  $1 - \mathbb{E}(e^{-pv}) \sim bp$  when  $a > 1$  and  $bp \rightarrow 0$ .

**3.2.2. Weibull distributions.** We assume in this section that the number of balls with a given color follows a Weibull distribution. In this case, we have the following result, which follows from a simple variable change and the expansion of  $\exp(-x^\beta)$  in power series of  $x^\beta$  or  $\exp(-px)$  in power series of  $x$ ; the proof is omitted.

**Proposition 4.** *If  $v$  has a Weibull distribution with skew parameter  $\beta$  and scale parameter  $\eta$ , then for  $0 < \beta < 1$*

$$(15) \quad \lim_{K \rightarrow +\infty} \mathbb{E}(W_{j+1}) = \frac{\beta}{j!} \sum_{n=0}^{\infty} \frac{(-1)^n}{(p\eta)^{(n+1)\beta}} \frac{\Gamma((n+1)\beta + j)}{n!}$$



and for  $\beta > 1$ ,

$$(16) \quad \lim_{K \rightarrow +\infty} \mathbb{E}(W_{j+1}) = \frac{(p\eta)^j}{j!} \sum_{n=0}^{\infty} \frac{(-p\eta)^n}{n!} \Gamma\left(\frac{(n+j)}{\beta} + 1\right).$$

Note that  $\mathbb{E}(W_j)$  can be written in the form

$$\mathbb{E}(W_j) = \frac{1}{j!} \frac{\beta}{(p\eta)^\beta} \int_0^\infty u^{j+\beta-1} e^{-u+tu^\beta} du$$

with  $t = -1/(p\eta)^\beta$ . The above integral is known in the literature as to be of the Faxen's type and can be expressed by means of Meijer  $G$ -function, when  $\beta$  is a rational number, see Abramowitz and Stegun [1].

Contrary to the case of Pareto distribution for the initial distribution of balls of a given color, there is no simple relations giving the parameters  $\beta$  and  $\eta$  from the mean values  $\mathbb{E}(W_j)$ ,  $j \geq 1$ . In fact, we shall prove in the following that  $\mathbb{P}(\tilde{v} \geq j)$  has also a Weibull tail. This eventually gives a means of identifying the parameters.

#### 4. POISSON APPROXIMATIONS

In the previous section, we have established bounds for the mean values of the random variables  $W_j$  and  $W_j^+$ . To obtain more information on their distributions, we intend to use Chen-Stein method. For a fixed environment (namely fixed values of the quantities  $v_i$  for  $i = 1, \dots, K$ ), these random variables appear as sums of non independent Bernoulli random variables. A preliminary analysis of the Bernoulli random variables appearing in the expression of  $W_j$  reveals that it seems not possible to invoke a monotonic coupling argument. It is well known (see [4] for details) that the situation is more favorable with the random variables  $W_j^+$  and we can specifically prove that if  $\mathcal{F}$  is the set  $\mathcal{F} = \{v_i, 1 \leq i \leq K\}$ , then the total number  $W_j^+$  of colors with at least  $j$  balls at the end of the trial satisfies the relation

$$(17) \quad \left\| \mathbb{P}(W_j^+ \in \cdot \mid \mathcal{F}) - \mathbb{P}(Q_{\mathbb{E}(W_j^+ \mid \mathcal{F})} \in \cdot) \right\|_{tv} \leq \mathbb{E} \left( 1 - \frac{\text{Var}(W_j^+ \mid \mathcal{F})}{\mathbb{E}(W_j^+ \mid \mathcal{F})} \right).$$

Indeed, given the random variables  $v_i$ , the model is equivalent to a standard urn and ball problem consisting of putting  $pV_i$  balls into  $K$  urns, a ball falling into urn  $i$  with probability  $p_i = v_i/V_i$ . The number of balls in urn  $i$  is the number of balls with color  $i$  in the original urn and ball problem. Even in the case when the quantities  $p_i$  are different, the variables  $I_{i,j}^+ \stackrel{\text{def}}{=} \mathbb{1}_{\{\tilde{v}_i \geq j\}}$  are negatively related so that Theorem 2 can be used. See Page 24 and Corolary 2.C.2 Page 26 of [4] for a definition and the main inequality in this domain. Chapter 6 of this reference is entirely devoted to related occupancy problems.

The rest of this section is devoted to the estimation of the bound in Equation (17). We first establish the following lemma.

**Lemma 1.** *For a fixed environment  $\mathcal{F} = \{v_i, 1 \leq i \leq K\}$ , the distance in total variation between the distribution of  $W_j^+$  and the Poisson distribution  $Q_{\mathbb{E}(W_k^+ \mid \mathcal{F})}$  satisfies the inequality*

$$(18) \quad \lim_{K \rightarrow +\infty} \left\| \mathbb{P}(W_j^+ \in \cdot \mid \mathcal{F}) - \mathbb{P}(Q_{\mathbb{E}(W_k^+ \mid \mathcal{F})} \in \cdot) \right\|_{tv} \leq \frac{m_{2,j}(p)}{m_j(p)} + \frac{p}{\mathbb{E}(v)} \frac{m'_j(p)^2}{m_j(p)},$$

where  $m_j(p)$  and  $m_{2,j}(p)$  are the first two moments of the random variable defined by

$$(19) \quad X_j(p) = \sum_{\ell \geq j} \frac{(pv)^\ell}{\ell!} e^{-pv},$$

and the prime sign denotes the derivative with respect to  $p$ .

*Proof.* For  $\mathcal{F}$  fixed, the number  $W_j$  of colors with  $j \leq pV$  balls at the end of the trial is such that

$$\mathbb{E}(W_j \mid \mathcal{F}) = \sum_{i=1}^K \binom{pV}{j} \left(\frac{v_i}{V}\right)^j \left(1 - \frac{v_i}{V}\right)^{pV-j}.$$

By using the fact that

$$\frac{1}{V} = \frac{1}{K\mathbb{E}(v)} + o\left(\frac{1}{K}\right) \quad \text{a.s.}$$

for large  $K$ , straightforward calculations show that

$$(20) \quad \begin{aligned} \mathbb{E}(W_j \mid \mathcal{F}) &= \sum_{i=1}^K \frac{(pv_i)^j}{j!} e^{-pv_i} \left(1 - \frac{j(j-1)}{2pK\mathbb{E}(v)} + \frac{2jv_i - pv_i^2}{2\mathbb{E}(v)K}\right) + o\left(\frac{1}{K}\right) \\ &= \sum_{i=1}^K \left( \frac{(pv_i)^j}{j!} e^{-pv_i} - \frac{p}{2\mathbb{E}(v)K} \frac{d^2}{dp^2} \left( e^{-pv_i} \frac{(pv_i)^j}{j!} \right) \right) + o\left(\frac{1}{K}\right). \end{aligned}$$

By summing up the terms above and by checking that the  $o\left(\frac{1}{K}\right)$  term remains valid, since the sum can be written as  $\sum_{i=1}^K f(v_i) e^{-pv_i} / K^2$ , where  $f$  is a polynomial, we have for  $j \geq 1$  and  $0 < p < 1$

$$\mathbb{E}(W_j^+ \mid \mathcal{F}) = \sum_{\ell \geq j} \mathbb{E}(W_\ell \mid \mathcal{F}) = \sum_{i=1}^K X_{i,j}(p) - \frac{p}{2\mathbb{E}(v)K} \sum_{i=1}^K X_{i,j}''(p) + o\left(\frac{1}{K}\right),$$

where

$$X_{i,j}(x) = \sum_{\ell \geq j} \frac{(xv_i)^\ell}{\ell!} e^{-xv_i}.$$

For the variance, if  $I_{i,j}$  is 1 if color  $i$  has exactly  $j$  balls at the end of the trial and 0 otherwise, then  $W_j = \sum_{i=1}^K I_{i,j}$  and, for  $j \neq \ell$ ,

$$\mathbb{E}(W_j W_\ell \mid \mathcal{F}) = \sum_{1 \leq i \neq m \leq K} \mathbb{E}(I_{i,j} I_{m,\ell} \mid \mathcal{F})$$

and

$$\mathbb{E}(W_j^2 \mid \mathcal{F}) = \mathbb{E}(W_j \mid \mathcal{F}) + \sum_{1 \leq i \neq m \leq K} \mathbb{E}(I_{i,j} I_{m,j} \mid \mathcal{F}).$$

For  $j, \ell$  such that  $j + \ell \leq pV$ ,

$$\mathbb{E}(I_{i,j} I_{m,\ell} \mid \mathcal{F}) = \frac{(pV)!}{j!\ell!(pV-j-\ell)!} \left(\frac{v_i}{V}\right)^j \left(\frac{v_m}{V}\right)^\ell \left(1 - \frac{v_i + v_m}{V}\right)^{pV-j-\ell}.$$

The quantity in the right hand side of the above equation can be expanded as

$$\frac{e^{-p(v_i+v_m)} p^{j+\ell} v_i^j v_m^\ell}{j!\ell!} - \frac{p}{2V} \frac{e^{-p(v_i+v_m)} v_i^j v_m^\ell}{j!\ell!} c_{i,m}(j, \ell) + o\left(\frac{1}{K}\right),$$

where

$$c_{i,m}(j, \ell) = p^{j+\ell-2}(j+\ell)(j+\ell-1) - 2(j+\ell)(v_i+v_m)p^{j+\ell-1} + (v_i+v_m)^2 p^{j+\ell}$$

is such that

$$\frac{e^{-p(v_i+v_m)} v_i^j v_m^\ell}{j! \ell!} c_{i,m}(j, \ell) = \frac{d^2}{dp^2} \frac{e^{-p(v_i+v_m)} v_i^j v_m^\ell}{j! \ell!}.$$

Since

$$(W_j^+)^2 = \left( \sum_{\ell \geq j} W_\ell \right)^2 = \sum_{\ell \neq k \geq j} W_k W_\ell + \sum_{\ell \geq j} W_\ell^2,$$

$$\begin{aligned} \mathbb{E}((W_j^+)^2 \mid \mathcal{F}) - \mathbb{E}(W_j^+ \mid \mathcal{F}) &= \sum_{1 \leq i \neq m \leq K} \sum_{\ell, k \geq j} \mathbb{E}(I_{i,k} I_{m,\ell} \mid \mathcal{F}) \\ &= \sum_{1 \leq i \neq m \leq K} \left( X_{i,j}(p) X_{m,j}(p) - \frac{p}{2\mathbb{E}(v)K} (X_{i,j} X_{m,j})''(p) \right) + o\left(\frac{1}{K}\right), \end{aligned}$$

and

$$1 - \frac{\text{Var}(W_j^+ \mid \mathcal{F})}{\mathbb{E}(W_j^+ \mid \mathcal{F})} = \frac{\mathbb{E}(W_j^+ \mid \mathcal{F}) - \mathbb{E}((W_j^+)^2 \mid \mathcal{F}) + \mathbb{E}(W_j^+ \mid \mathcal{F})^2}{\mathbb{E}(W_j^+ \mid \mathcal{F})}.$$

The right-hand side of this equation can be expanded as

$$\begin{aligned} &\frac{1}{\sum_{i=1}^K X_{i,j} + O(1)} \left( - \sum_{1 \leq i \neq m \leq K} X_{i,j}(p) X_{m,j}(p) \right. \\ &\quad \left. + \frac{p}{2\mathbb{E}(v)K} \sum_{1 \leq i \neq m \leq K} (X_{i,j} X_{m,j})''(p) + \left( \sum_{i=1}^K X_{i,j}(p) - \frac{p}{2\mathbb{E}(v)K} \sum_{i=1}^K X_{i,j}''(p) \right)^2 \right) \\ &\quad + o\left(\frac{1}{K}\right) \end{aligned}$$

which can be rewritten as

$$\begin{aligned} &\frac{1}{\sum_{i=1}^K X_{i,j} + O(1)} \left( \sum_{1 \leq i \leq K} X_{i,j}^2(p) \right. \\ &\quad \left. + \frac{p}{2\mathbb{E}(v)K} \left( \sum_{1 \leq i \neq m \leq K} (X_{i,j} X_{m,j})''(p) - 2 \sum_{i=1}^K X_{i,j}(p) \sum_{i=1}^K X_{i,j}''(p) \right) \right) + O(1) \end{aligned}$$

using that

$$\sum_{i \neq m} X_{i,j} X_{m,j} = \left( \sum_i X_{i,j} \right)^2 - \sum_i X_{i,j}^2.$$

By the law of large numbers, we have that, almost surely,

$$\begin{aligned} \lim_{K \rightarrow +\infty} \frac{1}{K} \sum_{i=1}^K X_{i,j}^2(p) &= \mathbb{E}(X_j^2(p)) = m_{2,j}(p), \\ \lim_{K \rightarrow +\infty} \frac{1}{K^2} \sum_{i \neq m} (X_{i,j} X_{m,j})''(p) &= (m_j^2)''(p), \end{aligned}$$

together with

$$\lim_{K \rightarrow +\infty} \frac{1}{K} \sum_{i=1}^K X_{i,j}(p) = m_j(p) \quad \text{and} \quad \lim_{K \rightarrow +\infty} \frac{1}{K} \sum_{i=1}^K X''_{i,j}(p) = m''_j(p).$$

Hence,

$$\begin{aligned} \lim_{K \rightarrow \infty} 1 - \frac{\text{Var}(W_j^+ | \mathcal{F})}{\mathbb{E}(W_j^+ | \mathcal{F})} &= \frac{m_{2,j}(p) + p[(m''_j(p)/2 - m_j(p)m''_j(p))/\mathbb{E}(v)]}{m_j(p)} \quad a.s. \\ &= \frac{m_{2,j}(p) + pm'_j(p)^2/\mathbb{E}(v)}{m_j(p)} \quad a.s. \end{aligned}$$

and the result follows.  $\square$

To illustrate the fact that the bound in Equation (18) is tight when  $p \rightarrow 0$  and  $v$  has finite moments of any order, let us note that, provided the corresponding moments are finite,

$$(21) \quad \lim_{p \rightarrow 0} \frac{m_j(p)}{p^j} = \frac{v^j}{j!}$$

Moreover,

$$\lim_{p \rightarrow 0} \frac{m_{2,j}(p)}{p^{2j}} = \frac{\mathbb{E}(v^{2j})}{j!^2} \quad \text{and} \quad \lim_{p \rightarrow 0} \frac{m'_j(p)}{p^{j-1}} = \frac{\mathbb{E}(v^j)}{(j-1)!}.$$

Thus, the limit when  $K$  tends to  $+\infty$  of the bound given by Equation (18) is equivalent to

$$\frac{jp^{j-1}}{(j-1)!} \frac{\mathbb{E}(v^j)}{\mathbb{E}(v)}$$

when  $p$  tends to 0. If  $j \geq 2$ , this term tends to 0 when  $p \rightarrow 0$ .

By using the above lemma, we are now able to state a limit result for the distribution of the random variables  $W_j^+$ .

**Proposition 5.** *The inequality*

$$(22) \quad \lim_{K \rightarrow +\infty} \sup_{y \in \mathbb{R}} \left| \mathbb{P} \left( \frac{W_j^+ - \mathbb{E}(W_j^+)}{\sqrt{\mathbb{E}(W_j^+)}} \leq y \right) - \int_{-\infty}^y \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \right| \leq \frac{m_{2,j}(p)}{m_j(p)} + \frac{p}{\mathbb{E}(v)} \frac{(m'_j(p))^2}{m_j(p)}$$

holds.

Thus, for  $j \geq 2$  and for small  $p$ , this gives the following approximation

$$W_j^+ \sim \mathbb{E}(W_j^+) + \sqrt{\mathbb{E}(W_j^+)},$$

where  $G$  is a standard normal random variable. It should be noted nevertheless that Equation (22) is almost a central limit result but because of the scaling in  $1/\sqrt{\mathbb{E}(W_j^+)}$  instead of  $1/\sqrt{\text{Var}(W_j^+)}$ , the bound in the right hand side is not 0 as  $K$  gets large but, according to the proof of Lemma 18, only an upper bound on the distance between  $\mathbb{E}(W_j^+)$  and  $\text{Var}(W_j^+)$ .

*Proof.* From Lemma 1, we have

$$\begin{aligned} \left\| \mathbb{P} \left( \frac{W_j^+ - \mathbb{E}(W_j^+)}{\sqrt{\mathbb{E}(W_j^+)}} \in \cdot \mid \mathcal{F} \right) - \mathbb{P} \left( \frac{Q_{\mathbb{E}(W_j^+ | \mathcal{F})} - \mathbb{E}(W_j^+ | \mathcal{F})}{\sqrt{\mathbb{E}(W_j^+ | \mathcal{F})}} \in \cdot \right) \right\|_{tv} \\ \leq \frac{m_{2,j}(p)}{m_j(p)} + \frac{p}{\mathbb{E}(v)} \frac{m'_j(p)^2}{m_j(p)}. \end{aligned}$$

From Equation (20), we have that

$$\lim_{K \rightarrow \infty} \frac{1}{K} \mathbb{E}(W_j^+ | \mathcal{F}) = \mathbb{E}(X_j(p)) = K \sum_{\ell \geq j} \mathbb{Q}_\ell = K m_j(p),$$

where the quantities  $\mathbb{Q}_\ell$  are defined in Proposition 1. In addition, from Corollary 1,  $\mathbb{E}(W_j^+) \sim K m_j(p)$  when  $K \rightarrow +\infty$ . The result then follows by applying the central limit theorem for Poisson distributions and by deconditioning with respect to  $\mathcal{F}$ .  $\square$

To conclude this section, let us notice that when balls are drawn with probability  $p$  independently of each other, we do not have to condition on the environment and we have

$$\left\| \mathbb{P}(W_j^+ \in \cdot) - \mathbb{P}(Q_{\mathbb{E}(W_j^+)} \in \cdot) \right\|_{tv} \leq \frac{\mathbb{E} \left( \sum_{k=j}^v \binom{v}{k} p^k (1-p)^{v-k} \mathbb{1}_{\{v \geq j\}} \right)^2}{\mathbb{E} \left( \binom{v}{j} p^j (1-p)^{v-j} \mathbb{1}_{\{v \geq j\}} \right)},$$

It is worth noting that the results are independent of the number of colors and that we do not need take  $K \rightarrow \infty$  to obtain a bound for the distance in total variation. In addition, when  $\mathbb{E}(W_j)$  become large, then it is possible to obtain a central limit-type approximation similar to Proposition 5.

## 5. COMPARISON WITH ORIGINAL DISTRIBUTIONS

**5.1. Uniform model.** In this section, we compare the distribution of the number  $\tilde{v}$  of balls drawn with a given color with that of the original number  $v$  of balls with a given color. We are in particular interested in giving a sense to the heuristic stating that  $v$  and  $\tilde{v}/p$  have distributions close to each other.

**Proposition 6.** *Under the condition that the random variable  $v$  has a Weibull or Pareto distribution, we have*

$$\lim_{j \rightarrow \infty} \lim_{K \rightarrow \infty} \frac{\mathbb{E}(W_j^+)}{K \mathbb{P}(v \geq j/p)} = 1.$$

*Proof.* From Corollary 1, we know that  $\mathbb{E}(W_j)/K \rightarrow \mathbb{Q}_j$  when  $K \rightarrow \infty$ . Since

$$\mathbb{Q}_j = \mathbb{E} \left( \frac{(pv)^j}{j!} e^{-pv} \right) = \sum_{\ell=1}^{\infty} \frac{(p\ell)^j}{j!} e^{-p\ell} \mathbb{P}(v = \ell),$$

we can show that if  $v$  has a Weibull or Pareto distribution, then  $\mathbb{Q}_j \sim \mathbb{P}(v = j/p)/p$  when  $j \rightarrow \infty$ . Indeed, the above sum can be rewritten as

$$\frac{1}{j!} \sum_{\ell=1}^{\infty} e^{f_j(\ell)} \mathbb{P}(v = \ell),$$

where  $f_j(\ell) = -p\ell + j \log(p\ell)$ , which attains its maximum at point  $j/p$  with  $f_j''(j/p) = -p^2/j$ . If the random variable  $v$  is Weibull or Pareto and  $j/p$  is sufficiently large, then  $\mathbb{P}(v = \ell)/\mathbb{P}(v = j/p) - 1 \sim 0$  uniformly on  $j$  for  $\ell$  in the neighborhood of  $j/p$ . It follows that

$$\mathbb{Q}_j \sim \frac{1}{j!} \mathbb{P}(v = j/p) e^{f_j(j/p)} \sum_{\ell=-\infty}^{\infty} e^{-\ell^2 \frac{p^2}{2j}}.$$

For  $a > 0$  converging to 0,

$$\begin{aligned} \sum_{\ell=-\infty}^{\infty} e^{-a\ell^2} &= \sum_{\ell=-\infty}^{\infty} \int_0^{+\infty} \mathbb{1}_{\{u > a\ell^2\}} e^{-u} du \sim 2 \int_0^{+\infty} \sqrt{\frac{u}{a}} e^{-u} du \\ &= 2 \int_0^{+\infty} \frac{u^2}{\sqrt{a}} e^{-u^2/2} du = \sqrt{\frac{\pi}{a}} \end{aligned}$$

and by Stirling formula  $j! \sim \sqrt{2\pi} j^{j+\frac{1}{2}} e^{-j}$  for large  $j$ , so that  $\mathbb{Q}_j \sim \mathbb{P}(v = j/p)/p$ . It is then easy to deduce that  $\sum_{\ell \geq j} \mathbb{Q}_j \sim \mathbb{P}(v \geq j/p)$  for large  $j$ .  $\square$

The above Proposition implies that  $\mathbb{P}(\tilde{v} \geq j)$  is such that  $\mathbb{P}(\tilde{v} \geq j) \sim \mathbb{P}(v \geq j/p)$  when the number of colors is large. This means that the tail of the distribution of the random variable  $v$  can be obtained by rescaling that of the number  $\tilde{v}$  of sampled balls with a given color. When  $v$  has a Pareto distribution, Equation (13) can still be used for large  $j$  to estimate the shape parameter  $a$ . The estimation of the probability  $1 - \mathbb{E}(e^{-pv})$  of sampling a color and the scale parameter  $b$  can also be estimated from the tail by using the expression of that probability as a function of  $b$  and  $a$  as in Equation (14). The same method applies for Weibull distributions.

**5.2. Probabilistic model.** From now on, we consider the probabilistic model and we establish stronger results on the distance between  $\mathbb{P}(\tilde{v} \geq j)$  and  $\mathbb{P}(v \geq j/p)$ , where  $\tilde{v}$  is the number of balls with a given color at the end of a trial. For this sampling mode, it was not possible to prove a result similar to Corollary 1, but Berry-Essen's theorem [6] can be used to establish a stronger result for the comparison between  $\tilde{v}$  and  $v$ . In [5], it is specifically proved that if we define the function  $h_j(x) = x^2/4p^2 \left( \sqrt{1 + 4jp/x^2} - 1 \right)^2$  for  $x \in \mathbb{R}$  and  $j > 0$ , then

$$\left| \mathbb{P}(\tilde{v} \geq j) - \mathbb{P}\left(v \geq h_j\left(\sqrt{p(1-p)}\mathcal{G}\right) \vee k\right) \right| \leq c\mathbb{E}\left(\frac{1}{\sqrt{v}} \mathbb{1}_{\{v \geq j\}}\right),$$

where  $\mathcal{G}$  is a standard Gaussian random variable, for real numbers  $a \vee b = \max(a, b)$ , and  $c = 3(p^2 + (1-p)^2)/\sqrt{p(1-p)}$ . For small  $p$ , the constant  $c \sim 3/\sqrt{p}$ . The above bound is very loose for small  $p$  and becomes accurate only for very large values of  $j$ . This is why we go further in this paper by establishing a tighter bound for the ratio  $\mathbb{P}(\tilde{v} \geq j)/\mathbb{P}(v \geq j/p)$ .

Let  $(B_n)$  be some sequence of i.i.d. Bernoulli random variables with parameter  $p$  and  $v$  some independent r.v. on  $\mathbb{N}$ . Take some  $\alpha \in ]1/2, 1[$ . Let  $\tilde{v} = \sum_{l=1}^v B_l$ .

**Theorem 3.** For  $\alpha \in (1/2, 1)$ , we have for all  $j \geq 1$

$$\frac{\mathbb{P}(\tilde{v} \geq j)}{\mathbb{P}(v \geq j/p)} = A(j) + B(j),$$

where

$$A_1(j) \leq A(j) \leq A_2(j)$$

with

$$A_1(j) = \left( 1 - \exp \left( - \frac{p}{2 \left( 1 + \left( \frac{j}{p} \right)^{\alpha-1} \right)} \left( \frac{j}{p} \right)^{2\alpha-1} \right) \right) \frac{\mathbb{P}(v \geq j/p + \lfloor (j/p)^\alpha \rfloor + 1)}{\mathbb{P}(v \geq j/p)},$$

$$A_2(j) = \frac{\mathbb{P}(v \geq j/p - \lfloor (j/p)^\alpha \rfloor)}{\mathbb{P}(v \geq j/p)},$$

and where  $B(j)$  is a positive quantity such that

$$B(j) \leq e^{-\frac{p}{2(1-p)} \left( \frac{j}{p} \right)^{2\alpha-1}} \frac{\mathbb{P}(v \geq j)}{\mathbb{P}(v \geq j/p)}.$$

*Proof.* We have

$$\mathbb{P}(\tilde{v} \geq j) = \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j \right) = T_1 + T_2,$$

where

$$T_1 = \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j, j \leq v \leq j/p - \lfloor (j/p)^\alpha \rfloor - 1 \right),$$

$$T_2 = \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j, j/p - \lfloor (j/p)^\alpha \rfloor \leq v \right).$$

Let us first recall the following inequality for the sum of independent Bernoulli random variables  $B_\ell$ ,  $\ell \geq 1$  [9]: for  $x \in [0, 1-p]$

$$(23) \quad \mathbb{P} \left( \sum_{\ell=1}^n B_\ell - np \geq nx \right) \leq e^{-\frac{nx^2}{A(x)}},$$

where

$$(24) \quad A(x) = 2p(1-p) + \frac{2}{3}x(1-2p) - \frac{2}{9}x^2.$$

It follows that for  $j \leq v \leq j/p$

$$\mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j \right) \leq e^{-\frac{(j-pv)^2}{vA\left(\frac{j-pv}{v}\right)}}.$$

It is easily checked that the function  $v \rightarrow vA\left(\frac{j}{v} - p\right)$  is increasing in the interval  $[j, j/p]$  and that for all  $v \in [j, j/p]$

$$vA\left(\frac{j}{v} - p\right) \leq 2j(1-p).$$

Hence, for  $v \in [j, j/p]$

$$\mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j \right) \leq e^{-\frac{(j-pv)^2}{2j(1-p)}}$$

and for  $v \in [j, j/p - \lfloor (j/p)^\alpha \rfloor - 1]$

$$\mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j \right) \leq e^{-\frac{p}{2(1-p)} \left( \frac{j}{p} \right)^{2\alpha-1}}.$$

This implies that

$$\begin{aligned} T_1 &\leq \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j, j \leq v \leq j/p - \lfloor (j/p)^\alpha \rfloor - 1 \right) \\ &\leq \mathbb{P} \left( \sum_{\ell=1}^{j/p - \lfloor (j/p)^\alpha \rfloor - 1} B_\ell \geq j \right) \mathbb{P}(v \geq j) \\ &= e^{-\frac{p}{2(1-p)} \left(\frac{j}{p}\right)^{2\alpha-1}} \mathbb{P}(v \geq j). \end{aligned}$$

For the term  $T_2$ , we first note that

$$T_2 \leq \mathbb{P}(v \geq j/p - \lfloor (j/p)^\alpha \rfloor).$$

Then, we clearly have

$$T_2 \geq \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \geq j, j/p + \lfloor (j/p)^\alpha \rfloor + 1 \leq v \right)$$

and then

$$\frac{T_2}{\mathbb{P}(v \geq j/p)} \geq \mathbb{P} \left( \sum_{\ell=1}^{j/p + \lfloor (j/p)^\alpha \rfloor + 1} B_\ell > j \right) \frac{\mathbb{P}(v \geq j/p + \lfloor (j/p)^\alpha \rfloor + 1)}{\mathbb{P}(v \geq j/p)}.$$

Chernoff bound implies for  $v = j/p + \lfloor (j/p)^\alpha \rfloor + 1$

$$\begin{aligned} \mathbb{P} \left( \sum_{\ell=1}^v B_\ell \leq j \right) &\leq \exp \left( -\frac{(pv - j)^2}{2pv} \right) \\ &\leq \exp \left( -\frac{p}{2 \left( 1 + \left( \frac{j}{p} \right)^{\alpha-1} \right)} \left( \frac{j}{p} \right)^{2\alpha-1} \right). \end{aligned}$$

It follows that

$$\begin{aligned} \frac{T_2}{\mathbb{P}(v \geq j/p)} &\geq \\ &\left( 1 - \exp \left( -\frac{p}{2 \left( 1 + \left( \frac{j}{p} \right)^{\alpha-1} \right)} \left( \frac{j}{p} \right)^{2\alpha-1} \right) \right) \frac{\mathbb{P}(v \geq j/p + \lfloor (j/p)^\alpha \rfloor + 1)}{\mathbb{P}(v \geq j/p)}. \end{aligned}$$

and the proof follows.  $\square$

The above result can be applied to specific distributions for  $v$ , namely Pareto and Weibull distributions, in order to show that the tails of the probability distribution functions of  $\tilde{v}$  and  $pv$  are the same. This is the analog of Proposition 6 for the probabilistic model.

**Corollary 2.** *If  $v$  has either*

(1) *a Pareto tail distribution with parameter  $a > 1$  such that for  $x \geq 0$ ,  $\mathbb{P}(v \geq x) = L(x)x^{-a}$  where  $L$  is a slowly varying function, i.e., for each  $t > 0$ ,*

$$\lim_{x \rightarrow +\infty} \frac{L(tx)}{L(x)} = 1;$$



or

(2) a Weibull tail distribution with  $\beta \in ]0, 1/2[$  such that for  $x \geq 0$ ,  $\mathbb{P}(v \geq x) = L(x)e^{-\delta x^\beta}$  for some  $\delta > 0$  and  $L$  a slowly varying function

then

$$\lim_{j \rightarrow +\infty} \left| \frac{\mathbb{P}(\tilde{v} \geq j)}{\mathbb{P}(v \geq j/p)} - 1 \right| = 0.$$

*Proof.* For (1),

$$\frac{\mathbb{P}(v \geq j)}{\mathbb{P}(v \geq j/p)} = \frac{L(j)}{L(j/p)} \frac{j^{-a}}{(j/p)^{-a}} = \frac{L(j)}{L(j/p)} p^a \xrightarrow{j \rightarrow +\infty} p^{-a}$$

and

$$\frac{\mathbb{P}(v \geq j/p + \epsilon(j/p)^\alpha)}{\mathbb{P}(v \geq j/p)} = \frac{L((j/p)(1 + \epsilon(j/p)^{\alpha-1}))}{L(j/p)} (1 + \epsilon(j/p)^{\alpha-1})^{-a}$$

which tends to 1 when  $j$  tends to  $+\infty$ . This implies that the quantities  $A_1(j)$  and  $A_2(j)$  appearing in Theorem 3 tends to 1 and  $B(j)$  tends to 0 when  $j \rightarrow \infty$ .

For (2),

$$\frac{\mathbb{P}(v \geq j)}{\mathbb{P}(v \geq j/p)} = \frac{L(j)}{L(j/p)} e^{-\delta j^\beta (1 - p^{-\beta})} \xrightarrow{j \rightarrow +\infty} 0$$

and it is straightforward that

$$\begin{aligned} \frac{\mathbb{P}(v \geq j/p + \epsilon(j/p)^\alpha)}{\mathbb{P}(v \geq j/p)} &= \frac{L(j/p(1 + \epsilon(j/p)^{\alpha-1}))}{L(j/p)} e^{-\delta(j/p + \epsilon(j/p)^\alpha)^\beta + \delta(j/p)^\beta} \\ &= \frac{L(j/p(1 + \epsilon(j/p)^{\alpha-1}))}{L(j/p)} e^{-\delta\beta\epsilon(j/p)^{\alpha+\beta-1}(1+o(1))} \end{aligned}$$

which tends to 1 if  $\alpha + \beta < 1$ . Let  $\beta \in ]0, 1[$ . It is sufficient to find  $\alpha \in ]1/2, 1[$  such that  $\alpha + \beta < 1$ . Necessarily  $1 - \beta > \alpha > 1/2$  thus  $\beta < 1/2$  and for such a  $\beta$ , such an  $\alpha$  exists.  $\square$

## 6. CONCLUDING REMARKS ON SAMPLING AND PARAMETER INFERENCE

We have established in this paper convergence results for the distribution of the number of balls with a given color under the assumption that there is a large number of colors in the urn, that the number of balls with a given color has a heavy tailed distribution independent of the color, and that only a small fraction  $p$  of the total number of ball is sampled. We have considered two ball sampling rules. The first one states that the probability of drawing a ball with a given color depends upon the relative contribution of the color to the total number of balls and that a drawn ball is immediately replaced into the urn. With the second rule, each ball is selected with probability  $p$  independently of the others. The two rules do not give the same results, even if they coincide when  $p \rightarrow 0$  (see [5] for details).

From a practical point of view, we have shown that it is possible to identify the original distribution of the number of balls with a given color by using the tail of the distribution of the number of balls with a given color drawn from the urn. A stronger result holds for Pareto when the number of colors is very large (see Proposition 3). This result is robust in practice because it does not rely on the asymptotics of the tail distribution (in Proposition 3 assertions hold for all  $j > a$ ).

The determination of the original number of balls per color is valid when the number of balls follows a unique distribution of Pareto or Weibull type. This could

be used in the context of packet sampling in the Internet. In practice, however, the number of packets in flows is in general not described by a unique “nice” distribution, but can only be locally approximated by a series of Pareto distributions (see [2] for a discussion). More sophisticated techniques are then necessary to get the original statistics of flows.

## REFERENCES

- [1] M. Abramowitz and I. Stegun, *Handbook of mathematical functions*, National Bureau of Standards, Applied Mathematics Series 55, 1972.
- [2] N. Antunes, Y. Chabchoub, C. Fricker, F. Guillemin, and P. Robert, *On the estimation of flow statistics via packet sampling in the Internet*, Submitted for publication.
- [3] S. Asmussen, C. Klüppelberg, and K. Sigman, *Sampling at subexponential times, with queueing application*, Stochastic Process. Appl. **79** (1999), 265–286.
- [4] A. D. Barbour, Lars Holst, and Svante Janson, *Poisson approximation*, The Clarendon Press Oxford University Press, New York, 1992, Oxford Science Publications.
- [5] Yousra Chabchoub, Christine Fricker, Fabrice Guillemin, and Philippe Robert, *Deterministic versus probabilistic packet sampling in the Internet*, Proceedings of ITC’20, June 2007.
- [6] W. Feller, *An introduction to probability theory, Theory and application*, vol. 2, Wiley, 1966.
- [7] S. Foss and D. Korshunov, *Sampling at a random time with a heavy-tailed distribution*, Markov Process. Related Fields **6** (2000), no. 4, 543–568.
- [8] Philippe Robert, *Réseaux et files d’attente: méthodes probabilistes*, Mathématiques et Applications, vol. 35, Springer-Verlag, Berlin, Octobre 2000.
- [9] A. Siegel, *Toward a usable theory of chernoff bounds for heterogeneous and partially dependent random variables*, Paper available at <http://cs.nyu.edu/faculty/siegel/HHf.pdf>.

(C. Fricker) INRIA-ROCQUENCOURT, RAP PROJECT, DOMAINE DE VOLUCEAU, 78153 LE CHESNAY, FRANCE

*E-mail address:* Christine.Fricker@inria.fr

*URL:* <http://www-c.inria.fr/twiki/bin/view/RAP/ChristineFricker>

(F. Guillemin) ORANGE LABS, F-22300 LANNION

*E-mail address:* Fabrice.Guillemin@orange-ftgroup.com

(Ph. Robert) INRIA-ROCQUENCOURT, RAP PROJECT, DOMAINE DE VOLUCEAU, 78153 LE CHESNAY, FRANCE

*E-mail address:* Philippe.Robert@inria.fr

*URL:* <http://www-rocq.inria.fr/~robert>