Asymptotics of the Overflow in Urn Models

Paweł Hitczenko

(based on joint work with R. Gouet and J. Wesołowski)

Analytic and Probabilistic Combinatorics
Banff, November 13–18, 2022

Set–Up

- Consider a collection (possibly infinite) of distinct containers in which balls are to be inserted. All containers have the same finite capacity r.
- Each arriving ball is to be placed in one of the containers, randomly (according to a given distribution) and independently of other balls. However, if the container selected for a given ball is already full, the ball lands in the overflow basket. We are interested in the number of balls in that basket as the number balls grows.

Relation to Existing Literature

- The notion of the overflow appeared, for example, in the context of collision resolution for hashing algorithms, see a discussion in section: "External searching" in **Knuth, vol. 3**..
- When $r = 1$ this is he number of balls falling in the occupied urns and is sometimes called the number of collisions and, when distribution of balls among urns is uniform, has been used e.g. to test the random number generators (**Knuth, vol. 2**).
- **Ramakrishna (1987)** and **Monahan (1987)** compute the probability that there is no overflow (under the uniformity assumption), and the estimation of the probability of unusually large overflow is in **Dupuis, Nuzman, Whiting (2004)**.
- As a byproduct of their methods **Hwang, Janson (2008)** gave sufficient conditions for the Poissonian limit when $r = 1$.

- For $n \geq 1$, let $X_{n,1}, \ldots, X_{n,n}$ be iid random variables with values in $M_n \subset \mathbb{N} := \{1, 2, \ldots\}$ and let $p_{n,m} = \mathbb{P}(X_{n,1} = m)$, $m \in M_n$, be the common distribution among the boxes for each of the n balls in the nth experiment.
- Let for any $n \in \mathbb{N}$, $k \in \{1, \ldots, n, n+1\}$ and $m \in M_n$

$$N_{n,k}(m) = \sum_{j=1}^{k-1} I_{\{X_{n,j}=m\}},$$

be the number of balls among first $k-1$ balls falling in the mth box.
- Let $r \geq 1$ be the (same) capacity of every container. Then

$$Y_{n,k} = \sum_{m \in M_n} I_{\{X_{n,k}=m\}} I_{\{N_{n,k}(m) \geq r\}} = I_{\{k\text{th ball is in overflow}\}}.$$

- Then, the size of the overflow, $V_{n,r}$, can be written as

$$V_{n,r} = \sum_{k=1}^{n} Y_{n,k}.$$

- We will be interested in the asymptotic distribution of $V_{n,r}$, as $n \to \infty$. We show that there are regimes related to $p_{n,m}$ under which the limiting distribution of $V_{n,r}$ (possibly standardized) is either Poisson or normal. These regimes are defined through the limiting behavior of the sequences

$$n\, p_n^* \qquad \text{and} \qquad n^{r+1} \sum_{m \in M_n} p_{n,m}^{r+1},$$

where $p_n^* = \sup_{m \in M_n} p_{n,m}$.

- **Notation:**

$$\mathbb{E}\ p_{n,X_n}^r := \sum_{m \in M_n} p_{n,m}^{r+1}.$$

(We'll also use $\mathbb{E}\ p_{X_n}^r$.)

Poissonian asymptotics

Theorem

Let $\mathrm{Pois}(\mu)$ *denote the Poisson distribution with parameter* $\mu \in (0, \infty)$. *If*

$$n^{r+1}\mathbb{E}\, p_{X_n}^r \to (r+1)!\mu$$

and

$$n\, p_n^* \to 0,$$

then $V_{n,r} \xrightarrow{d} \mathrm{Pois}(\mu)$.

Examples:

- The uniform case: $p_{n,j} = \frac{1}{m_n}$, for $j \in M_n = \{1, \ldots, m_n\}$, $m_n = \left\lfloor a n^{\frac{r+1}{r}} \right\rfloor$, $a > 0$. Then

$$np_n^* = \frac{n}{m_n} \to 0 \qquad \text{and} \qquad n^{r+1}\,\mathbb{E}\,p_{X_n}^r = \frac{n^{r+1}}{m_n^r} \to \frac{1}{a^r}.$$

Thus, $V_{n,r} \xrightarrow{d} \mathrm{Pois}(\mu)$, with $\mu = \frac{1}{a^r\,(r+1)!}$.

Examples:

- The uniform case: $p_{n,j} = \frac{1}{m_n}$, for $j \in M_n = \{1, \ldots, m_n\}$, $m_n = \left\lfloor a n^{\frac{r+1}{r}} \right\rfloor$, $a > 0$. Then

$$np_n^* = \frac{n}{m_n} \to 0 \qquad \text{and} \qquad n^{r+1}\,\mathbb{E}\,p_{X_n}^r = \frac{n^{r+1}}{m_n^r} \to \frac{1}{a^r}.$$

  Thus, $V_{n,r} \xrightarrow{d} \mathrm{Pois}(\mu)$, with $\mu = \frac{1}{a^r\,(r+1)!}$.

- The geometric case: $p_{n,j} = p_n(1-p_n)^j$, $j \geq 0$.Take $p_n = \frac{a}{n^{(r+1)/r}}$, $a > 0$. Then $np_n^* = np_n \to 0$ and

$$n^{r+1}\mathbb{E}\,p_{X_n}^r = \frac{(np_n)^{r+1}}{1-(1-p_n)^{r+1}} \to \frac{a^r}{r+1}$$

  Thus, $V_{n,r} \xrightarrow{d} \mathrm{Pois}(\mu)$, with $\mu = \frac{a^r}{(r+1)!(r+1)}$.

Main technical result:

Theorem (Beśka, Kłopotowski,Słomiński (1982))

Let $\{Y_{n,k},\ k = 1, \ldots, n;\ n \geq 1\}$ *be an array of non-negative random variables, adapted to a row-wise increasing array of* σ*-fields* $\{\mathcal{F}_{n,k},\ k = 1, \ldots, n;\ n \geq 1\}$*, and let* $\eta > 0$*. If*

$$\max_{1 \leq k \leq n} \mathbb{E}\left(Y_{n,k} | \mathcal{F}_{n,k-1}\right) \xrightarrow{\mathbb{P}} 0,$$

$$\sum_{k=1}^{n} \mathbb{E}\left(Y_{n,k} | \mathcal{F}_{n,k-1}\right) \xrightarrow{\mathbb{P}} \eta$$

and, for any $\epsilon > 0$*,*

$$\sum_{k=1}^{n} \mathbb{E}\left(Y_{n,k} I_{\{|Y_{n,k}-1|>\epsilon\}} | \mathcal{F}_{n,k-1}\right) \xrightarrow{\mathbb{P}} 0,$$

then $\sum_{k=1}^{n} Y_{n,k} \xrightarrow{d} \mathrm{Pois}(\eta)$.

Theorem (proof based on martingale CLT)

Assume that $\lambda := \limsup np_n^* < \infty$ *and* $n^{r+1}\mathbb{E}\, p_{X_n}^r \to \infty$. *Then*

$$\frac{V_{n,r} - \mathbb{E}\, V_{n,r}}{\sqrt{\mathbb{V}ar\, V_{n,r}}} \xrightarrow{d} \mathrm{N}(0,1),$$

$$\frac{\Gamma_\lambda(r+1)}{r!} \le \liminf \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \le \limsup \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} \le \frac{1}{(r+1)!},$$

$$\frac{e^{-2\lambda}}{(r+1)!} \le \liminf \frac{\mathbb{V}ar\, V_{n,r}}{n^{r+1}\,\mathbb{E}\, p_{X_n}^r} \le \limsup \frac{\mathbb{V}ar\, V_{n,r}}{n^{r+1}\,\mathbb{E}\, p_{X_n}^r} \le \frac{1}{r!}.$$

where, for $p > 0$ *and* $x \ge 0$, *we have set*

$$\Gamma_x(p) := \int_0^1 t^{p-1}\, e^{-xt}\, dt.$$

Note: $\lambda = 0$ implies $\lim \frac{\mathbb{E}\, V_{n,r}}{n^{r+1}\mathbb{E}\, p_{X_n}^r} = \frac{1}{(r+1)!}$ since $\Gamma_0(r+1) = \frac{1}{r+1}$.

Corollary

Assume that $np_n^* \to 0$ *and that there exists an* $r \in \{1, 2, \ldots\}$ *such that*

$$n^{r+1}\mathbb{E}\, p_{X_n}^r \to (r+1)!\mu.$$

Then

1. $\frac{V_{n,s} - \mathbb{E}\, V_{n,s}}{\sqrt{\mathbb{V}ar\, V_{n,s}}} \xrightarrow{d} \mathrm{N}(0,1)$, *for* $s \in \{1, \ldots, r-1\}$;
2. $V_{n,r} \xrightarrow{d} \mathrm{Pois}(\mu)$;
3. $V_{n,s} \xrightarrow{\mathbb{P}} 0$, *for* $s \in \{r+1, r+2, \ldots\}$.

Follows from: if $u < t$ then $n^{t+1}\mathbb{E}p_{X_n}^t \leq (np_n^*)^{t-u} n^{u+1}\mathbb{E}p_{X_n}^u$.

More on the asymptotics of the expected value

Assume that np_n^* is bounded and that $\lambda = \limsup np_n^* > 0$. Let $\mathcal{X}_n$ denote the set of distinct values among $\frac{p_{n,k}}{p_n^*}$, $k \in M_n$. Define random variables T_n, $n \geq 1$, as follows:

$$\mathbb{P}(T_n = x) = \frac{1}{\mathbb{E}\, p_{X_n}^r} \sum_{k \in K(x)} p_{n,k}^{r+1}, \quad x \in \mathcal{X}_n,$$

where, for $x \in \mathcal{X}_n$ we let $K(x) = \{k \in M_n : x = \frac{p_{n,k}}{p_n^*}\}$.

Definition

We say that the sequence $(X_n)_{n \geq 1}$ is in the class $\mathcal{T}(r)$ if the sequence $(T_n)_{n \geq 1}$ converges in distribution.

For $(X_n)_{n \geq 1} \in \mathcal{T}(r)$, if $\lim_{n \to \infty} np_n^*$ exists and is positive then

$$H(r, \lambda) := \lim_{n \to \infty} \frac{\mathbb{E}\, V_{n,r}}{n^{r+1} \mathbb{E}\, p_{X_n}^r}$$

also exists.

Specifically,

Theorem

Let $(X_n)_{n\geq 1}$ be in $\mathcal{T}(r)$ and $np_n^* \to \lambda > 0$. Then

$$H(r,\lambda) = \frac{1}{(r+1)!} \int {}_1F_1(r; r+2; -\lambda u)\, \nu_r(du),$$

where ${}_pF_q$ is the generalized hypergeometric function defined by

$${}_pF_q(a_1,\ldots,a_p; b_1,\ldots,b_q; z) = \sum_{k=0}^{\infty} \frac{(a_1)_k \cdots (a_p)_k}{(b_1)_k \cdots (b_q)_k} \frac{z^k}{k!},$$

where $(a)_0 = 1$ and $(a)_k = a(a+1)\cdots(a+k-1)$ for $k \geq 1$ and ν_r is the limiting distribution of (T_n).

Examples

- uniform distribution: $p_{n,j} = \frac{1}{m_n}$, $j \in M_n = \{1, \ldots, m_n\}$, $n \geq 1$. Assume that $\frac{n}{m_n} \to \lambda > 0$. Then, $T_n = 1$ $\mathbb{P}$-a.s., and thus $\nu_r = \delta_1$. Hence,

$$H(r, \lambda) = \frac{{}_1F_1(r; r+2; -\lambda)}{(r+1)!}.$$

- geometric distribution: $p_{n,j} = p_n(1 - p_n)^j$, $j \in M_n = \{0, 1, \ldots\}$, $n \geq 1$,(here $p_n^* = p_n$). Assume $np_n \to \lambda > 0$. Then $\nu_r(du) = (r+1)u^r \, I_{[0,1]}(u) \, du$ and

$$H(r, \lambda) = \frac{{}_2F_2(r, r+1; r+2, r+2; -\lambda)}{(r+1)!}.$$

- Riemann ζ distribution: Let $p_{n,j} = \frac{j^{-\alpha_n}}{\zeta(\alpha_n)}$, $j \in M_n = \{1, 2, \ldots\}$, and $\alpha_n > 1$, $n \geq 1$. Assume that $n(\alpha_n - 1) \to \lambda > 0$. Then

$$\nu_r = \sum_{k \geq 1} \frac{k^{-(r+1)}}{\zeta(r+1)} \, \delta_{1/k}.$$

and

$$H(r, \lambda) = \frac{1}{(r+1)!} \int_{\mathbb{R}} {}_1F_2(r; r+1, r+2; -\lambda x) \, \mu_r(dx),$$

where μ_r is the probability measure defined on $(0, \infty)$ by

$$\mu_r(dx) = \frac{1}{r! \zeta(r+1)} \, \frac{x^r}{e^x - 1} \, dx.$$

- Riemann ζ distribution: Let $p_{n,j} = \frac{j^{-\alpha_n}}{\zeta(\alpha_n)}$, $j \in M_n = \{1, 2, \ldots\}$, and $\alpha_n > 1$, $n \geq 1$. Assume that $n(\alpha_n - 1) \to \lambda > 0$. Then

$$\nu_r = \sum_{k \geq 1} \frac{k^{-(r+1)}}{\zeta(r+1)} \, \delta_{1/k}.$$

and

$$H(r, \lambda) = \frac{1}{(r+1)!} \int_{\mathbb{R}} {}_1F_2(r; r+1, r+2; -\lambda x) \, \mu_r(dx),$$

where μ_r is the probability measure defined on $(0, \infty)$ by

$$\mu_r(dx) = \frac{1}{r! \zeta(r+1)} \, \frac{x^r}{e^x - 1} \, dx.$$

Thank you!