ALEA, Lat. Am. J. Probab. Math. Stat. 11 (1), 197–208 (2014)

Some new aspects for the random coupon collector's problem

Aristides V. Doumas and Vassilis G. Papanicolaou

Department of Mathematics
National Technical University of Athens
Zografou Campus
157 80 Athens, Greece.
E-mail address: aris.doumas@hotmail.com

Department of Mathematics
National Technical University of Athens
Zografou Campus
157 80 Athens, Greece.
E-mail address: papanico@math.ntua.gr
URL: http://www.math.ntua.gr/~papanico/

Abstract. Let T_N be the number of coupons that a collector has to buy in order to find all N existing different coupons. The probabilities p_n (occurring frequencies) of the coupons are taken to be identically distributed random variables. We develop techniques of computing the asymptotics of the expectation of $\overline{T_N(T_N+1)}$, where $\overline{X}$ denotes average, given the p_n's. Using these asymptotics we derive the leading behavior of the expectation of $\mathcal{V}(T_N)$, where $\mathcal{V}(T_N)$ is the variance of T_N, given the p_n's (see Theorems 3.1 and 3.2). We also conjecture on the minimum of this quantity.

1. Introduction

The classic coupon collector's problem (CCP) concerns a population whose members are of N different *types* (ants, fish, words, viruses, etc). The members of the population are sampled independently with replacement and their types are recorded. For $1 \leq n \leq N$ we denote by p_n the probability that a member of the population is of type n, where $p_1+p_2+\cdots+p_N=1,\ \ p_n>0$. Let T_N, the number of trials it takes until all N types are detected (at least once). CCP pertains to the family of urn problems along with other famous problems, such as the birthday, or occupancy. This problem (in its simplest form) had appeared in W. Feller's classical work (Feller (1968)) and in some earlier relatively unknown works whose origin

Received by the editors May 25, 2013; accepted March 25, 2014.
2010 *Mathematics Subject Classification.* Primary 60C05; Secondary 60E10.
Key words and phrases. Random coupon collector's problem, asymptotics, Laplace integral, biology ecosystems.

can be traced back to De Moivre's treatise *De Mensura Sortis* of 1712 (see, e.g., Holst (1986)) and Laplace's pioneering work *Theorie Analytique de Probabilites* of 1812 (see Diaconis and Holmes (2002)). The literature in the case the sampling probabilities $p_1, p_2, ..., p_N$ are fixed parameters is extensive (see, e.g., Boneh and Hofri (1997), Brayton (1963), Doumas and Papanicolaou (2012a,b)).

We are motivated by the fact that in many applications, the sampling probabilities behave **more** like random variables rather than parameters. Consider, for example, a complex ecosystem inhabited by a great variety of species. Biologists who explore that environment want to detect all the different species of a certain type. For example, all the different species of fish in the Mediterranean sea. Observing a particular type of fish constitutes a detection of that species. In the species detection process, the probabilities are proportional to the size of the subpopulations of the different species. These subpopulations change randomly, and their exact sizes are typically unknown. We shall refer to such a version of the CCP as the *Random CCP* (RCCP). Note that this point of view is also in accordance to the Bayesian approach, namely to put a prior distribution on the probabilities (occurring frequencies); see, e.g., Diaconis and Holmes (2002).

In Papanicolaou et al. (1998) it is assumed that $p_n = a_n/A_N$, where a_n were taken to be i.i.d. random variables and $A_N = \sum_{n=1}^{N} a_n$, (here and in what follows the abbreviation i.i.d. means, as usual, independent and identically distributed). Asymptotic estimates for $E\left[\overline{T_N}\right]$ were obtained, where $\overline{T_N}$ denotes the average of T_N, given p_n, $n = 1, 2 \cdots, N$. The results of Papanicolaou et al. (1998) were improved in Holst (2001). In particular, the limiting distribution of T_N was obtained for a large class of priors.

In this paper, we present asymptotic results regarding the expectation of the variance of $\overline{T_N}$ for the i.i.d. case. We begin our analysis with some well known results for the classic CCP. The set-up of the RCCP completes this section. In Section 2 we discuss the case of independent sampling probabilities. The i.i.d. case is studied in Section 3. Our main results are presented in Theorems 3.1 and 3.2, regarding the asymptotics of the expectation of $\overline{T_N(T_N+1)}$, and the expectation of $\mathcal{V}(T_N)$ respectively (given the p_n's). General examples are also exhibited at the end of the paper.

1.1. *Some preliminary results regarding the classic CCP.* In the classic CCP the sampling probabilities $p_1, p_2, ..., p_N$ are fixed parameters satisfying

$$p_1 + p_2 + \cdots + p_N = 1, \qquad p_n > 0, \quad \text{for} \quad 1 \le n \le N.$$

In particular, if $\overline{T_N}$ denotes the average of T_N, it has been established that

$$\overline{T_N} = \sum_{k=1}^{N} (-1)^{k+1} \sum_{1 \le n_1 < \cdots < n_k \le N} \frac{1}{p_{n_1} + \cdots + p_{n_k}} = \int_0^\infty \left[1 - \prod_{n=1}^{N} \left(1 - e^{-p_n t} \right) \right] dt. \tag{1.1}$$

For a proof see, for example, Ross (1996). In a similar way, if $\overline{T_N(T_N+1)}$ denotes the second (rising) moment of T_N, it has been established (see, e.g., Brayton (1963))

that

$$\begin{aligned}\overline{T_N(T_N+1)} =& 2\sum_{m=1}^{N}(-1)^{m-1}\sum_{1\le n_1<\cdots<n_m\le N}\frac{1}{\left(p_{n_1}+\cdots+p_{n_m}\right)^2}\\ =& 2\int_0^\infty\left[1-\prod_{n=1}^{N}\left(1-e^{-p_n t}\right)\right]t\,dt. \end{aligned}\tag{1.2}$$

Therefore, for the variance of T_N we have

$$\mathcal{V}(T_N) := \overline{T_N(T_N+1)} - \overline{T_N} - \left(\overline{T_N}\right)^2. \tag{1.3}$$

1.2. *Randomization.* Let $\alpha = \{a_n\}_{n=1}^{\infty}$ be a sequence of (strictly) positive independent random variables with

$$\mu_n := E_\alpha[\,a_n\,],\quad \sigma_n^2 := V_\alpha[\,a_n\,] < \infty \quad\text{and}\quad g_n(t) := E_\alpha\left[\,e^{-ta_n}\,\right],\ t\ge 0, \tag{1.4}$$

where $E_\alpha[\,\cdot\,]$ stand for the expectation and $V_\alpha[\,\cdot\,]$ for the variance associated to the sequence α. For a given N, we can create sampling probabilities $p_1, p_2, ..., p_N$ by taking

$$p_n = \frac{a_n}{A_N},\qquad \text{where}\qquad A_N = \sum_{n=1}^{N} a_n. \tag{1.5}$$

Notice that each p_n, $1\le n\le N$, is a random variable which depends on N and α. This is the set-up of what we call *Random Coupon Collector's Problem.* Our main interest here is the study of the quantity $E_\alpha[\,\mathcal{V}(T_N)\,]$ as $N\to\infty$. From (1.3) we have

$$E_\alpha[\,\mathcal{V}(T_N)\,] = E_\alpha\left[\overline{T_N(T_N+1)}\right] - E_\alpha\left[\,\overline{T_N}\,\right] - E_\alpha\left[\left(\overline{T_N}\right)^2\right]. \tag{1.6}$$

For typographical convenience we set

$$Q_N := E_\alpha\left[\overline{T_N(T_N+1)}\right]\quad\text{and}\quad m_N := E_\alpha\left[\,\overline{T_N}\,\right]. \tag{1.7}$$

2. The general case of independent sampling probabilities

If we denote

$$W_N^\alpha := 2\sum_{m=1}^{N}(-1)^{m-1}\sum_{1\le j_1<\cdots<j_m\le N}\frac{1}{\left(a_{j_1}+\cdots+a_{j_m}\right)^2},$$

then, $W_N^{\lambda\alpha} = \lambda^{-2}W_N^\alpha$, where $\lambda>0$ and $\lambda\alpha = \{\lambda a_1, \lambda a_2, ...\}$. Therefore by (1.2) and (1.5),

$$Q_N = E_\alpha\left[W_N^{\alpha/A_N}\right] = E_\alpha\left[\,A_N^2 W_N^\alpha\,\right]. \tag{2.1}$$

Furthermore, as in (1.2), we have

$$W_N^\alpha = 2\int_0^\infty\left[1-\prod_{j=1}^{N}\left(1-e^{-a_j t}\right)\right]t\,dt,$$

hence (2.1) can be written as

$$Q_N = 2E_\alpha\left[\left(\sum_{n=1}^{N}a_n\right)^2\int_0^\infty\left\{1-\prod_{k=1}^{N}\left(1-e^{-a_k t}\right)\right\}t\,dt\right]. \tag{2.2}$$

Theorem 2.1. *Let $\alpha = \{a_n\}_{n=1}^{\infty}$ be a sequence of (strictly) positive independent random variables satisfying (1.4). Then for Q_N of (1.7) we have*

$$\begin{aligned} Q_N =& 2\sum_{j=1}^{N} g_j''(0) \int_0^{\infty} \left\{ 1 - \left(\frac{1 - g_j''(t)/g_j''(0)}{1 - g_j(t)} \right) \prod_{k=1}^{N} (1 - g_k(t)) \right\} t\, dt \\ &+ 4 \sum_{i<j}^{N} g_i'(0) g_j'(0) \\ &\times \int_0^{\infty} \left\{ 1 - \left(\frac{1 - g_i'(t)/g_i'(0)}{1 - g_i(t)} \right) \left(\frac{1 - g_j'(t)/g_j'(0)}{1 - g_j(t)} \right) \prod_{k=1}^{N} (1 - g_k(t)) \right\} t\, dt. \end{aligned} \tag{2.3}$$

Proof. Using the independence of the a_n's and Tonelli's Theorem (for switching expectation and integral) we get from (2.2)

$$\begin{aligned} Q_N =& 2\sum_{j=1}^{N} E_\alpha \left[a_j^2 \right] \int_0^{\infty} \left\{ 1 - \left(1 - \frac{E_\alpha \left[a_j^2 e^{-a_j t} \right]}{E_\alpha \left[a_j^2 \right]} \right) \prod_{\substack{k=1 \\ k \neq j}}^{N} [1 - g_k(t)] \right\} t\, dt \\ &+ 4 \sum_{i<j}^{N} E_\alpha \left[a_i \right] E_\alpha \left[a_j \right] \\ &\times \int_0^{\infty} \left\{ 1 - \left(1 - \frac{E_\alpha \left[a_i e^{-a_i t} \right]}{E_\alpha \left[a_i \right]} \right) \left(1 - \frac{E_\alpha \left[a_j e^{-a_j t} \right]}{E_\alpha \left[a_j \right]} \right) \prod_{\substack{k=1 \\ k \neq i,\, j}}^{N} [1 - g_k(t)] \right\} t\, dt. \end{aligned}$$

Now, observe that, from (1.4),

$$E_\alpha \left[a_n e^{-a_n t} \right] = -g_n'(t) \qquad \text{and} \qquad E_\alpha \left[a_n^2 e^{-a_n t} \right] = g_n''(t). \tag{2.4}$$

Notice that $g_n'(t)$ and $g_n''(t)$ are finite for $t > 0$. The result follows. ∎

Proposition 1. Let $F_n(x)$ be the distribution function of the random variable a_n and ε be any strictly positive number. If for some n we have

$$\int_0^{\varepsilon} \frac{dF_n(x)}{x^2} = \infty \qquad \text{or} \qquad E_\alpha \left[a_n^2 \right] = \infty,$$

then $Q_N = \infty$, for all $N \geq \max\{n, 2\}$.

Proof. We can assume, without loss of generality, that $n = 1$, and then observe that it is enough to prove the statement only for $N = 2$. From the first equality in (1.2) we get

$$Q_2 = 2E_\alpha \left[\frac{1}{p_1^2} + \frac{1}{p_2^2} \right] - 2,$$

or, by invoking (1.5) and the independence of a_1, a_2

$$Q_2 = 4\mu_2 E_\alpha \left[\frac{1}{a_1} \right] + 4\mu_1 E_\alpha \left[\frac{1}{a_2} \right] + 2E_\alpha \left[a_2^2 \right] E_\alpha \left[\frac{1}{a_1^2} \right] + 2E_\alpha \left[a_1^2 \right] E_\alpha \left[\frac{1}{a_2^2} \right] + 2. \tag{2.5}$$

Thus, if $E_\alpha\left[\,a_1^2\,\right]=\infty$, then $Q_2=\infty$. On the other hand

$$\int_0^\varepsilon \frac{dF_1(x)}{x^2}=\infty, \qquad \text{if and only if} \qquad E_\alpha\left[\,\frac{1}{a_1^2}\,\right]=\infty,$$

and therefore, if $\int_0^\varepsilon x^{-2}dF_1(x)=\infty$, (2.5) implies that $Q_2=\infty$. ■
Hence, if the a_n's have infinite variance but finite mean, Proposition 1 immediately implies that $Q_N=\infty$.

Remark 1. (i) Intuitively

$$\int_0^\varepsilon \frac{dF_n(x)}{x^2}=\infty \tag{2.6}$$

means that the values of a_n have a high concentration near 0. If in particular a_n has a density $f_n(x)$, such that $f_n(x)\geq kx+\mathrm{o}(x)$ as $x\to 0^+$, $k>0$, then (2.6) is valid. For example, if a_1 is exponentially distributed, or more generally, Gamma distributed with parameters $p\leq 2$, $a>0$ (i.e., $f_n(x)=a^p x^{p-1}e^{-ax}/\Gamma(p)$, $x>0$), then $Q_N=\infty$, for all $N\geq 2$.
(ii) Since

$$E_\alpha\left[\,\frac{1}{a_n^2}\,\right]=E_\alpha\left[\int_0^\infty t\,e^{-ta_n}dt\right]=\int_0^\infty t\,E_\alpha\left[\,e^{-ta_n}\,\right]dt=\int_0^\infty t\,g_n(t)\,dt,$$

we have that

$$\int_0^\varepsilon \frac{dF_n(x)}{x^2}=\infty, \qquad \text{if and only if} \qquad \int_0^\infty t\,g_n(t)\,dt=\infty.$$

(iii) Proposition 1 has a converse: If, for all a_n, $1\leq n\leq N$,

$$\int_0^\varepsilon \frac{dF_n(x)}{x^2}<\infty \qquad \text{and} \qquad E_\alpha\left[\,a_n^2\,\right]<\infty,$$

then $Q_N<\infty$. This implies (of course), that $m_N<\infty$.

3. The i.i.d. case

We now consider the case where the independent random variables a_n are also identically distributed with common distribution function $F(x)$. We set

$$\mu:=E_\alpha\left[\,a_n\,\right], \qquad g(t):=E_\alpha\left[\,e^{-ta_n}\,\right] \tag{3.1}$$

and, in view of Proposition 1 and Remark 1, we assume that

$$\mu^{(2)}:=E_\alpha\left[\,a_n^2\,\right]<\infty, \qquad \text{and} \qquad \int_0^\infty t\,g(t)dt=E_\alpha\left[\,\frac{1}{a_n^2}\,\right]<\infty, \tag{3.2}$$

which also implies

$$\int_0^\infty g(t)dt=E_\alpha\left[\,\frac{1}{a_n}\,\right]<\infty. \tag{3.3}$$

In this set-up, Theorem 2.1 immediately yields

$$\begin{aligned} Q_N=&2N\,(N-1)\,g'(0)^2\int_0^\infty \left\{1-\left[1-g'(t)/g'(0)\right]^2\left[1-g(t)\right]^{N-2}\right\}t\,dt \\ &+2Ng''(0)\int_0^\infty \left\{1-\left[1-g''(t)/g''(0)\right]\left[1-g(t)\right]^{N-1}\right\}t\,dt. \end{aligned} \tag{3.4}$$

In Papanicolaou et al. (1998) it has been shown that in the i.i.d. case the following results hold

$$m_N = -g'(0)N \int_0^\infty \left\{1 - [1 - g'(t)/g'(0)]\,[1 - g(t)]^{N-1}\right\} dt, \tag{3.5}$$

and as $N \to \infty$

$$m_N \sim -\mu N \sum_{n=2}^{N} \int_0^\varepsilon u e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]}, \quad \text{for any} \quad \varepsilon \text{ in } (0,1). \tag{3.6}$$

In order to compute $E_\alpha\left[\mathcal{V}(T_N)\right]$ we need to derive asymptotics for Q_N of (3.4). Let us first look at an example.

Example 1. Consider the (trivial) case of a_n deterministic, i.e. $a_n = c = \mu$, a.s.. Then by (3.5) and (3.4) we have

$$m_N = NH_N, \quad Q_N = N^2\left(H_N^2 + \sum_{m=1}^{N} \frac{1}{m^2}\right), \tag{3.7}$$

where $H_N = \sum_{m=1}^{N} 1/m$. Hence, (recall 1.6)

$$E_\alpha\left[\mathcal{V}(T_N)\right] = \mathcal{V}(T_N) = \frac{\pi^2}{6}N^2 - N\ln N - (\gamma+1)N + O\left(\frac{\ln N}{N}\right). \tag{3.8}$$

Formulas (3.7) and (3.8) are well known results.

Conjecture. For the i.i.d. case the quantity $E_\alpha\left[\mathcal{V}(T_N)\right]$ becomes minimum, when $g(t) = \exp(-ct)$, $c > 0$, i.e. when $a_n = c$, a.s.

We return now to formula (3.4). Set

$$I_1(N) := \int_0^\infty \left\{1 - [1 - g'(t)/g'(0)]^2\,[1 - g(t)]^{N-2}\right\} t\, dt, \tag{3.9}$$

$$I_2(N) := \int_0^\infty \left\{1 - [1 - g''(t)/g''(0)]\,[1 - g(t)]^{N-1}\right\} t\, dt. \tag{3.10}$$

Our task is to find asymptotics for $I_1(N)$, $I_2(N)$. First we deal with $I_1(N)$. We have

$$I_1(n) - I_1(n-1) = \int_0^\infty t\, g(t)\left[1 - \frac{g'(t)}{g'(0)}\right]^2 [1 - g(t)]^{n-3}\, dt, \tag{3.11}$$

If we sum (3.11) from $n = 3$ to N, we obtain

$$I_1(N) = I_1(2) + \sum_{n=3}^{N} J(n), \tag{3.12}$$

where

$$J(n) := \int_0^\infty t\, g(t)\left[1 - \frac{g'(t)}{g'(0)}\right]^2 [1 - g(t)]^{n-3}\, dt. \tag{3.13}$$

Observation 1. In view of (2.4), conditions (3.2), and the Dominated Convergence Theorem, the following properties for $g(t)$ are immediate:
(i) $g(t)$ is convex, monotone decreasing, with $g(0) = 1$ and $g(\infty) = 0$,
(ii) $\lim_{t\to\infty} g'(t) = \lim_{t\to\infty} g''(t) = \lim_{t\to\infty} t\, g(t) = \lim_{t\to\infty} t\, g'(t) = 0$.

Lemma 1.

$$I_1(2) < \infty.$$

Proof. By (3.9), (3.3), Observation 1, and integration by parts we have

$$\begin{aligned} I_1(2) &= \int_0^\infty \left\{1 - [1 - g'(t)/g'(0)]^2\right\} t\, dt \\ &= \frac{2}{g'(0)} \int_0^\infty t g'(t)\, dt - \frac{1}{g'(0)^2} \int_0^\infty t g'(t)^2\, dt < \infty. \quad \blacksquare \end{aligned}$$

Lemma 2. If μ and $g(t)$ are as in (3.1), then for $J(n)$ of (3.13) we have

$$\sum_{n=3}^{\infty} J(n) = \infty.$$

Proof. By (3.13)

$$\sum_{n=3}^{N+2} J(n) = \int_0^\infty t \left[1 - \frac{g'(t)}{g'(0)}\right]^2 \left\{1 - [1 - g(t)]^N\right\} dt.$$

Since $0 < g(t) < 1$ for all $t > 0$, monotone convergence and Lemma 1 imply

$$\begin{aligned} \sum_{n=3}^{\infty} J(n) &= \lim_N \int_0^\infty t \left[1 - \frac{g'(t)}{g'(0)}\right]^2 \left\{1 - [1 - g(t)]^N\right\} dt \\ &= \int_0^\infty t \left[1 - \frac{g'(t)}{g'(0)}\right]^2 dt = \infty. \quad \blacksquare \end{aligned}$$

Next, we compute the asymptotics of $J(n)$, as $n \to \infty$. By (3.13)

$$J(n) = \int_0^\infty t \left[1 - \frac{g'(t)}{g'(0)}\right]^2 g(t)\, e^{(n-3)\ln[1-g(t)]}\, dt.$$

Substituting $u = g(t)$ we get

$$J(n) = -\int_0^1 g^{-1}(u) \left[1 - \frac{g'(g^{-1}(u))}{g'(0)}\right]^2 u\, e^{(n-3)\ln(1-u)} \frac{du}{g'[g^{-1}(u)]}.$$

Using Laplace method for integrals (see, e.g., Bender and Orszag (1978), Sec. 6.4) we get

$$J(n) = -\int_0^\varepsilon g^{-1}(u) \left[1 - \frac{g'(g^{-1}(u))}{g'(0)}\right]^2 u\, e^{-nu} \frac{du}{g'[g^{-1}(u)]} \left[1 + O\left(\frac{1}{n}\right)\right]. \quad (3.14)$$

for any ε in $(0, 1)$. By expanding the bracket, the integral above becomes

$$\begin{aligned} &-\int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} \frac{du}{g'[g^{-1}(u)]} + \frac{2}{g'(0)} \int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} du \\ &\qquad - \frac{1}{g'(0)^2} \int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} g'(g^{-1}(u)) du. \quad (3.15) \end{aligned}$$

Using Observation 1 and the properties of g we see that only the first of the three above integrals contributes to the leading behavior of $J(n)$ of (3.14). Hence,

$$J(n) \sim -\int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} \frac{du}{g'[g^{-1}(u)]}, \quad \text{for any } \varepsilon \text{ in } (0, 1). \quad (3.16)$$

Next, we treat $I_2(N)$ of (3.10) as we treat $I_1(N)$. We get

$$I_2(N) = I_2(1) + \sum_{n=2}^{N} L(n), \tag{3.17}$$

where,

$$L(n) := \int_0^\infty t\, g(t) \left[1 - \frac{g''(t)}{g''(0)}\right] [1 - g(t)]^{n-2}\, dt. \tag{3.18}$$

By Observation 1 we have $I_2(1) < \infty$, and by imitating the proof of Lemma 2, we get $\sum_{n=2}^{\infty} L(n) = \infty$. Furthermore, Laplace method again gives

$$L(n) = -\int_0^\varepsilon g^{-1}(u) \left[1 - \frac{g''(g^{-1}(u))}{g''(0)}\right] u\, e^{(n-2)\ln(1-u)} \frac{du}{g'\left[g^{-1}(u)\right]} \left[1 + O\left(\frac{1}{n}\right)\right]. \tag{3.19}$$

Expanding the bracket in the above integral and using Observation 1 and the properties of g we have

$$L(n) \sim -\int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]}, \quad \text{for any } \varepsilon \text{ in } (0,1), \tag{3.20}$$

i.e. $J(n)$ and $L(n)$ have the same leading asymptotics as $n \to \infty$. The following lemma is an easy exercise:

Lemma 3. Let $\psi(n) > 0$, for $n = 1, 2, 3, ...$, and $\sum_{n=2}^{\infty} \psi(n) = \infty$. If $\phi(n) \sim \psi(n)$ as $n \to \infty$, then, for any fixed n_0,

$$\sum_{n=n_0}^{N} \phi(n) \sim \sum_{n=n_0}^{N} \psi(n) \text{ as } N \to \infty.$$

Thus, $I_1(N)$ and $I_2(N)$, of (3.9) and (3.10) respectively, have the same (!) leading asymptotics as $N \to \infty$. Furthermore, Lemma 3 allows us to substitute (3.16) in (3.12), and (3.20) in (3.17) respectively, in order to get the desired asymptotics:

Theorem 3.1. If Q_N is as in (3.4) and μ, $g(t)$ are as in (3.1), then

$$Q_N \sim -\mu^2 N^2 \sum_{n=3}^{N} \int_0^\varepsilon u\, e^{-nu}\, d\left[\left(g^{-1}(u)\right)^2\right], \quad \text{for any } \varepsilon \text{ in } (0,1) \tag{3.21}$$

as $N \to \infty$.
From (3.6) and (3.21) we conclude:

Theorem 3.2. Let $\overline{T_N}$ be as in (1.1) and m_N, Q_N, μ, and $g(t)$ are as given in (3.5), (3.4), and (3.1). For any ε in $(0,1)$, if $-A(N)$ is the leading behavior of

$$\sum_{n=3}^{N} \int_0^\varepsilon u\, e^{-nu}\, d\left[\left(g^{-1}(u)\right)^2\right]$$

(notice that $A(N) > 0$) and $B(N)$ is the leading behavior of

$$\sum_{n=2}^{N} \int_0^\varepsilon u\, e^{-nu} d\left[\left(g^{-1}(u)\right)\right],$$

then as $N \to \infty$,

$$E_\alpha\left[\mathcal{V}(T_N)\right] \sim \mu^2 N^2 \left[A(N) - B(N)^2\right] \tag{3.22}$$

provided

$$A(N) - B(N)^2 \neq o\left(A(N) + B(N)^2\right). \tag{3.23}$$

Example 2. Gamma distribution with parameters $p > 2$, $a > 0$ and $g(t) = (a/(a+t))^p$ (see, Remark 1, (iii)). We have

$$g^{-1}(t) = a\, t^{-1/p} - a \quad \text{and} \quad g'(g^{-1}(u)) = -\frac{p}{a}\, u^{1+\frac{1}{p}}.$$

Hence, for ε in $(0,1)$

$$\sum_{n=2}^{N} \int_0^{\varepsilon} u\, e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]} = \left(\frac{-a}{p}\right) \sum_{n=2}^{N} \int_0^{\varepsilon} u^{-1/p} e^{-nu}\, du.$$

Now substitute $nu = t$ and then by the definition of the Gamma function, and the Euler-Maclaurin summation formula one has

$$B(N) = (-a)\, \Gamma\left(1 - \frac{1}{p}\right) N^{1/p}. \tag{3.24}$$

In a similar way $A(N)$ of Theorem 3.2 becomes

$$A(N) = a^2\, \Gamma\left(1 - \frac{2}{p}\right) N^{2/p}. \tag{3.25}$$

Since the Gamma function is log-convex we have $\Gamma(1-2/p) > \Gamma(1-1/p)^2$. Thus, from (3.22) we have as $N \to \infty$

$$E_\alpha\left[\mathcal{V}(T_N)\right] \sim \mu^2 a^2 \left[\Gamma\left(1 - \frac{2}{p}\right) - \Gamma\left(1 - \frac{1}{p}\right)^2\right] N^{2+2/p}. \tag{3.26}$$

Remark 2. Condition (3.23) may not hold for certain distributions. In such cases one has to go further in the asymptotics of the quantity appearing in (3.16), i.e.,

$$\int_0^{\varepsilon} g^{-1}(u)\, u\, e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]}, \quad \text{for any } \varepsilon \text{ in } (0,1)$$

as illustrated by the following example.

Example 3. Let $g(t) \sim \exp\left(-ct^{\delta}\right)$ as $t \to \infty$, where $c > 0$ and $0 < \delta \leq 1$ ($g(t)$ cannot decay faster than an exponential). We furthermore assume that

$$\frac{d}{du}\left[g^{-1}(u)\right] \sim -\frac{1}{\delta c^{1/\delta}} \frac{(-\ln u)^{(1/\delta)-1}}{u} \qquad \text{as } u \to 0^+$$

(this is in agreement with the given asymptotics of $g(t)$ as $t \to \infty$). For any ε in $(0,1)$ we have

$$J_1(n) := -\int_0^{\varepsilon} g^{-1}(u)\, u\, e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]} = \frac{1}{\delta c^{2/\delta}} \int_0^{\varepsilon} e^{-nu} (-\ln u)^{(2/\delta)-1}\, du.$$

(notice that this is the quantity appearing in (3.16)). Substituting $nu = t$ and expanding the resulting power yields

$$J_1(n) = \frac{(\ln n)^{(2/\delta)-1}}{\delta\, c^{2/\delta} n} \int_0^{n\varepsilon} e^{-t} \left\{ 1 - \left(\frac{2}{\delta} - 1\right) \frac{\ln t}{\ln n} + \left(\frac{2}{\delta} - 1\right)\left(\frac{2}{\delta} - 2\right) \frac{\ln^2 t}{2 \ln^2 n} \left[1 + O\left(\frac{1}{\ln n}\right)\right]\right\} dt. \tag{3.27}$$

But (see, e.g., Boros and Moll (2004))

$$\int_0^\infty e^{-t} \ln t\, dt = \Gamma'(1) = -\gamma \quad \text{and} \quad \int_0^\infty e^{-t} \ln^2 t\, dt = \Gamma''(1) = \frac{\pi^2}{6} + \gamma^2,$$

where $\Gamma(\,\cdot\,)$ denotes the Gamma function and γ is the Euler's constant. Hence (3.27) yields (with exponentially small errors, which are clearly dominated by the error term of (3.27))

$$J_1(n) = \frac{1}{\delta\, c^{2/\delta}} \frac{(\ln n)^{(2/\delta)-1}}{n} + \frac{\gamma}{\delta\, c^{2/\delta}} \left(\frac{2}{\delta} - 1\right) \frac{(\ln n)^{(2/\delta)-2}}{n}$$

$$+ \frac{\left(\frac{2}{\delta} - 1\right)\left(\frac{\pi^2}{6} + \gamma^2\right)}{2\,\delta\, c^{2/\delta}} \left(\frac{2}{\delta} - 2\right) \frac{(\ln n)^{(2/\delta)-3}}{n} \left[1 + O\left(\frac{1}{\ln n}\right)\right].$$

Summing from $n = 3$ to N and applying the Euler-Maclaurin summation formula (see, e.g., Bender and Orszag (1978)) yields

$$\sum_{n=3}^{N} J_1(n) = \frac{1}{2\, c^{2/\delta}} (\ln N)^{2/\delta} + \frac{\gamma}{\delta\, c^{2/\delta}} (\ln N)^{(2/\delta)-1} + \frac{\left(\frac{2}{\delta} - 1\right)\left(\frac{\pi^2}{6} + \gamma^2\right)}{2\,\delta\, c^{2/\delta}} (\ln N)^{(2/\delta)-2} + o\,(\ln N)^{(2/\delta)-2}. \tag{3.28}$$

Set

$$J_2(n) := \int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu}\, du \quad \text{and} \quad J_3(n) := \int_0^\varepsilon g^{-1}(u)\, u\, e^{-nu} g'(g^{-1}(u))\, du$$

(these quantities are the second and the third summands appearing in (3.15)). In a similar way one has that

$$\sum_{n=3}^{N} J_2(n) \quad \text{and} \quad \sum_{n=3}^{N} J_3(n) \quad \text{stay bounded as } N \to \infty.$$

Starting from (3.19) we see that the asymptotics of $\sum_{n=3}^{N} L(n)$ are determined, again, by formula (3.28). Hence, we have detailed asymptotics for Q_N of (3.4). Let

$$J_0(n) := -\int_0^\varepsilon u\, e^{-nu} \frac{du}{g'\left[g^{-1}(u)\right]}$$

(this is the quantity appearing in (3.6)). If we treat $J_0(n)$ as we treated $J_1(n)$ we get

$$\sum_{n=2}^{N} J_0(n) = \frac{1}{c^{1/\delta}} (\ln N)^{1/\delta} + \frac{\gamma}{\delta\, c^{1/\delta}} (\ln N)^{(1/\delta)-1} + \frac{\left(\frac{1}{\delta}-1\right)\left(\frac{\pi^2}{6}+\gamma^2\right)}{2\,\delta\, c^{1/\delta}} (\ln N)^{(1/\delta)-2} + o\,(\ln N)^{(1/\delta)-2}. \tag{3.29}$$

Thus, we have detailed asymptotics for m_N of (3.5). In conclusion (1.6) yields

$$E_\alpha\left[\mathcal{V}(T_N)\right] \sim \frac{\pi^2}{6}\mu^2 N^2 \frac{(\ln N)^{(2/\delta)-2}}{\delta^2\, c^{2/\delta}}. \tag{3.30}$$

Notice that, if we take $\delta = 1$, then $g(t) = \exp(-ct)$ and this forces a_n to be deterministic, i.e. $a_n = c = \mu$ a.s, (see Example 1).

4. Concluding remark

The main topic of this paper was the asymptotics of the expectation of $\mathcal{V}(T_N)$, given the p_n's, in the case where the independent random variables a_n, are identically distributed with common distribution function (see formula (1.5)). Our approach was based on the asymptotics of the expectation of $\overline{T_N(T_N+1)}$ (given the p_n's). It is notable that if the random probability measure $\pi_N = (p_1, p_2 \cdots, p_N)$ is concentrated near $(1/N, 1/N \cdots, 1/N)$ for large N, then Theorem 3.2 is not enough to give asymptotics of $E_\alpha\left[\mathcal{V}(T_N)\right]$ and hence, one has to compute more terms in the asymptotic expansions of Q_N and m_N as illustrated by Example 3.

Acknowledgements

We thank the anonymous referee for carefully reading our manuscript and for making various constructive comments and suggestions.

References

C.M. Bender and S.A. Orszag. *Advanced mathematical methods for scientists and engineers*. McGraw-Hill Book Co., New York (1978). ISBN 0-07-004452-X. International Series in Pure and Applied Mathematics. MR538168.

A. Boneh and M. Hofri. The coupon-collector problem revisited—a survey of engineering problems and computational methods. *Comm. Statist. Stochastic Models* **13** (1), 39–66 (1997). MR1430927.

G. Boros and V. Moll. *Irresistible integrals*. Cambridge University Press, Cambridge (2004). ISBN 0-521-79636-9. Symbolics, analysis and experiments in the evaluation of integrals. MR2070237.

R.K. Brayton. On the asymptotic behavior of the number of trials necessary to complete a set with random selection. *J. Math. Anal. Appl.* **7**, 31–61 (1963). MR0158427.

P. Diaconis and S. Holmes. A Bayesian peek into Feller volume I. *Sankhyā Ser. A* **64** (3, part 2), 820–841 (2002). Special issue in memory of D. Basu. MR1981513.

A.V. Doumas and V.G. Papanicolaou. Asymptotics of the rising moments for the coupon collector's problem. *Electron. J. Probab.* **18**, 1–15 (2012a). DOI: 10.1214/EJP.v18-1746.

A.V. Doumas and V.G. Papanicolaou. The coupon collector's problem revisited: asymptotics of the variance. *Adv. in Appl. Probab.* **44** (1), 166–195 (2012b). MR2951551.

W. Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney (1968). MR0228020.

L. Holst. On birthday, collectors', occupancy and other classical urn problems. *Internat. Statist. Rev.* **54** (1), 15–27 (1986). MR959649.

L. Holst. Extreme value distributions for random coupon collector and birthday problems. *Extremes* **4** (2), 129–145 (2002) (2001). MR1893869.

V.G. Papanicolaou, G.E. Kokolakis and S. Boneh. Asymptotics for the random coupon collector problem. *J. Comput. Appl. Math.* **93** (2), 95–105 (1998). MR1638006.

S.M. Ross. *Initiation aux probabilités.* Presses Polytechniques et Universitaires Romandes, Lausanne (1996). ISBN 2-88074-327-3. With a preface by Peter Nüesch, Translated from the 4th (1994) English edition by Christian Hofer and Frédéric Dorsaz. MR1426831.