

Dixie cup problem in an interlacing process

Aristides V. Doumas*

*Dedicated to the memory of professor Miltos Makris.
Always a cadet, brother, and Best Friend.*

Abstract

The “double Dixie cup problem” of D.J. Newman and L. Shepp is a well-known variant of the coupon collector’s problem, where the object of study is the number of coupons that a collector has to buy in order to complete m sets of all N existing different coupons. In this paper we consider the case where the coupons distribution is a *mixture* of two different distributions, where the coupons from the first distribution are far rarer than the ones coming from the second. We apply a Poissonization technique, as well as well known results and techniques from our previous work, to derive the asymptotics (leading term) of the expectation of the above random variable as $N \rightarrow \infty$ for large classes of distributions. As it turns out, both distributions contribute to this result. The leading asymptotics of the rising moments of the aforementioned random variable are also discussed. We conclude by generalizing the problem to the case where the family of coupons is a mixture of j subfamilies.

Keywords. Urn problems; Double Dixie cup problem; Coupon collector’s problem; interlacing processes, asymptotics, Euler-Maclaurin summation formula, Schur concave functions.

MSC 2022 Mathematics Classification. 40-08; 40G99; 65A16.

1 Introduction

The “double Dixie cup problem” of D.J. Newman and L. Shepp [16] refers to a population whose members are of N different types (coupons).¹ For $j = 1, 2, \dots, N$, we denote by p_j the probability that a randomly chosen coupon is of type j , where $p_j > 0$ and $\sum_{j=1}^N p_j = 1$. The members of the population are sampled independently with replacement and their types are recorded. Let $\alpha = \{a_j\}_{j=1}^\infty$ be a sequence of strictly positive numbers. Then, for each positive

*National Technical University of Athens, Department of Mathematics, Zografou Campus, 157 80 Athens, GREECE, and Archimedes/Athena Research Center, GREECE. adou@math.ntua.gr, aris.doumas@hotmail.com

¹It is entertaining that the famous American girl group The Dixie Cups was active at the beginning of the 1960s, i.e., around the time of [16].

integer N one can create a probability measure $\pi_N = \{p_1, \dots, p_N\}$ on the set of types $\{1, \dots, N\}$ by taking

$$p_j = \frac{a_j}{A_N}, \text{ where } A_N = \sum_{j=1}^N a_j. \quad (1)$$

The number $T_m(N; \alpha)$ of trials needed until all N types are detected (at least m times) is, naturally, of interest. In this paper we consider the case where the sequence α is formed as the *interlacing* of two subsequences and derive the asymptotics (leading term) of the expectation of the random variable $T_m(N; \alpha)$ as $N \rightarrow \infty$.

Applications of the classic problem arise in many areas of science, such as computer science (search algorithms), physics, biology, ecology, earth and planetary sciences, economics and finance, as well as linguistics, demography, and the social sciences; see, e.g., [2]. The variant of the problem studied here is often applied in the same contexts, since there are cases where half of the coupons follow a different distribution (and are in fact rarer) than the other half. In cybersecurity, the Dixie cup problem can model the process of discovering all vulnerabilities in a network. Each unique vulnerability corresponds to a coupon, and each security scan to an attempt to collect one. Since not all vulnerabilities are equally likely, it is natural to distinguish two categories: common vulnerabilities, which are easily detected by standard scanning tools, and rare vulnerabilities, which are harder to uncover and follow a separate probability distribution. This framework allows one to predict the number of scans required to achieve a complete collection (the case where $m = 1$), i.e., to identify every vulnerability in the system. A similar application arises in epidemiology, where scientists track the emergence of virus strains within a population. Some strains are common, while others are rare or geographically localized, and the process of collecting samples to detect all strains can likewise be modeled by this problem. We now return to the above-mentioned random variable $T_m(N; \alpha)$. Its average w.r.t. to the sequence α will be denoted by $E[T_m(N; \alpha)]$. It is well known (see, e.g., [5]) that

$$E[T_m(N; \alpha)] = \int_0^\infty \left[1 - \prod_{j=1}^N \left(1 - S_m(p_j t) e^{-p_j t} \right) \right] dt, \quad (2)$$

where $S_m(y)$ is the m -th partial sum of e^y , namely

$$S_m(y) = \sum_{i=0}^{m-1} \frac{y^i}{i!}.$$

From now on we assume that N is even and consider the real sequences $\beta := \{b_j\}_{j=1}^\infty$ and $\delta := \{d_j\}_{j=1}^\infty$ such that

$$a_{2j-1} = b_j \quad \text{and} \quad a_{2j} = d_j, \quad j = 1, 2, \dots \quad (3)$$

For convenience we set

$$N := 2M. \quad (4)$$

Hence,

$$A_N = B_M + D_M, \quad (5)$$

where

$$B_M := \sum_{j=1}^M b_j \quad \text{and} \quad D_M := \sum_{j=1}^M d_j. \quad (6)$$

In words, the set of N coupons consists of two different sets, each one having M coupons with the corresponding coupon probabilities

$$q_j = \frac{b_j}{B_M} \quad \text{and} \quad \tilde{q}_j = \frac{d_j}{D_M}. \quad (7)$$

Notice that, when $M \rightarrow \infty$, then, of course, $N \rightarrow \infty$; thanks to (4). In this paper we consider *growing* sequences $\beta := \{b_j\}_{j=1}^\infty$, such that

$$b_j \rightarrow \infty \quad (8)$$

and *decaying* sequences $\delta := \{d_j\}_{j=1}^\infty$ of the form

$$d_j = \frac{1}{f(j)}, \quad (9)$$

where

$$f(x) > 0 \quad \text{and} \quad f'(x) > 0, \quad x > 0. \quad (10)$$

Furthermore we assume that $f(x)$ possesses three derivatives and satisfies the following conditions as $x \rightarrow \infty$:

$$(i) \ f(x) \rightarrow \infty, \quad (ii) \ \frac{f'(x)}{f(x)} \rightarrow 0, \quad (iii) \ \frac{f''(x)/f'(x)}{f'(x)/f(x)} = O(1), \quad (iv) \ \frac{f'''(x) f(x)^2}{f'(x)^3} = O(1). \quad (11)$$

In words $f(\cdot)$ belongs to the class of positive and strictly increasing $C^3(0, \infty)$ functions, which grow to ∞ (as $x \rightarrow \infty$) slower than exponentials, but faster than powers of logarithms. Examples of the growing sequences β for are: j^p , $p > 0$, e^{pj} , $p > 0$, $j!$, and for the sequences δ are: $j^{-p} (\ln j)^{-q}$, $p > 0$, $q \in \mathbb{R}$, e^{-j^r} , where, $0 < r < 1$. This work *also covers* the case where one of the two sequences β or δ is the uniform distribution. Recently and just for the case where $m = 1$, the scenario where the sequence α is the mixing of two subsequences one of which is constant (this corresponds to a uniform subcollection of coupons), while the other obeys the well-known *Zipf* law has been considered; we will come back to this case later. Hence, we may say that the coupons produced by the sequence δ are *far rarer* than the ones from the sequence β . We conclude this section by mentioning that $T_m(M; \beta)$ denotes the number of trials needed to obtain m complete sets of the M coupons generated by the sequence β via relation (7). In other words, in the Dixie cup problem concerning m complete sets of M different coupons, each coupon has probability $q_j = b_j/B_M$. (Similarly, one defines $T_m(M; \delta)$).

2 Asymptotics of $T_m(N; \alpha)$

Motivated by the seminal work of Holst (see, [12]) we consider a homogeneous Poisson process $\{N_t\}_{t \geq 0}$ with intensity one. Each event of this process is a collected coupon, and is classified as type j with probability p_j . Set

$$N_t := X_t + Y_t,$$

where $\{X_t\}_{t \geq 0}$ and $\{Y_t\}_{t \geq 0}$ are two independent Poisson processes, such that each Poisson event of them is a collected coupon coming from the sequences β and δ respectively, thanks to (5) and (6). (This Poissonization technique is also presented in detail in [17]). Recall that from our assumptions the coupons coming from the sequence δ are sufficiently rarer than the ones coming from the sequence β . Hence, as $N \rightarrow \infty$ with probability 1 the last coupon collected will be from the sequence δ . Let W_j denote the random variable representing the number of arrivals of the process $\{X_t\}_{t \geq 0}$ between two consecutive arrivals of $\{Y_t\}_{t \geq 0}$. Then, W_j 's are i.i.d random variables. Moreover, it is an easy exercise for one to see that they are geometrically distributed with parameter $\frac{D_M}{A_N}$, namely:

$$W_j \sim \text{Geo}\left(\frac{D_M}{A_N}\right), \quad j = 1, 2, \dots, \quad (12)$$

(where we have used (5)). Hence, one has

$$T_m(N; \alpha) = \sum_{j=1}^{T_m(M; \delta)} W_j, \quad \text{a.s.}, \quad N \rightarrow \infty.$$

It follows that

$$\begin{aligned} E[T_m(N; \alpha)] &= E[T_m(M; \delta)] E[W_j] \\ &= E[T_m(M; \delta)] \left(\frac{D_M + B_M}{D_M} \right) \\ &= E[T_m(M; \delta)] \left(\frac{A_N}{D_M} \right), \end{aligned} \quad (13)$$

where we have used Wald's lemma in the first equation above (since the finite mean random variable $T_m(M; \delta)$ is a stopping time with respect to the finite mean sequence $\{W_j\}_{j \in \mathbb{N}}$), the distribution of W_j 's in the second, and relation (5) in the last one. Since,

$$E[T_m(M; \delta)] = \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(\tilde{q}_j t) e^{-\tilde{q}_j t} \right) \right] dt,$$

and $\tilde{q}_j = \frac{d_j}{D_M}$ the change of variables $t = u D_M$ yields

$$E[T_m(M; \delta)] = D_M \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(d_j u) e^{-d_j u} \right) \right] du. \quad (14)$$

By invoking (14) in (13) one has proved the following

Theorem 1 *Let the sequence $\alpha = \{a_j\}_{j=1}^\infty$ is formed as the interlacing of two subsequences $\beta = \{b_j\}_{j=1}^\infty$ and $\delta = \{d_j\}_{j=1}^\infty$, as given in relation (3). Then, as $N = 2M \rightarrow \infty$ we have*

$$E[T_m(N; \alpha)] \sim (D_M + B_M) \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(d_j u) e^{-d_j u} \right) \right] du \quad (15)$$

where,

$$B_M := \sum_{j=1}^M b_j \quad \text{and} \quad D_M := \sum_{j=1}^M d_j,$$

provided that the coupons from the sequence δ are far rarer than the ones from the sequence β (see, (8)–(11)).

From here and in what follows, the asymptotic notation used in relation (15) has the meaning that $a_n \sim b_n$, as $n \rightarrow \infty$ means that $\lim_{n \rightarrow \infty} a_n/b_n = 1$. Relation (15) is the *key formula* to derive leading asymptotics for $E[T_m(N; \alpha)]$ as $N \rightarrow \infty$. It follows that our problem can be treated as *two separate problems*, namely gaining inside the order of magnitude of

$$A_N = D_M + B_M$$

and approximating the integral appearing in (15), i.e.:

$$L_1(M; \delta) := \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(d_j u) \right) e^{-d_j u} \right] du.$$

As we will see later, the number m *does not* appear in the leading term of the average as $M \rightarrow \infty$. Intitutevely, m is invisible in the leading order, since the first few coupons are easy to collect, while the last few missing coupon types (the “hard part”) dominates the total time.

Step I. The good news is that asymptotics of the quantity A_N (as $M \rightarrow \infty$, i.e., as $N \rightarrow \infty$) can be handled by existing powerful methods, such as the celebrated Euler-Maclaurin summation formula (see, e.g., [1]). In particular, the leading term in the asymptotic expansion of the sum $\sum_{j=1}^M l_j$ (for some sequence $\lambda := \{l_j\}_{j=1}^\infty$) depends on the behaviour of the series $\sum_{j=1}^\infty l_j$. If

$$C_\lambda := \sum_{j=1}^\infty l_j < \infty,$$

then

$$\sum_{j=1}^M l_j \sim C_\lambda, \quad M \rightarrow \infty,$$

while, in the case where

$$C_\lambda = \infty,$$

we have

$$\sum_{j=1}^M l_j \sim \int_1^M l(x) dx, \quad M \rightarrow \infty. \quad (16)$$

where $l(x)$ is the real function associated with the sequence $\{l_j\}_{j=1}^\infty$. To mention a few examples we have (as $M \rightarrow \infty$):

$$\sum_{j=1}^M j^p \sim \frac{M^{p+1}}{p+1}, \quad p > 0, \text{ (positive power law),}$$

and

$$\sum_{j=1}^M j^{-p} \sim \begin{cases} \frac{M^{1-p}}{1-p} & \text{if } p \in (0, 1), \\ \ln M & \text{if } p = 1, \\ \zeta(p) & \text{if } p > 1, \end{cases} \quad (\text{generalized Zipf law})$$

where $\zeta(\cdot)$ is the Riemann's Zeta function. Another interesting example is the one where $l_j = (\ln j)^{-p}$, $p > 0$. Here, $C_\lambda = \infty$. Hence, (16) and integration by parts yields

$$\sum_{j=2}^M l_j \sim \frac{M}{(\ln M)^p}, \quad M \rightarrow \infty.$$

In particular, for $p = 1$ we get the so-called *offset logarithmic integral* or *Eulerian logarithmic integral*, namely,

$$Li(M) := \int_2^M \frac{dt}{\ln t}$$

which is a very good approximation to the number of prime numbers less than M (as $M \rightarrow \infty$).

Let us now consider the case where $l_j = \frac{1}{j!}$, $j = 0, 1, \dots$. Since, $\sum_{j=0}^\infty \frac{1}{j!} = e$ we have

$$\sum_{j=0}^M \frac{1}{j!} \sim e, \quad M \rightarrow \infty.$$

Remark 1. Notice that there are cases where the Euler-Maclaurin summation formula is *not* effective. For example, if $l_j = e^{jc_j}$, where c_j is increasing and $\lim_{j \rightarrow \infty} c_j = \infty$. Then,

$$\sum_{j=1}^M e^{jc_j} \sim e^{Mc_M}, \quad M \rightarrow \infty. \quad (17)$$

In words, the last term in the above sum cancels all the previous terms. To prove (17) observe that

$$\sum_{j=1}^M e^{jc_j} \leq C e^{(M+1)c_M},$$

for some positive constant C and

$$C e^{(M+1)c_M} = o\left(e^{(M+1)c_{M+1}}\right), \quad M \rightarrow \infty,$$

and the desired result follows.

Having information for the quantity A_N is critical regarding our analysis for the asymptotics of relation (15).

Step II. On the other hand a substantial amount of information is known for the behaviour of the quantity $L_1(M; \delta)$. In particular, the following dichotomy holds (see, [5]):

$$\begin{aligned} \text{(Case I)} \quad \lim_{M \rightarrow \infty} L_1(M; \delta) < \infty, \quad \text{iff} \quad \sum_{j=1}^{\infty} e^{-d_j \xi} < \infty \quad \text{for some } \xi > 0, \\ \text{(Case II)} \quad \lim_{M \rightarrow \infty} L_1(M; \delta) = \infty, \quad \text{iff} \quad \sum_{j=1}^{\infty} e^{-d_j \xi} = \infty \quad \text{for all } \xi > 0. \end{aligned}$$

(Case II) has been studied for *rich classes* of sequences, namely those which satisfy conditions ((9)–(11)). Then, (see, [4], [5], and, [6])

$$L_1(M; \delta) \sim f(M) \ln \left(\frac{f(M)}{f'(M)} \right), \quad M \rightarrow \infty. \quad (18)$$

In fact the first five terms of the asymptotic expansion of $L_1(M; \delta)$ are known as $M \rightarrow \infty$. Moreover, these classes were extended in the case where

$$d_j = (\ln j)^{-p}, \quad p > 0,$$

see, [7] for details. At the end of the paper we present a table including the leading term (asymptotically) for the quantities $\sum_{j=1}^M l_j$, $L_1(M; \lambda)$, and, $E[T_m(M; \lambda)]$, for several special distributions.

Conclusion. Under conditions (8)–(11), and having formula (15) (of Theorem 1), as well as powerful tools as the Euler-Maclaurin summation formula, and tangible results for *rich* classes of distributions (see, relation (18) and Table 1), the leading asymptotics of $E[T_m(N; \alpha)]$ in the case where the sequence α is the “union” of two subsequences β and δ is easy to compute (as $N \rightarrow \infty$). To obtain these asymptotics one has: (i) to deduce which one of the subsequences β and δ produces the rare set of coupons, (ii) to compute leading asymptotics for the quantity $D_M + B_M$ (see relation (15) and apply the methods described in Step

1 above), and finally, (iii) complete the work by using the asymptotic behaviour of the quantity $L_1(M; \delta)$ (see, e.g., Table 1 below), where δ is the subsequence of the rare coupons as we have already indicated. Let us also note, that both distributions contribute to our final answer (see, Subsection 2.1 below).

Remark 2. Let $r \geq 1$, be an integer. Set

$$T_m(N; \alpha)^{(r)} := T_m(N; \alpha) (T_m(N; \alpha) + 1) (T_m(N; \alpha) + 2) \cdots (T_m(N; \alpha) + r - 1).$$

Similarly, for the r -th rising moment of the random variable $T_m(N; \alpha)$, that is, $E[T_m(N; \alpha)^{(r)}]$, one has (see, [4], [5])

$$E[T_m(N; \alpha)^{(r)}] = r \int_0^\infty \left[1 - \prod_{j=1}^N \left(1 - S_m(p_j t) e^{-p_j t} \right) \right] t^{r-1} dt, \quad r = 1, 2, \dots,$$

and easily get the leading asymptotics of these moments as $N \rightarrow \infty$. We will give closure to this section with the following (kind of) general

Proposition 2 *Suppose that we fix the distribution of the half of the coupons. Then the quantity $E[T_m(N; \alpha)^{(r)}]$ becomes minimum when the other half of the coupons are uniformly distributed.*

Proof. Without loss of generality we may assume that the distribution of the coupons is a mixture of the two distributions generating from the sequences c_j and b_j . In view of (2) we have

$$\begin{aligned} E[T_m(N; \alpha)^{(r)}] &= r \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m\left(\frac{b_j}{A_N} t\right) \right) e^{-\frac{b_j}{A_N} t} \prod_{j=1}^M \left(1 - S_m\left(\frac{c_j}{A_N} t\right) \right) e^{-\frac{c_j}{A_N} t} \right] t^{r-1} dt \\ &= r A_N^r \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(b_j t) \right) e^{-b_j t} \prod_{j=1}^M \left(1 - S_m(c_j t) \right) e^{-c_j t} \right] t^{r-1} dt. \end{aligned}$$

Let us fix the distribution generated by the sequence c_j and set

$$h(b_1, b_2, \dots, b_M) := \sum_{j=1}^M \ln \left[\left(1 - S_m(b_j t) \right) e^{-b_j t} \right].$$

Clearly, the function h is symmetric under all permutations of its variables. Moreover, it is not hard to see that, for all $i, j = 1, 2, \dots, M$, with $i \neq j$:

$$(b_i - b_j) \left(\frac{\theta h(b_1, \dots, b_M)}{b_i} - \frac{\theta h(b_1, \dots, b_M)}{b_j} \right) \leq 0.$$

Hence, h is a Schur concave function. It follows that h attains its maximum value, when all b_j 's (which are strictly positive) are equal, (see, for details [15], pp.84 and 413). This completes the proof. ■

For a concrete instance of Proposition 2, we refer the reader to Figure 2, where it has been assumed that the total number of coupons ranges from 20 to 200. The distribution of the odd coupons is assumed to be fixed; in particular, in view of relation (3), we take $a_{2j-1} = j$. As for the distribution of the even coupons, it is sequentially assumed that they follow either the uniform distribution ($a_{2j} = 1$), the Zipf distribution ($a_{2j} = j^{-2}$), or a positive power law ($a_{2j} = j^3$). The number of trials for one complete set ($m = 1$) of coupons is smaller when ($a_{2j} = 1$).

2.1 Examples

Here we illustrate the previous arguments with a few characteristic examples. Notice, that the number m *does not* appear in the leading term of the average $E[T_m(N; \alpha)]$ as $N \rightarrow \infty$.

Example 1. If half of the coupons are coming from the *positive power law* distribution with (positive) parameter r and the rest of the coupons are coming from the *Zipf law*, then,

$$\beta := \{b_j\}_{j=1}^{\infty} = j^r, \quad r > 0, \quad \delta := \{d_j\}_{j=1}^{\infty} = j^{-p}, \quad p > 0.$$

In words, the rare coupons are coming from the *Zipf* distribution. Hence, from formulae (8–11) and the results from Table 1 we get (recall that $N = 2M$)

$$E[T_m(N; \alpha)] \sim \frac{N^{r+p+1}}{2^{r+p+1}(r+1)} \ln N, \quad N \rightarrow \infty.$$

Example 2. Let us consider the case of *exponential - uniform* problem. Then, clearly

$$\beta := \{b_j\}_{j=1}^{\infty} = e^{pj}, \quad p > 0, \quad \delta := \{d_j\}_{j=1}^{\infty} = 1.$$

The sequence $\delta := \{d_j\}_{j=1}^{\infty}$ does not satisfy conditions (9–(11)). Hence, we can not apply Theorem 1. Though, using relation (13) and the well known results (from Table 1) one immediately has

$$E[T_m(N; \alpha)] \sim \frac{e^{p(M+1)}}{e^p - 1} \ln M, \quad M \rightarrow \infty.$$

Example 3. The *uniform versus Zipf distribution* scenario. This problem (for the case of $m = 1$) has been studied extensively in ([9]), where the authors derived the first and second moment of $T_m(N; \alpha)$ up to the fifth and sixth (asymptotically) term in order to obtain the corresponding variance as $M \rightarrow \infty$,

and the distribution, as well. Following the notation of our paper the rare coupons are coming from the Zipf law. We have

$$E[T_m(N; \alpha)] \sim \frac{N^{p+1} \ln N}{2^{p+1}}, \quad N \rightarrow \infty, \quad (19)$$

in accordance with the results of [9]. Figure 1 illustrates the case where half of the coupons are uniformly distributed, while the rest of the coupons follow the standard Zipf distribution. Simulations using Python present $E[T_m(N; \alpha)]$ in comparison with the theoretical result (see relation (19)).

Example 4. Suppose now that half of the coupons are coming from the *logarithmic law* with parameter $r > 0$ and the rest of the coupons are coming from the *Zipf law*. In this case one has:

$$\delta := \{d_j\}_{j=1}^{\infty} = j^{-p}, \quad p \geq 1, \quad \beta := \{b_j\}_{j=2}^{\infty} = (\ln j)^{-r}, \quad r > 0.$$

Hence, from formula (15) and the results from Table 1 we get (since $N = 2M$)

$$E[T_m(N; \alpha)] \sim \frac{N^{p+1}}{2^{p+1} (\ln N)^{r-1}}, \quad N \rightarrow \infty. \quad (20)$$

We will give closure with the following

Example 5. Consider the classic coupon collector for the sequences $\beta := \{b_j\}_{j=0}^{\infty} = e^{pj}$ and $\delta := \{d_j\}_{j=0}^{\infty} = e^{-pj}$, where $p > 0$, and observe that the elements of these sequences are proportional to each other. Hence, they produce the same coupon probabilities. This follows from the fact that for each M , if we let $c_M = e^{-pM}$, then $\{b_j : 0 \leq j \leq M\} = \{c_M d_j : 0 \leq j \leq M\}$. Thus,

$$\begin{aligned} E[T_m(M; \beta)] &= E[T_m(M; \delta)] \\ &\sim C \frac{e^{p(M+1)}}{e^p - 1}, \quad M \rightarrow \infty, \end{aligned}$$

where C is a constant. The sequence δ does not satisfy conditions (ii) and (iii) of relation (11). Yet by applying relation (13) and the well known results (from Table 1) one immediately has (the expected asymptotic result)

$$E[T_m(N; \alpha)] \sim C \frac{e^{p(2M+1)}}{e^p - 1}, \quad M \rightarrow \infty. \quad (21)$$

3 Generalization

Let us now consider the case where the family of coupon probabilities is the *interlacing of three subfamilies* one of which consists of far rather coupons (in the sense of Section 1), than the ones of the other two families. For example, and following the notation of Section 1 we assume that $N = 3M$ and consider Zipf,

exponential, and uniform distributions, namely the sequences $\delta := \{d_j\}_{j=1}^\infty$, $\beta := \{b_j\}_{j=0}^\infty$, and, $\eta := \{n_j\}_{j=1}^\infty$, such that

$$\begin{aligned} a_{3j} &= b_j = e^{qj}, \quad q > 0, \quad j = 1, 2, \dots, \\ a_{3j+1} &= d_j = j^{-p}, \quad p > 1, \quad j = 1, 2, \dots, \\ a_{3j-1} &= n_j = 1, \quad j = 1, 2, \dots. \end{aligned}$$

Hence, $A_N = B_M + D_M + H_M$, where

$$D_M := \sum_{j=1}^M d_j, \quad B_M := \sum_{j=0}^{M-1} b_j \quad \text{and} \quad H_M := \sum_{j=1}^M n_j.$$

Clearly, the rare coupons come from the sequence δ , i.e., the Zipf distribution. Using similar arguments as in Section 2 one arrives at the following result

$$E[T_m(N; \alpha)] = (D_M + B_M + H_M) \int_0^\infty \left[1 - \prod_{j=1}^M \left(1 - S_m(d_j u) e^{-d_j u} \right) \right] du. \quad (22)$$

Again, the leading term of the average comes from the integral above, as well as, from the leading term of the quantity: $D_M + B_M + H_M$, which may be evaluated by the Euler-Maclaurin summation formula, and the rest of the tools presented in Section 2. For example, in our case one has

$$E[T_m(N; \alpha)] \sim \frac{e^{q(M+1)}}{e^q - 1} M^p \ln M, \quad M \rightarrow \infty,$$

where $N = 3M$. Observe that only the exponential and the Zipf distributions contribute to the leading term of $E[T_m(N; \alpha)]$, whereas the number of coupon sets m has no effect on the result.

Concluding Remarks. If our family of coupons is the interlacing of j sub-families, where j is any positive integer, our approach may easily be generalized with similar arguments. Let us also mention that by mimicking this approach one may derive the leading term of the moments of the r.v. (see, relation (13)). For example, for the second moment of the random variable $T_m(N; \alpha)$ one has

$$\begin{aligned} E[T_m(N; \alpha)^2] &= E \left[\sum_{j=1}^{T_m(M; \delta)} W_j^2 \right] + 2E \left[\sum_{i < j}^{T_m(M; \delta)} W_i W_j \right] \\ &= E[T_m(M; \delta)] E[W_j^2] + E[T_m(M; \delta) (T_m(M; \delta) - 1)] E[W_j]^2, \end{aligned}$$

and following the same steps one has the (leading) asymptotics as $M \rightarrow \infty$ (see, relation (12)). The limiting distribution of $T_m(N; \alpha)$ will be determined mainly by the distribution of the rare coupons. However, in order to obtain the appropriate normalization for $T_m(N; \alpha)$, one must delve deeper into the

asymptotics of the first and second moments. In our recent work [9], where half of the coupons are uniformly distributed and the other half follow the standard Zipf law, and in the case $m = 1$, the appropriately normalized random variable $T_m(N; \alpha)$ converges in distribution to a Gumbel random variable. To the best of our knowledge, this is the only published result concerning this variant of the coupon collector problem. We conjecture that, for the classes of distributions studied here, the result under the proper normalization will be the same.

Let us also mention that the distribution in the classic version of the problem (the uniform distribution) was first presented by P. Erdős and A. Rényi [10], and, by embedding a Poisson process, by L. Holst [12]. This work and its generalizations to many covering contexts are studied in the paper by Godbole, Grubb, and Kay [11], where interesting distributional corollaries are provided. Variants of the double Dixie cup problem, like the one presented here, could be applied to several other areas for future research, such as quantum search algorithms, cryptographic engineering, and combinatorial probability (see, e.g., [14], [13], and [18]).

4 Table of results and simulation

Table 1 below presents leading asymptotics for rich classes of distributions and has been used in the examples exhibited in Section 2. Hence, one may compute easily the leading behaviour of the random variable $E[T_m(N; \alpha)]$ of Theorem 1. Figure 1 exhibits repeated simulations in Python, in the case where half of the coupons follow uniform distribution, while the rest come from the standard Zipf distribution. Simulation versus theoretical results are presented for several values of the total number of coupons N . In Figure 2 one inspects the validity of Proposition 2, in the case where the distribution of the odd coupons is fixed. We have assumed that the even coupons follow uniform distribution (Case 1), power law (Case 2), and Zipf law (Case 3). $E[T_m(N; \alpha)]$ is minimized when the even coupons are uniformly distributed.

law $\lambda := \{l_j\}_{j=1}^\infty$	type	$\sum_{j=1}^M l_j$ (leading term)	$L_1(M; \lambda)$ (leading term)	$E[T_m(M; \lambda)]$ (leading term)	Ref.
1	uniform	M	$\ln M$	$M \ln M$	[10]
j^r	positive power law	$\frac{M^{r+1}}{r+1}$	constant	$C \frac{M^{r+1}}{r+1}$	[5], [3]
j^{-p}	Zipf law $0 < p < 1$	$\frac{M^{1-p}}{1-p}$	$M^p \ln M$	$\frac{M \ln M}{1-p}$	[5], [3], [9]
j^{-1}	std Zipf $p = 1$	$\ln M$	$M \ln M$	$M (\ln M)^2$	[5], [3], [9]
j^{-p}	Zipf law $p > 1$	$\zeta(p)$	$M^p \ln M$	$\zeta(p) M^p \ln M$	[5], [3], [9]
$(\ln j)^{-r}$	logarithmic	$\frac{M}{(\ln M)^r}$	$(\ln M)^{r+1}$	$M \ln M$	[7]
e^{pj}	exponential	$\frac{e^{p(M+1)}}{e^p - 1}$	constant	$C \frac{e^{p(M+1)}}{e^p - 1}$	[5], [3], [4]
e^{-pj}	exponential	$\frac{1}{1-e^{-p}}$	Example 5	$C \frac{e^{p(M+1)}}{e^p - 1}$	[5], [3], [4]
$e^{ic_j},$ $c_j \nearrow \infty$	super exponential	e^{Mc_M}	constant	$C e^{Mc_M}$	[5], [6], [4]
$j!$	factorial	$M!$	constant	$C e M!$	[5], [3], [4]
$\frac{1}{j!}$	reciprocal factorial	e	$C_1 M!$	$C M!$	[5], [3], [4]

Table 1: Leading asymptotics for $\sum_{j=1}^M l_j$, $L_1(M; \lambda)$, and, $E[T_m(M; \lambda)]$. The parameters p and r are strictly positive. As usual, we use C and C_1 to denote constants in any equation.

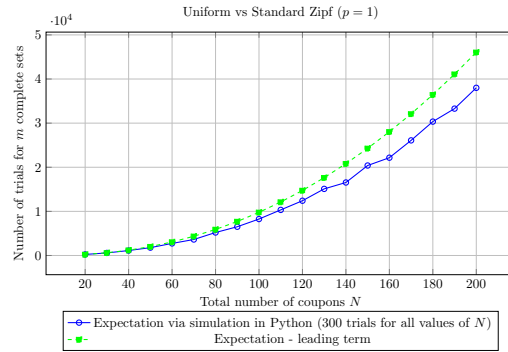


Figure 1: Comparison of simulated vs theoretical expectations

Acknowledgements. The author has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

References

- [1] C.M. Bender and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*, Springer-Verlag, New York, 1999.
- [2] A. Boneh and M. Hofri, The coupon collector problem revisited—A survey of engineering problems and computational methods, *Comm. Statist. Stochastic Models* **13**, (1997), no. 1, 39–66.
- [3] A.V. Doumas and V.G. Papanicolaou, The Coupon Collector’s Problem Revisited: Asymptotics of the Variance, *Adv. Appl. Prob.* **44** (1) (2012) 166–195.
- [4] A.V. Doumas and V.G. Papanicolaou, Asymptotics of the rising moments for the Coupon Collector’s Problem, *Electron. J. Probab.* **Vol. 18**, (Article no. 41) (2012) 1–15.
- [5] A.V. Doumas and V.G. Papanicolaou, The Coupon Collector’s Problem Revisited: Generalizing the Double Dixie Cup Problem of Newman and Shepp, *ESAIM: Probability and Statistics*, **20**, (2016) 367–399.
- [6] A.V. Doumas and V.G. Papanicolaou, Sampling from a Mixture of Different Groups of Coupons, *Acta Mathematica Sinica*, English Series 36 **12**, (2020) 1357–1383.
- [7] A.V. Doumas and V.G. Papanicolaou, The logarithmic Zipf law in a general urn problem, *ESAIM: Probability and Statistics*, **24**, (2020) 275–293.
- [8] A.V. Doumas and V.G. Papanicolaou, The Siblings of the Coupon Collector Problem, *Theory of Probability and Its Applications*, **62**, (3)(2018) 2 444–470.
- [9] A.V. Doumas and V.G. Papanicolaou, Uniform versus Zipf distribution in a mixing collection process, *Statistics and Probability letters*, 155 (2019) 1–7.
- [10] P. Erdős and A. Rényi, On a classical problem of probability theory, *Magyar. Tud. Akad. Mat. Kutató Int. Közl.*, **6**, (1961), 215–220.
- [11] A. Godbole, T. Grubb, K Han, and, B KaY, Threshold Progressions in a Variety of Covering and Packing Contexts, *Journal of Combinatorics* **13**, (2022), 303–331.
- [12] L. Holst, On Birthday, Collectors’, Occupancy and Other Classical Urn Problems, *International Statist. Review* **54**, (1986), 15–27.
- [13] A. Jana, A. Nath, G. Paul, G. et al., Differential fault analysis of NORX using variants of coupon collector problem., *J Cryptogr Eng* **12**, 433–459 (2022). <https://doi.org/10.1007/s13389-022-00285-y>.
- [14] M. Kessler, D. Alonso, and, P. Sánchez, Determination of the number of shots for Grover’s search algorithm, *EPJ Quantum Technol.*, **10**, 47 (2023).
- [15] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, New York, 1979.
- [16] D.J. Newman and L. Shepp, The double Dixie cup problem, *Amer. Math. Monthly* **67** (1960) 58–61. MR0120672

- [17] S. Ross, *Introduction to Probability Models*, Tenth Edition, Elsevier Inc., Burlington, MA, 2010.
- [18] Saunders, John C., The number of k -tons in the coupon collector problem, *J. Appl. Probab.* 60 (2023), no. 3, 723–736.