Optimization results for a generalized coupon collector problem

Emmanuelle Anceaume, Yann Busnel, Ernst Schulte-Geers, Bruno Sericola

Optimization results for a generalized coupon collector problem *

Emmanuelle Anceaume[1], Yann Busnel[2,4], Ernst Schulte-Geers[3], Bruno Sericola[4]

[1] CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France
[2] Ensai, Campus de Ker-Lann, BP 37203, 35172 Bruz, Cedex, France
[3] BSI, Godesberger Allee 185-189, 53175 Bonn, Germany
[4] Inria, Campus de Beaulieu, 35042 Rennes Cedex, France

Abstract

We study in this paper a generalized coupon collector problem, which consists in analyzing the time needed to collect a given number of distinct coupons that are drawn from a set of coupons with an arbitrary probability distribution. We suppose that a special coupon called the null coupon can be drawn but never belongs to any collection. In this context, we prove that the almost uniform distribution, for which all the non-null coupons have the same drawing probability, is the distribution which stochastically minimizes the time needed to collect a fixed number of distinct coupons. Moreover, we show that in a given closed subset of probability distributions, the distribution with all its entries, but one, equal to the smallest possible value is the one, which stochastically maximizes the time needed to collect a fixed number of distinct coupons. An computer science application shows the utility of these results.

Keywords

Coupon collector problem; Optimization; Schur-convex functions.

1 Introduction

The coupon collector problem is an old problem, which consists in evaluating the time needed to get a collection of different objects drawn randomly using a given probability distribution. This problem has given rise to a lot of attention from researchers in various fields since it has applications in many scientific domains including computer science and optimization, see [2] for several engineering examples.

More formally, consider a set of n coupons, which are drawn randomly one by one, with replacement, coupon i being drawn with probability p_i. The classical coupon collector problem is to determine the expectation or the distribution of the number of coupons that need to be drawn from the set of n coupons to obtain the full collection of the n coupons. A large number of papers have been devoted to the analysis of asymptotics and limit distributions of this distribution when n tends to infinity, see [4] or [8] and the references therein. In [3], the authors obtain new formulas concerning this distribution and they also provide simulation techniques to compute it as well as analytic bounds of it. The asymptotics of the rising moments are studied in [5].

We suppose in this paper that $p = (p_1, \ldots, p_n)$ is not necessarily a probability distribution, *i.e.*, we suppose that $\sum_{i=1}^{n} p_i \leq 1$ and we define $p_0 = 1 - \sum_{i=1}^{n} p_i$. This means that there is a null coupon, denoted by 0, which is drawn with probability p_0, but that does not belong to the collection. We are interested, in this setting, in the time needed to collect c different coupons

*This work was partially funded by the French ANR project SocioPlug (ANR-13-INFR-0003), and by the DeSceNt project granted by the Labex CominLabs excellence laboratory (ANR-10-LABX-07-01).

among coupons $1, \ldots, n$, when a coupon is drawn, with replacement, at each discrete time $1, 2, \ldots$ among coupons $0, 1, \ldots, n$. This time is denoted by $T_{c,n}(p)$ for $c = 1, \ldots, n$. Clearly, $T_{n,n}(p)$ is the time needed to get the full collection. The random variable $T_{c,n}(p)$ has been considered in [9] in the case where the drawing probability distribution is uniform. The expected value $\mathrm{E}[T_{c,n}(p)]$ has been obtained in [6] when $p_0 = 0$. Its distribution and its moments have been obtained in [1] using Markov chains.

In this paper, we prove that the almost uniform distribution, denoted by v and defined by $v = (v_1, \ldots, v_n)$ with $v_i = (1 - p_0)/n$, where p_0 is fixed, is the distribution which stochastically minimizes the time $T_{c,n}(p)$ such that $p_0 = 1 - \sum_{i=1}^{n} p_i$. This result was expressed as a conjecture in [1] where it is proved that the result is true for $c = 2$ and for $c = n$ extending the sketch of the proof proposed in [2] to the case $p_0 > 0$. It has been also proved in [1] that the result is true for the expectations, that is that $\mathrm{E}[T_{c,n}(u)] \leq \mathrm{E}[T_{c,n}(v)] \leq \mathrm{E}[T_{c,n}(p)]$.

We first consider in Section 2, the case where $p_0 = 0$ and then we extend it to one of $p_0 > 0$. We show moreover in Section 3, that in a given closed subset of probability distributions, the distribution with all its entries, but one, equal to the smallest possible value is the one which stochastically maximizes the time $T_{c,n}(p)$. This work is motivated by the worst case analysis of the behavior of streaming algorithms in network monitoring applications as shown in Section 4.

2 Distribution minimizing the distribution of $T_{c,n}(p)$

Recall that $T_{c,n}(p)$, with $p = (p_1, \ldots, p_n)$, is the number of coupons that need to be drawn from the set $\{0, 1, 2, \ldots, n\}$, with replacement, till one first obtains a collection with c different coupons, $1 \leq c \leq n$, among $\{1, \ldots, n\}$, where coupon i is drawn with probability p_i, $i = 0, 1, \ldots, n$.

The distribution of $T_{c,n}(p)$ has been obtained in [1] using Markov chains and is given by

$$\mathrm{P}\{T_{c,n}(p) > k\} = \sum_{i=0}^{c-1} (-1)^{c-1-i} \binom{n-i-1}{n-c} \sum_{J \in S_{i,n}} (p_0 + P_J)^k, \tag{1}$$

where $S_{i,n} = \{J \subseteq \{1, \ldots, n\} \mid |J| = i\}$ and, for every $J \subseteq \{1, \ldots, n\}$, P_J is defined by $P_J = \sum_{j \in J} p_j$. Note that we have $S_{0,n} = \emptyset$, $P_\emptyset = 0$ and $|S_{i,n}| = \binom{n}{i}$.

This result also shows as expected that the function $\mathrm{P}\{T_{c,n}(p) > k\}$, as a function of p, is symmetric, which means that it has the same value for any permutation of the entries of p.

We recall that if X and Y are two real random variable then we say that X is stochastically smaller (resp. larger) than Y, and we write $X \leq_{\mathrm{st}} Y$ (resp. $Y \leq_{\mathrm{st}} X$), if $\mathrm{P}\{X > t\} \leq \mathrm{P}\{Y > t\}$, for all real numbers t. This stochastic order is also referred to as the strong stochastic order.

We consider first, in the following subsection, the case where $p_0 = 0$.

2.1 The case $p_0 = 0$

This case corresponds the fact that there is no null coupon, which means that all the coupons can belong to the collection. We thus have $\sum_{i=1}^{n} p_i = 1$. For all $n \geq 1$, $i = 1, \ldots, n$ and $k \geq 0$, we denote by $N_i^{(k)}$ the number of coupons of type i collected at instants $1, \ldots, k$. It is well-known that the joint distribution of the $N_i^{(k)}$ is a multinomial distribution, *i.e.*, for all $k_1, \ldots, k_n \geq 0$ such that $\sum_{i=1}^{n} k_i = k$, we have

$$\mathrm{P}\{N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n\} = \frac{k!}{k_1! \cdots k_n!} p_1^{k_1} \cdots p_n^{k_n}. \tag{2}$$

We also denote by $U_n^{(k)}$ the number of distinct coupon's types, among $1, \ldots, n$, already drawn at instant k. We clearly have, with probability 1, $U_n^{(0)} = 0$, $U_n^{(1)} = 1$ and, for $i = 0, \ldots, n$,

$$\mathrm{P}\{U_n^{(k)} = i\} = \sum_{J \in S_{i,n}} \mathrm{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J\}.$$

Moreover, it is easily checked that,

$$T_{c,n}(p) > k \Longleftrightarrow U^{(k)} < c.$$

We then have

$$\begin{aligned} \mathrm{P}\{T_{c,n}(p) > k\} &= \mathrm{P}\{U_n^{(k)} < c\} \\ &= \sum_{i=0}^{c-1} \mathrm{P}\{U_n^{(k)} = i\} \\ &= \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \mathrm{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J\}. \end{aligned}$$

Using Relation (2), we obtain

$$\mathrm{P}\{T_{c,n}(p) > k\} = \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \sum_{\underline{k} \in E_{k,J}} k! \prod_{j \in J} \frac{p_j^{k_j}}{k_j!}, \tag{3}$$

where $E_{k,J}$ is the set of vectors defined by

$$E_{k,J} = \{\underline{k} = (k_j)_{j \in J} \mid k_j > 0, \text{ for all } j \in J \text{ and } K_J = k\},$$

with $K_J = \sum_{j \in J} k_j$.

Theorem 1 *For all $n \geq 2$ and $p = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i = 1$, and for all $c = 1, \ldots, n$, we have $T_{c,n}(p') \leq_{\mathrm{st}} T_{c,n}(p)$, where $p' = (p_1, \ldots, p_{n-2}, p'_{n-1}, p'_n)$ with $p'_{n-1} = \lambda p_{n-1} + (1-\lambda)p_n$ and $p'_n = (1-\lambda)p_{n-1} + \lambda p_n$, for all $\lambda \in [0,1]$.*

Proof. The result is trivial for $c = 1$, since we have $T_{1,n}(p) = 1$, and for $k = 0$ since both terms are equal to 1. We thus suppose now that $c \geq 2$ and $k \geq 1$. The fact that $k \geq 1$ means in particular that the term $i = 0$ in Relation (3) is equal to 0. We then have

$$\mathrm{P}\{T_{c,n}(p) > k\} = \sum_{i=1}^{c-1} \sum_{J \in S_{i,n}} \sum_{\underline{k} \in E_{k,J}} k! \prod_{j \in J} \frac{p_j^{k_j}}{k_j!}. \tag{4}$$

To simplify the notation, we denote by $T_i(p)$ the ith term of this sum, that is

$$T_i(p) = \sum_{J \in S_{i,n}} \sum_{\underline{k} \in E_{k,J}} k! \prod_{j \in J} \frac{p_j^{k_j}}{k_j!}. \tag{5}$$

For $i = 1$, we have $S_{1,n} = \{\{1\}, \ldots, \{n\}\}$ and $E_{k,\{j\}} = \{k\}$. The term $T_1(p)$ is thus given by

$$T_1(p) = \sum_{j=1}^{n} p_j^k.$$

Consider now the term T_i for $i \geq 2$. We split the set $S_{i,n}$ into four subsets depending on wether the indices $n-1$ and n belong to its elements. More precisely, we introduce the partition of the set $S_{i,n}$ in the four subsets $S_{i,n}^{(1)}$, $S_{i,n}^{(2)}$, $S_{i,n}^{(3)}$ and $S_{i,n}^{(4)}$ defined by

$$\begin{aligned} S_{i,n}^{(1)} &= \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ with } n-1 \in J \text{ and } n \notin J\}, \\ S_{i,n}^{(2)} &= \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ with } n-1 \notin J \text{ and } n \in J\}, \\ S_{i,n}^{(3)} &= \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ with } n-1 \in J \text{ and } n \in J\}, \\ S_{i,n}^{(4)} &= \{J \subseteq \{1, \ldots, n\} \mid |J| = i \text{ with } n-1 \notin J \text{ and } n \notin J\}. \end{aligned}$$

These subsets can also be written as $S^{(1)}_{i,n} = S_{i-1,n-2} \cup \{n-1\}$, $S^{(2)}_{i,n} = S_{i-1,n-2} \cup \{n\}$, $S^{(3)}_{i,n} = S_{i-2,n-2} \cup \{n-1, n\}$, and $S^{(4)}_{i,n} = S_{i,n-2}$. Using these equalities, the term T_i of Relation (5) becomes

$$\begin{aligned} T_i(p) &= \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in E_{k,J\cup\{n-1\}}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \frac{p_{n-1}^{k_{n-1}}}{k_{n-1}!} \\ &+ \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in E_{k,J\cup\{n\}}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \frac{p_n^{k_n}}{k_n!} \\ &+ \sum_{J \in S_{i-2,n-2}} \sum_{\underline{k} \in E_{k,J\cup\{n-1,n\}}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \frac{p_{n-1}^{k_{n-1}}}{k_{n-1}!} \frac{p_n^{k_n}}{k_n!} \\ &+ \sum_{J \in S_{i,n-2}} \sum_{\underline{k} \in E_{k,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right). \end{aligned}$$

Let us now introduce the sets $L_{k,J}$ defined by

$$L_{k,J} = \{\underline{k} = (k_j)_{j\in J} \mid k_j > 0, \text{ for all } j \in J \text{ and } K_J \le k\}.$$

Using these sets and, to clarify the notation, setting $k_{n-1} = \ell$ and $k_n = h$ when needed, we obtain

$$\begin{aligned} T_i(p) &= \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in L_{k-1,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \frac{p_{n-1}^{k-K_J}}{(k-K_J)!} \\ &+ \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in L_{k-1,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \frac{p_n^{k-K_J}}{(k-K_J)!} \\ &+ \sum_{J \in S_{i-2,n-2}} \sum_{\underline{k} \in L_{k-2,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \sum_{\ell>0,h>0,\ell+h=k-K_J} \frac{p_{n-1}^{\ell}}{\ell!} \frac{p_n^{h}}{h!} \\ &+ \sum_{J \in S_{i,n-2}} \sum_{\underline{k} \in E_{k,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right), \end{aligned}$$

which can also be written as

$$\begin{aligned} T_i(p) &= \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in L_{k-1,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \left(p_{n-1}^{k-K_J} + p_n^{k-K_J}\right) \\ &+ \sum_{J \in S_{i-2,n-2}} \sum_{\underline{k} \in L_{k-2,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) (p_{n-1} + p_n)^{k-K_J} \\ &- \sum_{J \in S_{i-2,n-2}} \sum_{\underline{k} \in L_{k-2,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right) \left(p_{n-1}^{k-K_J} + p_n^{k-K_J}\right) \\ &+ \sum_{J \in S_{i,n-2}} \sum_{\underline{k} \in E_{k,J}} k! \left(\prod_{j\in J} \frac{p_j^{k_j}}{k_j!}\right). \end{aligned}$$

We denote these four terms respectively by $A_i(p)$, $B_i(p)$, $C_i(p)$ and $D_i(p)$. We thus have, for $i \ge 2$, $T_i(p) = A_i(p) + B_i(p) - C_i(p) + D_i(p)$. We have already shown that $T_1(p) = A_1(p) + D_1(p)$, so we

set $B_1(p) = C_1(p) = 0$. We then have

$$\begin{aligned} \mathrm{P}\{T_{c,n}(p) > k\} &= \sum_{i=1}^{c-1} T_i(p) \\ &= A_{c-1}(p) + \sum_{i=1}^{c-2}(A_i(p) - C_{i+1}(p)) + \sum_{i=2}^{c-1} B_i(p) + \sum_{i=1}^{c-1} D_i(p). \end{aligned} \tag{6}$$

For $i \geq 1$, we have

$$\begin{aligned} A_i(p) - C_{i+1}(p) &= \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in L_{k-1,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j \in J} \frac{p_j^{k_j}}{k_j!} \right) \left(p_{n-1}^{k-K_J} + p_n^{k-K_J} \right) \\ &\quad - \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in L_{k-2,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j \in J} \frac{p_j^{k_j}}{k_j!} \right) \left(p_{n-1}^{k-K_J} + p_n^{k-K_J} \right) \\ &= \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in E_{k-1,J}} \frac{k!}{(k-K_J)!} \left(\prod_{j \in J} \frac{p_j^{k_j}}{k_j!} \right) \left(p_{n-1}^{k-K_J} + p_n^{k-K_J} \right). \end{aligned}$$

By definition of the set $E_{k-1,J}$, we have $K_J = k-1$ in the previous equality. This gives

$$A_i(p) - C_{i+1}(p) = \sum_{J \in S_{i-1,n-2}} \sum_{\underline{k} \in E_{k-1,J}} k! \left(\prod_{j \in J} \frac{p_j^{k_j}}{k_j!} \right) (p_{n-1} + p_n).$$

Using the fact that the function $x \longmapsto x^s$ is convex on interval $[0,1]$ for every non negative integer s, we have

$$\begin{aligned} {p'}_{n-1}^{k-K_J} + {p'}_n^{k-K_J} &= (\lambda p_{n-1} + (1-\lambda)p_n)^{k-K_J} + ((1-\lambda)p_{n-1} + \lambda p_n)^{k-K_J} \\ &\leq \lambda p_{n-1}^{k-K_J} + (1-\lambda)p_n^{k-K_J} + (1-\lambda)p_{n-1}^{k-K_J} + \lambda p_n^{k-K_J} \\ &= p_{n-1}^{k-K_J} + p_n^{k-K_J}, \end{aligned}$$

and in particular $p'_{n-1} + p'_n = p_{n-1} + p_n$. It follows that

$$\begin{aligned} &A_{c-1}(p') \leq A_{c-1}(p), \\ &A_i(p') - C_{i+1}(p') = A_i(p) - C_{i+1}(p), \\ &B_i(p') = B_i(p), \\ &D_i(p') = D_i(p), \end{aligned}$$

and from (6) that $\mathrm{P}\{T_{c,n}(p') > k\} \leq \mathrm{P}\{T_{c,n}(p) > k\}$, which concludes the proof. ■

The function $\mathrm{P}\{T_{c,n}(p) > k\}$, as a function of p, being symmetric, this theorem can easily be extended to the case where the two entries p_{n-1} and p_n of p are any $p_i, p_j \in \{p_1, \ldots, p_n\}$, with $i \neq j$.

In fact, we have shown in this theorem that for fixed n and k, the function of p, $\mathrm{P}\{T_{c,n}(p) \leq k\}$, is a Schur-convex function, that is, a function that preserves the order of majorization. See [7] for more details on this subject.

Theorem 2 *For every $n \geq 1$ and $p = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i = 1$, and for all $c = 1, \ldots, n$, we have $T_{c,n}(u) \leq_{\mathrm{st}} T_{c,n}(p)$, where $u = (1/n, \ldots, 1/n)$ is the uniform distribution.*

Proof. To prove this result, we apply successively and at most $n-1$ times Theorem 1 as follows. We first choose two different entries of p, say p_i and p_j such that $p_i < 1/n < p_j$ and next to define p'_i and p'_j by

$$p'_i = \frac{1}{n} \text{ and } p'_j = p_i + p_j - \frac{1}{n}.$$

This leads us to write $p'_i = \lambda p_i + (1-\lambda)p_j$ and $p'_j = (1-\lambda)p_i + \lambda p_j$, with

$$\lambda = \frac{p_j - 1/n}{p_j - p_i}.$$

From Theorem 1, vector p' obtained by taking the other entries equal to those of p, *i.e.*, by taking $p'_\ell = p_\ell$, for $\ell \neq i, j$, is such that $\mathrm{P}\{T_{c,n}(p') > k\} \leq \mathrm{P}\{T_{c,n}(p) > k\}$. Note that at this point vector p' has at least one entry equal to $1/n$, so repeating at most $n-1$ this procedure, we get vector u, which concludes the proof. ■

To illustrate the steps used in the proof of this theorem, we take the following example. Suppose that $n = 5$ and $p = (1/16, 1/6, 1/4, 1/8, 19/48)$. In a first step, taking $i = 4$ and $j = 5$, we get

$$p^{(1)} = (1/16, 1/6, 1/4, 1/5, 77/240).$$

In a second step, taking $i = 2$ and $j = 5$, we get

$$p^{(2)} = (1/16, 1/5, 1/4, 1/5, 69/240).$$

In a third step, taking $i = 1$ and $j = 3$, we get

$$p^{(3)} = (1/5, 1/5, 9/80, 1/5, 69/240).$$

For the fourth and last step, taking $i = 3$ and $j = 5$, we get

$$p^{(4)} = (1/5, 1/5, 1/5, 1/5, 1/5).$$

2.2 The case $p_0 > 0$

We consider now the case where $p_0 > 0$. We have $p = (p_1, \ldots, p_n)$ with $p_1 + \cdots + p_n < 1$ and $p_0 = 1 - (p_1 + \cdots + p_n)$. Recall that in this case $T_{c,n}(p)$ is the time or the number of steps needed to collect a subset of c different coupons among coupons $1, \ldots, n$. Coupon 0 is not allowed to belong to the collection. As in the previous subsection we denote, for $i = 0, 1, \ldots, n$ and $k \geq 0$, by $N_i^{(k)}$ the number of coupons of type i collected at instants $1, \ldots, k$. We then have for all $k_0, k_1, \ldots, k_n \geq 0$ such that $\sum_{i=0}^n k_i = k$, we have

$$\mathrm{P}\{N_0^{(k)} = k_0, N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n\} = \frac{k!}{k_0! k_1! \cdots k_n!} p_0^{k_0} p_1^{k_1} \cdots p_n^{k_n},$$

which can also be written as

$$\begin{aligned} \mathrm{P}\{N_0^{(k)} &= k_0, N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n\} \\ &= \binom{k}{k_0} p_0^{k_0} (1-p_0)^{k-k_0} \frac{(k-k_0)!}{k_1! \cdots k_n!} \left(\frac{p_1}{1-p_0}\right)^{k_1} \cdots \left(\frac{p_n}{1-p_0}\right)^{k_n}. \end{aligned} \tag{7}$$

Note that $p/(1-p_0)$ is a probability distribution since

$$\frac{1}{1-p_0} \sum_{i=1}^n p_i = 1,$$

so summing over all the $k_1, \ldots, k_n \geq 0$ such that $\sum_{i=1}^n k_i = k - k_0$, we get, for all $k_0 = 0, \ldots, k$,

$$\mathrm{P}\{N_0^{(k)} = k_0\} = \binom{k}{k_0} p_0^{k_0} (1-p_0)^{k-k_0}. \tag{8}$$

From (7) and (8) we obtain, for all $k_1, \ldots, k_n \geq 0$ such that $\sum_{i=1}^n k_i = k - k_0$,

$$\mathrm{P}\{N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n \mid N_0^{(k)} = k_0\} = \frac{(k-k_0)!}{k_1! \cdots k_n!} \left(\frac{p_1}{1-p_0}\right)^{k_1} \cdots \left(\frac{p_n}{1-p_0}\right)^{k_n}. \tag{9}$$

Recall that $U_n^{(k)}$ is the number of coupon's types among $1, \ldots, n$ already drawn at instant k. We clearly have, with probability 1, $U_n^{(0)} = 0$, and, for $i = 0, \ldots, n$,

$$\mathrm{P}\{U_n^{(k)} = i\} = \sum_{J \in S_{i,n}} \mathrm{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J\}.$$

Moreover, we have as in Subsection 2.1,

$$T_{c,n}(p) > k \Longleftrightarrow U_n^{(k)} < c.$$

and so

$$\begin{aligned}\mathrm{P}\{T_{c,n}(p) > k \,|\, N_0^{(k)} = k_0\} \\ = \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \mathrm{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J \mid N_0^{(k)} = k_0\}.\end{aligned}$$

Using Relation (9), we obtain

$$\mathrm{P}\{T_{c,n}(p) > k \mid N_0^{(k)} = k_0\} = \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \sum_{\underline{k} \in E_{k-k_0,J}} (k-k_0)! \left(\prod_{j \in J} \frac{\left(\frac{p_j}{1-p_0}\right)^{k_j}}{k_j!} \right), \tag{10}$$

where the set $E_{k,J}$ has been defined in Subsection 2.1.

Theorem 3 *For every $n \geq 1$ and $p = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i < 1$, and for all $c = 1, \ldots, n$, we have $T_{c,n}(u) \leq_{\mathrm{st}} T_{c,n}(v) \leq_{\mathrm{st}} T_{c,n}(p)$, where $u = (1/n, \ldots, 1/n)$, $v = (v_1, \ldots, v_n)$ with $v_i = (1-p_0)/n$ and $p_0 = 1 - \sum_{i=1}^n p_i$.*

Proof. From Relation (3) and Relation (10), we obtain, for all $k_0 = 0, \ldots, k$,

$$\mathrm{P}\{T_{c,n}(p) > k \mid N_0^{(k)} = k_0\} = \mathrm{P}\{T_{c,n}(p/(1-p_0)) > k - k_0\}. \tag{11}$$

Using (8) and unconditioning, we obtain

$$\mathrm{P}\{T_{c,n}(p) > k\} = \sum_{\ell=0}^{k} \binom{k}{\ell} p_0^{\ell} (1-p_0)^{k-\ell} \mathrm{P}\{T_{c,n}(p/(1-p_0)) > k - \ell\}. \tag{12}$$

Since $p/(1-p_0)$ is a probability distribution, applying Theorem 2 to this distribution and observing that $u = v/(1-p_0)$, we get

$$\begin{aligned}
\mathrm{P}\{T_{c,n}(p) > k\} &= \sum_{\ell=0}^{k} \binom{k}{\ell} p_0^{\ell} (1-p_0)^{k-\ell} \mathrm{P}\{T_{c,n}(p/(1-p_0)) > k - \ell\} \\
&\geq \sum_{\ell=0}^{k} \binom{k}{\ell} p_0^{\ell} (1-p_0)^{k-\ell} \mathrm{P}\{T_{c,n}(u) > k - \ell\} \\
&= \sum_{\ell=0}^{k} \binom{k}{\ell} p_0^{\ell} (1-p_0)^{k-\ell} \mathrm{P}\{T_{c,n}(v/(1-p_0)) > k - \ell\} \\
&= \mathrm{P}\{T_{c,n}(v) > k\},
\end{aligned}$$

where the last equality follows from (12). This proves the second inequality.

To prove the first inequality, observe that $\mathrm{P}\{T_{c,n}(p/(1-p_0)) > \ell\}$ is decreasing with ℓ. This leads, using (12), to

$$\mathrm{P}\{T_{c,n}(p) > k\} \geq \mathrm{P}\{T_{c,n}(p/(1-p_0)) > k\}.$$

Taking $p = v$ in this inequality gives

$$\mathrm{P}\{T_{c,n}(v) > k\} \geq \mathrm{P}\{T_{c,n}(u) > k\},$$

which completes the proof. ■

In fact, we have shown in the proof of this theorem and more precisely in Relation (11) and using Theorem 2 that for fixed n, k and k_0, the function of p, $\mathrm{P}\{T_{c,n}(p) \leq k \mid N_0^{(k)} = k_0\}$ is a Schur-convex function, that is, a function that preserves the order of majorization. In particular, from (12), $\mathrm{P}\{T_{c,n}(p) \leq k\}$ is also a Schur-convex function, even when $p_0 > 0$. See [7] for more details on this subject.

3 Distribution maximizing the distribution of $T_{c,n}(p)$

We consider in this section the problem of stochastically maximizing the time $T_{c,n}(p)$ when p varies in a closed subset of dimension n. In the previous section the minimization was made on the set $\mathcal{A}$ defined, for every n and for all $p_0 \in [0, 1)$, by

$$\mathcal{A} = \{p = (p_1, \ldots, p_n) \in (0,1)^n \mid p_1 + \cdots + p_n = 1 - p_0\}.$$

According to the application described in the Section 4, we fix a parameter $\theta \in (0, (1-p_0)/n]$ and we are looking for distributions p which stochastically maximizes the time $T_{c,n}(p)$ on the subset $\mathcal{A}_\theta$ of $\mathcal{A}$ defined by

$$\mathcal{A}_\theta = \{p = (p_1, \ldots, p_n) \in \mathcal{A} \mid p_j \geq \theta, \text{ for every } j = 1, \ldots, n\}.$$

The solution to this problem is given by the following theorem. We first introduce the set $\mathcal{B}_\theta$ defined by the distributions of $\mathcal{A}_\theta$ with all their entries, except one, are equal to θ. The set $\mathcal{B}_\theta$ has n elements given by

$$\mathcal{B}_\theta = \{(\gamma, \theta, \ldots, \theta), (\theta, \gamma, \theta, \ldots, \theta), \ldots, (\theta, \ldots, \theta, \gamma)\},$$

where $\gamma = 1 - p_0 - (n-1)\theta$. Note that since $\theta \in (0, (1-p_0)/n]$, we have $1 - p_0 - (n-1)\theta \geq \theta$ which means that $\mathcal{B}_\theta \subseteq \mathcal{A}_\theta$.

Theorem 4 *For every $n \geq 1$ and $p = (p_1, \ldots, p_n) \in \mathcal{A}_\theta$ and for all $c = 1, \ldots, n$, we have $T_{c,n}(p) \leq_{\mathrm{st}} T_{c,n}(q)$, for every $q \in \mathcal{B}_\theta$.*

Proof. Since $\mathrm{P}\{T_{c,n}(p) > k\}$ is a symmetric function of p, $\mathrm{P}\{T_{c,n}(q) > k\}$ has the same value for every $q \in \mathcal{B}_\theta$, so we suppose that $q_\ell = \theta$ for every $\ell \neq j$ and $q_j = \gamma$.

Let $p \in \mathcal{A}_\theta \setminus \mathcal{B}_\theta$ and let i be the first entry of p such that $i \neq j$ and $p_i > \theta$. We then define the distribution $p^{(1)}$ as $p_i^{(1)} = \theta$, $p_j^{(1)} = p_i + p_j - \theta > p_j$ and $p_\ell^{(1)} = p_\ell$, for $\ell \neq i, j$. This leads us to write $p_i = \lambda p_i^{(1)} + (1-\lambda) p_j^{(1)}$ and $p_j = (1-\lambda) p_i^{(1)} + \lambda p_j^{(1)}$, with

$$\lambda = \frac{p_j^{(1)} - p_i}{p_j^{(1)} - p_i^{(1)}} = \frac{p_j - \theta}{p_j - \theta + p_i - \theta} \in [0, 1).$$

From Theorem 1, we get $\mathrm{P}\{T_{c,n}(p) > k\} \leq \mathrm{P}\{T_{c,n}(p^{(1)}) > k\}$. Repeating the same procedure from distribution $p^{(1)}$ and so on, we get, after at most $n-1$ steps, distribution q, that is $\mathrm{P}\{T_{c,n}(p) > k\} \leq \mathrm{P}\{T_{c,n}(q) > k\}$, which completes the proof. ■

To illustrate the steps used in the proof of this theorem, we take the following example. Suppose that $n = 5$, $\theta = 1/20$, $p_0 = 1/10$ and $q = (1/20, 1/20, 1/20, 7/10, 1/20)$, which means that $j = 4$. Suppose moreover that $p = (1/16, 1/6, 1/4, 1/8, 71/240)$. We have $p \in \mathcal{A}_\theta$ and $q \in \mathcal{B}_\theta$. In a first step, taking $i = 1$ and since $j = 4$, we get

$$p^{(1)} = (1/20, 1/6, 1/4, 11/80, 71/240).$$

In a second step, taking $i = 2$ and since $j = 4$, we get

$$p^{(2)} = (1/20, 1/20, 1/4, 61/240, 71/240).$$

In a third step, taking $i = 3$ and since $j = 4$, we get

$$p^{(3)} = (1/20, 1/20, 1/20, 109/240, 71/240).$$

For the fourth and last step, taking $i = 5$ and since $j = 4$, we get

$$p^{(4)} = (1/20, 1/20, 1/20, 7/10, 1/20) = q.$$

4 Application to the detection of distributed denial of service attacks

A Denial of Service (DoS) attack tries to progressively take down an Internet resource by flooding this resource with more requests than it is capable to handle. A Distributed Denial of Service (DDoS) attack is a DoS attack triggered by thousands of machines that have been infected by a malicious software, with as immediate consequence the total shut down of targeted web resources (*e.g.*, e-commerce websites). A solution to detect and to mitigate DDoS attacks is to monitor network traffic at routers and to look for highly frequent signatures that might suggest ongoing attacks. A recent strategy followed by the attackers is to hide their massive flow of requests over a multitude of routes, so that locally, these requests do not appear as frequent, while globally each of these requests represent a portion θ of the network traffic. The term "iceberg" has been recently introduced [10] to describe such an attack as only a very small part of the iceberg can be observed at each single router. The approach adopted to defend against such new attacks is to rely on multiple routers that locally monitor their network traffic, and upon detection of potential icebergs, inform a monitoring server that aggregates all the monitored information to accurately detect icebergs. Now to prevent the server from being overloaded by all the monitored information, routers continuously keep track of the c (among n) most recent distinct requests. These requests locally represent at least a fraction θ of the local stream. Once collected, each router sends them to the server, and throw away all the requests i that appear with a probability p_i smaller than θ. The sum of these small probabilities is represented by probability p_0. Parameter c is dimensioned so that the frequency at which all the routers send their c last requests is low enough to enable the server to aggregate all of them and to trigger a DDoS alarm when needed. This amounts to compute the distribution of the time $T_{c,n}(p)$ needed to collect c distinct requests among the n ones. Theorem 3 shows that the distribution p that stochastically minimizes the time $T_{c,n}(p)$ is the almost uniform distribution v. This means that if locally each router receives a stream in which all the frequent requests (that is, those whose probability of occurrence is greater than or equal to θ) occur with same probability, then the complementary distribution of the time needed to locally complete a collection of c distinct requests is minimized. As a consequence the delay between any two interactions between a router and the server is minimized.

Another important aspect of DDoS detection applications is their ability to bound the detection latency of global icebergs, that is the maximal time that elapses between the presence of a global iceberg at some of the routers and its detection at the server. This can be implemented through a timer that will fire if no communication has been triggered between a router and the server, which happens if locally a router has received less than c requests. Dimensioning such a timer amounts to determine the distribution of the maximal time it takes for a router to collect c distinct requests. Theorem 4 shows that the distributions p that stochastically maximizes this time $T_{c,n}(p)$ are all the distributions $q \in \mathcal{B}_\theta$.

References

[1] Anceaume, E., Busnel, Y. and Sericola, B. (2015). New results on a generalized coupon collector problem using Markov chains. *J. Appl. Prob.* **52(2)**.

[2] Boneh, A. and Hofri, M. (1997). The coupon-collector problem revisited - A survey of engineering problems and computational methods. *Stochastic Models* **13(1)**, pp. 39–66.

[3] Brown, M., Peköz, E. A. and Ross, S. M. (2008). Coupon collecting. *Probability in the Engineering and Informational Sciences* **22**, pp. 221–229.

[4] Doumas, A. V. and Papanicolaou, V. G. (2012). The coupon collector's problem revisited: asymptotics of the variance. *Adv. Appl. Prob.* **44(1)**, pp. 166–195.

[5] Doumas, A. V. and Papanicolaou, V. G. (2012). Asymptotics of the rising moments for the coupon collector's problem. *Electronic Journal of Probability* **18(41)**, pp. 1–15.

[6] Flajolet, P., Gardy, D. and Thimonier, L (1992). Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics* **39**, pp. 207–229.

[7] Marshall, A. W. and Olkin, I. (1981). *Inequalities via majorization – An introduction*, Technical Report No. 172, Department of Statistics, Stanford University, California, USA.

[8] Neal, P. (2008). The generalised coupon collector problem. *J. Appl. Prob.* **45(3)**, pp. 621–629.

[9] Rubin, H. and Zidek, J. (1965). *A waiting time distribution arising from the coupon collector's problem*, Technical Report No. 107, Department of Statistics, Stanford University, California, USA.

[10] Zhao, Q and Lall, A. and Ogihara, M. and Xu, J. (2010). *Global iceberg detection over Distributed Streams*, Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE), Long Beach, CA, USA, March 1-6.