# 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,* *CNRS*

YANN BUSNEL,** *Ensai*

ERNST SCHULTE-GEERS,*** *BSI*

BRUNO SERICOLA,**** *Inria*

**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.

*Keywords:* Coupon collector problem; Optimization; Schur-convex functions

2010 Mathematics Subject Classification: Primary 60C05
Secondary 60J05

* Postal address: CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France
** Postal address: Ensai, Campus de Ker-Lann, BP 37203, 35172 Bruz, Cedex, France
*** Postal address: BSI, Godesberger Allee 185-189, 53175 Bonn, Germany
**** Postal address: Inria, Campus de Beaulieu, 35042 Rennes Cedex, France

## 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 [3] 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 moments 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 such analysis when $n$ tends to infinity; see [4] and the references therein.

We suppose in this paper that $\boldsymbol{p} = (p_1, \ldots, p_n)$ 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 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 $T_{c,n}(\boldsymbol{p})$ for $c = 1, \ldots, n$. Clearly, $T_{n,n}(\boldsymbol{p})$ is the time needed to get the full collection. The random variable $T_{c,n}(\boldsymbol{p})$ has been considered in [7] in the case where the drawing probability distribution is uniform. The expected value $\mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$ has been obtained in [5] 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, defined by $\boldsymbol{v} = (v_1, \ldots, v_n)$ with $v_i = (1 - p_0)/n$, where $p_0$ is fixed, is the distribution which stochastically minimizes the time $T_{c,n}(\boldsymbol{p})$ when $p_0 = 1 - \sum_{i=1}^n p_i$. 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 [3] to the case $p_0 > 0$. It has been proved in [1] that the result is true for the expectations, that is that $\mathbb{E}\{T_{c,n}(\boldsymbol{u})\} \leq \mathbb{E}\{T_{c,n}(\boldsymbol{v})\} \leq \mathbb{E}\{T_{c,n}(\boldsymbol{p})\}$, where $\boldsymbol{u} = (1/n, \ldots, 1/n)$ is the uniform distribution.

We first consider in Section 2, the case where $p_0 = 0$ and then we extend it to

the one where $p_0 > 0$. 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 $T_{c,n}(\boldsymbol{p})$. This work is motivated by the worst case analysis of the behavior of streaming algorithms in network monitoring applications as shown in [2].

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

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

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

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 that the function $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}$, as a function of $\boldsymbol{p}$, is symmetric, meaning that it has the same value for any permutation of the entries of $\boldsymbol{p}$.

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

### 2.1. The case $p_0 = 0$

This case corresponds to 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

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

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

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

It is easily checked that $T_{c,n}(\boldsymbol{p}) > k \Longleftrightarrow U^{(k)} < c$. We then have using Relation (2),

$$\begin{aligned}
\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} &= \mathbb{P}\{U_n^{(k)} < c\} = \sum_{i=0}^{c-1} \mathbb{P}\{U_n^{(k)} = i\} \\
&= \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \mathbb{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J\} \\
&= \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \sum_{\boldsymbol{k} \in E_{k,J}} k! \prod_{j \in J} \frac{p_j^{k_j}}{k_j!}, \qquad (3)
\end{aligned}$$

where $E_{k,J} = \{\boldsymbol{k} = (k_j)_{j \in J} \mid k_j > 0, \text{ for all } j \in J \text{ and } K_J = k\}$, with $K_J = \sum_{j \in J} k_j$.

**Theorem 1.** *For all $n \geq 2$ and $\boldsymbol{p} = (p_1, \ldots, p_n) \in (0,1)^n$ with $\sum_{i=1}^n p_i = 1$, and for all $c = 1, \ldots, n$, we have $T_{c,n}(\boldsymbol{p}') \leq_{\mathrm{st}} T_{c,n}(\boldsymbol{p})$, where $\boldsymbol{p}' = (p_1, \ldots, p_{n-2}, p'_{n-1}, p'_n)$ 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 $T_{1,n}(\boldsymbol{p}) = 1$ for all $\boldsymbol{p}$. Moreover, we have $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > 0\} = 1$ for all $\boldsymbol{p}$. We thus suppose now that $c \geq 2$ and $k \geq 1$. The fact that $k \geq 1$ implies that the term $i = 0$ in Relation (3) is equal to 0. We then have

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

To simplify the notation, we denote by $T_i(\boldsymbol{p})$ the $i$th term of this sum, that is

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

For $i = 1$, we have $S_{1,n} = \{\{1\}, \ldots, \{n\}\}$ and $E_{k,\{j\}} = \{k\}$, thus, $T_1(\boldsymbol{p}) = \sum_{j=1}^n p_j^k$. For $i \geq 2$ we introduce the following partition of the set $S_{i,n}$.

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

These subsets can also be written as $S_{i,n}^{(1)} = S_{i-1,n-2} \cup \{n-1\}$, $S_{i,n}^{(2)} = S_{i-1,n-2} \cup \{n\}$,

$S_{i,n}^{(3)} = S_{i-2,n-2} \cup \{n-1, n\}$, and $S_{i,n}^{(4)} = S_{i,n-2}$. The term $T_i(\boldsymbol{p})$ of Relation (5) becomes

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

We introduce the sets $L_{k,J} = \{\boldsymbol{k} = (k_j)_{j \in J} \mid k_j > 0, \text{ for all } j \in J \text{ and } K_J \leq k\}$. To clarify the notation, setting $k_{n-1} = \ell$ and $k_n = h$ when needed, we obtain

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

which can also be written as

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

We denote these four terms respectively by $A_i(\boldsymbol{p})$, $B_i(\boldsymbol{p})$, $C_i(\boldsymbol{p})$ and $D_i(\boldsymbol{p})$. We thus have, for $i \geq 2$, $T_i(\boldsymbol{p}) = A_i(\boldsymbol{p}) + B_i(\boldsymbol{p}) - C_i(\boldsymbol{p}) + D_i(\boldsymbol{p})$. We have already shown that $T_1(\boldsymbol{p}) = A_1(\boldsymbol{p}) + D_1(\boldsymbol{p})$, so we set $B_1(\boldsymbol{p}) = C_1(\boldsymbol{p}) = 0$. We then have

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} \quad = \quad A_{c-1}(\boldsymbol{p}) + \sum_{i=1}^{c-2}(A_i(\boldsymbol{p}) - C_{i+1}(\boldsymbol{p})) + \sum_{i=2}^{c-1} B_i(\boldsymbol{p}) + \sum_{i=1}^{c-1} D_i(\boldsymbol{p}). \quad (6)$$

For $i \geq 1$, we obtain

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

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

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

The function $x \longmapsto x^s$ being convex on interval $[0,1]$ for every $s \in \mathbb{N}$, 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 $A_{c-1}(\boldsymbol{p}') \leq A_{c-1}(\boldsymbol{p})$, $A_i(\boldsymbol{p}') - C_{i+1}(\boldsymbol{p}') = A_i(\boldsymbol{p}) - C_{i+1}(\boldsymbol{p})$, $B_i(\boldsymbol{p}') = B_i(\boldsymbol{p})$, $D_i(\boldsymbol{p}') = D_i(\boldsymbol{p})$, and from (6) that $\mathbb{P}\{T_{c,n}(\boldsymbol{p}') > k\} \leq \mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}$, which concludes the proof.

The function $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}$ being symmetric, this theorem easily extends 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 $\boldsymbol{p}$, $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) \leq k\}$ is a Schur-convex function, that is, a function that preserves the order of majorization; see [6] for more details on this subject.

**Theorem 2.** *For all $n \geq 1$, $\boldsymbol{p} \in (0,1)^n$ and $c = 1, \ldots, n$ we have $T_{c,n}(\boldsymbol{u}) \leq_{\mathrm{st}} T_{c,n}(\boldsymbol{p})$.*

*Proof.* We apply successively and at most $n-1$ times Theorem 1 as follows. We first choose two different entries of $\boldsymbol{p}$, say $p_i$ and $p_j$ such that $p_i < 1/n < p_j$ and then we define $p'_i$ and $p'_j$ by $p'_i = 1/n$ and $p'_j = p_i + p_j - 1/n$. 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 $\boldsymbol{p}'$, which is obtained by taking the other entries equal to those of $\boldsymbol{p}$, i.e. by taking $p'_\ell = p_\ell$, for $\ell \neq i, j$, is such that $\mathbb{P}\{T_{c,n}(\boldsymbol{p}') > k\} \leq \mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\}$. Note that at this point vector $\boldsymbol{p}'$ has at least one entry equal to $1/n$, so repeating at most $n-1$ this procedure, we get vector $\boldsymbol{u}$, which concludes the proof.

### 2.2. The case $p_0 > 0$

We consider now the case where $p_0 > 0$. We have $\boldsymbol{p} = (p_1, \ldots, p_n)$ with $\sum_{i=1}^n p_i < 1$ and $p_0 = 1 - \sum_{i=1}^n p_i$. Recall that in this case $T_{c,n}(\boldsymbol{p})$ 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. The number $N_i^{(k)}$ of coupons of type $i$ collected at instants $1, \ldots, k$ follows the binomial distribution with parameters $k$ and $p_i$. Moreover, for all $k_0, k_1, \ldots, k_n \geq 0$ such that $\sum_{i=0}^n k_i = k$, we have

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

It follows that for all $k_1, \ldots, k_n \geq 0$ such that $\sum_{i=1}^n k_i = k - k_0$,

$$\mathbb{P}\{N_1^{(k)} = k_1, \ldots, N_n^{(k)} = k_n \mid N_0^{(k)} = k_0\} = \frac{(k-k_0)!}{k_1! \cdots k_n!} \left(\frac{p_1}{1-p_0}\right)^{k_1} \cdots \left(\frac{p_n}{1-p_0}\right)^{k_n}. \tag{7}$$

As in Subsection 2.1, we have $T_{c,n}(\boldsymbol{p}) > k \Longleftrightarrow U_n^{(k)} < c$ and so using (7), we obtain

$$\begin{aligned}
&\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k \mid N_0^{(k)} = k_0\} \\
&\qquad = \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \mathbb{P}\{N_u^{(k)} > 0,\ u \in J \text{ and } N_u^{(k)} = 0,\ u \notin J \mid N_0^{(k)} = k_0\} \\
&\qquad = \sum_{i=0}^{c-1} \sum_{J \in S_{i,n}} \sum_{\boldsymbol{k} \in E_{k-k_0,J}} (k-k_0)! \left( \prod_{j \in J} \frac{\left(\frac{p_j}{1-p_0}\right)^{k_j}}{k_j!} \right).
\end{aligned} \tag{8}$$

**Theorem 3.** *For all $n \geq 1$, $\boldsymbol{p} \in (0,1)^n$ with $\sum_{i=1}^n p_i < 1$, and for all $c = 1, \ldots, n$, we have $T_{c,n}(\boldsymbol{u}) \leq_{\mathrm{st}} T_{c,n}(\boldsymbol{v}) \leq_{\mathrm{st}} T_{c,n}(\boldsymbol{p})$,*

*Proof.* From Relation (3) and Relation (8), we obtain, for all $k_0 = 0, \ldots, k$,

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k \mid N_0^{(k)} = k_0\} = \mathbb{P}\{T_{c,n}(\boldsymbol{p}/(1-p_0)) > k - k_0\},$$

and unconditioning,

$$\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} = \sum_{\ell=0}^{k} \binom{k}{\ell} p_0^{\ell}(1-p_0)^{k-\ell}\mathbb{P}\{T_{c,n}(\boldsymbol{p}/(1-p_0)) > k-\ell\}. \tag{9}$$

Since $\boldsymbol{p}/(1-p_0)$ is a probability distribution, applying Theorem 2 to this distribution, observing that $\boldsymbol{u} = \boldsymbol{v}/(1-p_0)$ and applying (9) to $\boldsymbol{v}$, we obtain

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

This proves the second inequality. To prove the first one, observe that $\mathbb{P}\{T_{c,n}(\boldsymbol{p}/(1-p_0)) > \ell\}$ is decreasing with $\ell$. This leads, using (9), to

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

Taking $\boldsymbol{p} = \boldsymbol{v}$ gives $\mathbb{P}\{T_{c,n}(\boldsymbol{v}) > k\} \geq \mathbb{P}\{T_{c,n}(\boldsymbol{u}) > k\}$, which completes the proof.

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

We fix a parameter $\theta \in (0, (1-p_0)/n]$ and we are looking for the distributions $\boldsymbol{p}$ which stochastically maximize the time $T_{c,n}(\boldsymbol{p})$ on the set $\mathcal{A}_\theta$ defined by

$$\mathcal{A}_\theta = \{\boldsymbol{p} \in (0,1)^n \mid \sum_{i=1}^{n} p_i = 1-p_0 \text{ and } p_j \geq \theta, \text{ for every } j = 1,\ldots,n\}.$$

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 $\theta$. The set $\mathcal{B}_\theta$ has $n$ elements given by

$$\mathcal{B}_\theta = \{(\gamma,\theta,\ldots,\theta),(\theta,\gamma,\theta,\ldots,\theta),\ldots,(\theta,\ldots,\theta,\gamma)\}, \text{ with } \gamma = 1-p_0-(n-1)\theta.$$

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

*Proof.* By symmetry, $\mathbb{P}\{T_{c,n}(\boldsymbol{q}) > k\}$ has the same value for every $\boldsymbol{q} \in \mathcal{B}_\theta$, so we suppose that $q_\ell = \theta$ for every $\ell \neq j$ and $q_j = \gamma$. Let $\boldsymbol{p} \in \mathcal{A}_\theta \setminus \mathcal{B}_\theta$ and let $i$

be the first entry of $\boldsymbol{p}$ such that $i \neq j$ and $p_i > \theta$. We define $\boldsymbol{p}^{(1)}$ as $p_i^{(1)} = \theta$, $p_j^{(1)} = p_i + p_j - \theta > p_j$ and $p_\ell^{(1)} = p_\ell$, for $\ell \neq i, j$. Thus $p_i = \lambda p_i^{(1)} + (1-\lambda) p_j^{(1)}$ and $p_j = (1-\lambda) p_i^{(1)} + \lambda p_j^{(1)}$, with $\lambda = (p_j - \theta)/(p_j - \theta + p_i - \theta) \in [0, 1)$. From Theorem 1, we get $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} \leq \mathbb{P}\{T_{c,n}(\boldsymbol{p}^{(1)}) > k\}$. Repeating the same procedure from distribution $\boldsymbol{p}^{(1)}$ and so on, we get, after at most $n-1$ steps, distribution $\boldsymbol{q}$, that is $\mathbb{P}\{T_{c,n}(\boldsymbol{p}) > k\} \leq \mathbb{P}\{T_{c,n}(\boldsymbol{q}) > k\}$, which completes the proof.

## 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] Anceaume, E., Busnel, Y., Rivetti, N. and Sericola, B. (2015). Identifying global icebergs in distributed streams. *Research Report.*

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

[4] Doumas, A. V. and Papanicolaou, V. G. (2012). Asymptotics of the rising moments for the coupon collector's problem. *Electron. J. Prob.* **18**, pp. 1–15.

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

[6] Marshall, A. W., Olkin, I. and Arnold, B. C. (2011). Inequalities : Theory of Majorization and Its Applications. 2nd. edn. Springer-Verlag, New York.

[7] 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.