Slowest and fastest coupon collectors

Tipaluck Krityakierne\textsuperscript{1}  Thotsaporn Aek Thanatipanonda\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand
\textsuperscript{2}Science Division, Mahidol University International College, Nakhon Pathom, Thailand

Abstract

In the coupon collector’s problem, every cereal box contains one coupon from a collection of \(n\) distinct coupons, each equally likely to appear. The goal is to find the expected number of boxes a player needs to purchase to complete the whole collection. In this work, we extend the classical problem to \(k\) players, and find the expected number of boxes required for the slowest and fastest players to complete the whole collection. The probability that a particular player is the slowest or fastest player to finish will also be touched upon.

1 Prologue

During a lecture for the course \textit{Generic Skills for Research in Mathematics} given to first-year graduate and advanced undergraduate students, we presented some classic results of the coupon collector’s problem to students:

“\textit{You buy cereals in order to collect coupons that come with it. The upcoming collection has \(n\) collectible coupons. Each cereal box contains one coupon. Assume that every type of coupons is equally likely to appear. What is the expected number of boxes you need to buy until you have a complete set of \(n\) coupons?}”

By recalling the mean of the geometric distribution, the answer to the well-known problem above is \(\mathbb{E}[X] = nH(n)\), where \(H(n)\) is the \(n\)-th harmonic number. An approximate solution to the coupon collector’s problem is

\[
\mathbb{E}[X] \approx n \log n + \gamma n + \frac{1}{2},
\]

where \(\gamma \approx 0.5772156649\) is the Euler–Mascheroni constant.

To promote deeper learning skills, we posed a couple of questions to the class.
Question 1 What is the expected number of boxes you need to buy to complete two sets of coupons?

Question 2* Now, instead of you alone trying to collect two sets of coupons, let’s say that you and your brother independently buy a box of cereal everyday with the goal to each collect the complete set of $n$ coupons. What is the expected number of boxes required for

2.1 the slower player (who last completed a collection);
2.2 the faster player (who first completed a collection)?

The first question, known as the double dixie cup problem, was solved by Newman and Shepp in 1960. Their results, published in American Mathematical Monthly, showed that the expected number of boxes needed to complete $m$ sets of coupons is $n (\log n + (m - 1) \log \log n + O(1))$ [3]. The note [1] gave an extensive review on approaches for solving the classical problem, and established some interesting results regarding multiple collections.

Generalization of the problem to a two-player game was also studied previously, in different contexts. For example, the probability that the faster player was never behind at any intermediate stage of the play has been investigated in [2]. However, to our surprise, it seems the second question has never been considered in the existing literature (and apparently none of the students in class knew how to solve it). We thus take this opportunity to present some novel results that directly answer this question.

2 Main results

We will solve a generalized version of the expected maximum time for two players who are still missing $s$ and $t$ coupons, respectively. Once a result for two players has been reached, we will consider the expected maximum time for $k$ players, followed by the expected minimum time for $k$ players. We conclude this section by providing answers to the questions posed in the classroom.

2.1 Two players: the slower one

Given $0 \leq s, t \leq n$, let $X_1(s)$ and $X_2(t)$ be random variables representing the number of boxes the first player (who are still missing $s$ coupons) and second player (missing $t$ coupons) need to open, until they each collect all $n$ coupons. Then,

$$M(s, t) := \mathbb{E}[\max\{X_1(s), X_2(t)\}]$$
is the expected number of boxes required for the \textit{slower player} to collect a complete set of $n$ coupons.

Using the law of total expectation, conditioning on whether a player found a new coupon type in the next box or not, we can write a recurrence relation:

$$M(s,t) = \left(\frac{s}{n}\right)\left(1 - \frac{1}{n}\right)M(s-1,t) \quad \text{first player (found a new coupon)}$$

$$+ \left(1 - \frac{s}{n}\right)\left(\frac{t}{n}\right)M(s,t-1) \quad \text{second player}$$

$$+ \left(\frac{s}{n} \cdot \frac{t}{n}\right)M(s-1,t-1) \quad \text{both}$$

$$+ \left(1 - \frac{s}{n}\right)\left(1 - \frac{t}{n}\right)M(s,t) \quad \text{neither}$$

$$+ 1 \quad \text{(1 more box has been opened)}$$

(1)

with the initial condition $M(0,0) = 0$ and $M(s,t) = 0$ if $s$ or $t < 0$.

**Theorem 1.** The solution of (1) is

$$M(s,t) = nH(s+t) - \frac{(H(s+t) - 1)st}{(s+t)(s+t-1)} + o(1).$$

**Proof.** The solution can be easily verified by substitution once we manage to guess a formula, which was derived by experimental method in our case.

### 2.2 Multiple players: how slow is the slowest one?

Suppose now that there are $k$ players, where player $i$ is still missing $s_i$ coupons. The recurrence (1) above can be extended and solved for the expected maximum time for $k$ players, whose solution turns out to have a surprisingly simple leading term.

**Theorem 2.** Let $M(s_1, s_2, \ldots, s_k) := E[\max\{X_1(s_1), \ldots, X_k(s_k)\}]$ be the expected number of boxes required for the slowest player to collect all $n$ coupons. Then,

$$M(s_1, s_2, \ldots, s_k) = nH(S) - \frac{(H(S) - 1)\sum_{i<j} s_is_j}{S(S-1)} + o(1),$$

where $S := \sum_{i=1}^{k} s_i$.

**Proof.** Again, the solution can be verified by substitution.
2.3 How fast is the fastest player?

Let us denote by

\[ m(s_1, s_2, \ldots, s_k) := E\left[ \min\{X_1(s_1), \ldots, X_k(s_k)\} \right] \]

the expected number of boxes required for the fastest player to collect all \( n \) coupons. While we could have derived the expression of the expected minimum time using the recurrence relation, it turns out that the task is much simpler if we do not.

**Corollary 3.** The expected minimum time for \( k \) players, \( m(s_1, s_2, \ldots, s_k) \), is given by

\[
\sum_{i=1}^{k} M(s_i) - \sum_{i<j} M(s_i, s_j) + \sum_{i<j<l} M(s_i, s_j, s_l) + \cdots + (-1)^{k-1} M(s_1, s_2, \ldots, s_k)
\]

\[ = n \left( \sum_{i=1}^{k} H(s_i) - \sum_{i<j} H(s_i + s_j) + \sum_{i<j<l} H(s_i + s_j + s_l) + \cdots + (-1)^{k-1} H \left( \sum_{i=1}^{k} s_i \right) \right) + o(n). \]

**Proof.** Retaining only the leading terms, the result follows immediately from the maximum-minimum identity and the linearity of expectation:

\[
\min\{X_1, \ldots, X_k\} = \sum_{i=1}^{k} X_i - \sum_{i<j} \max\{X_i, X_j\} + \sum_{i<j<l} \max\{X_i, X_j, X_l\}
\]

\[ + \cdots + (-1)^{k-1} \max\{X_1, \ldots, X_k\}. \]

\[ \square \]

2.4 Answers to Question 2

Return to the two-player problem where both players start with empty hands. We are now in a position to provide answers to the expected number of boxes required for the slower and faster players to collect the whole collection of \( n \) coupons:

\[ M(n, n) = nH(2n) - \frac{(H(2n) - 1)n^2}{2n(2n - 1)} + o(1) \]

\[ = n \log n + (\log 2 + \gamma)n - \frac{n \log n}{2(2n - 1)} - \frac{n(\log 2 + \gamma - 1)}{2(2n - 1)} + 1 \frac{4}{4} + o(1). \]

\[ m(n, n) = 2nH(n) - M(n, n) \]

\[ = n \log n - (\log 2 - \gamma)n + \frac{n \log n}{2(2n - 1)} + \frac{n(\log 2 + \gamma - 1)}{2(2n - 1)} + 3 \frac{4}{4} + o(1). \]
Figure 1 shows the graphs of the expected numbers of boxes required for the slowest player $M(n,n,\ldots,n)$ and fastest player $m(n,n,\ldots,n)$ to complete the whole collection of $n$ coupons, where the number of players ranges from $k = 1, 2, \ldots, 40$.

![Figure 1: The expected number of boxes required for the slowest (top 40 lines) and fastest players (bottom 40 lines) to complete the whole collection of $n$ coupons. Each line represents a different number of $k$ players: the smallest value ($k = 1$) is blue and the largest ($k = 40$) red. The darkest line in the middle is a separator line corresponding to $k = 1$ player, for which the slowest and fastest players are the same person.](image)

3  Continuous counterpart of a discrete-time coupon collectors’ problem

The leading term, $nH(S)$, in the solution of $M(s_1,s_2,\ldots,s_k)$ is shockingly simple. We try to provide insight into this term using a continuous-time counterpart for the coupon collector’s problem.
Let us start with one player, who is missing $s$ coupons. Through the concept of interarrival times of an inhomogeneous counting process, let $W_1 \sim \exp \left( \lambda_1 = \frac{s}{n} \right)$ be the time of the first arrival of the coupon. Similarly, let $W_i \sim \exp \left( \lambda_i = \frac{s - i + 1}{n} \right)$ be the interarrival time (elapsed time) between the $(i - 1)$th and the $i$th arrivals, for $i = 2, \ldots, s$. It follows that the expected completion time for this particular player satisfies
\[
E[X(s)] = E \left[ \sum_{i=1}^{s} W_i \right] = \sum_{i=1}^{s} \frac{1}{\lambda_i} = \sum_{i=1}^{s} \frac{n}{s - i + 1} = nH(s).
\]
The concept of interarrival times can be extended to find the expected maximum time for the $k$ coupon collectors’ problem. Assume that player $j$ is still missing $s_j$ coupons. Let $T_1$ be the time of the first arrival of coupon, regardless of which player finds it. Recall a classical property that the minimum of independent exponential random variables is again exponential with the rate parameter equals to the sum of the rates. Then, $T_1 \sim \exp \left( \lambda_1 = \frac{S}{n} \right)$, where $S = s_1 + s_2 + \cdots + s_k$.

In addition, let $T_i$ be the interarrival times between the $(i - 1)$th and the $i$th arrivals of the coupon, regardless of which player finds the coupon. By the independence of interarrival times and the player who finds the coupon, $T_i \sim \exp \left( \lambda_i = \frac{S - i + 1}{n} \right)$. Finally, the completion time of the slowest player is simply
\[
E[\max\{X_1(s_1), \ldots, X_k(s_k)\}] = E \left[ \sum_{i=1}^{S} T_i \right] = \sum_{i=1}^{S} \frac{1}{\lambda_i} = \sum_{i=1}^{S} \frac{n}{S - i + 1} = nH(S).
\]
We often hear people say “Continuous problems tend to be easier to solve than discrete ones.” Here, things simplify as two events cannot occur at the same time, and it does not matter which player finds a next new coupon as the rate parameter of the counting process is based solely on the total number of coupons still missing at that time.

### 3.1 A continuous-time recurrence relation

By dropping the terms in the original discrete-time recurrence relation, e.g. see [1] for $k = 2$, that corresponds to “multiple players finding a new coupon in the next box” and retaining the leading terms (in $n$), we obtain a continuous-time recurrence relation for $E[\max\{X_1(s_1), \ldots, X_k(s_k)\}]$:
\[
M(s_1, s_2, \ldots, s_k) = \sum_{j=1}^{k} \left( \frac{s_j}{S} \right) M(s_1, s_2, \ldots, s_j - 1, \ldots, s_k) + \frac{n}{S}, \quad (2)
\]
where $S = \sum_{i=1}^{k} s_i$.

Since the rate parameter of a new arrival is $\lambda = \frac{S}{n}$, the term $\frac{n}{S}$ in (2) represents the mean arrival time of a new coupon (regardless of which player finds it). Moreover, by recalling another classical property of the exponential distribution concerning the probability of $j$th random variable being smallest among others, the term $\frac{s_j}{S}$ in (2) is the probability that player $j$ is the one who finds the next new coupon, as one would expect.

4 Who is the slowest or fastest player?

In this final section, we will discuss the probability of being the slowest or fastest player to finish. The solutions are again surprisingly simple.

Starting with two players, let $P_1(s,t)$ denote the probability that the first player finishes after the second player, i.e. $X_1(s) = \max\{X_1(s), X_2(t)\}$. Then, we can write a recurrence relation:

$$
P_1(s,t) = \frac{s}{n} \left( 1 - \frac{t}{n} \right) P_1(s-1,t) + \frac{t}{n} \left( 1 - \frac{s}{n} \right) P_1(s,t-1)$$

$$+ \left( \frac{t}{n} \cdot \frac{s}{n} \right) P_1(s-1,t-1) + \left( 1 - \frac{s}{n} \right) \left( 1 - \frac{t}{n} \right) P_1(s,t),$$

with the initial conditions $P_1(1,0) = 1$, $P_1(0,1) = 0$, and $P_1(s,t) = 0$ if $s$ or $t < 0$.

In a similar fashion, we can set up a more general recurrence relation for $k > 2$ players, whose solution is given in the next proposition.

**Proposition 4.** Let $P_1(s_1, s_2, \ldots, s_k)$ be the probability that the first player is slowest among the $k$ players to collect the whole set of $n$ coupons. Then,

$$P_1(s_1, s_2, \ldots, s_k) = \frac{s_1}{\sum_{i=1}^{k} s_i} + o(1).$$

**Proof.** The solution can be verified by substitution. \qed

4.1 A combinatorial interpretation for a continuous-time model

We now give a combinatorial interpretation for the term $\frac{s_1}{\sum_{i=1}^{k} s_i}$ in the probability $P_1(s_1, s_2, \ldots, s_k)$. 7
For simplicity, let us consider $k = 3$ players: the number of combinations where the first player finishes last is \( \binom{s_1-1+s_2+s_3}{s_1-1,s_2,s_3} \), and the probability of each combination is \( \frac{s_1!s_2!s_3!}{(s_1+s_2+s_3)!} \) (following from our earlier discussion that \( \sum_{i=1}^{k} s_i \) is the probability that player \( j \) is the one who finds the next new coupon). Thus,

\[
P(\text{the first player is slowest}) = \frac{s_1!s_2!s_3!}{(s_1+s_2+s_3)!} \cdot \binom{s_1-1+s_2+s_3}{s_1-1,s_2,s_3}
\]

\[
= \frac{s_1}{s_1+s_2+s_3}.
\]

Last but not least, following the same maximum-minimum identity we applied before, results for the probability of being the fastest player can be derived.

**Corollary 5.** Let \( Q_1(s_1, s_2, \ldots, s_k) \) denote the probability that the first player is the fastest player to finish, i.e. \( X_1(s_1) = \min\{X_1(s_1), X_2(s_2), \ldots, X_k(s_k)\} \). Then,

\[
Q_1(s_1, s_2, \ldots, s_k) = 1 - \sum_{1<i}^{s_1} \frac{s_1}{s_1+s_i} + \sum_{1<i<j}^{s_1} \frac{s_1}{s_1+s_i+s_j} - \sum_{1<i<j<l}^{s_1} \frac{s_1}{s_1+s_i+s_j+s_l} + \cdots + (-1)^{k-1} \frac{s_1}{s_1+\cdots+s_k} + o(1).
\]

**References**

[1] Ferrante, Marco, and Monica Saltalamacchia. *The coupon collector’s problem.* Materials matemàtiques (2014): 0001-35.

[2] Myers, Amy N., and Herbert S. Wilf. *Some new aspects of the coupon collector’s problem.* SIAM review 48.3 (2006): 549-565.

[3] Newman, Donald J., and Lawrence Shepp. *The double dixie cup problem.* The American Mathematical Monthly 67.1 (1960): 58-61.