The Slowest Coupon Collector's Problem

Tipaluck Krityakierne and Thotsaporn Aek Thanatipanonda

Abstract

In the classical coupon collector's problem, every box of breakfast cereal 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 who compete with one another to be the first to collect the whole collection. We find the expected numbers of boxes required for the slowest and fastest players to finish the game. The odds of a particular player being the slowest or fastest player will also be touched upon. The solutions will be discussed from both the tractable algebraic techniques as well as the probability point of views.

Keywords: coupon collector's problem; fastest player; slowest player; multiple players.

1 Prologue

The coupon collector's problem is a classical mathematics problem that shows up in a number of courses, from probability theory, simulations, programming, to name a few. The classic version of the problem can be described as follows.

"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 (see for example [1, p. 225]). 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.

The problem has been extended to a scenario where the player has to collect multiple sets of coupons. This problem, 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 $n(\log n + (m-1)\log\log n + \mathcal{O}(1))$ [4]. In [5], Zeilberger found the generating function for the expected number of types of cards of which the player has exactly i copies at the end. The note [2] 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 has also been studied previously. For example, the probability that the faster player was never behind at any intermediate stage of the play has been investigated in [3].

In this work, we extend the problem to k players who compete with one another in collecting the coupons. To our surprise, generalizations of the coupon collector's problem in this direction seem to have never been addressed in the literature. We thus take this opportunity to present and contribute some novel results. Notably, using algebraic recurrence relations and difference equations as the tools, our main theorem finds the expected number of boxes required for the slowest player to collect the whole collection of n coupons. We further investigate the problem from a probability point of view, which allows us to provide full insight into the recurrence relation and the obtained solution.

1.1 A two-player scenario: the slower one

As a warm up, we consider a generalized version of the expected *maximum* time for two players who are still missing s and t coupons, respectively. To be more precise, 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,

$$M(s,t) := \mathbb{E}\left[\max\{X_1(s), X_2(t)\}\right]$$

is the expected number of boxes required for the *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:

$$\begin{aligned} M(s,t) = &\left(\frac{s}{n}\right)\left(1-\frac{t}{n}\right) M(s-1,t) \quad \text{first player (found a new coupon)} \\ &+ \left(1-\frac{s}{n}\right)\left(\frac{t}{n}\right) M(s,t-1) \quad \text{second player} \\ &+ \left(\frac{s}{n}\cdot\frac{t}{n}\right) M(s-1,t-1) \quad \text{both} \\ &+ \left(1-\frac{s}{n}\right)\left(1-\frac{t}{n}\right) M(s,t) \quad \text{neither} \\ &+ 1 \quad \text{(1 more box has been opened)}, \end{aligned} \tag{1}$$

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

For the two-player scenario, the recurrence relation takes a vector argument $[s,t]$. The initial condition $M(0,0) = 0$ means that the game has ended since both players completed the whole set of coupons. Of course, we will not consider the case when one of the arguments s or t is negative, so we assign a zero value whenever this happens.

1.2 A one-player scenario

The above recurrence can be simplified to get a recurrence for the classical one-player scenario:

$$\begin{aligned} M(s) = &\left(\frac{s}{n}\right) M(s-1) \quad \text{the player found a new coupon} \\ &+ \left(1-\frac{s}{n}\right) M(s) \quad \text{the player did not find a new coupon} \\ &+ 1 \quad \text{(1 more box has been opened)}, \end{aligned}$$

with the initial condition $M(0) = 0$ and $M(s) = 0$ if $s < 0$. The reader can quickly verify that $M(s) = nH(s)$ indeed satisfies this recurrence together with the initial condition given, consistent with the established result of the classical coupon collector's problem.

Our goal is to generalize the recurrence (1) to $k > 2$ players and come up with a general strategy for solving it. Before proceeding to a more detailed explanation, let us end this section with the main theorem of this paper.

Theorem 1. *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,*

$$M(s_1, s_2, \ldots, s_k) = nH(S) - \frac{(H(S)-1)\sum_{i<j} s_i s_j}{S(S-1)} + \mathcal{O}\left(\frac{1}{n}\right), \quad \textit{where } S = \sum_{i=1}^{k} s_i.$$

In the next section, we will introduce several important tools and concepts along the way during the course of proving the theorem.

2 Proof of the theorem

We first give an algebraic proof of the theorem, and then in the next section we will provide an alternative proof (from the probability view point) for the leading term, which gives additional insight into the recurrence and the obtained solution.

2.1 Recurrence relation for the slowest player

We have seen that the recurrence for two players was set up after each player has opened one more box, and checked whether or not they found a new coupon. Suppose now that there are k players, where player i is still missing s_i coupons. The same idea is applied to obtain a recurrence relation for the number of boxes required for the slowest player in the k-player scenario.

$$M\left(s_1, s_2, \ldots, s_k\right) = \sum_{I \subseteq \{1,2,\ldots,k\}} \underbrace{\left[\prod_{j \in I} \frac{s_j}{n}\right] \left[\prod_{j \notin I} \left(1 - \frac{s_j}{n}\right)\right] M\left(V_I\right)}_{\text{players in } I \text{ found a new coupon}} + 1, \tag{2}$$

with the initial condition $M(0, \ldots, 0) = 0$ and $M(s_1, s_2, \ldots, s_k) = 0$ if at least one of $s_i < 0$.

The meaning of the notation in (2) is as follows. Let I be the set of index (possibly empty) of the players who found a new coupon. For each I, the probability that this event happens is $\left[\prod_{j \in I} \frac{s_j}{n}\right] \left[\prod_{j \notin I} \left(1 - \frac{s_j}{n}\right)\right]$. V_I represents the updated vector argument after the players in I found a new coupon, that is,

$$V_I := [s_1 - \delta_I(1), s_2 - \delta_I(2), \ldots, s_k - \delta_I(k)],$$

where $\delta_I(j) = 1$ if $j \in I$ and 0 otherwise. In particular, we write $V_{\{\}}$, when no players found a new coupon, and $V_{\{j\}}$ when only the player j found a new coupon.

2.2 Solutions via difference equations

To solve the recurrence (2) for the first two leading terms, we shall reformulate the solution as a difference equation. First, we expand out (2) to get

$$\begin{aligned} M(s_1, s_2, \ldots, s_k) = 1 &+ \prod_{j=1}^{k} \left(1 - \frac{s_j}{n}\right) M(s_1, s_2, \ldots, s_k) \\ &+ \sum_{j=1}^{k} \frac{s_j}{n} \left[\prod_{i \neq j} \left(1 - \frac{s_i}{n}\right)\right] M(s_1, \ldots, s_j - 1, \ldots, s_k) \\ &+ \sum_{i<j} \frac{s_i s_j}{n^2} \left[\prod_{l \notin i,j} \left(1 - \frac{s_l}{n}\right)\right] M(s_1, \ldots, s_i - 1, \ldots, s_j - 1, \ldots, s_k) + \ldots. \end{aligned} \tag{3}$$

The coefficients of this recurrence indicate that the solution must be in the form:

$$M(s_1, s_2, \ldots, s_k) = nA_1 + A_0 + \frac{1}{n}A_{-1} + \frac{1}{n^2}A_{-2} + \ldots, \tag{4}$$

where $A_i, \quad i \leq 1$ is a function of $s_1, s_2, \ldots, s_k$.

Having a solution written in the form (4), the proof of Theorem 1 boils down to the problem of identifying A_1 and A_0 (which are the coefficients of the first two leading terms of the solution). This also explains the remainder term $\mathcal{O}\left(\frac{1}{n}\right)$ in the theorem.

2.3 Setting up difference equations

To set up a difference equation for A_1, plug the assumed form (4) into (3) and equate the resulting constant terms (which correspond to the leading term) on both sides of (3):

$$A_0 = 1 + A_0 - \sum_{j=1}^{k} \frac{s_j}{n} nA_1(V_{\{\}}) + \sum_{j=1}^{k} \frac{s_j}{n} nA_1(V_{\{j\}}).$$

(For simplicity of notation, we omit the full vector argument $V_{\{\}} = [s_1, s_2, \ldots, s_k]$ of A_i when there is no ambiguity.)

Simplifying the above equation, we obtain the difference equation of A_1

$$\sum_{j=1}^{k} s_j \left(A_1(V_{\{\}}) - A_1(V_{\{j\}})\right) = 1, \tag{5}$$

along with the initial condition $A_1(s, 0, 0, \ldots, 0) = H(s)$, obtained from the one-player scenario. Note that (5) is a first order difference equation as $V_{\{\}} = [s_1, s_2, \ldots, s_k]$

and $V_{\{j\}} = [s_1, s_2, \ldots, s_j - 1, \ldots, s_k]$. Although we will soon explain how we come up with the solution, the reader may quickly check that $A_1(V) = H(S)$, where $S = \sum_{i=1}^{k} s_i$, satisfies the difference equation (5) and the initial condition.

Next, we find a difference equation of A_0. After plugging in (4), we equate the coefficients of $\frac{1}{n}$ (which is the second leading term) on both sides of (3):

$$\frac{1}{n}A_{-1} = -\frac{\sum_{j=1}^{k} s_j}{n}A_0 + \frac{1}{n}A_{-1} + \sum_{i<j}\frac{s_i s_j}{n^2}nA_1 + \frac{\sum_{j=1}^{k} s_j}{n}A_0(V_{\{j\}}) - \sum_{i<j}\frac{s_i s_j}{n^2}nA_1(V_{\{i\}}) - \sum_{i<j}\frac{s_i s_j}{n^2}nA_1(V_{\{j\}}) + \sum_{i<j}\frac{s_i s_j}{n^2}nA_1(V_{\{i,j\}}).$$

Simplifying the above equation, we obtain the difference equation of A_0

$$\sum_{j=1}^{k} s_j \cdot \big(A_0(V_{\{\}}) - A_0(V_{\{j\}})\big) = \sum_{i<j} s_i s_j \cdot \big(A_1(V_{\{\}}) - A_1(V_{\{i\}}) - A_1(V_{\{j\}}) + A_1(V_{\{i,j\}})\big).$$

We can further simplify the above equation by substituting $A_1(V) = H(S)$, and the difference equation of A_0 becomes

$$\sum_{j=1}^{k} s_j \cdot \big(A_0(V_{\{\}}) - A_0(V_{\{j\}})\big) = \sum_{i<j} s_i s_j \left(\frac{1}{S} - \frac{1}{S-1}\right), \tag{6}$$

together with the initial condition $A_0(s, 0, 0, \ldots, 0) = 0$. This condition is due to the absence of the other terms except the leading term, $nH(s)$, in the solution of the classical one-player scenario. Again, (6) is a first order difference equation.

2.4 Solving difference equations

The difference equations (5) and (6) that we are dealing with are a discrete version of first order linear partial differential equations. In particular, both difference equations take the following form

$$x_1\frac{\partial u}{\partial x_1} + x_2\frac{\partial u}{\partial x_2} + \cdots + x_k\frac{\partial u}{\partial x_k} = f(x_1, x_2, \ldots, x_k),$$

in the first quadrant ($x_i > 0$).

We digress momentarily to discuss the following proposition which gives a solution to a new family of PDEs, and will be used to come up with a "good guess" (solution) for our difference equations.

Proposition 2. *Let $M, N \geq 0$ and P be a multivariate polynomial where the degree of each monomial is N. Let $X = \sum_{j=1}^{k} x_j$. Then, the solution of*

$$\sum_{j=1}^{k} x_j \frac{\partial u}{\partial x_j} = (\ln X)^M \cdot \frac{P(x_1, x_2, \ldots, x_k)}{X^N} \tag{7}$$

is

$$u(x_1, x_2, \ldots, x_k) = \left(\frac{(\ln X)^{M+1}}{M+1} + C \right) \cdot \frac{P(x_1, x_2, \ldots, x_k)}{X^N},$$

for any constant C.

Proof. We solve this by the method of characteristics. Suppose $u = f(x_1, x_2, \ldots, x_k)$ is a differentiable function of $x_1, \ldots, x_k$, where each x_i is parameterized as a function of t. Then, by the chain rule, u is a differentiable function of t and

$$\frac{du}{dt} = \sum_{i=1}^{k} \frac{\partial f}{\partial x_i} \frac{dx_i}{dt}.$$

In order to find the solution, we solve

$$\frac{dx_i}{dt} = x_i, \quad 1 \leq i \leq k, \qquad \text{and} \qquad \frac{du}{dt} = (\ln X)^M \frac{P(x_1, x_2, \ldots, x_k)}{X^N}.$$

Then, $x_i = c_i e^t$ where c_i are constants. Substitute this into $\dfrac{du}{dt}$ to get

$$\frac{du}{dt} = \ln \left(\sum_i c_i e^t \right)^M \frac{P(c_1, c_2, \ldots, c_k)}{\left(\sum_i c_i\right)^N}.$$

The last equality holds because P is a multivariate polynomial where the degree of each monomial is N. The final step is to integrate both sides of the differential equation to find:

$$\begin{aligned} u &= \frac{P(c_1, c_2, \ldots, c_k)}{\left(\sum_i c_i\right)^N} \int \ln \left(\sum_i c_i e^t \right)^M dt \\ &= \frac{P(c_1, c_2, \ldots, c_k)}{\left(\sum_i c_i\right)^N} \left(\frac{\ln \left(\sum_i c_i e^t\right)^{M+1}}{M+1} + C \right) \\ &= \frac{P(x_1, x_2, \ldots, x_k)}{X^N} \left(\frac{(\ln X)^{M+1}}{M+1} + C \right), \quad \text{for any constant } C. \end{aligned}$$

□

We will now make use of a more readily available solution to this family of PDEs to obtain a solution for our difference equations (5) and (6).

For (5), after comparing the target (5) to the PDE (7), we apply the proposition with $M = 0$ and $N = 0$. The obtained solution $u = \ln(X)+C$ reminds us of the Harmonic number in a discrete version. Thus, our guess is $\tilde{A}_1(s_1, s_2, \ldots, s_k) = H(S) + C$. Of course, we have to verify that this solution satisfies (5), which obviously does. Moreover, the initial condition $A_1(0, 0, \ldots, 0) = 0$ implies that the constant $C = 0$.

Therefore, the particular solution to (5), which is the coefficient of our leading term solution, is

$$A_1(s_1, s_2, \ldots, s_k) = H(S).$$

The target (6) suggests us to apply the proposition with $M = 0$ and $N = 2$, which gives the solution $u = -(\ln(X) + C)\dfrac{\sum_{i<j} x_i x_j}{X^2}$. Thus, a guess for the discrete analogue is

$$\tilde{A}_0(s_1, s_2, \ldots, s_k) = -\frac{(H(S) + C)\sum_{i<j} s_i s_j}{S(S-1)}. \tag{8}$$

(Notice the difference between the denominators $S(S-1)$ and X^2 first arising in the target (6) and the PDE (7), and later appearing again in their solutions.)

In order to verify that this guess is indeed the solution of (6), there is one tricky calculation, which will be dealt with in the next lemma.

Lemma 3. *Assume $\tilde{A}_0(s_1, s_2, \ldots, s_k)$ as in (8). Then,*

$$\sum_{j=1}^{k} s_j \tilde{A}_0(V_{\{j\}}) = -\frac{H(S-1)+C}{(S-1)}\left(\sum_{i<j} s_i s_j\right).$$

Proof. Consider vectors $V_{\{\}} = [s_1, s_2, \ldots, s_k]$ and $V_{\{j\}} = [s'_1, s'_2, \ldots, s'_k] = [s_1, s_2, \ldots, s_j - 1, \ldots, s_k]$. Then,

$$\sum_{j=1}^{k} s_j \tilde{A}_0(V_{\{j\}}) = \sum_{j=1}^{k} s_j \left[-\frac{H(S-1)+C}{(S-1)(S-2)} \cdot \sum_{i<l} s'_i s'_l\right] = -\frac{H(S-1)+C}{(S-1)(S-2)} \cdot \sum_{j=1}^{k} s_j \sum_{i<l} s'_i s'_l,$$

and

$$\sum_{i<l} s'_i s'_l = \sum_{i<l} s_i s_l - \sum_{i=1}^{k} s_i + s_j.$$

Therefore,

$$\sum_{j=1}^{k} s_j \sum_{i<l} s'_i s'_l = S \sum_{i<l} s_i s_l - S^2 + \sum_{j=1}^{k} s_j^2 = (S-2)\sum_{i<l} s_i s_l,$$

and the result is immediate. □

We are now ready to verify the solution of (6). Substitute our guess (8) into the l.h.s. of (6), and use the lemma to obtain

$$\sum_{j=1}^{k} s_j \left(\tilde{A}_0(V_{\{\}}) - \tilde{A}_0(V_{\{j\}}) \right) = -\frac{(H(S)+C)\sum_{i<j} s_i s_j}{S-1} + \frac{(H(S-1)+C)\sum_{i<j} s_i s_j}{(S-1)}$$
$$= \frac{\sum_{i<j} s_i s_j}{S-1} \left(H(S-1) - H(S) \right) = \sum_{i<j} s_i s_j \left(\frac{1}{S} - \frac{1}{S-1} \right),$$

and so (8) is indeed a solution of (6).

The unique value of C can be determined by making sure that the initial condition $A_0(1,0,\ldots,0) = 0$ is satisfied. In particular, in order to make this point a removable singularity, it is necessary that $H(1) + C = 0$, and so $C = -1$.

Thus, the particular solution of (6), which is the coefficient of our second leading term solution, is

$$A_0(s_1, s_2, \ldots, s_k) = -\frac{(H(S)-1)\sum_{i<j} s_i s_j}{S(S-1)}.$$

Having obtained the closed-form formula for A_1 and A_0, Theorem 1 has been verified.

The next corollary, which is an immediate corollary of Theorem 1, provides the solution to our original problem when all the k players start with empty hands.

Corollary 4.

$$M(n, n, \ldots, n) \approx nH(kn) - \frac{(H(kn)-1)(k-1)n}{2(kn-1)}.$$

2.5 Remarks on the remainder

The formula given in Theorem 1 becomes more precise as n increases. For example, with $n = 30$, the exact value of $M(30,30,30)$ computed numerically from (2) is 151.0692567, while Theorem 1 gives 151.1009707. The general formula for $M(30,30,30)$ computed from (2) with symbolic n is

$$M(30,30,30) = 5.082570603n - 1.376147394 - \frac{0.9078106927}{n} - \frac{1.231342610}{n^2}$$
$$- \frac{2.159152821}{n^3} - \frac{4.180289796}{n^4} - \frac{7.669304559}{n^5} - \frac{7.488122252}{n^6}$$
$$+ \ldots.$$

3 Second proof of the theorem: Insight into the leading term

In this section, we give an alternative proof for the leading term solution from the probability point of view.

Consider the coupon collector's problem in a continuous-time setting. 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 ith 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 a 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 ith 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

$$E[\max\{X_1(s_1), \ldots, X_k(s_k)\}] = E\left[\sum_{i=1}^{S} T_i\right] = \sum_{i=1}^{S} \frac{1}{\lambda_i} = \sum_{i=1}^{S} \frac{n}{S-i+1} = nH(S). \quad (9)$$

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.

One can write a recurrence relation for the continuous-time setting as

$$M(s_1, s_2, \ldots, s_k) = \sum_{j=1}^{k} \left(\frac{s_j}{S}\right) M(s_1, s_2, \ldots, s_j - 1, \ldots, s_k) + \frac{n}{S}, \quad (10)$$

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

An interpretation of the recurrence relation is now given. Since the rate parameter of a new arrival is $\lambda = \dfrac{S}{n}$, the last term $\dfrac{n}{S}$ 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 jth random variable being smallest among others , the term $\dfrac{s_j}{S}$ is the probability that player j is the one who finds the next new coupon, as one would expect. While no such explanations can be given when we solved the difference equations in the discrete-time setting, the continuous-time setting allows us to gain full insight into the recurrence relation and the solution we already obtained.

Last but not least, the fact that the solution (9) coincides with the leading term solution of the discrete-time recurrence is not a mere happenstance. In fact, the recurrence relation for the leading term solution can be obtained by dropping those terms in (2) which correspond to "multiple players finding a new coupon in the next box". As a result, we arrive at precisely the same recurrence (10).

4 Miscellaneous topics

The final section contains a miscellaneous selection of results related to the k coupon collectors' problem.

4.1 The fastest player

The expected number of boxes required for the *fastest player* to complete the whole collection turns out to be a corollary of our main theorem.

Corollary 5. *The expected number of boxes required for the fastest player to collect all n coupons, $\mathbb{E}\left[\min\{X_1(s_1), \ldots, X_k(s_k)\}\right]$, is given by*

$$\sum_{i=1}^{k} M(s_i) - \sum_{i<j} M(s_i, s_j) + \sum_{i<j<l} M(s_i, s_j, s_l) + \cdots + (-1)^{k-1} M\left(s_1, s_2, \ldots, s_k\right)$$
$$= n\left(\sum_{i=1}^{k} H(s_i) - \sum_{i<j} H(s_i + s_j) + \sum_{i<j<l} H(s_i + s_j + s_l) + \cdots + (-1)^{k-1} H\left(\sum_{i=1}^{k} s_i\right)\right) + \mathcal{O}(1).$$

Proof. Retaining only the leading term, the result follows immediately from the

maximum-minimum identity :

$$\min\{X_1,\ldots,X_k\} = \sum_{i=1}^{k} X_i - \sum_{i<j} \max\{X_i,X_j\} + \sum_{i<j<l} \max\{X_i,X_j,X_l\} + \cdots + (-1)^{k-1} \max\{X_1,\ldots,X_k\},$$

and the linearity of expectation. The remainder is $\mathcal{O}(1)$ as we keep only the leading term in the solution. □

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 $k = 1,2,\ldots,40$.

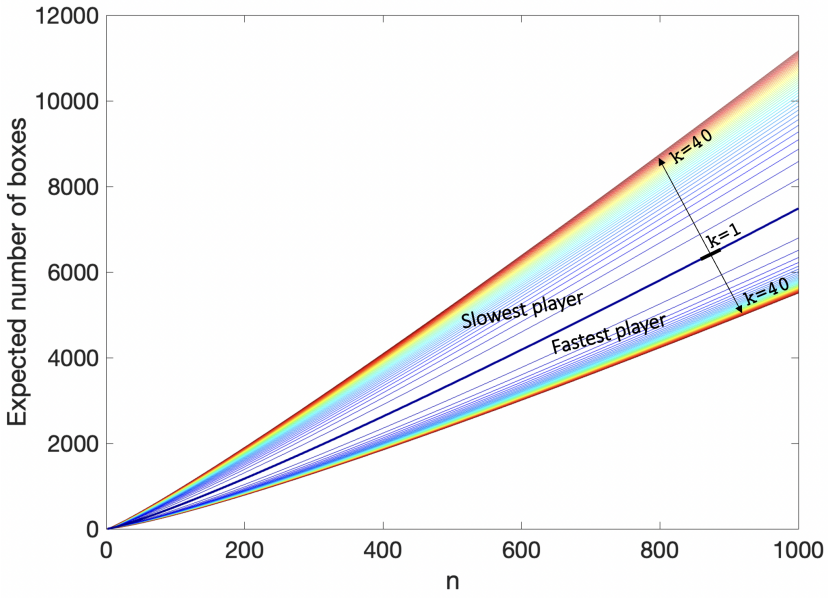

Figure 1: The expected number of boxes required for the slowest (top 40 lines) and fastest players (bottom 40 lines) to complete the whole collection of n coupons. Each line represents a different number of k players: the smallest value ($k = 1$) is blue and the largest ($k = 40$) red. The darkest line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

4.2 Probability of being the slowest player

We start with the probability of being the slowest player. Let $P_1(s_1,s_2,\ldots,s_k)$ be the probability that the *first player* is the last person to complete the whole collection,

i.e. $X_1(s_1) = \max\{X_1(s_1), X_2(s_2), \ldots, X_k(s_k)\}$. Then, we can write a recurrence

$$P_1(s_1, s_2, \ldots, s_k) = \sum_{I \subseteq \{1,2,\ldots,k\}} \underbrace{\left[\prod_{j \in I} \frac{s_j}{n}\right] \left[\prod_{j \notin I} \left(1 - \frac{s_j}{n}\right)\right]}_{\text{players in } I \text{ found a new coupon}} P_1(V_I), \tag{11}$$

with the initial conditions $P_1(s_1, 0, \ldots, 0) = 1$ if $s_1 \geq 0$, and $P_1(0, s_2, \ldots, s_k) = 0$ if some of $s_i > 0$, and $P_1(s_1, s_2, \ldots, s_k) = 0$ if at least one of $s_i < 0$.

The absence of $+1$ term in this recurrence as compared to (2) leads to the solution of the form:

$$P_1(s_1, s_2, \ldots, s_k) = B_0 + \frac{1}{n} B_{-1} + \frac{1}{n^2} B_{-2} + \ldots, \tag{12}$$

where $B_i, \;\; i \leq 0$ is a function of $s_1, s_2, \ldots, s_k$.

The next proposition finds the leading term solution B_0. Having the solution written in the form (12) explains the remainder term $\mathcal{O}\left(\frac{1}{n}\right)$ in the proposition.

Proposition 6. *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,*

$$P_1(s_1, s_2, \ldots, s_k) = \frac{s_1}{\sum_{i=1}^k s_i} + \mathcal{O}\left(\frac{1}{n}\right).$$

Proof. Following the same procedure as in the proof of Theorem 1, one may prove this statement by means of an algebraic recurrence relation. Nevertheless, we will alternatively prove the leading term solution using a combinatorial interpretation through the continuous-time framework. The number of combinations where the first player finishes last (i.e. the last coupon is found by the first player) is $\binom{s_1 - 1 + s_2 + s_3 + \cdots + s_k}{s_1 - 1, s_2, s_3, \ldots, s_k}$, and the probability of each combination is $\frac{s_1! s_2! s_3! \ldots s_k!}{(\sum_{i=1}^k s_i)!}$ (following from our discussion in a continuous-time setting that $\frac{s_j}{\sum_{i=1}^k s_i}$ is the probability that player j is the one who finds the next new coupon.

Thus,

$$\begin{aligned} P(\text{the first player is slowest}) &= \frac{s_1! s_2! s_3! \ldots s_k!}{(\sum_{i=1}^k s_i)!} \cdot \binom{s_1 - 1 + s_2 + s_3 + \cdots + s_k}{s_1 - 1, s_2, s_3, \ldots, s_k} \\ &= \frac{s_1}{\sum_{i=1}^k s_i}. \end{aligned}$$

This completes the proof. □

4.3 Remarks on the remainder of probability

The formula given in Proposition 6 is more precise as n increases. For example, with $n = 50$, the exact value of $P_1(24, 22, 14)$ computed numerically from the recurrence is 0.4039306738, while Proposition 6 gives 0.40. The general formula for $P_1(24, 22, 14)$ computed from the recurrence with symbolic n is

$$P_1(24,22,14) = 0.4 + \frac{0.1904262018}{n} + \frac{0.2926113116}{n^2} + \frac{0.6072298683}{n^3} + \frac{1.461461046}{n^4} + \frac{3.965909505}{n^5} + \dots.$$

To conclude this work, the probability of being the fastest player, whose result is a corollary to Proposition 6, will now be discussed.

Corollary 7. *Let $Q_1(s_1, s_2, \dots, s_k)$ denote the probability that the first player is the fastest player to finish, i.e. $X_1(s_1) = \min\{X_1(s_1), X_2(s_2), \dots, X_k(s_k)\}$. Then,*

$$Q_1(s_1, s_2, \dots, s_k) = 1 - \sum_{1<i} \frac{s_1}{s_1 + s_i} + \sum_{1<i<j} \frac{s_1}{s_1 + s_i + s_j} - \sum_{1<i<j<l} \frac{s_1}{s_1 + s_i + s_j + s_l} + \dots + (-1)^{k-1} \frac{s_1}{s_1 + \dots + s_k} + \mathcal{O}\left(\frac{1}{n}\right).$$

Proof. The proof is a straightforward application of the inclusion-exclusion principle and Proposition 6. □

References

[1] Feller, W., 1967. *An introduction to probability theory and its applications.* Wiley series in probability and mathematical statistics, 3rd edn.(Wiley, New York, 1968).

[2] Ferrante, M. and Saltalamacchia, M., 2014. The coupon collector's problem. *Materials matemàtics*, pp.1-35.

[3] Myers, A.N. and Wilf, H.S., 2006. Some new aspects of the coupon collector's problem. *SIAM review*, 48(3), pp.549-565.

[4] Newman, D.J. and Shepp L., 1960. The double dixie cup problem. *The American Mathematical Monthly*, 67(1), pp.58-61.

[5] Zeilberger, D., 2001. How many singles, doubles, triples, etc. should the coupon collector expect?. *Unpublished manuscript available at Prof. Zeilberger's website.*