Optimal Strategy of Coupon Subset Collection when Each Package Contains Half of the Coupons

Chengfang Fang[1] Ee-Chien Chang[2]

[1]fang.chengfang@huawei.com
Huawei International
[2]changec@comp.nus.edu.sg
School of Computing
National University of Singapore

Abstract

The coupon subset collection is a generalization of the classical coupon collection, where instead of selecting (with replacement) a single coupon, a subset of at most k coupons (known as a "package") is selected in each round. In this paper, we study how to design the collection of packages and assign probabilities to the packages, so as to minimize the expected number of rounds to collect all n distinct coupons. When k divides n, a seemingly optimal strategy is to choose a collection of non-intersecting packages, and assign equal probability to each package in the collection. We prove the optimality of this strategy when the size of the package is half the number coupons, that is, $n = 2k$.

Keywords: Coupon Subset Collection, Probabilistic Method, Distribute Algorithms, Experimental Design

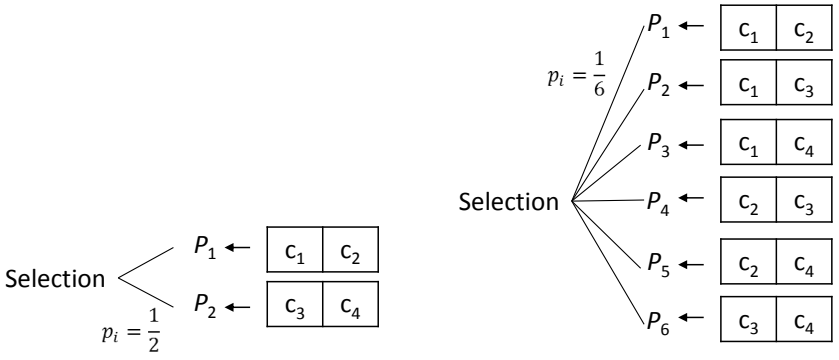
1. Introduction

Coupon subset collection is a generalization of the classical coupon collection. Let us represent a set of n distinct coupons as $C_n = \{1, 2, \ldots, n\}$,

and let $\mathbb{P} = \{P_1, \ldots, P_m\}$ be a pool of subsets of C_n, where $|P_i| \leq k$ for each i, and $\cup_i P_i = C_n$. Let us call an element in $\mathbb{P}$ a package. The package P_i is assigned a positive real number p_i for each i where $\sum_i p_i = 1$. During the process of *coupon subset collection* [1], a package is selected with replacement in a round, where the package P_i is selected with probability p_i for each i. The selection process ceases when *all* the n distinct coupons have been collected. Coupon subset collection is a generalization of the classical coupon collection [2]. In the classical case, each package contains only a single coupon, whereas here, we allow packages with multiple but at most k coupons.

We want to find a strategy, which is a set $\mathbb{P}$ and the associated probabilities, that minimizes the expected number of rounds. There are many potential applications, for instances, in distributed computing [3] and experimental design[4]. To illustrate the subtlety of this problem, let us consider the following two strategies shown in Figure 1 for $n = 4$ and $k = 2$. Strategy 1 consists of two non-intersecting packages $P_1 = \{1, 2\}$ and $P_2 = \{3, 4\}$ with equal probability; whereas Strategy 2 consists of all 6 possible pairs of coupons, and each with equal probability. We can calculate that with Strategy 1, the coupon subset collection process is expected to cease in 3 rounds, whereas 3.8 rounds is expected with Strategy 2. It seems that Strategy 1 is optimal, and in general, when k divides n, the optimal strategy is to choose a pool of n/k non-intersecting packages of size k, and assign equal probability to the packages.

Although the strategy is simple, proving its optimality is challenging, even for special cases. When $k = 1$, (i.e. the classical coupon collection



(a) Strategy 1 (b) Strategy 2

Figure 1: Two possible Strategies.

problem), it was suggested by Boneh et al. [5] that the global minimum to the process occurs only in the *equally likely case*, that is, the probability of a coupon being collected is $1/n$. The conjecture was proven many years later by Boneh et al. [6]. When $k > 1$, the problem has an interesting twist. It was conjectured [7] that when $k = 2$ and n is even, a global minimum occurs in a *pairing case*, i.e. equal probabilities are assigned to a pool of non-intersecting packages of size 2. Caron et al. [8] proved special cases of the conjecture for $n = 4, k = 2$ and $n = 6, k = 2$. To the best of our knowledge, the above two cases are the only cases known to be optimal. One might be tempted to prove the optimality using convexity argument, however, this is not at all clear as pointed out by Boneh et al. [6], page 5: "The difficulty was that the common expressions for $E[T(p)]$ (the expected time to collect all coupons) are not convex in the components of p (the probabilities assigned to the coupons)". We refer the readers to Section 3 for more works on this problem and its variants.

In this paper, we study cases when $n = 2k$ for any k, and prove that the strategy of assigning equal probability to two non-intersecting packages is optimal, and it is the unique optimal up to permutation of the coupons (Corollary 3). Our main result (Theorem 2) in fact gives a stronger result, showing that the strategy of having 2 non-intersecting packages has "stochastic dominance" over any strategy with intersecting packages, with respect to the outcome that all the coupons are collected within t rounds for any t. Our intermediate result (Lemma 1) gives an optimality property of non-intersecting packages for any k,n where k divides n, which could be useful in other works.

2. Main Result

Notations. Given n and k, let us define a strategy $\mathbb{S}$ to be a set of tuple $\{\langle P_1, p_1\rangle \ldots, \langle P_m, p_m\rangle\}$, where the package P_i is a subset of $C_n = \{1, 2, \ldots, n\}$ of size at most k, and the non-zero p_i is the probability that P_i is selected. When the packages in a strategy form a partition of C_n, i.e. there is no common coupon in any two packages, we say that the strategy is *non-intersecting*; otherwise we say that it is *intersecting*. When $n = 2k$, let us denote $\mathbb{S}_2$ be the strategy $\{\langle P_1, 0.5\rangle, \langle P_2, 0.5\rangle\}$ where $P_1 = \{1, 2, \ldots, k\}$, and $P_2 = \{k+1, \ldots, n\}$. Note that a non-intersecting strategy that consists of two packages with equal probability can always be written as $\mathbb{S}_2$, after applying some permutation of the coupons. In general, when $n = rk$, let us denote $\mathbb{S}_r$ be the strategy containing r non-intersecting packages, each with equal probability.

Outline of proof. We first show that when k divides n, i.e. $n = rk$ for some r, the strategy $\mathbb{S}_r$ is optimal with respect to the expected number of distinct coupons collected during the first t rounds for any t (Lemma 1). This is shown using linearity of expectation and Jensen's inequality, together with the classical result by Boneh et al. when $k = 1$ [6]. However, note that Lemma 1 does not immediately imply the minimality on the expected number of rounds required to collect all n distinct coupons. Theorem 2 closes the gap. We prove by contradiction that, when $r = 2$, if an strategy is intersecting, the probability that we collect all the coupons at round t with the strategy is strictly less than that with $\mathbb{S}_2$ (Theorem 2). In other words, $\mathbb{S}_2$ has stochastic dominance over any intersecting strategy. From Theorem 2, it is easy to show that the $\mathbb{S}_2$ must be the unique optimal in minimizing the expected number of rounds required (Corollary 3).

Let $N_{\widetilde{\mathbb{S}},t}$ denotes the number of distinct coupons collected within the first t rounds with a strategy $\widetilde{\mathbb{S}}$. We have the following Lemma:

Lemma 1. *When $n = rk$ the non-intersecting strategy $\mathbb{S}_r$ has the maximum expected number of distinct coupons at any round t, that is,*

$$\mathbb{E}[N_{\mathbb{S}_r,t}] \geq \mathbb{E}[N_{\widetilde{\mathbb{S}},t}],$$

for any strategy $\widetilde{\mathbb{S}}$ and any t.

Proof of Lemma 1. Note that if $\widetilde{\mathbb{S}}$ contains a package with size less than k, we can always construct another $\widetilde{\mathbb{S}}'$ by adding more coupons to that package, and yet $\mathbb{E}[N_{\widetilde{\mathbb{S}}',t}] \geq \mathbb{E}[N_{\widetilde{\mathbb{S}},t}]$. Therefore, Without loss of generality, we can assume that the packages in $\widetilde{\mathbb{S}}$ are of size exactly k.

Let us consider the given $\widetilde{\mathbb{S}}$, and let $E_{t,i} = 1$ if the coupon i is collected within t rounds, and $E_{t,i} = 0$ otherwise. Let e_i be the probability that the coupon i is collected in a round (since the selections in different rounds are independent, hence the probability e_i is the same in every round).

Note that $N_{\widetilde{\mathbb{S}},t} = E_{t,1} + \ldots + E_{t,n}$. By linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] &= \mathbb{E}[E_{t,1} + \ldots + E_{t,n}] \\
&= \mathbb{E}[E_{t,1}] + \ldots + \mathbb{E}[E_{t,n}] \\
&= \sum_{i=1}^{n} \left(1 - (1 - e_i)^t\right).
\end{aligned}$$

Recall that all packages are of size k, hence it must be the case that $\sum_i e_i = k$. In addition, the polynomial $f(x) = (1-x)^t$ is convex on $[0,1]$ when $t = 0$ or $t \geq 1$. Therefore, by Jensen's inequality, we have

$$\sum_{i=1}^{n} \left(1 - (1 - e_i)^t\right) \leq \sum_{i=1}^{n} \left(1 - \left(1 - \frac{k}{n}\right)^t\right) = \mathbb{E}[N_{\mathbb{S}_r,t}],$$

where the inequality holds with equality if and only if $e_i = k/n$ for all i. Therefore, we have

$$\mathbb{E}[N_{\mathbb{S}_r,t}] \geq \mathbb{E}[N_{\widetilde{\mathbb{S}},t}], \tag{1}$$

for any strategy $\widetilde{\mathbb{S}}$ as required.

□

Theorem 2. *When $n = 2k$ the non-intersecting strategy $\mathbb{S}_r$ has strictly higher probability of collecting all n distinct coupons within t round than an intersecting strategy $\widetilde{\mathbb{S}}$, that is,*

$$\Pr(N_{\mathbb{S}_r,t} = n) > \Pr(N_{\widetilde{\mathbb{S}},t} = n),$$

for any intersecting strategy and any $t \geq 2$.

PROOF OF THEOREM 2. Similar to Lemma 1, if $\widetilde{\mathbb{S}}$ contains a package with size less than k, we can always construct another $\widetilde{\mathbb{S}}'$ by adding more coupons to that package, and yet $\widetilde{\mathbb{S}}'$ is intersecting and the expected number of rounds to collect all n coupons does not increase. Therefore, without loss of generality, we can assume that all packages in $\widetilde{\mathbb{S}}$ are of size exactly k.

Let us consider any strategy $\widetilde{\mathbb{S}}$ that is intersecting and has only size k packages. Since $\widetilde{\mathbb{S}}$ is intersecting, there exists two packages P_i and P_j in $\widetilde{\mathbb{S}}$ and a positive integer x_0, such that $x_0 = |P_i \cup P_j|$ and $n/2 < x_0 < n$. Recall that the probability assigned to a package is non-zero by definition, hence, $\Pr(N_{\widetilde{\mathbb{S}},t} = x_0) > 0$ for $t \geq 2$.

By the definition of expectation, we have

$$\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] = \sum_{x=1}^{n} x \Pr\left(N_{\widetilde{\mathbb{S}},t} = x\right).$$

Since the packages in $\widetilde{\mathbb{S}}$ have size k, $\Pr(N_{\widetilde{\mathbb{S}},t} < k) = 0$ for $t \geq 1$. In other words, after one round, it is not possible to collect strictly less than k different coupons. Therefore,

$$\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] = \frac{n}{2} \Pr\left(N_{\widetilde{\mathbb{S}},t} = \frac{n}{2}\right) + \ldots + n \Pr\left(N_{\widetilde{\mathbb{S}},t} = n\right).$$

With Strategy $\mathbb{S}_2$, the number of distinct coupons collected can only be $n/2$ or n. Thus,

$$\mathbb{E}[N_{\mathbb{S}_2,t}] = \frac{n}{2} \Pr\left(N_{\mathbb{S}_2,t} = \frac{n}{2}\right) + n \Pr\left(N_{\mathbb{S}_2,t} = n\right).$$

As the sum of probabilities over all possible outcomes of $N_{\widetilde{\mathbb{S}},t}$ is 1, we have

$$\Pr\left(N_{\widetilde{\mathbb{S}},t} = \frac{n}{2}\right) + \ldots + \Pr\left(N_{\widetilde{\mathbb{S}},t} = n-1\right) = \Pr\left(N_{\mathbb{S}_2,t} = \frac{n}{2}\right) + \Pr\left(N_{\mathbb{S}_2,t} = n\right) - \Pr\left(N_{\widetilde{\mathbb{S}},t} = n\right). \tag{2}$$

We now show by contradiction that $\Pr(N_{\mathbb{S}_2,t} = n) > \Pr(N_{\widetilde{\mathbb{S}},t} = n)$ when $t \geq 2$. Suppose this is not the case, i.e. $\Pr(N_{\widetilde{\mathbb{S}},t} = n) = \Pr(N_{\mathbb{S}_2,t} = n) + \delta$ for some non-negative number δ, then we have

$$\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] = \frac{n}{2}\Pr\left(N_{\widetilde{\mathbb{S}},t} = \frac{n}{2}\right) + \ldots + (n-1)\Pr\left(N_{\widetilde{\mathbb{S}},t} = n-1\right) + n\left(\Pr\left(N_{\mathbb{S}_2,t} = n\right) + \delta\right).$$

Since all $\Pr\left(N_{\widetilde{\mathbb{S}},t} = i\right)$ are non-negative for $i = 1, \ldots, n$, and $\Pr\left(N_{\widetilde{\mathbb{S}},t} = x_0\right) > 0$, we have the following strict inequality:

$$\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] > \frac{n}{2}\left(\Pr\left(N_{\widetilde{\mathbb{S}},t} = \frac{n}{2}\right) + \ldots + \Pr\left(N_{\widetilde{\mathbb{S}},t} = n-1\right) + \delta\right) + n\Pr\left(N_{\mathbb{S}_2,t} = n\right). \tag{3}$$

Substituting equation (2) into inequality (3), we have

$$\mathbb{E}[N_{\widetilde{\mathbb{S}},t}] > \frac{n}{2}\Pr\left(N_{\mathbb{S}_2,t} = \frac{n}{2}\right) + n\Pr\left(N_{\mathbb{S}_2,t} = n\right) = \mathbb{E}[N_{\mathbb{S}_2,t}].$$

However, this contradicts the result from Lemma 1 (i.e. inequality (1)). Thus we have $\Pr(N_{\mathbb{S}_2,t} = n) > \Pr(N_{\widetilde{\mathbb{S}},t} = n)$, and Theorem 2 holds as desired. □

Theorem 2 implies that any strategy that minimizes the expected number of rounds must be non-intersecting. For any non-intersecting strategy, we can treat each package as a new single coupon, and reduce the collection process

essentially to the classical coupon collection. Applying the well-known results by Boneh et al. [6, p. 12] which dictates that the optimal (with respect to the classical coupon collection) strategy occurs only in the equal likely cases, we can see that the number of packages must be 2, and the assigned probability must be 1/2. Thus, we have the following Corollary:

Corollary 3. *When $n = 2k$ the strategy $\mathbb{S}_2$ is the unique optimal solution, up to permutation of C_n.*

3. Related Works

There are extensive amount of works on classic coupon collection (see [6] for a survey), and coupon subset collection. Some of the works focus on determining the expected number of rounds for a given strategy. Barton et al. [9] considered the classic coupon collection, and gave a general expression for it. Stadje [1] studied a special case of coupon subset collection with an assumption that the number of coupons in a subset has a hypergeometric distribution. Subsequently, Adler et al. [10] studied a more general setting and gave bounds on the expected number of rounds required to complete the collection. They also gave three simulators for estimating the numbers.

There are also efforts on constructing the optimal strategy. For the classical problem, it was suggested by Boneh et al. [5] that the global minimum to the classic coupon collection problem occurs only in the equally likely case. Caron et al. [8] proved the above statement for $n = 4$ and $n = 6$, and they also showed that the equally likely case is a strong local minimum for any n. Subsequently, this conjecture was proved by Boneh et al. [6].

For coupon subset collection, it was conjectured by Caron et al. [7] that when $k = 2$ and n is even, a global minimum occurs in the pairing case, i.e. the strategy $\mathbb{S}_{n,2} = \{\langle P_1, 2/n\rangle, \ldots, \langle P_{n/2}, 2/n\rangle\}$, where each $P_i = \{2i-1, 2i\}$ contains a pair of coupons. Later Caron et al. [8] proved special cases of the conjecture for $n = 4$ and $n = 6$.

There are many applications on constructing the optimal strategy. Yu et al. [3] showed that the coupon subset collection can be applied to ensure the data availability with multiple backups in different machines. The goal is to maximize the success probability of recovering all data files. In this case, the data files are the coupons and the machines are packages. They studied a case when each machine stores two files, and the machines have the same failure probability. Ewens [4] considered experimental design where each "coupon package" are a set of alleles and showed that the problem can be formulated as coupon collection problem. He considered a case when the sampling size is small compared to the size of the population (so it can be approximated by sampling with replacement).

The problem Yu et al. [3] considered is a variant where the packages are obtained *without* replacement, and all packages have the same probability of being selected. They showed that in this variant, when $k = 2$ and when the number of packages collected is equal to n, the optimal strategy is the pairing case. Note that in the setting where packages are selected without replacement, the pool $\mathbb{P}$ can contain repeated packages. They also showed that the problem of calculating the exact success probability of recovering all data files for any given strategy is #P-hard problem; and they also gave an upper bound and a lower bound on the probability for any strategy.

4. Discussion and Conclusion

The expected number of rounds taken by $\mathbb{S}_2$ is 3.

A natural question to ask next is the generalization of the proof to any n and k where k divides n. Since Lemma 1 holds in the general cases, the challenge is to generalize Theorem 2. Extending current proof of Theorem 2 is challenging as the common expressions for the expected time to collect all coupons are not convex in the components of p [6].

[1] W. Stadje, The collector's problem with group drawings, Advances in Applied Probability (1990) 866–882.

[2] G. Green, A note on certain probabilities, Journal of the Institute of Actuaries (1928) 289–295.

[3] H. Yu, P. B. Gibbons, Optimal inter-object correlation when replicating for availability, Distributed Computing (2009) 367–384.

[4] W. J. Ewens, The sampling theory of selectively neutral alleles, Theoretical population biology (1972) 87–112.

[5] A. Boneh, A. Golan, Constraints' redundancy and feasible region boundedness by random feasible point generator (RFPG), Third European Congress on Operations Research (1979) 22–31.

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

[7] R. Caron, J. McDonald, A new approach to the analysis of random methods for detecting necessary linear inequality constraints, Mathematical programming (1989) 97–102.

[8] R. Caron, M. Hlynka, J. McDonald, On the best case performance of hit and run methods for detecting necessary constraints, Mathematical programming (1992) 233–249.

[9] D. Barton, F. David, Combinatorial chance, Griffin, London (1962) 181–182.

[10] I. Adler, S. M. Ross, The coupon subset collection problem, Journal of Applied Probability (2001) 737–746.