收集策略分析：优惠券收集问题的计算视角

Hadas Abraham†、**Ido Feldman**†和**Eitan Yaakobi**† †以色列理工学院亨利与玛丽莲计算机科学学院，以色列海法。电子邮箱：hadasabraham@campus.technion.ac.il、ido19311@gmail.com、yaakobi@cs.technion.ac.il

arXiv:2601.12351v1 [cs.DS] 18 Jan 2026

摘要

优惠券收集问题（CCP）是一个著名的组合问题，旨在估算集齐 n 种不同优惠券类型所需的随机抽取次数。该问题的各种推广已应用于众多工程领域。然而，实际应用往往受到计算难题的阻碍，尤其是难以得出矩和分布的数值结果。本文提出三种算法，用于求解最一般形式的 CCP：在任意抽取概率下收集优惠券，目标是从总数为 n的 k 种优惠券中，集齐某个子集的 t 个副本数。第一种算法提供基准模型，用于计算收集过程的期望、方差和二阶矩。第二种算法利用基准模型的构造，在均匀抽取分布下，以关于 n 的多项式时间计算相同的数值；第三种算法则推广到任意一般抽取分布。所有算法均采用专门设计的马尔可夫模型来应对计算难题，确保精确计算收集过程的期望和方差。算法实现采用基于马尔可夫模型框架的动态规划方法，并据此分析其时间复杂度。

I. 引言

经典的 优惠券收集问题（*CCP*） [1], [2] 假设有 n 种不同类型的优惠券，所关注的问题是要收集多少张优惠券，才能集齐每种类型各一张。众所周知，如果优惠券是均匀随机抽取的（允许重复），那么至少各获得一张不同类型优惠券所需优惠券数量的期望大约为 $n\log n$。CCP 在计算复杂度、学习模型、概率算法以及其他各种生物科学和工程应用中具有广泛的应用 [3]。此外，它在从系统故障分析到计算机科学、优化等科学领域的最优抽样方法等应用中发挥着关键作用。由于 CCP 在解决复杂工程和概率问题方面用途广泛，因此已在这些领域得到深入研究 [4]。

文献中还探讨了 CCP 的若干其他变体，例如考察收集到的优惠券数量（k）、每种已收集优惠券的副本数（t），以及抽取每种优惠券的概率分布 [2]，[5]–[10]。本文聚焦于 CCP 的一**般版本**，研究在任意抽取概率下，抽取多少张优惠券才能收集到 k 种不同优惠券各 t 个副本。事实上，这一一般设定下抽取优惠券数量的期望值已在 [2], [4], [11]–[13] 中求得（见公式（1））。然而，即使 n 的取值适中，计算上也无法实际求出该公式，k, t 原因在于计算溢出、舍入误差和时间复杂度。例如， [14], 中指出：“看来只有在少数情况下，而且对于相当小的 n，才会明确计算出期望值或概率”； [14] 中也报告称：“情况似乎糟糕到这样的程度：一个应用 CCP 的领域中的研究者抱怨说，一旦 n 超过大约 30，是否知道这些概率都无关紧要——因为无论如何都无法据此进行计算。”

本文的主要贡献在于提出高效方法，精确计算收集过程的期望和方差，从而克服这些主要缺陷。具体而言，当概率分布为均匀分布时，对于任意 t，$k > 1$，根据公式 (2) 计算精确期望的复杂度为指数级，且为 n；而我们提出的算法的复杂度为 $O(n^t)$。

由于精确计算存在数值方面的挑战，渐近估计具有重要意义。已有多项研究 [8], [13], [15]，[16] 得到了均匀抽取概率下 classicCCP 情形的一些渐近结果。然而，对于任意抽取概率，只有在 $\boldsymbol{p}$ 由若干参数刻画的少数特定情形下，才得到了渐近估计。因此，据我们所知，在 CCP 的一般设置中，目前没有高效的方法能够显式计算，或者说估计收集过程的期望。

本文结构安排如下：第 II 节给出定义和问题陈述，并讨论 CCP 的一些相关既有结果。第 III 节介绍一般 CCP 的基础算法，并分析其时间复杂度。随后，第 IV 节提出一种适用于均匀分布下的一般 CCP 的算法，其时间复杂度为 $O(n^t)$。接着，第 V 节将该算法扩展到任意概率分布，并分析其时间复杂度。第 VI 节评估并比较这些算法。最后，第 VII 节总结研究结果并作出结论。

II. 定义、问题陈述、相关工作

A. 定义与符号说明

设 n 为不同优惠券的数量， k 为要收集的不同优惠券的目标数量， t 为获得一张优惠券所需的副本数。概率向量 $\boldsymbol{p} \in \mathbb{R}^n$ 表示抽取概率，其中 p_i 是发放优惠券 i 的概率。为分析收集过程，我们将其表示为一个离散时间马尔可夫链。该过程中的每个状态都是一个长度为 n 的向量，其中第 i 个值表示已收集的第 i 种优惠券的副本数。

由于我们只需要每种收集到的优惠券各有 t 个副本数，因此无需跟踪大于 t的数值。这也意味着马尔可夫链是有限的。该马尔可夫模型正式定义如下。

定义1。优惠券收集过程的马尔可夫模型由以下内容组成 :

1) 令 X_b 为一个随机变量，表示抽取 b 张“跟踪”优惠券后收集过程所处的状态。2) 令 $\mathcal{S}_{(n,t)}$ 为状态集合，使得 $\mathcal{S}_{(n,t)} = \{\boldsymbol{s} = (s_1, \ldots, s_n) \mid \forall i \in [n]\colon s_i \leq t\}$，其中，对于每个 $1 \leq i \leq n$， s_i表示从第 i种优惠券中抽取的副本数。对于 $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$，令 $\widehat{\boldsymbol{s}} \triangleq \sum_{i=1}^{n} s_i$为状态 $\boldsymbol{s}$下已收集的跟踪优惠券总数。3) 将至少有 t 个副本数的优惠券定义为完整优惠券，将少于 t 个副本数的优惠券定义为未完整优惠券。令 $F(\boldsymbol{s}) \triangleq \{i \mid s_i = t\}$ 表示状态 $\boldsymbol{s}$中的完整优惠券集合。将层 $L(h)$定义为所有包含 h张完整优惠券的状态的集合，即 $L(h) \triangleq \{\boldsymbol{s} \mid |F(\boldsymbol{s})| = h\}$。4) 该马尔可夫链的终止状态是 $L(k)$中的所有状态，因为该过程在首次到达某个包含 k 张至少有 t个副本数的优惠券的状态时结束。5) 如果 $\forall i \in [n]$ 成立，则称状态 $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$ 可从状态 $\boldsymbol{s}'$ 到达。此外， $A(\boldsymbol{s})$是状态 $\boldsymbol{s}$的直接前驱集合，即 $A(\boldsymbol{s}) = \{\boldsymbol{s}' | \boldsymbol{s}$ 可从 $\boldsymbol{s}'$ 到达，且 $\widehat{\boldsymbol{s}} - \widehat{\boldsymbol{s}'} = 1\}$。6) 对于所有 $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$，令 $D_{\boldsymbol{s}}$ 为一个随机变量，表示从收集过程开始到到达状态 $\boldsymbol{s}$所需的抽取次数。7) 令 M 表示该马尔可夫链的转移矩阵，其中，对于两个状态 $\boldsymbol{s}'$ 和 $\boldsymbol{s}$， $M_{\boldsymbol{s}',\boldsymbol{s}} = \Pr(X_b = \boldsymbol{s} | X_{b-1} = \boldsymbol{s}')$。定义 $M^{(n)}_{\boldsymbol{s}',\boldsymbol{s}} = \Pr(X_n = \boldsymbol{s} | X_0 = \boldsymbol{s}')$，表示从 $\boldsymbol{s}'$ 到 $\widehat{\ }$收集到额外“跟踪”优惠券的概率，即在状态 $\boldsymbol{s}$中优惠券的组成（也就是 $\widehat{\boldsymbol{s}} - \widehat{\boldsymbol{s}'} = n$条）。 n步转移概率矩阵定义为 $M^{(n)} \triangleq (M^{(n)}_{\boldsymbol{s}',\boldsymbol{s}})$。转移矩阵 M和 $M^{(n)}$被分析并定义为：在抽取到未完整优惠券的条件下转移到新状态的条件概率。

优惠券收集过程满足马尔可夫性质，即对于一组 b状态，记为 $\boldsymbol{s}_0$， $\boldsymbol{s}_1$，……， $\boldsymbol{s}_{b-1}$：$\Pr(X_b = \boldsymbol{s}_b | X_0 = \boldsymbol{s}_0$，……，$X_{b-1} = \boldsymbol{s}_{b-1}) = \Pr(X_b = \boldsymbol{s}_b | X_{b-1} = \boldsymbol{s}_{b-1})$。为简便起见，我们将初始状态记为 $\boldsymbol{s}_0$（即每种优惠券的副本数为零），且$\Pr(X_0 = \boldsymbol{s}_0) = 1$。在任意状态 $\boldsymbol{s}$，存在两种可能情形：要么抽到已有 t 副本数的完整优惠券而停留在当前状态，要么收集到未完整优惠券并进入新状态。

下一条断言给出了直接由定义1得出的基本性质。

断言1. 给定 $\boldsymbol{s}'$， $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$，则有

1）若 $\boldsymbol{s}' \notin A(\boldsymbol{s})$，则$M_{\boldsymbol{s}',\boldsymbol{s}} = 0$。2） $M^{(n)}_{\boldsymbol{s}'',\boldsymbol{s}} = \sum_{\boldsymbol{s}' \in A(\boldsymbol{s})} M^{(n-1)}_{\boldsymbol{s}'',\boldsymbol{s}'} M_{\boldsymbol{s}',\boldsymbol{s}}$，且若 $\widehat{\boldsymbol{s}} - \widehat{\boldsymbol{s}''} \neq n$，则 $M^{(n)}_{\boldsymbol{s}'',\boldsymbol{s}} = 0$。3）对于 $\boldsymbol{s}'$， $\boldsymbol{s} \in L(k)$，有$M_{\boldsymbol{s}',\boldsymbol{s}} = M_{\boldsymbol{s},\boldsymbol{s}'} = 0$。4）对所有 $\boldsymbol{s}' \in A(\boldsymbol{s})$，有 $I(\boldsymbol{s}', \boldsymbol{s}) = i$ 当且仅当 $s_i - s'_i = 1$，这表示从状态 $\boldsymbol{s}'$ 转移到状态 $\boldsymbol{s}$时所收集的优惠券。

下一个定义将引入辅助变量，用于分析收集过程。

定义2。 给定抽取概率向量 $\boldsymbol{p}$，对于每个状态 $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$，我们定义 1) $p^F(\boldsymbol{s}) := \sum_{i \in F(\boldsymbol{s})} p_i$: 抽取到完整优惠券的概率。2) $d(\boldsymbol{s})$ 是一个随机变量，用于描述从状态 $\boldsymbol{s}$ 转移到新状态所需的抽取次数。注意， $d(\boldsymbol{s})$ 服从几何分布，也就是说，$d(\boldsymbol{s}) \sim$ 几何分布$(1 - p^F(\boldsymbol{s}))$。

例1。给定 $n = 3$， $k = 2$， $t = 2$， $\boldsymbol{p} = (0.2,\ 0.3,\ 0.5)$，**它**

1) $\mathcal{S}_{(3,2)} = \{\boldsymbol{s} = (s_1, s_2, s_3) \mid \forall i \in [3]\colon s_i \leq 2\}$2) 对于状态 $\boldsymbol{s} = (1, 2, 0)$，优惠券 2 是完整优惠券，而 1、3 是未完整优惠券。此外， $F(\boldsymbol{s}) = \{i \mid s_i = 2\} = \{2\}$ 和 $\widehat{\boldsymbol{s}} = 3$, $p^F(\boldsymbol{s}) = 0.3$, $\boldsymbol{s}' = (0, 2, 0) \in A(\boldsymbol{s})$ 和 $d(\boldsymbol{s}) \sim$几何分布$(1 - 0.3)$。3) 终止状态为 $L(2) = \{(2, 2, 0), (2, 2, 1), (0, 2, 2), (1, 2, 2), (2, 0, 2), (2, 1, 2)\}$

B. 问题陈述

在优惠券收集过程的一般版本中，n 优惠券按照概率向量 $\boldsymbol{p}$有放回地抽取，目标是收集 k 种不同优惠券，每种至少抽取 t 次。令 $\nu_t^{\boldsymbol{p}}(n,k)$ 表示所需的抽取次数；当 $\boldsymbol{p}$均匀时，记作 $\nu_t(n,k)$。本文探讨以下问题：

问题 1. 给定 n $n \geq k \geq 1$ 和 $\boldsymbol{p}$，计算精确的期望 $\mathbb{E}[\nu_t^{\boldsymbol{p}}(n,\ k)]$ 和方差 $Var(\nu_t^{\boldsymbol{p}}(n,\ k))$.

C. 相关工作

经典优惠券收集问题（即 $t=1$）最早由 Feller [1], 研究，当时它被称为迪克西杯问题。对于 n 种等概率的优惠券，收集齐所有优惠券所需抽取次数的期望为 $\mathbb{E}[\nu_1(n,\ k=n)]=nH_n=n\log\ n+\gamma n+\mathcal{O}(1)$，其中 H_n是第 n 个调和数，$\gamma\sim 0.577$ 是欧拉–马歇罗尼常数。更一般地，根据 [2], $\mathbb{E}[\nu_1(n,\ k)]=n(H_n-H_{n-k})$；当 $n\to\infty$（在本例中，存在 $0<a<1$，使得对于足够大的 n，$k<an$）时，可近似为 $\mathbb{E}[\nu_1(n,\ k)]\approx n\log(\frac{n}{n-k})$。

对于一般CCP（$t>1$）而言，该问题与经典 urn 模型 [11], [13]密切相关。在拥有 n 个urn、抽取概率为 p $=$（p_1，.....，p_n）的情况下，Newman [13] 证明，当 $n\to\infty$ 且服从均匀分布时，$\mathbb{E}[\nu_t(n,\ n)]=n\log\ n+n(t-1)\log\log\ n+nC_t+o(n)$，其中 C_t 是取决于 t的常数。Erd˝os 和 Rényi [11] 进一步证明，$\nu_t(n,\ n)$ 高度集中于其均值附近：经过 $n\log\ n+n(t-1)\log\log\ n+nx$ 次抽取后，所有优惠券e^{-x}至少出现 t 次的概率，对于足够大的 n，收敛于 $e^{-\frac{}{(t-1)!}}$。在附加约束下对非均匀分布的扩展研究见 [17], , Flajolet 等人 [2] 则建立了任意离散分布下的结果，并证明要从 n 张优惠券中获得至少 t 份 k 所需的预期抽取次数为

$$\mathbb{E}[\nu_t^{\boldsymbol{p}}(n,k)]=\sum_{q=0}^{k-1}\int_0^\infty[u^q]\prod_{i=1}^{n}\Big(e_{t-1}(p_iv)+u\cdot\big(e^{p_iv}-e_{t-1}(p_iv)\big)\Big)e^{-v}dv \tag{1}$$

其中 $e_t(x)=\sum_{i=0}^{t}\frac{x^i}{i}!$ 并且对于多项式 $Q(u)$，$[u^q]Q(u)$ 是 $Q(u)$ 中 u^q 的系数。若 $\boldsymbol{p}$ 是均匀分布，则利用 [2] 的结果可得

$$\mathbb{E}[\nu_t(n,k)]=\sum_{q=0}^{k-1}\binom{n}{q}\left[\int_0^\infty\Big(e_{t-1}(\frac{v}{n})\Big)^{n-q}\Big(e^{\frac{v}{n}}-e_{t-1}(\frac{v}{n})\Big)^{q}e^{-v}dv\right]. \tag{2}$$

可见，在实际应用中，式（1）和式（2）中的表达式及其渐近行为都不易计算或处理。此外，据我们所知，显式计算累积分布函数$P[\nu_t^{\boldsymbol{p}}(n,\ k)>m]$ 仍是一个尚未解决的问题。再者，[4], [14] 指出：“对式（1）、式（2）中求和项的积分进行计算，工作量极其庞大，而且要处理 2^n 振荡项.....除非 k 是一个非常小的数——或同样地，非常接近 n——否则，基本上没有更简单的方式来表示这个真正复杂的组合量。”

先前的研究已针对均匀分布和非均匀分布下的 CCP，给出了若干极限结果和近似方法。Holst [8] 推导了 $\nu_t(n,\ k)$ 的分布，并针对 $k=n$ 在 $n\to\infty$ 时的情形，给出了其期望和累积概率的渐近界。文献 [18] 还给出了在抽取总次数受到限制时，与 CCP 相关的随机变量的其他极限结果。文献 [7], [19], [20] 进一步推导了 $\nu_t^{\boldsymbol{p}}(n,\ k=n)$ 的矩、期望和方差的渐近表达式。对于非均匀分布，Boneh 和 Hofri [4] 表明，可以在线性时间内计算出准确的近似值 t，而精确计算则需要指数时间。

一些研究使用马尔可夫链分析 CCP 的变体 [5], [21], [22], , 通常基于诸如$n\to\infty$ 和/或 $k=n$之类的渐近假设。概率通过反复乘以转移矩阵来计算，但状态空间呈指数级增长，使这种方法计算开销大，且容易出现数值不稳定。

III. 一般 CCP 算法

本节介绍用于直接计算一般 CCP 的期望和方差的基础算法（**BA**）。虽然概念上很简单，但为了完整起见，我们仍将其列出；据我们所知，文献中尚未出现过该算法。蒙特卡洛模拟是另一种方法，但既不精确，也无法扩展。

该算法遵循第 II-A 节中介绍的基础马尔可夫模型定义。对马尔可夫链中的每个状态 $\boldsymbol{s}\in\mathcal{S}_{(n,t)}$，都会计算这些值，具体包括：

1) 转移概率 $M_{\boldsymbol{s}_0,\ \boldsymbol{s}}^{(\widehat{\boldsymbol{s}})}=\Pr(X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}\mid X_0=\boldsymbol{s}_0)$。**2)** 到达 $\boldsymbol{s}$所需抽取次数的期望和二阶矩，分别记为 $\mathbb{E}[D_{\boldsymbol{s}}\mid X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}]$ 和$\mathbb{E}[D_{\boldsymbol{s}}^2\mid X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}]$。

下一个命题定义转移矩阵的值。

命题 2. 给定 $\boldsymbol{s}\in\mathcal{S}_{(n,t)}$，$\boldsymbol{s}'\in A(\boldsymbol{s})$，有

1) $M_{\boldsymbol{s}',\boldsymbol{s}} = \frac{p_{I(\boldsymbol{s}',\boldsymbol{s})}}{1-p^F(\boldsymbol{s}')}$.
2) $M^{(\widehat{\boldsymbol{s}})}_{\boldsymbol{s}_0,\boldsymbol{s}} = \sum_{\boldsymbol{s}'\in A(\boldsymbol{s})} M^{(\widehat{\boldsymbol{s}'})}_{\boldsymbol{s}_0,\boldsymbol{s}'} \cdot M_{\boldsymbol{s}',\boldsymbol{s}}$.

接下来的引理分别计算达到状态 $\boldsymbol{s}$ 所需抽取次数的期望和二阶矩。

引理1。 给定状态 $\boldsymbol{s}\in\mathcal{S}_{(n,t)}$，有

$$\mathbb{E}[D_{\boldsymbol{s}}|X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}] = \sum_{\boldsymbol{s}'\in A(\boldsymbol{s})} \frac{M^{(\widehat{\boldsymbol{s}'})}_{\boldsymbol{s}_0,\boldsymbol{s}'}\cdot M_{\boldsymbol{s}',\boldsymbol{s}}}{M^{(\widehat{\boldsymbol{s}})}_{\boldsymbol{s}_0,\boldsymbol{s}}} \cdot \left(\mathbb{E}[D_{\boldsymbol{s}'}|X_{\widehat{\boldsymbol{s}'}}=\boldsymbol{s}'] + \frac{1}{1-p^F(\boldsymbol{s}')}\right).$$

证明概要。 给定 $X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}$，到达状态 $\boldsymbol{s}$ 所需抽取次数的期望，计算方法是对所有祖先状态 $\boldsymbol{s}'\in A(\boldsymbol{s})$ 求和：将其转移至 $\boldsymbol{s}$ 的转移概率，乘以到达 $\boldsymbol{s}'$ 的期望与离开该状态的期望之和，即 $\frac{1}{1-p^F(\boldsymbol{s}')}$。

□

引理2。 给定状态 $\boldsymbol{s}\in\mathcal{S}_{(n,t)}$，有

$$\mathbb{E}[D^2_{\boldsymbol{s}}|X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}] = \sum_{\boldsymbol{s}'\in A(\boldsymbol{s})} \frac{M^{(\widehat{\boldsymbol{s}'})}_{\boldsymbol{s}_0,\boldsymbol{s}'}\cdot M_{\boldsymbol{s}',\boldsymbol{s}}}{M^{(\widehat{\boldsymbol{s}})}_{\boldsymbol{s}_0,\boldsymbol{s}}} \cdot \left(\mathbb{E}[D^2_{\boldsymbol{s}'}|X_{\widehat{\boldsymbol{s}'}}=\boldsymbol{s}'] + \frac{2}{1-p^F(\boldsymbol{s}')}\cdot\mathbb{E}[D_{\boldsymbol{s}'}|X_{\widehat{\boldsymbol{s}'}}=\boldsymbol{s}'] + \frac{1+p^F(\boldsymbol{s}')}{(1-p^F(\boldsymbol{s}'))^2}\right).$$

证明。 给定 $X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}$，到达状态 $\boldsymbol{s}$ 所需抽取次数的二阶矩，计算方法是对所有祖先状态 $\boldsymbol{s}'\in A(\boldsymbol{s})$ 求和：将其转移至 $\boldsymbol{s}$ 的转移概率，乘以经由 $\boldsymbol{s}'$ 到达 $\boldsymbol{s}$ 的二阶矩。注意，到达状态 $\boldsymbol{s}'$ 所需的抽取次数记为 A，而从 $\boldsymbol{s}'$ 转移至 $\boldsymbol{s}$ 所需的抽取次数记为 B；这两个随机变量相互独立，因此可得：$\mathbb{E}[(A+B)^2] = \mathbb{E}[A^2] + 2\mathbb{E}[A]\cdot\mathbb{E}[B] + \mathbb{E}[B^2]$。此外，由于 $d(\boldsymbol{s})\sim$服从参数为 $1-p^F(\boldsymbol{s})$ 的几何分布，二阶矩为 $\mathbb{E}[d(\boldsymbol{s})^2] = \frac{1+p^F(\boldsymbol{s})}{(1-p^F(\boldsymbol{s}))^2}$。综合以上结果即可得到引理所述结论。

□

回顾 II-B 节中的 $\nu^{\boldsymbol{p}}_t(n,\ k)$。最后，下面的定理给出了计算一般CCP收集过程的期望和方差的公式。

定理1。 给定 $n,\ k,\ t,\ \boldsymbol{p}$，有

1) $\mathbb{E}[\nu^{\boldsymbol{p}}_t(n,k)] = \sum_{\boldsymbol{s}\in L(k)} M^{(\widehat{\boldsymbol{s}})}_{\boldsymbol{s}_0,\boldsymbol{s}}\cdot\mathbb{E}[D_{\boldsymbol{s}}|X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}]$.
2) $\mathbb{E}[\nu^{\boldsymbol{p}}_t(n,k)^2] = \sum_{\boldsymbol{s}\in L(k)} M^{(\widehat{\boldsymbol{s}})}_{\boldsymbol{s}_0,\boldsymbol{s}}\cdot\mathbb{E}[D^2_{\boldsymbol{s}}|X_{\widehat{\boldsymbol{s}}}=\boldsymbol{s}]$.
3) $Var(\nu^{\boldsymbol{p}}_t(n,k)) = \mathbb{E}[\nu^{\boldsymbol{p}}_t(n,k)^2] - \mathbb{E}[\nu^{\boldsymbol{p}}_t(n,k)]^2$.

算法 1 是一种动态规划方法的伪代码算法，它使用定理1中给出的变量。注意，对于层 k 中的任意终止状态，可抽取的优惠券最大数量恰好为 k张优惠券，其中有 t 个副本，另有 $n-k$ 张优惠券各有 $t-1$ 个副本，即 $kt+(n-k)(t-1)$。

BA 的时间复杂度由马尔可夫链中的边数推导得出。用 $V_{BA}(n,\ k,\ t)$ 表示 **BA** 中马尔可夫链对应的顶点数。类似地，令 $E_{BA}(n,\ k,\ t)$ 表示 **BA** 马尔可夫链中的边数。下面的定理给出了所有 $n,\ k,\ t$ 对应的顶点集和边集的大小。

定理 2。 给定 $n,\ k,\ t$，有

1) $V_{BA}(n,k,t) = \sum_{h=0}^{k}\binom{n}{h}\cdot V_{BA}(n-h,n-h,t-1) = \sum_{h=0}^{k}\binom{n}{h}\cdot t^{n-h}$.
2) $E_{BA}(n,k,t) = nt\cdot V_{BA}(n-1,k-1,t)$.

证明。 1）设 h 为每个状态中的完整优惠券数量，因此 h 的值介于 0 和 k之间。对于给定的 h，有 $\binom{n}{h}$ 种选取 h 个完整优惠券索引的方式。其余每个索引对应的优惠券数量必须介于 0 和 $t-1$ 之间，因此有 t 种选择。

2）马尔可夫链中的每条边由以下方式确定：首先选取抽到的未完整优惠券及其对应的副本数，共有 nt 种可能的选择。在其余 $n-1$ 张优惠券中，最多可以有 $k-1$ 张完整优惠券，以 $V_{BA}(n-1,\ k-1,\ t)$表示。请注意，更新一张未完整优惠券时，不可能存在 k 张完整优惠券。

□

请注意，如果 $k=n$，那么我们得到 $V_{BA}(n,\ n,\ t)=(t+1)^n$，且 $E_{BA}(n,\ n,\ t)=nt(t+1)^{n-1}$。下一个推论给出了 **BA** 的马尔可夫链中状态数和边数的上界与下界。

推论 1. 给定 n,k,t，有

算法 1 BA期望与二阶矩计算

初始化： $current_layer \leftarrow [\boldsymbol{s}_0]$ $E \leftarrow 0$ # 优惠券收集过程的期望 $E2 \leftarrow 0$ # 优惠券收集过程的二阶矩 $TS \leftarrow []$ # 终止状态集合 $new_layer \leftarrow []$ **对于** $iter \leftarrow 0$ 到 $(k \cdot t + (n-k) \cdot (t-1))$ **执行** $new_layer \leftarrow$从 $current_layer$创建可达状态 $new_layer \leftarrow$根据 $current_layer$中的祖先更新新状态的累积概率、期望和二阶矩 $TS \leftarrow new_layer$中的终止状态 $E \leftarrow E + \sum P(s) \cdot E(s)_{s \in TS}$ $E2 \leftarrow E2 + \sum P(s) \cdot E2(s)_{s \in TS}$ $current_layer \leftarrow new_layer \setminus TS$ **结束循环** **返回：** $E, E2$

1) $t^n = V_{BA}(n,n,t-1) \le V_{BA}(n,k,t) \le V_{BA}(n,n,t) = (t+1)^n$, 2) $n(t-1)^n < n(t-1) \cdot t^{n-1} = E_{BA}(n,n,t-1) \le E_{BA}(n,k,t) \le E_{BA}(n,n,t) = nt \cdot (t+1)^{n-1} < n(t+1)^n$. 因此， $V_{BA}(n,k,t) = O((t+1)^n)$，且 $E_{BA}(n,k,t) = O((t+1)^n \cdot n)$，这就是 BA 的时间复杂度 .

IV. 均匀分布的 CCP 算法

本文的主要目标是能够在 CCP 的一般版本中，高效计算抽取过程的期望。第 III 节的方法提供了一种动态规划算法，但随着 n 增大，该模型中的状态数量呈指数级增长。本节介绍 **UDA**（均匀分布算法），并对马尔可夫模型进行修改后的分析；当 $\boldsymbol{p}$ 为均匀分布时（即 $p_i = \frac{1}{n}, 1 \le i \le n$），状态数量会减少。因此，计算时间显著缩短，并且关于 n呈多项式级别。该方法基于合并马尔可夫链中的对称状态。

在收集过程中，由于抽到每种优惠券的概率相同，因此给定状态的任意排列在概率和期望方面都等价（例如，状态 (0, 1, 1)、(1, 0,1) 和 (1, 1,0) 的分析方式相同）。为了利用这种对称性，我们将这些状态归并为一个代表状态，称为重数向量状态。每个状态向量都可以表示为一个重数向量，该向量表示具有固定副本数的优惠券数量。副本数介于 0 和 t 之间，令 x_i 表示具有 i 个副本的优惠券数量。具体而言，状态 $\boldsymbol{s} = (s_1, \ldots, s_n)$ 可表示为重数向量状态：$(x_0, x_1, \ldots, x_t)$。此表示满足 $\sum_{i=0}^{t} x_i = n$。记 $mv(\boldsymbol{s})$ 为 $\boldsymbol{s}$ 的重数向量表示（例如，给定 $\boldsymbol{s} = (0,0,0,1,1,2)$，则 $mv(\boldsymbol{s}) = (x_0 = 3, x_1 = 2, x_2 = 1, x_3 = 0)$），并记 $MV(\mathcal{S}_{(n,t)})$ 为 $\mathcal{S}_{(n,t)}$ 的所有重数向量状态构成的集合。

回想 $d(\boldsymbol{s})$，由于 $\boldsymbol{p}$是均匀分布，因此对所有 $\boldsymbol{s} \in L(h)$ $p^F(\boldsymbol{s}) = \frac{h}{n}$，都有此性质。因此， $d(\boldsymbol{s}) \sim$服从几何分布 Geom $\left(\frac{n-h}{n}\right)$ 且$\mathbb{E}[d(\boldsymbol{s})] = \frac{n}{n-h}$。注意，给定状态 $mv(\boldsymbol{s}) = (x_0, x_1, \ldots\ldots, x_t) \in L(h)$，可得 $x_t = h$，因为 x_t表示完整优惠券的数量（即拥有 t 个副本数）。这表明终止状态为 $L(k) = \{\boldsymbol{x} \mid x_t = k\}$。下一个断言将定义转移矩阵的取值以及可达状态的数量（单次抽取）。在此情形下， $I(\boldsymbol{x}', \boldsymbol{x}) = i$ 读入一个拥有 i 个副本数的优惠券，从而使其副本数增加一个，进而产生从 $\boldsymbol{x}'$ 到 $\boldsymbol{x}$的转移。

断言3。给定 $\boldsymbol{s}' \in \mathcal{S}_{(n,t)}$, $\boldsymbol{x}' = mv(\boldsymbol{s}') = (x'_0, x'_1, \ldots\ldots, x'_t) \in MV(\mathcal{S}_{(n,t)})$，有 1) 从 s' 出发一步可达状态集合的大小为 $|\{\boldsymbol{x} = mv(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}| = \sum_{i=0}^{t-1} \mathbb{I}_{\{x'_i \ge 0\}}$。 2) $p^F(\boldsymbol{s}') = \frac{x'_t}{n}$。 3) 对所有 $\boldsymbol{x} \in \{\boldsymbol{x} = mv(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}$： $0 \le I(\boldsymbol{x}', \boldsymbol{x}) \le t-1$和 $M_{\boldsymbol{x}', \boldsymbol{x}} = {}^{x'_t}_{I(\boldsymbol{x}', \boldsymbol{x})}\, {}^{n-x'_t}$。 4) 对所有 $\boldsymbol{x} \in \{\boldsymbol{x} = mv(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}$： $\boldsymbol{x} = (x_0, \ldots, x_{I(\boldsymbol{x}', \boldsymbol{x})} - 1, x_{I(\boldsymbol{x}', \boldsymbol{x})+1} + 1, x_t)$。

示例2. 给定 $\boldsymbol{s}' = (0,0,0,1,1,2) \in \mathcal{S}_{(6,3)}$, $\boldsymbol{x}' = mv(\boldsymbol{s}') = (x'_0 = 3, x'_1 = 2, x'_2 = 1, x'_3 = 0) \in MV(\mathcal{S}_{(6,3)})$

- $I(\boldsymbol{x}', \boldsymbol{x}) = 0$： $\boldsymbol{x} = (x_0 = 2, x_1 = 3, x_2 = 1, x_3 = 0)$。

- $I(\boldsymbol{x}', \boldsymbol{x}) = 1$： $\boldsymbol{x} = (x_0 = 3, x_1 = 1, x_2 = 2, x_3 = 0)$。
- $I(\boldsymbol{x}', \boldsymbol{x}) = 2$： $\boldsymbol{x} = (x_0 = 3, x_1 = 2, x_2 = 0, x_3 = 1)$。

从状态 $\boldsymbol{x}'$ 出发，有三种可能的转移，每种转移都会改变一个索引；例如， $I(\boldsymbol{x}', \boldsymbol{x}) = 0$ 会使具有一个副本的优惠券数量增加 1，并使没有副本的优惠券数量减少 1。

对值 $M_{s_0}^{(\widehat{s})}$, s, $\mathbb{E}[D_{\boldsymbol{s}}|X_{\widehat{s}} = s]$, $\mathbb{E}[D_{\boldsymbol{s}}^2|X_{\widehat{s}} = s]$ 的其余分析与引理1、引理2 和定理1 中的分析相同，但将重数向量状态 $MV(\mathcal{S}_{(n,t)})$ 作为状态集合。以下将按照第 III 节中的讨论，分析 **UDA** 的时间复杂度。类似地，令 $V_{UDA}(n, k, t)$ 和 $E_{UDA}(n, k, t)$ 分别表示 **UDA** 马尔可夫链中的顶点数和边数。我们首先分析情形 $k = n$，这是一般复杂度分析的基础。

引理3。 给定 n和 t，有

1) $V_{UDA}(n, n, t) = \binom{n+t}{t}$.
2) $E_{UDA}(n, n, t) = t \cdot \binom{n+t-1}{t} = n \cdot \binom{n+t-1}{n}$.

证明。 1）根据重数向量状态的定义，状态数等于方程 $x_0 + x_1 + \cdots + x_t = n$在非负整数上的解的数量，该数量为 $\binom{n+t}{t}$。

2）我们根据 x_i （在重数向量状态中）在每条入边上发生变化的取值来计算入边的数量。假设一条入边将重数向量 $(x_0, x_1, \ldots, x_t)$ 中 x_i的值改变为 $1 \le i \le t$，因此该值增加1。因此，选项数等于方程 $x_0 + x_1 + \cdots + x_t = n$在非负整数上的解的数量，其中 $x_i \ge 1$，该数量为 $\binom{n+t-1}{t}$。

□

定理3。 给定 n, k, t，有

1) $V_{UDA}(n, k, t) = V_{UDA}(n, n, t) - V_{UDA}(n-(k+1)\ n-(k+1)\ t)$ $= \binom{n+t}{n} - \binom{n+t-(k+1)}{n-\ k+1\ \ ()}$。 2) $E_{UDA}(n, k, t) = t \cdot \binom{n+t-1}{t} - t \cdot \binom{n+t-k-1}{t}$。

证明。 1) 类似于引理3的证明，我们在约束 $x_t \le k$ 下计算状态的数量。这等于 $V_{UDA}(n, n, t)$ 的值，再减去满足 $x_t \ge k+1$ 的解的数量，即 $\binom{n+t-(k+1)}{n-\ k+1\ \ ()}$，最终得到 $\binom{n+t}{n} - \binom{n+t-(k+1)}{n-\ k+1\ \ ()}$。

2) 我们按照引理3的证明方法计算入边，但排除所有来自高于第 k 层的状态的边（因为第 k 层的状态是终止状态）。假设一条入边使 $1 \le i \le t$ 对应的 x_i 值改变 1。因此，我们减去方程 $x_0 + x_1 + \cdots + x_t = n$ 的解的数量，其中 $x_t \ge k$，而这个数量为 $\binom{n+t-k-1}{t}$ 。

□

下一个推论 给出了 **UDA** 马尔可夫链中状态数和边数的上下界 .

推论 2。 给定 n， k， t，有

1) $\binom{n+t-1}{n} \le V_{UDA}(n, k, t) \le \binom{n+t}{n}$。 2) $n \cdot \binom{n+t-2}{n} \le E_{UDA}(n, k, t) \le n \cdot \binom{n+t-1}{n}$。因此，$V_{UDA}(n, k, t) = O(n^t)$，且 $E_{UDA}(n, k, t) = O(n^t)$，这就是 **UDA** 的时间复杂度。

V. 利用对称性的任意分布 CCP 算法

第 III 节中的方法确实提供了一种动态规划算法，可以处理从每种 n 优惠券中抽取 $\boldsymbol{p}$ 的任意分布，但与第 IV 节中具有多项式时间复杂度的均匀方法相比，其时间复杂度呈指数增长。本节将介绍 **DPSA** （抽取概率对称算法），并展示如何利用具有相同抽取概率的优惠券之间的对称性，以与第 IV 节相同的方式减少计算时间。这是更常见的情况，即存在多个类型的优惠券，且每种类型具有相同的概率。例如，有两种类型的优惠券，每种类型对应不同的概率。

定义 3。 为了分析抽取概率向量 $\boldsymbol{p}$ 中的对称性，记以下符号： s:

1） G 是概率向量 $\boldsymbol{p}$ 中不同取值的数量，即 $G = |\{p_i \in \boldsymbol{p} \mid 1 \le i \le n\}|$。 2） q_1、.....、 q_G 是不同取值，即 $\{p_i \in \boldsymbol{p} \mid 1 \le i \le n\}$，其中 $\forall 1 \le g \le G-1$： $q_g > q_{g+1} > 0$。 3）对于 $1 \le g \le G$：$C_g \triangleq \{i | p_i = q_g, p_i \in \boldsymbol{p}\}$ 是第 g 组优惠券。注意， $\sum_{g=1}^{G} |C_g| = n$。

该马尔可夫链中的每个状态 $\boldsymbol{s} \in \mathcal{S}_{(n,t)}$ 都表示为一个向量，其中包含如第 IV 节所述的 G 重数向量。我们令 $x_{(g,i)}$ 表示第 g 组中抽取次数为 i 的优惠券数量。具体而言，状态 $\boldsymbol{s} = (s_1, \ldots, s_n)$ 表示为

G $x_{(1,0)}\ x_{(1,1)}\ \cdots\ x_{(1,t)}\ \cdots\ x_{(G,0)}\ x_{(G,1)}\ \cdots\ x_{(G,t)}$ $\forall 1 \le$ 重数向量状态： $\Big((,,),(,,)\Big)$。

$g \le G$ $\sum_{i=0}^{t} x_{(g,i)} = |C_g|$ $v_g = x_{(g,0)}\ x_{(g,1)}\ \cdots\ x_{(g,t)}$ $\boldsymbol{s}$

此表示满足： 。为方便起见，记：(,,)，于是可表示为

$(v_1 \ldots, v_G)$。将 $mv_G(\boldsymbol{s})$ 记为 $\boldsymbol{s}$ 的重数向量表示，将 $MV_G(\mathcal{S}_{(n,t)})$ 记为所有 $\mathcal{S}_{(n,t)}$ 的重数向量状态的集合。注意，给定状态 $\boldsymbol{s} \in L(h)$，可知 $h = \sum_{g=1}^{G} x_{(g,t)}$。由于对于所有 $1 \le g \le G, x_{(g,t)}$ 表示完整优惠券的数量（即拥有 t 副本数的优惠券）。这表明终止状态为 $L(k) = \left\{ \boldsymbol{x} \;\middle|\; \sum_{g=1}^{G} x_{(g,t)} = k \right\}$。

回忆 $d(\boldsymbol{s})$，由于 $\boldsymbol{p}$ 是均匀分布，因此对所有 $\boldsymbol{s} \in L(h)$ $p^F(\boldsymbol{s}) = \frac{h}{n}$ 均成立。因此， $d(\boldsymbol{s})$ ~服从几何分布 $\mathrm{Geom}\left(\frac{n-h}{n}\right)$，且 $\mathbb{E}[d(\boldsymbol{s})] = \frac{n}{n-h}$。注意，给定状态 $mv(\boldsymbol{s}) = (x_0, x_1, \ldots, x_t) \in L(h)$，可得 $x_t = h$，因为 x_t 表示完整优惠券的数量（即拥有 t 个副本数）。这表明终止状态为 $L(k) = \{\boldsymbol{x} \mid x_t = k\}$。下一个命题定义转移矩阵的取值以及可达状态的数量（单次抽取）。在这种情况下， $I(\boldsymbol{x}', \boldsymbol{x}) = i$ 读到一张拥有 i 个副本数的优惠券后，会增加一个副本数，从而引发从 $\boldsymbol{x}'$ 到 $\boldsymbol{x}$ 的转移。

下一个命题定义转移矩阵的取值以及可达状态的数量（单次抽取）。在这种情况下，$I(\boldsymbol{x}', \boldsymbol{x}) = (g, i)$ 从组 g 中抽到一张拥有 i 个副本数的优惠券后，会增加一个副本数，从而引发从 $\boldsymbol{x}'$ 到 $\boldsymbol{x}$ 的转移。

命题4. 给定 $\boldsymbol{s}' \in \mathcal{S}_{(n,t)}, \boldsymbol{x}' = mv_G(\boldsymbol{s}') = \left((x'_{(1,0)}, x'_{(1,1)}, \ldots, x'_{(1,t)}) \ldots, (x'_{(G,0)}, x'_{(G,1)}, \ldots, x'_{(G,t)})\right) \in MV_G(\mathcal{S}_{(n,t)})$ 成立

1) $|\{\boldsymbol{x} = mv_G(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}| = \sum_{g=1}^{G} \sum_{i=0}^{t-1} \mathbb{I}_{\{x'_{(g,i)} > 0\}}$. 2) $p^F(\boldsymbol{s}') = \sum_{g=1}^{G} x'_{(g,t)} \cdot q_g$. 3) $\forall \boldsymbol{x} \in \{\boldsymbol{x} = mv_G(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}$: $I(\boldsymbol{x}', \boldsymbol{x}) \in \{(1 \le g \le G, 0 \le i \le t-1)\}$ 且不失一般性 $M_{\boldsymbol{x}', \boldsymbol{x}} = \frac{x'_{(g,i)} \cdot q_g}{1 - p^F}$, ( ). 4) $\forall \boldsymbol{x} \in \{\boldsymbol{x} = mv_G(\boldsymbol{s}) \mid \boldsymbol{s}' \in A(\boldsymbol{s})\}$: $\boldsymbol{x} = \left(v_1, \ldots, (x'_{(g,0)}, \ldots, x'_{I(\boldsymbol{x}', \boldsymbol{x})} - 1, x'_{I(\boldsymbol{x}', \boldsymbol{x})+1} + 1, x'_{(g,t)}) \ldots, v_G\right)$.

对数值 $M_{s_0}^{(\widehat{s})}$, $\mathbb{E}[D_{\boldsymbol{s}} | X_{\widehat{s}} = s]$, $\mathbb{E}[D_{\boldsymbol{s}}^2 | X_{\widehat{s}} = s]$ 的其余分析与引理1、引理2和定理1相同，但将 G 重数向量状态 $MV_G(\mathcal{S}_{(n,t)})$ 作为状态集合。以下分析第 III 节所讨论的 **DPSA** 的时间复杂度。同样，令 $V_{DPSA}(n, k, t)$ 和 $E_{DPSA}(n, k, t)$ 分别表示 **DPSA** 马尔可夫链中的顶点数和边数。为方便起见，我们定义 $c_g := |C_g|$。

引理4。 给定 n, t 和 $\boldsymbol{p}$，有 $V_{DPSA}^{\boldsymbol{p}}(n, n, t) = \prod_{g=1}^{G} \binom{c_g + t}{t}$。

证明。 根据 G重数向量状态的定义，并将引理3的证明推广可知，状态数等于每个集合 g 的解的数量之积，即 $\binom{c_g+t}{t}$。 □

引理 5. 给定 n, t和 $\boldsymbol{p}$，有 $E_{DPSA}^{\boldsymbol{p}}(n, n, t) = t \cdot \left(\sum_{g=1}^{G} \frac{c_g}{c_g + t}\right) \cdot \left(\prod_{g=1}^{G} \binom{c_g + t}{t}\right)$。

证明。 我们根据每条入边上发生变化的 $x_{(g,i)}$ （在 G重数向量状态中）的取值来统计入边。假设一条入边改变了 $x_{(i,g)}$ 的取值，其中 $(1 \le g \le G, \; 1 \le i \le t)$，且该取值在重数向量 v_g 中增加 1。因此，选项数等于对每个 $g' \ne g$，方程 $(x_{(g',0)} + x_{(g',1)} + \cdots + x_{(g',t)}) = c_{g'}$ 在非负整数范围内的解的数量，与方程 $(x_{(g,0)} + x_{(g,1)} + \cdots + x_{(g,t)}) = c_g$ 的解的数量的乘积，其中 $x_{(g,i)} \ge 1$。因此我们得到 $\sum_{g=1}^{G} t \cdot \binom{c_g + t - 1}{t} \cdot \left(\prod_{g'=1, \; g' \ne g}^{G} \binom{c'_g + t}{t}\right)$。由于

$\binom{c_g+t-1}{t} = \binom{c_g+t}{t} \cdot \frac{c_g}{c_g+t}$ we obtain $t \cdot \left(\sum_{g=1}^{G} \frac{c_g}{c_g+t}\right) \cdot \left(\prod_{g=1}^{G} \binom{c_g+t}{t}\right)$.

□

可以验证，在均匀分布的情况下， $G = 1$ 和 $c_1 = n$的公式分别等于 V_{UDA} 和 E_{UDA}；而在没有可利用的对称性的情况下，$G = n$ 和 $c_g = 1$的公式分别等于 V_{BA} 和 E_{BA}。

The 接下来 引理给出了**DPSA** 马尔可夫链中状态数和边数的上界 .

引理 6. 给定 n, t 和 $\boldsymbol{p}$ ，有

1) $V_{DPSA}^{\boldsymbol{p}}(n, n, t) \le \binom{\frac{n}{G} + t}{t}^G$,

2) $E_{DPSA}^{\boldsymbol{p}}(n, n, t) \le t \cdot \frac{n}{\frac{n}{G} + t} \cdot \binom{\frac{n}{G} + t}{t}^G$,

其中 G 是概率向量 $\boldsymbol{p}$中不同取值的数量。

证明。 1) 给定 $V_{DPSA}^{\boldsymbol{p}}(n, n, t)$，根据引理4可得，

$$\log\left(\prod_{g=1}^{G} \binom{c_g + t}{t}\right) = \sum_{g=1}^{G} \log\left(\binom{c_g + t}{t}\right) = \sum_{g=1}^{G} \log\left(\frac{(c_g + t)!}{t! \cdot c_g!}\right) =$$

$$= \sum_{g=1}^{G} \left(\log\left(\frac{(c_g + t)!}{c_g!}\right) - \log(t!)\right) = \left(\sum_{g=1}^{G} \sum_{y=1}^{t} \log(c_g + y)\right) - G \cdot \log(t!)$$

注意到 $f\ x\ =\sum_{y=1}^{t}\quad x+y$ $\left(\sum_{y=1}^{t}\quad x+y\right)''=()\log()$ 是一个凹函数，因为其二阶导数为负，即 $\overline{\log()_{()}}$。因此，应用詹森不等式可得，$\log()$ $\sum_{g=1}^{G}\sum_{y=1}^{t}\quad c_g+y$ $\left(\sum_{y=1}^{t}{}_{x+y}^{1}\right)'=\sum_{y=1}^{t}{}_{x+y^{\,2}}^{-1}<0$ $G\cdot\sum_{y=1}^{t}\quad\binom{n}{G}+y\Big)$ $\sum_{g=1}^{G}c_g=n$ $\leq$ log，因为 。由此可得，

$$\log\left(\prod_{g=1}^{G}\binom{c_g+t}{t}\right)=\left(\sum_{g=1}^{G}\sum_{y=1}^{t}\log(c_g+y)\right)-G\cdot\log(t!)\leq G\cdot\left(\sum_{y=1}^{t}\log\Big(\frac{n}{G}+y\Big)-\log(t!)\right)=$$
$$=G\cdot\left(\log\left(\frac{(\frac{n}{G}+t)!}{(\frac{n}{G})!}\right)-\log(t!)\right)=G\cdot\log\left(\binom{\frac{n}{G}+t}{t}\right)=\log\left(\left(\frac{n}{G}+t\right)^{G}\right)$$

由于 $\log(\cdot)$ 单调递增，我们得到 $V_{DPSA}^{\boldsymbol{p}}(n,\ n,\ t)=\prod_{g=1}^{G}\binom{c_g+t}{t}\leq\binom{n}{G}+t^{G}$。

2) 根据引理 5 中的 $E_{DPSA}^{\boldsymbol{p}}(n,\ n,\ t)=t\cdot\left(\sum_{g=1}^{G}\frac{c_g}{c_g+t}\right)\cdot\left(\prod_{g=1}^{G}\binom{c_g+t}{t}\right)$，应用詹森不等式可得，

$\sum_{g=1}^{G}{}_{c_g+t}^{c_g}\leq G\cdot{}_{\frac{n}{G}+t}^{\frac{n}{G}}={}_{\frac{n}{G}+t}^{n}$ $x>0$ $f\ x\ =\ {}_{x+t}^{x}$ $\left({}_{x+t}^{x}\right)''=$因为对于任意值，$()$ 都 $\left({}_{x+t^{\,2}}^{t}\right)'={}_{x+t^{\,4}}^{-2t\cdot x+t}<0$ 是凹函数$_{()()()}$。因此，

$$E_{DPSA}^{\boldsymbol{p}}(n,n,t)=t\cdot\left(\sum_{g=1}^{G}\frac{c_g}{c_g+t}\right)\cdot\left(\prod_{g=1}^{G}\binom{c_g+t}{t}\right)\leq t\cdot\frac{n}{\frac{n}{G}+t}\cdot\left(\binom{\frac{n}{G}+t}{t}\right)^{G}$$

□

注释 1. 当所有集合大小相同时（$\binom{n}{G}$），引理 6 给出的上界是紧的。

VI. Comparison and Evaluation

本节将比较**DPSA** 在若干抽样概率分布下的时间复杂度，并评估**UDA** 和**DPSA**的性能。定义以下分布：

1）“c 集合”分布，其中 $G=c$ 和$c_g=\binom{n}{c}$。2）“c大小集合”分布，其中 $G=\binom{n}{c}$ 和 $c_g=c$。3）“近似均匀分布”，其中 $G=2$ 和 $c_1=n-1$，$c_2=1$。4）“0 对称性”分布，其中 $G=n$、 $c_g=1$。5）均匀分布，其中 $G=1$、 $c_1=n$。

表 I 按时间复杂度比较了上述不同分布。顶点和边使用**DPSA**计算，该方法利用概率向量中的对称性。例如，对于近似均匀分布，**DPSA** 可实现 $O(n^t)$ 的复杂度；而不采用**DPSA** 方法时，复杂度则为指数级。请注意，数值 $q_1,\ldots,q_G$ 与计算复杂度分析无关。

表 I：不同分布下**DPSA**马尔可夫链的顶点数和边数对比，该对比反映了算法时间复杂度。

分布	$V_{DPSA}(n,\ n,\ t)$	$E_{DPSA}(n,\ n,\ t)$
0 对称性	$(t+1)^n$	$t\cdot n\cdot(t+1)^{n-1}$
c大小的组	$\binom{c+t}{t}^{\frac{n}{c}}$	$t\cdot\frac{n}{c+t}\cdot\binom{c+t}{t}^{\frac{n}{c}}$
c组	$\binom{\frac{n}{c}+t}{t}^{c}$	$t\cdot\frac{n}{\frac{n}{c}+t}\cdot\binom{\frac{n}{c}+t}{t}^{c}$
近似均匀分布	$(t+1)\cdot\binom{n-1+t}{t}$	$t\cdot\left(\frac{1}{t+1}+\frac{n-1}{n-1+t}\right)\cdot\left((t+1)\cdot\binom{n-1+t}{t}\right)$
均匀	$\binom{n+t}{t}$	$t\cdot\binom{n+t-1}{t}$

接下来，我们通过将所提出的**UDA** 和**DPSA** 的解析结果与蒙特卡罗模拟[1]进行比较，对其进行定量评估。图1展示了**UDA**的计算效率：对于 $k = n$ 和 $t = 3$，以往被认为在计算上不可行的精确结果，如今可在五秒内得出。图2给出了**DPSA**的类似结果：对于 $k = n$ 和 $t = 2$，在考虑两种大小相同的优惠券类型 $0.5n$ 、其组概率总和分别为 $q_1 = 0.3 \cdot 0.5n$ 和 $q_2 = 0.7 \cdot 0.5n$ 的情况下，可在五秒内完成精确计算。

对于这两种方法，我们将其运行时间和准确度与使用 $10^3, 10^4,$ 和 10^5 次迭代的蒙特卡洛模拟进行比较。在所有情况下，**UDA** 和**DPSA** 的运行时间都始终优于模拟。虽然进行 10^3 次迭代的模拟能够达到相近的执行时间，但其误差率最高，这凸显了基于模拟的方法在准确性与效率之间固有的权衡，也进一步体现了所提出的解析方法的优势。

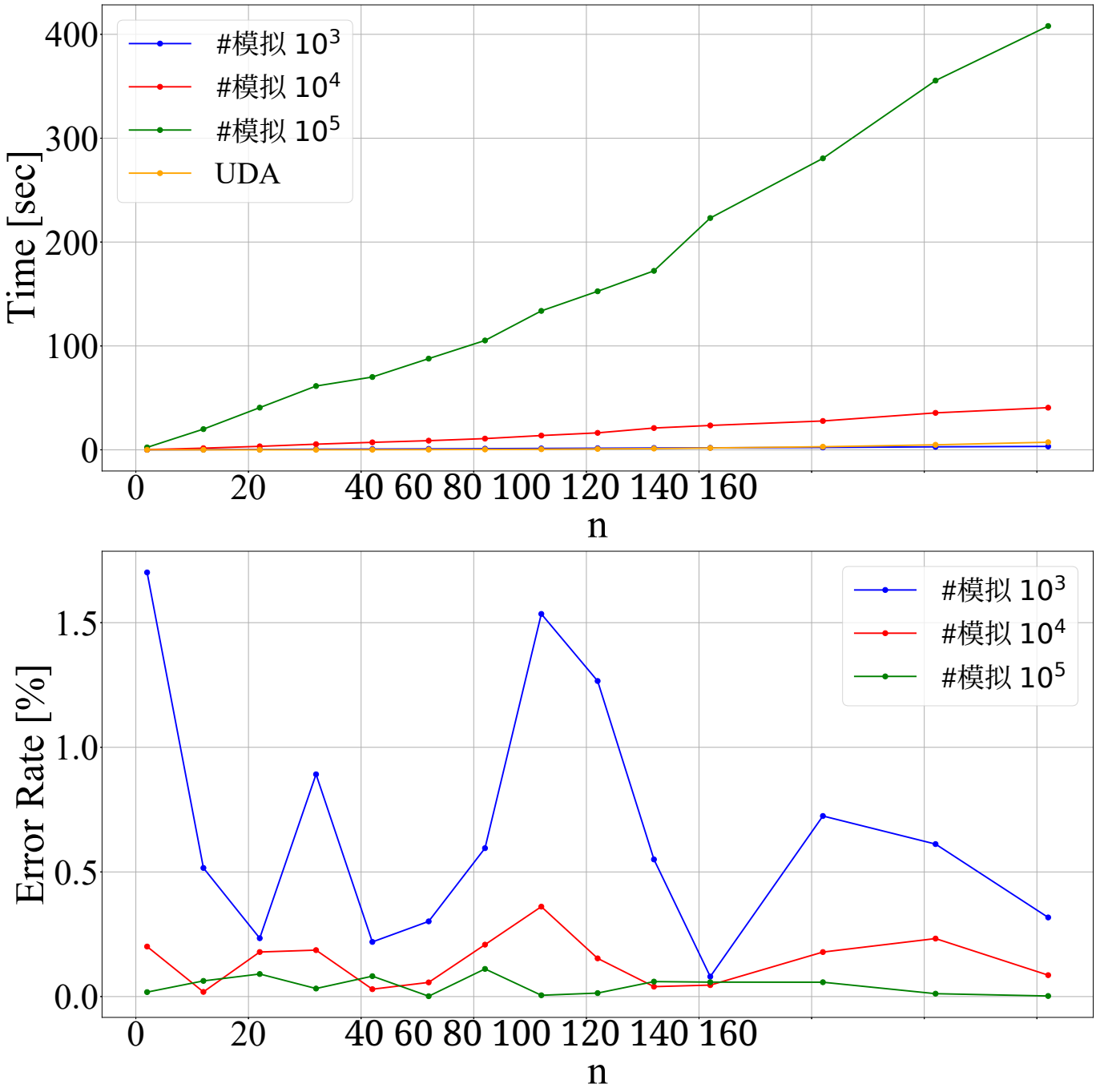

图 1：评估**UDA** 与蒙特卡罗模拟在计算时间和误差率方面的性能。图中展示了随着 n 增大时的计算时间，其中 $k = n$ ，且对于 $t = 3$。

[1]蒙特　卡罗模拟通过反复从概率模型中抽取随机变量，对不确定结果进行估计　　1.

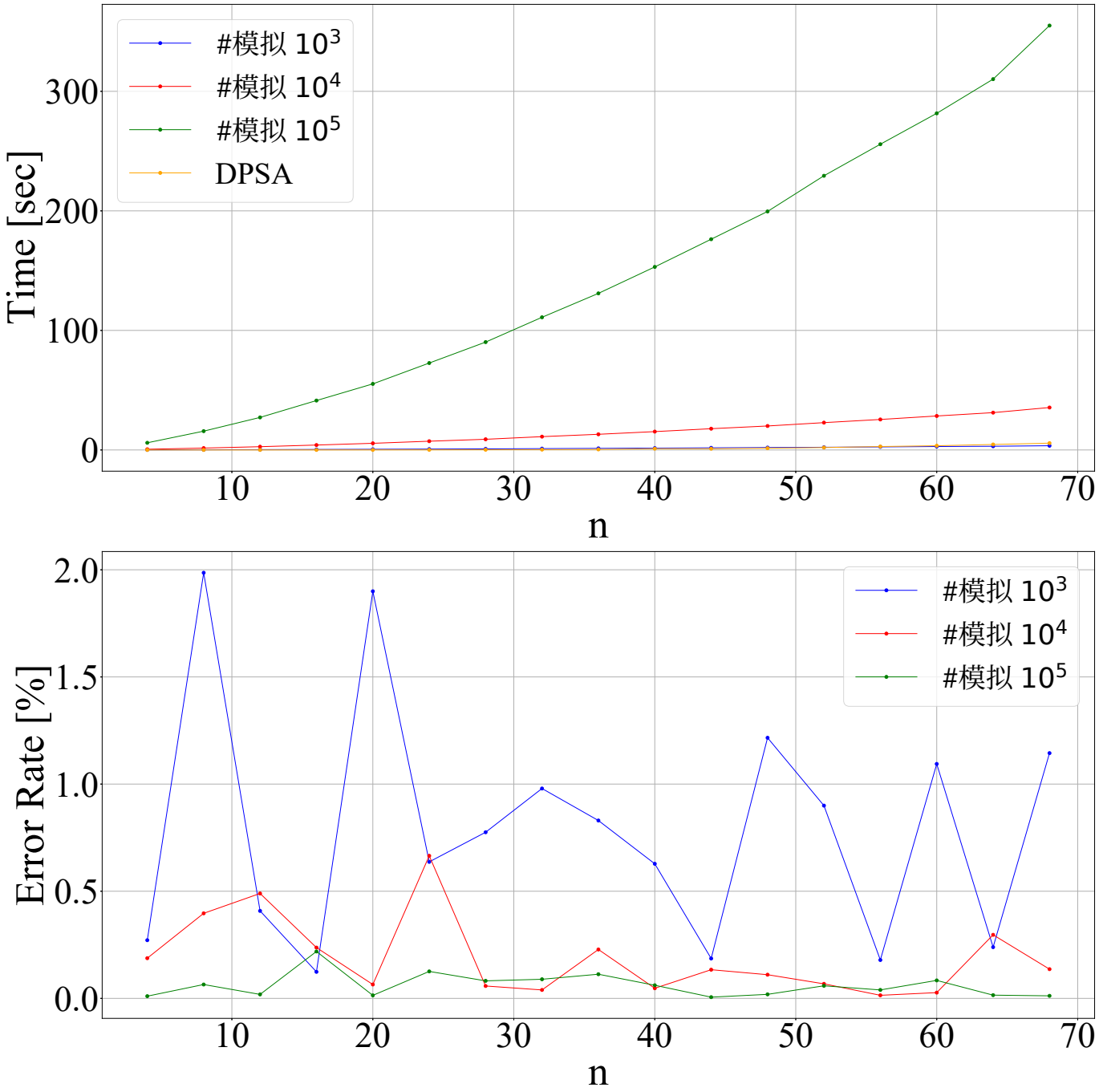

图2：评估**DPSA** 与蒙特卡罗模拟在计算时间和误差率方面的性能。随着 n 的值增大，计算时间如图所示，其中 $k = n$、 $t = 2$ 和 $g = 2$ 每组大小为 $0.5n$ 和 $q_1 = 0.3 \cdot 0.5n$， $q_2 = 0.7 \cdot 0.5n$。

VII. 结论与未来工作

在本研究中，我们对一般集邮者问题进行了全面建模，从而能够显式计算收集过程的期望、方差和二阶矩。我们提出了两种算法：**UDA** 和**DPSA**，分别用于计算均匀分布和任意分布下的这些值。这些算法能够在多项式时间内对均匀分布进行精确计算。此外，我们还将这一能力扩展到任意概率分布，以计算精确值；同时，将均匀分布方法推广到具有对称性的概率向量，从而将计算复杂度从指数级降至多项式级。未来的工作将重点推导关键量的闭式表达式，以减少对迭代马尔可夫链计算的依赖。

参考文献

[1] W. Feller，《概率论及其应用导论》，第 2 卷，第 81 卷。John Wiley & Sons，1991 年。[2] P. Flajolet、D. Gardy 和 L. Thimonier，“生日悖论、优惠券收集者问题、缓存算法与自组织搜索”，《离散应用数学》，第 39 卷，第 3 期，第 207–229 页，1992 年。[3] S. Boneh 和 V. G. Papanicolaou，“优惠券收集问题的一般渐近估计”，《计算与应用数学杂志》，第 67 卷，第 2 期，第 277–289 页，1996 年。[4] A. Boneh 和 M. Hofri，“重新探讨优惠券收集问题——工程问题与计算方法综述”，《随机模型》，第 13 卷，第 1 期，第 39–66 页，1997 年。[5] E. Anceaume、Y. Busnel 和 B. Sericola，“利用马尔可夫链研究广义优惠券收集问题的新结果”，《应用概率杂志》，第 52 卷，第 2 期，第 405–418 页，2015 年。[6] P. Berenbrink 和 T. Sauerwald，“加权优惠券收集问题及其应用”，载于《国际计算与组合数学会议》，第 449–458 页，Springer，2009 年。[7] A. V. Doumas 和 V. G. Papanicolaou，“重新探讨优惠券收集问题：方差的渐近性质”，《应用概率进展》，第 44 卷，第 1 期，第 166–195 页，2012 年。[8] L. Holst，“关于生日问题、优惠券收集者问题、占位问题及其他经典抽球问题”，《国际统计评论／Revue Internationale de Statistique》，第 15–27 页，1986 年。[9] P. Neal，“广义优惠券收集问题”，《应用概率杂志》，第 45 卷，第 3 期，第 621–629 页，2008 年。[10] H. Von Schelling，“不等概率下的优惠券收集”，《美国数学月刊》，第 61 卷，第 5 期，第 306–311 页，1954 年。[11] P. Erd˝os 和 A. Rényi，“关于概率论的一个经典问题”，《Magyar Tud. Akad. Mat. Kutató Int. Közl》，第 6 卷，第 1 期，第 215–220 页，1961 年。[12] M. S. Klamkin 和 D. J. Newman，“生日惊奇问题的扩展”，《组合理论杂志》，第 3 卷，第 3 期，第 279–282 页，1967 年。[13] D. J. Newman，“双 Dixie 杯问题”，《美国数学月刊》，第 67 卷，第 1 期，第 58–61 页，1960 年。[14] A. Boneh 和 M. Hofri，“重新探讨优惠券收集问题”，1989 年。[15] L. E. Baum 和 P. Billingsley，“优惠券收集问题的渐近分布”，《数学统计年鉴》，第 36 卷，第 6 期，第 1835–1839 页，1965 年。[16] S. Janson，“若干序贯占位问题的极限定理”，《应用概率杂志》，第 20 卷，第 3 期，第 545–553 页，1983 年。[17] R. K. Brayton，“关于随机选择下完成一个集合所需试验次数的渐近行为”，《数学分析与应用杂志》，第 7 卷，第 1 期，第 31–61 页，1963 年。[18] L. Flatto，“与抽球模型相关的若干随机变量的极限定理”，《概率年鉴》，第 927–934 页，1982 年。[19] A. Doumas 和 V. Papanicolaou，“优惠券收集问题上升阶乘矩的渐近性质”，2013 年。[20] A. V. Doumas 和 V. G. Papanicolaou，“重新探讨优惠券收集问题：推广 Newman 和 Shepp 的双 Dixie 杯问题”，《ESAIM：概率与统计》，第 20 卷，第 367–399 页，2016 年。[21] H. Abraham、R. Gabrys 和 E. Yaakobi，“覆盖所有碱基：DNA 测序效率的新阶段”，载于《2024 年 IEEE 国际信息论研讨会（ISIT）》，第 464–469 页，IEEE，2024 年。[22] I. Preuss、B. Galili、Z. Yakhini 和 L. Anavy，“组合式 DNA 存储系统的测序覆盖度分析”，《IEEE 分子、生物与多尺度通信汇刊》，第 10 卷，第 2 期，第 297–316 页，2024 年。