Coupon collecting

Markov Chains and Mixing Times

Shengbo Dong (董晟渤)

February 26, 2024

Introduction

In probability theory, the coupon collector's problem refers to mathematical analysis of "collect all coupons and win" contests.

Introduction

In probability theory, the coupon collector's problem refers to mathematical analysis of "collect all coupons and win" contests.

Introduction

In probability theory, the coupon collector's problem refers to mathematical analysis of "collect all coupons and win" contests.

Problem

How many coupons must we obtain so that our collection contains all n types?

Markov Chain

Our Model

Let X_t denote the number of different types of coupon represented among our first t coupons.

Markov Chain

Our Model

Let X_t denote the number of different types of coupon represented among our first t coupons.

- $X_0 = 0$;
- $\mathbb{P}(X_{t+1} = k+1 | X_t = k) = 1 - \frac{k}{n} = \frac{n-k}{n}$;
- $\mathbb{P}(X_{t+1} = k | X_t = k) = \frac{k}{n}$.

Markov Chain

Our Model

Let X_t denote the number of different types of coupon represented among our first t coupons.

- $X_0 = 0$;
- $\mathbb{P}(X_{t+1} = k + 1 | X_t = k) = 1 - \frac{k}{n} = \frac{n-k}{n}$;
- $\mathbb{P}(X_{t+1} = k | X_t = k) = \frac{k}{n}$.

Classifying the States

- Absorbing state: n;
- Essential state: n;
- Communicating class: $\{n\}$.

Expectation

Let $\tau = \inf\{t \geq 0 : X_t = n\}$ be the number of coupons collected when our collection contains all n types.

Expectation

Let $\tau = \inf\{t \geq 0 : X_t = n\}$ be the number of coupons collected when our collection contains all n types.

Theorem (Proposition 2.3, MCMT)

$\mathbb{E}(\tau) = nH_n$, *where* $H_n := \sum_{k=1}^{n} \frac{1}{k}$.

Expectation

Let $\tau = \inf\{t \geq 0 : X_t = n\}$ be the number of coupons collected when our collection contains all n types.

Theorem (Proposition 2.3, MCMT)

$\mathbb{E}(\tau) = nH_n$, *where* $H_n := \sum_{k=1}^{n} \frac{1}{k}$.

Proof. Let $\tau_k = \inf\{t \geq \tau_{k-1} : X_t = k\}$ be the total number of coupons when the collection first contains k different coupons. Then

$$\tau = \tau_n = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1}).$$

Next we analyse the distribution of each $\tau_k - \tau_{k-1}$.

Lemma

$$\tau_k - \tau_{k-1} \sim \mathcal{G}\left(\frac{n-k+1}{n}\right).$$

Lemma

$$\tau_k - \tau_{k-1} \sim \mathcal{G}\left(\frac{n-k+1}{n}\right).$$

Proof. After collecting $k-1$ types, there are $n-k+1$ types missing from the collection. Each missing coupon has the same probability

$$1 - \frac{k-1}{n} = \frac{n-k+1}{n}$$

to be collected. □

Lemma

$$\tau_k - \tau_{k-1} \sim \mathcal{G}\left(\frac{n-k+1}{n}\right).$$

Proof. After collecting $k-1$ types, there are $n-k+1$ types missing from the collection. Each missing coupon has the same probability

$$1 - \frac{k-1}{n} = \frac{n-k+1}{n}$$

to be collected. □

Recall (Geometric Distribution)

Let $X \sim \mathcal{G}(p)$, then

- Distribution: $\mathbb{P}(X = k) = p(1-p)^{k-1}$, $k \geq 1$;
- Expectation: $\mathbb{E}(X) = \frac{1}{p}$;
- Variance: $\mathrm{var}(X) = \frac{1-p}{p^2} \leq \frac{1}{p^2}$.

Thus

$$\mathbb{E}(\tau) = \sum_{k=1}^{n} \mathbb{E}\left(\tau_k - \tau_{k-1}\right) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k} = nH_n. \quad \square$$

Thus

$$\mathbb{E}(\tau) = \sum_{k=1}^{n} \mathbb{E}\left(\tau_k - \tau_{k-1}\right) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k} = nH_n. \quad \square$$

Recall

Let $\gamma_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n$, then

- $\{\gamma_n\}$ decreases;
- $\{\gamma_n\}$ is bounded and $0 < \gamma_n \leq 1$.
- $\gamma_n \downarrow \gamma \approx 0.577$.

Here γ is called the Euler constant.

Thus

$$\mathbb{E}(\tau) = \sum_{k=1}^{n} \mathbb{E}\left(\tau_k - \tau_{k-1}\right) = n \sum_{k=1}^{n} \frac{1}{n-k+1} = n \sum_{k=1}^{n} \frac{1}{k} = nH_n. \quad \square$$

Recall

Let $\gamma_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n$, then

- $\{\gamma_n\}$ decreases;
- $\{\gamma_n\}$ is bounded and $0 < \gamma_n \le 1$.
- $\gamma_n \downarrow \gamma \approx 0.577$.

Here γ is called the Euler constant.

We have $\left| \sum_{k=1}^{n} \frac{1}{k} - \log n \right| \le 1$, and $|\mathbb{E}(\tau) - n \log n| \le n$.

Large Deviation

τ is unlikely to be much larger than its expected value.

Theorem (Proposition 2.4, MCMT)

For any $c > 0$, $\mathbb{P}(\tau > \lceil n \log n + cn \rceil) \leq \exp(-c)$.

Large Deviation

τ is unlikely to be much larger than its expected value.

Theorem (Proposition 2.4, MCMT)

For any $c > 0$, $\mathbb{P}(\tau > \lceil n \log n + cn \rceil) \leq \exp(-c)$.

Proof. Let A_i be the event that the coupon i does not appear among the first $\lceil n \log n + cn \rceil$ coupons. Observe first that

$$\mathbb{P}(\tau > \lceil n \log n + cn \rceil) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}\left(A_i\right).$$

Large Deviation

τ is unlikely to be much larger than its expected value.

Theorem (Proposition 2.4, MCMT)

For any $c > 0$, $\mathbb{P}(\tau > \lceil n \log n + cn \rceil) \leq \exp(-c)$.

Proof. Let A_i be the event that the coupon i does not appear among the first $\lceil n \log n + cn \rceil$ coupons. Observe first that

$$\mathbb{P}(\tau > \lceil n \log n + cn \rceil) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

In each trial, the probability of not drawing coupon i is $1 - \dfrac{1}{n}$, so

$$\mathrm{RHS} = \sum_{i=1}^{n} \left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} = n\left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil}.$$

Now we use the inequality $1 + x \leq \exp(x)$ with $x = -\frac{1}{n}$ to get

$$1 - \frac{1}{n} \leq \exp\left(-\frac{1}{n}\right),$$

and $\lceil n \log n + cn \rceil \geq n \log n + cn$, thus

$$\text{RHS} = n\left(1 - \frac{1}{n}\right)^{\lceil n\log n + cn \rceil} \leq n \exp\left(-\frac{n \log n + cn}{n}\right) = \exp(-c). \quad \square$$

Now we use the inequality $1 + x \leq \exp(x)$ with $x = -\frac{1}{n}$ to get

$$1 - \frac{1}{n} \leq \exp\left(-\frac{1}{n}\right),$$

and $\lceil n \log n + cn \rceil \geq n \log n + cn$, thus

$$\text{RHS} = n\left(1 - \frac{1}{n}\right)^{\lceil n \log n + cn \rceil} \leq n \exp\left(-\frac{n \log n + cn}{n}\right) = \exp(-c). \quad \square$$

Remark

When $c \to \infty$,

$$\mathbb{P}(\tau > \lceil n \log n + cn \rceil) \leq \exp(-c) \to 0.$$

Limit Theorem

General Model

Let T_n be the time we spend to collect n different coupons.

- $\mathbb{E}(T_n) = n \sum_{k=1}^{n} \frac{1}{k} \sim n \log n;$
- $\mathrm{var}(T_n) \leq n^2 \sum_{k=1}^{n} \frac{1}{(n-k+1)^2} = n^2 \sum_{k=1}^{n} \frac{1}{k^2}.$

Limit Theorem

General Model

Let T_n be the time we spend to collect n different coupons.

- $\mathbb{E}(T_n) = n\sum_{k=1}^{n}\frac{1}{k} \sim n\log n;$
- $\mathrm{var}(T_n) \leq n^2\sum_{k=1}^{n}\frac{1}{(n-k+1)^2} = n^2\sum_{k=1}^{n}\frac{1}{k^2}.$

Recall (Basel problem)

$$\sum_{k=1}^{n}\frac{1}{k^2} \to \frac{\pi^2}{6}.$$

Theorem (Example 2.2.7, PTE)

$\frac{T_n}{n \log n} \to 1$ *in probability.*

Theorem (Example 2.2.7, PTE)

$\frac{T_n}{n \log n} \to 1$ *in probability.*

Proof. Since $\frac{\mathrm{var}(T_n)}{(\mathbb{E}(T_n))^2} \to 0$, we have

$$\frac{T_n - n \log n}{n \log n} \to 0 \quad \text{in probability.} \qquad \square$$

Theorem (Example 2.2.7, PTE)

$\frac{T_n}{n \log n} \to 1$ *in probability.*

Proof. Since $\frac{\mathrm{var}(T_n)}{(\mathbb{E}(T_n))^2} \to 0$, we have

$$\frac{T_n - n \log n}{n \log n} \to 0 \quad \text{in probability.} \qquad \square$$

Theorem (Extension of previous bounds)

$\frac{T_n - n \log n}{n} \Rightarrow \eta$, *where* $\mathbb{P}(\eta \le c) = \exp(-\exp(-c))$.

Theorem (Example 2.2.7, PTE)

$\frac{T_n}{n \log n} \to 1$ *in probability.*

Proof. Since $\frac{\text{var}(T_n)}{(\mathbb{E}(T_n))^2} \to 0$, we have

$$\frac{T_n - n\log n}{n \log n} \to 0 \quad \text{in probability.} \quad \square$$

Theorem (Extension of previous bounds)

$\frac{T_n - n\log n}{n} \Rightarrow \eta$, *where* $\mathbb{P}(\eta \le c) = \exp(-\exp(-c))$.

Based on this theorem, we have

$$\mathbb{P}\left(\frac{T_n - n\log n}{n} \ge c\right) = \mathbb{P}(T_n \ge n\log n + cn) \to 1 - \exp(-\exp(-c)).$$

Thanks for listening!