[AI2613 Lecture 10] Poisson Approximation

May 4, 2023

1 Coupon Collector Problem with Non-Uniform Coupons

Recall the coupon collector problem we met many times in this course: If each box of a brand of cereals contains a coupon which is chosen from n different types uniformly at random, then we need to buy nH_n boxes in expectation to collect all types of coupons.

In this lecture, we generalize the setting by allowing the non-uniformity. Suppose that each purchase yields a coupon of type j w.p. p_j for $j \in [n]$ and the coupon types contained in different boxes are independent, where $\sum_{j=1}^{n} p_j = 1$. Let N_j be the first time that we get type j. Then N_j follows the geometric distribution with parameter p_j. Let N be the number of purchases until all n types of coupons are collected, that is, $N = \max_{j\in[n]} N_j$. We would like to compute $\mathbf{E}[N]$ to see how many times of purchases is needed in expectation. However, it is not easy to compute the expected value of $\max_{j\in[n]} N_j$ since N_j's are not independent.

1.1 Coupon Collector Problem with Poisson Draw

We consider a variation of the coupon collector problem where the coupons are collected with Poisson draw. That is, each arrival of the Poisson process with rate 1 brings a coupon and the probability of the coupon being of type j is p_j. Note that this process is different from the ordinary coupon collector problem since the arrival time is random.

Recall the thinning of Poisson process we discussed in the last lecture. Let $X_j(t)$ be the number of type j coupons we collect in time $[0,t]$ with Poisson draw. Then $\{X_j(t)\}$ is a thinning of the process, meaning that $\{X_j(t)\}$ is a Poisson process with rate p_j and $X_j(t)$ is independent of $X_i(t)$ for $i \neq j$. For $j \in [n]$, let $T_j \triangleq \min\{t \mid X_j(t) = 1\}$ be the first time that type j coupon appears. Obviously, T_j is the same as $\tau_j(1)$ [1] and $T_j \sim \mathrm{Exp}(p_j)$.

[1] Here $\tau_j(1)$ denotes the time gap between the arrival of the customers with coupon j.

To determine the time of collecting all kinds of coupons, we need to compute $\mathbf{E}[T]$ where $T = \max_{j\in[n]} T_j$. The following proposition to compute expectation is useful.

Proposition 1 *Let X be a non-negative random variable.*

- *If X is discrete and $X \in \mathbb{N}$, then $\mathbf{E}[X] = \sum_{t=1}^{\infty} \mathbf{Pr}[X \geq t]$.*
- *If X is continuous, then $\mathbf{E}[X] = \int_0^{\infty} \mathbf{Pr}[X \geq t]\, dt$.*

Proof.

- When X is discrete, we apply the double counting trick:

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{s=1}^{\infty} s\mathbf{Pr}[X=s] = \sum_{s=1}^{\infty}\sum_{t=1}^{s}\mathbf{Pr}[X=s] \\
&= \sum_{t=1}^{\infty}\sum_{s=t}^{\infty}\mathbf{Pr}[X=s] = \sum_{t=1}^{\infty}\mathbf{Pr}[X\ge t].
\end{aligned}
$$

- When X is continuous,

$$
\begin{aligned}
\mathbf{E}[X] &= \mathbf{E}\left[\int_0^X 1\,dt\right] = \mathbf{E}\left[\int_0^\infty \mathbf{1}[X\ge t]\,dt\right] \\
&\overset{(\heartsuit)}{=} \int_0^\infty \mathbf{E}\left[\mathbf{1}[X\ge t]\right]\,dt = \int_0^\infty \mathbf{Pr}[X\ge t]\,dt,
\end{aligned}
$$

where $(\heartsuit)$ comes from the *Fubini's theorem*.

□

Note that for any $t \in \mathbb{R}_{\ge 0}$,

$$
\mathbf{Pr}[T\ge t] = 1-\mathbf{Pr}[T<t] = 1-\prod_{j=1}^{n}\mathbf{Pr}\left[T_j<t\right] = 1-\prod_{j=1}^{n}\left(1-e^{-p_jt}\right).
$$

By the continuous version of Proposition 1, we have

$$
\mathbf{E}[T] = \int_0^\infty \mathbf{Pr}[T\ge t]\,dt = \int_0^\infty 1-\prod_{j=1}^{n}\left(1-e^{-p_jt}\right)dt.
$$

That is, we need a time of $\int_0^\infty \left(1-e^{-p_jt}\right)dt$ in expectation to collect all kinds of coupons.

1.2 *Standard Coupon Collector Problem*

Then we relate the result we obtained in the previous section on the coupon collector with Poisson draw to the standard coupon collector problem by the technique of coupling. Specifically, let τ_i denote the time gap between the $i-1$-th and the i-th arrival. Imagine the standard version as one customer coming with a coupon in hand with constant time gap between arrivals. We couple the two process by letting the i-th arrival in the Poisson version carry the same type of coupon with the i-th arrival in the ordinary version.

Recall that N is the number of purchases until all n types of coupons are collected in the standard coupon collector problem. Then we have $T = \sum_{i=1}^{N}\tau_i$. Note that $\tau_i \sim \mathrm{Exp}(1)$ and $\mathbf{E}[\tau_i] = 1$. If N is a constant, we can deduce $\mathbf{E}[N] = \mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right] = \mathbf{E}[T]$ directly. However, N is a random variable and thus the summation and expectation are not guaranteed to be exchangeable. To show the validity of $\mathbf{E}[N]\,\mathbf{E}[\tau_i] = \mathbf{E}\left[\sum_{i=1}^{N}\tau_i\right]$ in this case, we make use of the Wald's equation introduced before.

Theorem 2 (Wald's Equation) *Let $X_1, X_2, \ldots$ be n i.i.d. random variables that $\mathbf{E}[|X_1|] < \infty$. Let T be a stopping time that $\mathbf{E}[T] < \infty$. Then we have $\mathbf{E}\left[\sum_{t=1}^{T} X_t\right] = \mathbf{E}[T]\,\mathbf{E}[X_1]$.*

It is easy to verify that $\mathbf{E}[\tau_i] = 1 < \infty$ and $\mathbf{E}[N] < \infty$ in our case. So applying the Wald's equation, we have $\mathbf{E}[N]\,\mathbf{E}[\tau_i] = \mathbf{E}\left[\sum_{i=1}^{N} \tau_i\right]$ and sequentially

$$\mathbf{E}[N] = \mathbf{E}[T] = \int_0^\infty 1 - \prod_{j=1}^{n} \left(1 - e^{-p_j t}\right) dt. \tag{1}$$

Then we go back to the coupon collector problem with uniform coupons for sanity check. Let $x = e^{-\frac{t}{n}}$. If $p_j = \frac{1}{n}$ for any $j \in [n]$, we have

$$\begin{aligned}
\mathbf{E}[N] &= \int_0^\infty 1 - \prod_{j=1}^{n} \left(1 - e^{-p_j t}\right) dt \\
&= n \int_0^\infty 1 - (1-x)^n \, d\log x \\
&= n \int_0^\infty \frac{1}{x} - \frac{(1-x)^n}{x} dx \\
&= n \int_0^\infty \sum_{k=1}^{n} \frac{(1-x)^{k-1}}{x} - \frac{(1-x)^k}{x} dx \\
&\overset{(\heartsuit)}{=} n \sum_{k=1}^{n} \int_0^\infty (1-x)^{k-1} dx \\
&= n \sum_{k=1}^{n} \frac{1}{k} = nH_n,
\end{aligned}$$

where the $(\heartsuit)$ follows from the *Fubini's theorem.* This verifies Equation (1) when the types of coupons are uniform.

2 *Balls-into-Bins*

Recall the balls-into-bins problem where we throw m identical balls into n bins. For $i \in [n]$, let X_i be the number of balls in the i-th bin. Then we have $X_i \sim \mathrm{Binom}(m, \frac{1}{n})$ and $\mathbf{E}[X_i] = \frac{m}{n}$. This model can be used to describe the scheme of the hash table. To avoid frequent collision when mapping the keys into slots, it is natural for us to be concerned about the value of $\max_{i \in [n]} X_i$. However, we are faced with the difficulty that X_i's are not independent when computing the distribution of $\max_{i \in [n]} X_i$. It turns out that one can use independent Poisson variables to approximate the distribution. First we have:

Theorem 3 *The distribution of $(X_1, X_2, \ldots, X_n)$ is the same as that of $(Y_1, Y_2, \ldots, Y_n)$ on condition that $\sum_{i=1}^{n} Y_i = m$ where $Y_i \sim \mathrm{Pois}(\lambda)$ are independent Poisson random variables with an arbitrary rate λ.*

Proof. Given $(a_1, a_2, \dots, a_n) \in \mathbb{N}^n$ and $\sum_{i=1}^n a_n = m$, we have

$$\mathbf{Pr}\left[(X_1, X_2, \dots, X_n) = (a_1, a_2, \dots, a_n)\right] = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}. \tag{2}$$

And

$$\begin{aligned}
&\mathbf{Pr}\left[(Y_1, Y_2, \dots, Y_n) = (a_1, a_2, \dots, a_n) \;\middle|\; \sum_{i=1}^n Y_i = m\right] \\
&= \frac{\mathbf{Pr}\left[(Y_1, Y_2, \dots, Y_n) = (a_1, a_2, \dots, a_n) \wedge \sum_{i=1}^n Y_i = m\right]}{\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n \mathbf{Pr}\left[Y_i = a_i\right]}{\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]} \\
&= \frac{\prod_{i=1}^n e^{-\lambda} \frac{\lambda^{a_i}}{a_i!}}{e^{-\lambda n} \frac{(\lambda n)^m}{m!}} = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!},
\end{aligned}$$

which equals to the RHS of Equation (2). □

Furthermore, we can deduce the following corollary from Theorem 3.

Corollary 4 *Let $f : \mathbb{N}^n \to \mathbb{N}$ be an arbitrary function and $Y_1, Y_2, \dots, Y_n$ be n independent Poisson random variables with rate $\lambda = \frac{m}{n}$. Then we have*

$$\mathbf{E}\left[f(X_1, X_2, \dots, X_n)\right] \le e\sqrt{m} \cdot \mathbf{E}\left[f(Y_1, Y_2, \dots, Y_n)\right].$$

Proof. By the law of total probability, we have

$$\begin{aligned}
\mathbf{E}\left[f(Y_1, Y_2, \dots, Y_n)\right] &= \sum_{k=0}^{\infty} \mathbf{E}\left[f(Y_1, Y_2, \dots, Y_n) \;\middle|\; \sum_{i=1}^n Y_i = k\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = k\right] \\
&\ge \mathbf{E}\left[f(Y_1, Y_2, \dots, Y_n) \;\middle|\; \sum_{i=1}^n Y_i = m\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right] \\
&= \mathbf{E}\left[f(X_1, X_2, \dots, X_n)\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right].
\end{aligned}$$

Note that $\sum_{i=1}^n Y_i \sim \mathrm{Pois}(m)$, then we have

$$\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right] = e^{-m} \frac{m^m}{m!} > \frac{1}{e\sqrt{m}},$$

where the inequality comes from the *Stirling's formula.* □

We can see from the proof of Corollary 4 that the choice of $\lambda = \frac{m}{n}$ is to maximize $\mathbf{Pr}\left[\sum_{i=1}^n Y_i = m\right]$.

Equipped with Corollary 4, we have the following theorem to bound $X = \max_{i \in [n]} X_i$.

Theorem 5 (Max Load) *When $m = n$, we have $X = \Theta\left(\frac{\log n}{\log \log n}\right)$ w.p. $1 - o(1)$.*

Proof. First we prove the upper bound, that is, there exists a constant c_1 such that $\mathbf{Pr}\left[X \ge \frac{c_1 \log n}{\log \log n}\right] = o(1)$. Let $k = \frac{c_1 \log n}{\log \log n}$ for brevity. By union bound, we have

$$\begin{aligned}\mathbf{Pr}[X \geq k] &= \mathbf{Pr}[\exists i \in [n], X_i \geq k] \leq \sum_{i=1}^{n} \mathbf{Pr}[X_i \geq k] \\ &= n \cdot \mathbf{Pr}[X_1 \geq k] \leq n \cdot \binom{n}{k} \frac{1}{n^k} \leq n \cdot \left(\frac{en}{k}\right)^k \frac{1}{n^k} = n \cdot \left(\frac{e}{k}\right)^k.\end{aligned}$$

Note that

$$\begin{aligned}k \log k &= \frac{c_1 \log n}{\log \log n} \cdot (\log \log n - \log \log \log n + \log c_1) \\ &> c_1 \log n \left(1 - \frac{\log \log \log n}{\log \log n}\right) > \frac{c_1}{2} \log n.\end{aligned}$$

Letting $c = 6$, we have that

$$\log n + k - k \log k < -\log n.$$

Thus, $\mathbf{Pr}[X \geq k] \leq n \cdot \left(\frac{e}{k}\right)^k < \frac{1}{n} = o(1)$ for $c_1 = 6$.

Then we prove the lower bound. Again let $g = \frac{c_2 \log n}{\log \log n}$ for a constant c_2. Let $f(X_1, X_2, \ldots, X_n) \triangleq \mathbf{1}[X < g] = \mathbf{1}[\max_{i \in [n]} X_i < g]$. Then by Corollary 4,

$$\begin{aligned}\mathbf{Pr}[X < g] &= \mathbf{E}[f(X_1, X_2, \ldots, X_n)] \\ &\leq e\sqrt{n} \cdot \mathbf{E}[f(Y_1, Y_2, \ldots, Y_n)] \\ &= e\sqrt{n} \cdot \mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right]. \qquad (3)\end{aligned}$$

By the definition of Y_i in Corollary 4, we have

$$\begin{aligned}\mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right] &= (\mathbf{Pr}[Y_1 \leq g])^n = (1 - \mathbf{Pr}[Y_1 > g])^n \\ &\leq (1 - \mathbf{Pr}[Y_1 = g + 1])^n = \left(1 - \frac{1}{(g+1)!e}\right)^n \leq e^{-\frac{n}{(g+1)!e}}\end{aligned}$$

Note that

$$\begin{aligned}\log(g+1)! &= \sum_{i=1}^{g+1} \log i < \int_1^{g+2} \log x \, dx \\ &= (g+2)\log(g+2) - g - 1 \leq (g+2)\log g - g + 3 \\ &= \frac{c_2 \log n + 2 \log \log n}{\log \log n} (\log \log n - \log \log \log n + \log c_2) - \frac{c_2 \log n}{\log \log n} + 3 \\ &\leq c_2 \log n - \log \log n - 2.\end{aligned}$$

Letting $c_2 = 1$, we have $\log(g+1)! \leq \log n - \log \log n - 2$ and sequentially

$$e(g+1)! \leq \frac{n}{e \log n}.$$

Thus,

$$\mathbf{Pr}\left[\max_{i \in [n]} Y_i < g\right] \leq e^{-\frac{n}{(g+1)!e}} \leq e^{-e \log n} = n^{-e}.$$

Combining with Equation (3), we have $\mathbf{Pr}\left[X < \frac{\log n}{\log \log n}\right] \leq e\sqrt{n} \cdot n^{-e} = o(1)$.

□