AI2613 随机过程 2020–2021 春季学期

Lecture 9 – Poisson Process (III)

2021 年 4 月 26 日

Lecturer: 张驰豪 *Scribe*: 杨宽

1 Review of Poisson Approximation

Review the model of *balls-into-bins*. This model can capture many important problems, such as the *birthday paradox*, the *coupon collector problem*. Specifically, suppose that we have m balls and put them into n bins independently and uniformly at random. Let X_i be the number of balls into the i-th bin. Then we are interested in the distribution of $(X_1, \ldots, X_n)$. The following result shows that the distribution of $(X_1, \ldots, X_n)$ is exactly the same as *independent* Poisson variables conditioned on the summation.

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

Moreover, in the last lecture we proved the following corollary, which transform the expectation of dependent binomial random variables to the expectation of independent Poissons without any conditioning, and we also used it to analyze the *maxload* problem.

Corollary 2. *Let $f : \mathbb{N}^n \to \mathbb{N}$ be an arbitrary function, $Y_1, Y_2, \ldots, Y_n$ be n independent Poisson random variables with rate $\lambda = m/n$, i.e., $Y_i \sim \mathsf{Pois}(m/n)$. Then we have*

$$\mathbb{E}\big[f(X_1, X_2, \ldots, X_n)\big] \le \mathrm{e}\sqrt{m} \cdot \mathbb{E}\big[f(Y_1, Y_2, \ldots, Y_n)\big] .$$

Remark. If f is a monotone function, then the factor e can be improved to 2.

Remark. When we apply this corollary, we usually let f be an indicator function of some bad event $\mathcal{B}(X_1, \ldots, X_n)$. Then

$$\mathbf{Pr}[\mathcal{B}(X_1, \ldots, X_n)] = \mathbb{E}\big[f(X_1, \ldots, X_n)\big] \le \mathrm{e}\sqrt{m} \cdot \mathbb{E}\big[f(Y_1, \ldots, Y_n)\big] = \mathbf{Pr}[\mathcal{B}(Y_1, \ldots, Y_n)] .$$

For example, see our proof of the maxload problem in the last lecture.

Theorem 3 (Maxload). *Assume that $m = n$ in the balls-into-bins model, and let $X = \max X_i$. Then there exists two constant $c_1, c_2 > 0$ such that*

$$\mathbf{Pr}\left[c_1 \cdot \frac{\log n}{\log\log n} < X < c_2 \cdot \frac{\log n}{\log\log n}\right] = 1 - o(1/n).$$

In the last lecture, we showed that there exists two constant $c_1, c_2 > 0$ such that

$$\mathbf{Pr}\left[X \le c_1 \cdot \frac{\log n}{\log\log n}\right] = o(1/n),$$

and

$$\mathbf{Pr}\left[X \ge c_2 \cdot \frac{\log n}{\log\log n}\right] = o(1/n).$$

The remaining part of the last lecture is to prove Theorem 1. Here we give a proof.

Proof of Theorem 1. We first give the distribution of $(X_1, \ldots, X_n)$. Note that the number of all possible ways to put balls into bins is n^m, and the number of the permutations of a multi-set of a_1 Z_1s, a_2 Z_2s, ..., and a_n Z_ns is

$$\binom{a_1 + a_2 + \cdots + a_n}{a_1, a_2, \ldots, a_n} \triangleq \frac{(a_1 + a_2 + \cdots + a_n)!}{a_1! a_2! \cdots a_n!}.$$

Hence for all $a_1, a_2, \ldots, a_n \in \mathbb{N}$ s.t. $\sum a_i = m$, we have

$$\mathbf{Pr}[X_1 = a_1, X_2 = a_2, \ldots, X_n = a_n] = \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}.$$

Next, we show that $(Y_1, \ldots, Y_n)$ has the same distribution.

$$\begin{aligned}
\mathbf{Pr}\left[Y_1 = a_1, Y_2 = a_2, \ldots, Y_n = a_n \mid \sum Y_i = m\right] &= \frac{\mathbf{Pr}[Y_1 = a_1, Y_2 = a_2, \ldots, Y_n = a_n]}{\mathbf{Pr}[\sum Y_i = m]} \\
&= \frac{\prod_{i=1}^n \mathbf{Pr}[Y_i = a_i]}{\mathbf{Pr}[\sum Y_i = m]} \\
&= \frac{\prod_{i=1}^n \mathrm{e}^{-\lambda} \cdot \frac{\lambda^{a_i}}{a_i!}}{\mathrm{e}^{-n\lambda} \cdot \frac{(n\lambda)^m}{m!}} \\
&= \frac{1}{n^m} \cdot \frac{m!}{a_1! a_2! \cdots a_n!}.
\end{aligned}$$

□

2 Non-Homogeneous Poisson Process

We introduced the transformation of conditioning and proved the following result in the last lecture. The next part of today's lecture will be based on it.

Theorem 4. *Conditioned on $N(t) = n$, the vector of arrival times $(T_1, T_2, \ldots, T_n)$ has the same distribution as $(V_1, V_2, \ldots, V_n)$, where $V_1, V_2, \ldots, V_n$ are sampled independently and uniformly at random from $[0, t]$ and then rearranged in the increasing order.*

Recall the *thinning* of the Poisson process. A thinning requires that the randon variables Y_i associated to arrivals are *independent and identically distributed.* Let p_j denote the probability that $Y_i = j$. Then we have $\sum_j p_j = 1$ and $N_j(t) \sim \text{Pois}(p_j \lambda t)$. Now the question is, what is $N_j(t)$ if p_j are not constants?

Then $N_j(t)$ may not be Poisson processes any longer. However, if we view $N_j(t)$ as random variables, then we have the following theorem.

Theorem 5. *Let $Y_i \in \{1, 2, \ldots, k\}$ be random variables and $p_1, \ldots, p_k$ are nonnegative functions such that $\sum_{j=1}^{k} p_j = 1$ (namely, $\forall\, s$, $\sum_{j=1}^{k} p_j(s) = 1$). Assume that $\mathbf{Pr}\big[Y_i = j\big] = p_j(s)$ if the i-th arrival is at time s. Then the number of arrivals with $Y_i = j$ before time t, denoted by $N_j(t)$, are independent Poisson distributed with mean*

$$\mathbb{E}\big[N_j(t)\big] = \lambda \int_0^t p_j(s)\, \mathrm{d}s\,.$$

Example 1 (Queueing Theory ($M/G/\infty$ Poisson Queue)). $M/G/\infty$ is a term in *queueing theory.* Here we give a brief explanation. The first character "M" stands for *Markovian* (lack of memory), which means that customers arrive as a Poisson process with rate λ. The second one, "G", stands for *general service times*, that is, we assume that the i-th customer requires some service time s_i, where s_i are independent and have a cumulative distribution G (i.e., $G(t) = \mathbf{Pr}[\text{service time} \le t]$). The final symbol "$\infty$" indicates there are infinite many servers.

Then we are interested in the distributions of the following two random variables:

1. $X(t)$: the number of customers completed services before time t;
2. $Y(t)$: the number of customers being served at time t.

We partition customers arrive before time t into two types:

Type 1. Customers have completed services before time t. The number is $X(t)$.

Type 2. Customers are being served at time t. The number is $Y(t)$.

Note that the number of servers are infinite. So the type of customers arrived are determined by the distribution G. Let $p_i(s)$ be the probability of the customer arrived at time t to be type i. It is easy to see that

$$p_1(s) = G(t - s)\,, \text{ and}$$
$$p_2(s) = 1 - G(t - s)\,.$$

To simplify our statement, let $\overline{G}(t) = 1 - G(t)$. Then we have

$$\mathbb{E}[X(t)] = \mathbb{E}[N_1(t)] = \lambda \int_0^t G(t-s)\,\mathrm{d}s = \lambda \int_0^t G(r)\,\mathrm{d}r\,, \text{ and}$$
$$\mathbb{E}[Y(t)] = \mathbb{E}[N_2(t)] = \lambda \int_0^t \overline{G}(t-s)\,\mathrm{d}s = \lambda \int_0^t \overline{G}(r)\,\mathrm{d}r\,.$$

Example 2. Suppose someone drives a car $\mathcal{C}$ through a road of length ℓ at speed x. We also assume that there are other cars entering into this section of road according to a Poisson process with rate λ, and all vehicles are travelling at a constant speed with distribution G. A car can overtake a slower moving car without any loss of speed. Our goal is to choose x to minimize the number of overtakes (including overtaking and being overtaken).

Given x, the time of car $\mathcal{C}$ passing the road is $t_0 = \ell/x$. If the car enter the section of road at time s, then it is on the road in time period $[s, s+t_0]$. Suppose there is another car coming into the road at a random time S. Its speed X is sampled from the distribution with cumulative distribution function G. Then the passing time is $T = \ell/X$. Let F be the distribution of T, that is, $F(t) = \mathbf{Pr}[T < t] = \mathbf{Pr}[X > \ell/t] = \overline{G}(\ell/t)$.

We consider the following two types of cars:

Type 1: cars overtaking $\mathcal{C}$. Then type-1 cars enter the road at time $t > s$ and exit at time $t+T < s+t_0$.

Type 2: cars overtaken by $\mathcal{C}$. Then type-2 cars enter the road at time $t < s$ and exit at time $t + T > s + t_0$.

Let $p(t)$ be the probability that a car arriving at time t encouters car $\mathcal{C}$ (overtaking or being overtaken) on the road. Then we have

$$p(t) = \begin{cases} \mathbf{Pr}[t+T > s+t_0] = \mathbf{Pr}[T > s+t_0-t] = \overline{F}(s+t_0-t), & \text{if } t < s; \\ \mathbf{Pr}[t+T < s+t_0] = \mathbf{Pr}[T < s+t_0-t] = F(s+t_0-t), & \text{if } s < t\,. \end{cases}$$

Remark. To simplify our statement, we assume that F is defined over $\mathbb{R}$ and $F(t) = 0$ if $t \le 0$.

Hence, we conclude that

$$\begin{aligned} \mathbb{E}[N_{\text{encounter}}(s+t_0)] &= \lambda\Big(\int_0^s \overline{F}(s+t_0-t)\,\mathrm{d}t + \int_t^\infty F(s+t_0-t)\,\mathrm{d}t\Big) \\ &= \lambda\Big(\int_{t_0}^{s+t_0} \overline{F}(t)\,\mathrm{d}t + \int_0^{t_0} F(t)\,\mathrm{d}t\Big)\,. \end{aligned}$$

Note that our goal is to determine x, namely, to find $x^* = \arg\min_{t_0=\ell/x} \mathbb{E}[N_{\text{encounter}}(s+t_0)]$. Since $\mathbb{E}[N_{\text{encounter}}(s+t_0)]$ is a function only depending on t_0 (given F and sufficiently large s), we compute the derivative as follows. Let $H(t_0) = \mathbb{E}[N_{\text{encounter}}(s+t_0)] = \lambda\big(\int_{t_0}^{s+t_0} \overline{F}(t)\,\mathrm{d}t + \int_0^{t_0} F(t)\,\mathrm{d}t\big)$.

Then

$$\begin{aligned}\frac{\mathrm{d}H}{\mathrm{d}t_0} &= \lambda \frac{\mathrm{d}}{\mathrm{d}t_0}\Big(\int_0^{s+t_0} \overline{F}(t)\,\mathrm{d}t - \int_0^{t_0} \overline{F}(t)\,\mathrm{d}t + \int_0^{t_0} F(t)\,\mathrm{d}t\Big) \\ &= \lambda\big(\overline{F}(s+t_0) - \overline{F}(t_0) + F(t_0)\big)\,.\end{aligned}$$

Since sufficiently large s implies that $\overline{F}(s+t_0) \approx 0$, we roughly have that

$$\frac{\mathrm{d}H}{\mathrm{d}t_0} = \lambda(F(t_0) - \overline{F}(t_0)) = \lambda(2F(t_0) - 1)\,.$$

It is clear that $H(t_0)$ achieves the minimum at $t_0 = t^*$ where $F(t^*)$ is roughly $1/2$. Thus $x^* = \ell / t^*$ is roughly $\ell / G^{-1}(1/2)$.

Example 3 (Spread of HIV). We would like to track the number of HIV infections. Suppose that the spread of HIV follows a Poisson process with an unknown rate λ. Then the number of individuals infected with HIV in time t has a Poisson distribution rate $t\lambda$ and has independent increments. Let $G(t)$ be the distribution of the incubation times. Namely, we assume that the time from when an individual becomes infected until symptoms of the disease appear is a random variable having distribution G. We further assume that G is known and the incubation times of different infected individuals are independent.

We also consider the following two infected individuals:

Type 1: individuals who have shown symptoms of the disease by time t.

Type 2: individuals who are infected but have not shown symptoms of the disease by time t.

Denote by $N_i(t)$ the number of type-i individuals. Note that the probability of an individual infected at time s being type-1 is $G(t-s)$. So we have

$$\begin{aligned}\mathbb{E}[N_1(t)] &= \lambda \int_0^t G(t-s)\,\mathrm{d}s = \lambda \int_0^t G(r)\,\mathrm{d}r\,, \text{ and} \\ \mathbb{E}[N_2(t)] &= \lambda \int_0^t \overline{G}(t-s)\,\mathrm{d}s = \lambda \int_0^t \overline{G}(r)\,\mathrm{d}r\,.\end{aligned}$$

However, the question here is that both λ and G are unknown. Since we can assume that the number of individuals showing symptoms are known, the following calculation gives an estimation $\hat{\lambda}$ of λ.

Let $\hat{n_1}$ be the number of individuals showing symptoms. We assume that $\hat{n_1} \approx \mathbb{E}[N_1]$ is an estimation of $\mathbb{E}[N_1]$. So we have

$$\hat{n_1} = \lambda \int_0^t G(r)\,\mathrm{d}r\,,$$

and let

$$\hat{\lambda} = \frac{\hat{n}_1}{\lambda \int_0^t G(r)\,\mathrm{d}r}.$$

Then we can estimate $\mathbb{E}[N_2(t)]$ by letting

$$\hat{n}_2 = \hat{\lambda} \int_0^t \overline{G}(r)\,\mathrm{d}r = \hat{n}_1 \cdot \frac{\int_0^t \overline{G}(r)\,\mathrm{d}r}{\int_0^t G(r)\,\mathrm{d}r}$$

In fact, all of the above examples are so-called *non-homogeneous Poisson processes*. We now formally give the definition.

Definition 4 (Non-homogeneous Poisson Process). We say $\{N(t)\colon t \geq 0\}$ is a *non-homogeneous Poisson process* with rate $\lambda(t)$ if

1. $N(0) = 0$;
2. $N(t)$ has independent increments;
3. $N(t) - N(s)$ has a Poisson distribution with rate $\int_s^t \lambda(r)\,\mathrm{d}r$.

Finally, we are ready to prove Theorem 5.

Proof of Theorem 5. We prove it by compute the distribution straightforwardly.
Given $n_1, n_2, \ldots, n_k$ and time t, conditioned on $N(t) = \sum n_i$, we can view the process in Theorem 5 as follows: we first generate $\sum n_i$ arrivals according to a Poisson process and next for each arrival we independently generate a uniform random number to determine its type.
Let $n = \sum_i n_i$, and $U_1, U_2, \ldots, U_n$ are uniformly and independently sampled from $[0, t]$. We also let $Y_1, Y_2, \ldots, Y_n$ are indpendent random variables associated with $U_1, U_2, \ldots, U_n$ respectively, where $\mathbf{Pr}\big[Y_i = j\big] = p_j(U_i)$. Applying Theorem 4, we have

$$\begin{aligned}
&\mathbf{Pr}[N_1(t) = n_1, N_2(t) = n_2, \ldots, N_k(t) = n_k] \\
&= \mathbf{Pr}[N_1(t) = n_1, N_2(t) = n_2, \ldots, N_k(t) = n_k \mid N(t) = n] \cdot \mathbf{Pr}[N(t) = n] \\
&= \mathbf{Pr}[m_1 = n_1, m_2 = n_2, \ldots, m_k = n_k] \cdot \mathbf{Pr}[N(t) = n]\,,
\end{aligned}$$

where m_i is the number of js such that $Y_j = i$. Since $U_1, U_2, \ldots, U_k$ are independently and identically distributed, it follows that $Y_1, Y_2, \ldots, Y_k$ have the same distribution. Moreover, the probability q_i of the event $Y_j = i$ is given by

$$q_i = \frac{1}{t} \cdot \int_0^t p_i(s)\,\mathrm{d}s.$$

Hence, we have

$$\mathbf{Pr}[m_1 = n_1, m_2 = n_2, \ldots, m_k = n_k] = \frac{n!}{n_1! n_2! \cdots n_k!} \cdot q_1^{n_1} q_2^{n_2} \cdots q_k^{n_k},$$

and consequently

$$\begin{aligned}
&\mathbf{Pr}[N_1(t) = n_1, N_2(t) = n_2, \ldots, N_k(t) = n_k] \\
&= \mathbf{Pr}[m_1 = n_1, m_2 = n_2, \ldots, m_k = n_k] \cdot \mathbf{Pr}[N(t) = n] \\
&= \frac{n!}{n_1! n_2! \cdots n_k!} \cdot q_1^{n_1} q_2^{n_2} \cdots q_k^{n_k} \cdot \mathrm{e}^{-\lambda t} \cdot \frac{(\lambda t)^n}{n!} \\
&= \prod_{i=1}^{k} \mathrm{e}^{q_i \lambda t} \cdot \frac{(q_i \lambda t)^{n_i}}{n_i!},
\end{aligned}$$

which completes our proof. □

Remark. In a non-homogeneous Poisson process, although the number of type-i arrivals in units of time still have a Poisson distribution, the time intervals between two consecutive type-i arrivals DO NOT have exponential distributions any longer. In fact, the time intervals may NOT be independent.