MATH4240 Tutorial 11 Notes

A birth-and-death process is a Markov jump process where from state x in one jump, only transition to $x-1$ and to $x+1$ are possible. This is just a continuous-time analogue of the discrete-time birth-and-death chain we are already familiar with, and the rate matrix and the embedded transition matrix (assuming no absorbing states) take the form

$$D = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\mu_1+\lambda_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\mu_2+\lambda_2) & \lambda_2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \qquad Q = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots \\ \frac{\mu_1}{\mu_1+\lambda_1} & 0 & \frac{\lambda_1}{\mu_1+\lambda_1} & 0 & \dots \\ 0 & \frac{\mu_2}{\mu_2+\lambda_2} & 0 & \frac{\lambda_2}{\mu_2+\lambda_2} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Remark 1. Recall that we (typically) assume that a Markov jump process is *non-explosive*, that is, $\lim_n \tau_n = \infty$ (with probability 1) on jump time τ_n. However, if the state space is infinite, for a generic birth-and-death process it *may not be true*. In particular, you can show the following *Reuter's criterion*[1]: on state space $S = \mathbb{N}$, the process is non-explosive iff

$$\sum_{n=1}^{\infty} \sum_{k=0}^{n} \prod_{i=1}^{k} \frac{1}{\lambda_n} \frac{\mu_{n+1-i}}{\lambda_{n-i}} = \infty$$

With pure-birth process ($\mu_n = 0$), this is $\sum 1/\lambda_n = \infty$ (which is also a sufficient condition in the general case). Obviously, Poisson process (and typical queues we will be working on) satisfies this condition, but it is easy to construct one that does not.

1 Branching

Consider a cluster of particles, each individually after a random life span that is exponentially distributed with parameter λ will split into some number of offspring with pmf $f(k)$. (By memoryless property, we may assume $f(1) = 0$).

I believe the case $f(2) = p$ and $f(0) = 1-p$ is already discussed in the lecture, and the jump matrix and rate matrix are given as

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 1-p & 0 & p & 0 & \dots \\ 0 & 1-p & 0 & p & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots \\ (1-p)\lambda & -\lambda & p\lambda & 0 & \dots \\ 0 & 2(1-p)\lambda & -2\lambda & 2p\lambda & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

That is, $\lambda_0 = 0$, and $\lambda_n = np\lambda$, $\mu_n = n(1-p)\lambda$ for $n \geq 1$. The case where a Poisson arrival of immigrants with rate α is also considered, in which case the birth rate is simply $\tilde{\lambda}_n = \lambda_n + \alpha$ (and 0 is no longer absorbing).

Let us deviate from the framework of birth-and-death process and consider the case where a particle may have more than 2 offspring (that is, $q_{x,y} > 0$ for some $y > x+1$).

With the same derivation as in the lecture, the jump matrix and the rate matrix should be

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ f(0) & 0 & f(2) & f(3) & \dots \\ 0 & f(0) & 0 & f(2) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots \\ \lambda f(0) & -\lambda & \lambda f(2) & \lambda f(3) & \dots \\ 0 & 2\lambda f(0) & -2\lambda & 2\lambda f(2) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

[1] See Brémaud, Thm. 4.5.

Of course, except the case that is discussed during the lecture ($f(0)+f(2)=1$), the process is no longer a birth-and-death process, and the backward/forward equation may not be solvable. However, under *certain* assumption[2], we can solve for the process, and it is non-explosive. For simplicity, we will assume that everything is nice enough.

For this process, the backward equation reads

$$\frac{\mathrm{d}}{\mathrm{d}t}P_{ij}(t)=\sum_k D_{ik}P_{kj}(t)=-i\lambda P_{ij}(t)+i\lambda\sum_{k\geq i-1}f(k-i+1)P_{kj}(t)$$

At $i=1$,

$$\frac{\mathrm{d}}{\mathrm{d}t}P_{1n}=-\lambda P_{1n}+\lambda\sum_{k\geq 0}f(k)P_{kn}$$

Consider the generating function $F(t,X)=\sum_k P_{1k}(t)X^k$. We have

$$\partial_t F(t,X)=-\lambda F(t,X)+\lambda\sum_k f(k)\sum_n P_{kn}(t)X^n$$

Note that $\sum_n P_{kn}(t)X^n$ is the generating function of the distribution with k initial particles, and as the evolution of the particles are independent, we have $\sum_n P_{kn}(t)X^n=F_1^k$, so

$$\partial_t F=-\lambda F+\lambda\Phi(F)$$

where $\Phi(X)=\sum f(n)X^n$ is the generating function of the offspring distribution f. Solving this with initial condition $F(t=0)=X^3$, we can obtain $P_{1n}(t)=\frac{1}{n!}(\partial_X)^n F(X=0)$.

In the case of *binary fission / Yule process* $f(2)=1$ ($\Phi(X)=X^2$), we have

$$\partial_t F=\lambda F(F-1)$$

which has solution

$$F(t,X)=e^{-\lambda t}X/(1-(1-e^{-\lambda t})X)=e^{-\lambda t}X\sum_k(1-e^{-\lambda t})^k X^k$$

In the case that is covered in lecture $f(2)=p$, $f(0)=1-p$, we have $\Phi(X)=(1-p)+pX^2$ and so

$$\partial_t F=-\lambda F+\lambda(1-p+pF^2)=\lambda(F-1)(pF-(1-p))$$

which can be solved explicitly[4] as

$$F(t,X)=(\xi+(1-\xi-\eta)X)/(1-\eta X)=(\xi+(1-\xi-\eta)X)\sum_k\eta^k X^k$$

where $\xi=1-e^{-\rho}/W$, $\eta=1-1/W$, $W=e^{-\rho}(1+(1-p)\lambda\int_0^t e^{\rho}(\tau)\,\mathrm{d}\tau)$, $\rho=\lambda(1-2p)t$, and so $P_{10}=\xi$ and $P_{1n}=(1-\xi)(1-\eta)\eta^{n-1}$ for $n\geq 1$.

The major issue is that the equation $\partial_t F=\lambda(\Phi(F)-F)$ is typically **hard** to solve, but if you are able to solve the equation, you can obtain lots of information on the distribution.

2 Queuing

The typical model of queuing process we will be working with is M/M/c/K queue with *Markovian* interarrival time, *Markovian* service time, c parallel servers, and capacity K. In the special case $K=\infty$ (unlimited capacity), this is a M/M/c queue.

It should be noted that the following relation always holds, whether the process is Markov or not:

Theorem 2.1 (Little's law)**.** *In a queue system, if λ is the average arrival rate, W is the average time a customer spends in the system, and L is the average number of customers in the system, then assuming λ, W are finite,*

$$L=\lambda W$$

[2]For example, $\int_{1-\epsilon}^1\frac{1}{\Phi(s)-s}\,\mathrm{d}t=\infty$ for all $\epsilon>0$, or more typically just $\Phi'(1)=\sum nf(n)<\infty$, where Φ is the generating function of the offspring distribution.

[3]Assuming some regularity conditions are satisfied, so that the solution is well-behaved.

[4]See Harris, Ch. V.7, or Athreya & Ney, Ch. III.5.5.

With Little's law, we can easily show some results.

Theorem 2.2 (Utilization law)**.** *Consider a M/M/1 queue with arrival rate λ and service rate μ, with $\mu > \lambda$. Then the long-run proportion ρ of time where the server is busy (**offered load**) is*

$$\rho = \lambda/\mu$$

(Compare this with the M/M/∞ result covered / being covered in lecture, which states that $\lim_t P_{xy}(t) = e^{-\rho}\rho^y/y!$ and so the probability that *some* servers are busy is $1 - \lim_t P_{x0}(t) = 1 - e^{-\rho} = \rho - \rho^2/2 + \ldots$.)

Proof. Consider the system consisting (only) of the server. Then

- the average a customer spends on the system (being served by the server) is $W = 1/\mu$
- the average number of customer in the system is $L = \rho \cdot 1 + (1-\rho) \cdot 0 = \rho$
- as $\lambda < \mu$, we should expect that there is no (long-time) accumulation of customers in the queue, and so the rate customers get to be served should be the same as the arrival rate λ[5]

By Little's law,

$$\rho = L = \lambda W = \lambda/\mu$$

□

You can also show this by solving for the stationary distribution, but note that Little's law does not require e.g. the service time to be exponentially distributed, in which case

$$\rho = \lambda E\,(\text{service time})$$

Another important property of such queue is the *PASTA* property, that is *Poisson Arrivals See Time Averages*. More precisely,[6]

Theorem 2.3. *Suppose in a queue the arrival is a Poisson process. Let*

- *$p_n(t) = P(X_t = n)$ be the probability that there are n people in the queue at time t*
- *$a_n(t)$ be the probability that an arrival at time t sees n people in the queue*

then $\pi_n(t) = a_n(t)$.

That is, the distribution that an outside observer sees is the same as the distribution that someone joining the queue (at Poisson time) sees.

It should be noted that each arrival affects X_t, and PASTA does not hold if the arrival is not Poisson (e.g. deterministic arrival once every 2 minutes, deterministic service of 1 minutes).

Proof. Consider a short time interval $(t, t+\delta]$. By memoryless property, the event that an arrival happens in $(t, t+\delta]$ is independent of the current number of people X_t in the system, no matter how the service time distributed, that is

$$P(\text{arrival at } (t,t+\delta] \,|\, X_t = n) = P(\text{arrival at } (t,t+\delta])$$

So

$$\begin{aligned}
a_n(t) &= \lim_{\delta\to 0} P(X_t = n \,|\, \text{arrival at } (t,t+\delta]\,) \\
&= \lim_{\delta\to 0} P(\text{arrival at } (t,t+\delta] \,|\, X_t = n)\, \frac{P(X_t = n)}{P(\text{arrival at } (t,t+\delta])} \\
&= \lim_{\delta\to 0} P(\text{arrival at } (t,t+\delta])\, \frac{P(X_t = n)}{P(\text{arrival at } (t,t+\delta])} \\
&= p_n(t)
\end{aligned}$$

□

Typically, we consider the long-time average of these quantities, the long-time distribution P_n of the queue and the long-time distribution A_n that an arrival sees, in which case

$$P_n = A_n$$

[5] This actually requires a bit more justification (that the queue is a *stable queue*), which we are omitting here.

[6] See Stewart, p.394.