Markov Jump Processes and Queueing Theory
2021-5-10
BFN/bfn

Markov Jump Processes (continuous time Markov chains) Section 6.6

We consider a stochastic process $\{X(t); t \geq 0\}$ with $X(t) \in \{0, 1, \ldots, N\}$

$$
\begin{aligned}
\mathbb{P}\{X(t+h) = j | X(t) = i\} &= q_{ij}h + o(h) \\
\mathbb{P}\{X(t+h) \neq i | X(t) = i\} &= \left(\sum_{j=0, j\neq i}^{N} q_{ij} \right) h + o(h) = q_i h + o(h), \quad q_i = \sum_{j=0, j\neq i}^{N} q_{ij} \\
\mathbb{P}\{X(t+h) = i | X(t) = i\} &= 1 - q_i h + o(h) \\
P_{ij}(t) &= \mathbb{P}\{X(t+s) = j | X(s) = i\} \\
P'_{ij}(t) &= \sum_{k=0, k\neq j}^{N} P_{ik}(t) q_{kj} - P_{ij}(t) q_j \\
&= \sum_{k=0}^{N} P_{ik}(t) q_{jk}, \quad q_{jj} = -q_j
\end{aligned}
$$

On matrix form ($\boldsymbol{P}(t) = ||P_{ij}(t)||$)

$$
\boldsymbol{P}'(t) = \boldsymbol{P}(t)\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}(t)
$$

For scalar a the equation

$$
p'(t) = ap(t) \Leftrightarrow p(t) = \exp(at) = \sum_{n=0}^{\infty} \frac{(at)^n}{n!}
$$

Define now the matrix function $\boldsymbol{H}(t) = \sum_{n=0}^{\infty} \frac{(\boldsymbol{A}t)^n}{n!}$. It is relatively easy to see, that the sum converges for all t for finite N ($\boldsymbol{A}$ being a finite dimensional matrix). Consider

$$
\frac{\mathrm{d}\boldsymbol{H}(t)}{\mathrm{d}t} = \sum_{n=1}^{\infty} \frac{\boldsymbol{A}^n}{n!} n t^{n-1} = \boldsymbol{A} \sum_{n=1}^{\infty} \frac{(\boldsymbol{A}t)^{n-1}}{(n-1)!} = \boldsymbol{A}\boldsymbol{H}(t) = \boldsymbol{H}(t)\boldsymbol{A}
$$

We define

$$
\exp(\boldsymbol{A}t) = e^{\boldsymbol{A}t} = \sum_{n=0}^{\infty} \frac{(\boldsymbol{A}t)^n}{n!}
$$

to get

$$\boldsymbol{P}(t) \quad = \quad e^{\boldsymbol{A}t}$$

We can check that $e^{\boldsymbol{A}t}$ satisfies Chapman-Kolmogorov

$$\boldsymbol{P}(s+t) \quad = \quad \boldsymbol{P}(s)\boldsymbol{P}(t), \quad \text{elementwise} P_{ij}(s+t) = \sum_{k=0}^{N} P_{ik}(s)P_{kj}(t)$$

Sequence S_i of sojourn times with $W_n = \sum_{i=0}^{n-1} S_i$, the time of the nth state change. We can view the process as a sequence of sojourn times.

$$X_n \quad = \quad X(W_n)$$

the state entered after the n'th state change, an embedded Markov chain

$$\begin{aligned}
\mathbb{P}\{S_n > t | X_n = i\} &= e^{-q_i t} = e^{q_{ii} t} \\
\mathbb{P}\{X_{n+1} = j | X_n = i\} &= \frac{q_{ij}}{q_i} = \frac{q_{ij}}{\sum_{k=0, k\neq i}^{N} q_{ik}} \\
\mathbb{P}\{X_{n+1} = i | X_n = i\} &= 0
\end{aligned}$$

Assume irreducibility of $\{X_n; n \geq 0\}$, then we have an invariant distribution of $\{X(t); t \geq 0\}$

$$\begin{aligned}
P_{ij}(t) &\overset{t\to\infty}{\to} \pi_j \\
\boldsymbol{\pi} &= (\pi_0, \pi_1, \ldots, \pi_N)
\end{aligned}$$

$$\boldsymbol{P}(t) \overset{t\to\infty}{\to} = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi} \\ \vdots \\ \boldsymbol{\pi} \end{pmatrix}$$

$$\begin{aligned}
\boldsymbol{P}'(t) &= \boldsymbol{P}(t)\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}(t) \\
\mathbf{0} &= \boldsymbol{\pi}\boldsymbol{A} \\
\pi_j q_j &= \sum_{i=0, i\neq j}^{N} \pi_i q_{ij}, \quad \text{global balance equations} \\
\pi_j q_j &= \pi_i q_{ij}, \quad \text{local balance equations}
\end{aligned}$$

Local balance implies global balance. When local balance equations are satisfied the Markov process is *reversible* $((X(t_1), X(t_2)) \overset{\text{dist}}{=} (X(t-t_1), X(t-t_2)), t > t_2 > t_1)$.

Relation between invariant distribution of X_n and $X(t)$.

$$\theta_j = \lim_{n\to\infty} \mathbb{P}\{X_n = j | X_0 = i\}$$

(NB! X_n could be a periodic chain, so a bit of care is needed here, to establish the existence of the limit). Now with $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_N)$

$$\boldsymbol{\theta} = \boldsymbol{\theta} \begin{pmatrix} 0 & \frac{q_{01}}{q_0} & \frac{q_{02}}{q_0} & \frac{q_{03}}{q_0} & \cdots & \frac{q_{0,N-1}}{q_0} & \frac{q_{0,N-2}}{q_0} \\ \frac{q_{10}}{q_1} & 0 & \frac{q_{12}}{q_1} & \frac{q_{13}}{q_1} & \cdots & \frac{q_{1,N-1}}{q_1} & \frac{q_{1,N}}{q_1} \\ \frac{q_{20}}{q_2} & \frac{q_{21}}{q_2} & 0 & \frac{q_{23}}{q_2} & \cdots & \frac{q_{2,N-1}}{q_2} & \frac{q_{2,N}}{q_2} \\ \frac{q_{30}}{q_3} & \frac{q_{31}}{q_3} & \frac{q_{32}}{q_3} & 0 & \cdots & \frac{q_{3,N-1}}{q_3} & \frac{q_{3,N}}{q_3} \\ \vdots & \vdots & \vdots & \vdots & \ddots\ \ddots\ \ddots & \vdots & \vdots \\ \frac{q_{N-1,0}}{q_{N-1}} & \frac{q_{N-1,1}}{q_{N-1}} & \frac{q_{N-1,2}}{q_{N-1}} & \frac{q_{N-1,3}}{q_{N-1}} & \cdots & 0 & \frac{q_{N-1,N}}{q_{N-1}} \\ \frac{q_{N0}}{q_N} & \frac{q_{N1}}{q_N} & \frac{q_{N2}}{q_N} & \frac{q_{N3}}{q_N} & \cdots & \frac{q_{N,N-1}}{q_N} & 0 \end{pmatrix}$$

with

$$\boldsymbol{\Delta}(\boldsymbol{a}) = \begin{pmatrix} a_0 & 0 & 0 & \ldots & 0 \\ 0 & a_1 & 0 & \ldots & 0 \\ 0 & 0 & a_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & a_N \end{pmatrix}$$

i.e. a diagonal matrix with the entries of the vector $\boldsymbol{a}$ on the diagonal. With

$$\boldsymbol{q} = \begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ \vdots \\ q_N \end{pmatrix}$$

We get

$$\begin{aligned} \boldsymbol{\theta} &= \boldsymbol{\theta}\left(\boldsymbol{I} + \boldsymbol{\Delta}(\boldsymbol{q})^{-1}\boldsymbol{A}\right) \\ \mathbf{0} &= \boldsymbol{\theta}\boldsymbol{\Delta}(\boldsymbol{q})^{-1}\boldsymbol{A} \end{aligned}$$

As $\boldsymbol{\pi}$ is the unique left eigenvector of $\boldsymbol{A}$ corresponding to the eigenvalue 0 we conclude

$$\begin{aligned} \boldsymbol{\pi} &= K \cdot \boldsymbol{\theta}\boldsymbol{\Delta}(\boldsymbol{q})^{-1} \\ \boldsymbol{\theta} &= C \cdot \boldsymbol{\pi}\boldsymbol{\Delta}(\boldsymbol{q}) = \frac{\boldsymbol{\pi}\boldsymbol{\Delta}(\boldsymbol{q})}{\boldsymbol{\pi}\boldsymbol{q}} \end{aligned}$$

Queueing - Chapter 9, sections 9.1-9.3

Queueing processes, Section 9.1

1. Input (arrival) process (S_n inter arrival times)
2. Service time distribution
3. Queueing discipline

Performance parameters

1. Queue length process/distribution/moments
2. Server utilisation
3. System throughput
4. (Customer) waiting time distribution/quantiles/moments
5. Probability of service/blocking

Let $\{X(t); t \geq 0\}$ be the queue length process (number of customers in the system at time t)

$$\begin{array}{rcl} L & = & \mathbb{E}(X(t)), \quad \text{Assuming limiting/invariant distribution} \\ \lambda & = & \mathbb{E}(S_n)^{-1}, \quad \text{arrival rate} \\ W & & \text{Mean time spent in system by "typical" customer} \end{array}$$

Little's law

$$\begin{array}{rcl} L & = & \lambda W, \quad \text{True for any subsystem} \\ L_0 & = & \lambda W_0, \quad \text{e.g. considering only the queue proper (not servers)} \end{array}$$

Kendall notation: $A/B/c$ compact way to describe queueing systems

A Describes the arrival process

B Describes the service (process) time distribution

c Is the number of servers, finite/infinite number

Examples of A/B

M Markov: Poisson arrival process/independent exponential service times

G General arrival/service time, sometimes GI for A to stress independence (renewal process)

Er Erlang distribution

H_n Hyper-exponential with n terms

Ph Phase type distribution

Sometimes elaborated like $M/G/c/K$ (Poisson arrival process, general (independent) service times, c servers and waiting buffer with K positions.

Markovian arrivals and exponential service times, Section 9.2

The systems in this section, and many if not all basic queueing models can be seen as birth-and-death processes The $M/M/1$ system

$$\begin{aligned}
\pi_k\lambda &= \pi_{k+1}\mu \\
\pi_k &= \left(\frac{\lambda}{\mu}\right)^k \pi_0 \\
\pi_0 \sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k &= 1, \quad \text{normalisation} \\
\pi_k &= \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k, \quad \lambda < \mu, k = 0, 1, \ldots \\
L &= \mathbb{E}(X(t)) = \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k = \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{\lambda}{\mu-\lambda}
\end{aligned}$$

Either use that the distribution is geometric or

$$\begin{aligned}
\sum_{k=0}^{\infty} k\pi_k &= \sum_{k=0}^{\infty} k\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k = \left(1-\frac{\lambda}{\mu}\right)\sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k\left(\frac{\lambda}{\mu}\right)^k = \left(1-\frac{\lambda}{\mu}\right)\sum_{k=0}^{\infty}\sum_{j=0}^{k-1}\left(\frac{\lambda}{\mu}\right)^k \\
&= \left(1-\frac{\lambda}{\mu}\right)\sum_{j=0}^{\infty}\sum_{k=j+1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k = \left(1-\frac{\lambda}{\mu}\right)\sum_{j=0}^{\infty}\sum_{k=j+1}^{\infty}\left(\frac{\lambda}{\mu}\right)^{j+(k-j-1)+1} \\
&= \left(1-\frac{\lambda}{\mu}\right)\frac{\lambda}{\mu}\sum_{j=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^j\sum_{h=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^h = \frac{\lambda}{\mu-\lambda}
\end{aligned}$$

The stochastic process $Y(t) = (X(t)-1)^+ = \max(X(t)-1, 0)$ is the process of customers in the queue (is it Markovian?)

Introduce $\delta(t) = I_{X(t)>0}$ so $Y = (X(t)-1)\delta(t)$

$$\begin{aligned} L_0 &= \mathbb{E}(Y(t)) = \mathbb{E}\left((X(t)-1)^+\right) = \mathbb{E}(X(t)\delta(t) - \delta(t)) \\ &= \mathbb{E}(X(t)) - \mathbb{E}(\delta(t)) = L - \mathbb{P}\{X(t) > 0\} \\ &= \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda}{\mu}\frac{\lambda}{\mu-\lambda} \\ W &= \frac{L}{\lambda} = \frac{1}{\mu-\lambda} \end{aligned}$$

We will consider the time in system T of a generic customer, i.e. a customer, that arrives to a system governed by the stationary distribution.

The distribution of customers at an arbitrary arrival will be given by π_k due to the *PASTA* (Poisson Arrivals See Time Averages)/Wikipedia: The Arrival Theorem.

The distribution is derived in the Pinsky and Karlin without transforms. Next week we will see that the distribution of T can be seen as an instance of a more general results. Here we will use a transform based approach.

Define the Laplace transform of a non-negative random variable as

$$L_X(\theta) = \mathbb{E}\left(e^{-\theta X}\right) = \int_0^\infty e^{-\theta x} f(x)\mathrm{d}x, \quad \text{provided } X \text{ is continuous with density } f(x)$$

1. We assume first come, first served, i.e. the arriving customers gets to the back of the queue.
2. The arbitrary customer finds N customers on arrival
3. $\mathbb{P}\{N = k\} = \pi_k$
4. If $N > 0$ one customer will be serviced, the remaining service time is exponential with rate μ, due to the memoryless property of the exponential distribution (and strictly due to the strong Markov property)
5. If $N > 0$ there will be $N-1$ customers in the queue, that needs to be served before the arriving customer. Each of these service times are exponential with rate μ, they are all independent, and independent for the remaining service time of the customer under service.
6. The arriving customer will be served due to an exponential service time with rate μ independent of the previous mentioned times.
7. In conclusion $T = \sum_{k=1}^{N+1} Y_i, \quad Y_i \sim \exp(\mu)$

$$
\begin{aligned}
L_{Y_i}(\theta) &= \int_0^\infty e^{-\theta y}\mu e^{-\mu y}\mathrm{d}y = \frac{\mu}{\mu+\theta} \\
M &= N+1 \sim \text{geo}\left(\frac{\lambda}{\mu}\right) \\
\phi_M(s) &= \mathbb{E}\left(s^M\right) = \sum_{k=1}^\infty s^k\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{k-1} = \frac{\left(1-\frac{\lambda}{\mu}\right)s}{\left(1-s\frac{\lambda}{\mu}\right)} = \frac{(\mu-\lambda)s}{\mu-s\lambda} \\
L_T(\theta) &= \mathbb{E}\left(e^{-\theta T}\right) = \mathbb{E}\left[\mathbb{E}\left(e^{-\theta T}\middle|\, M\right)\right] = \mathbb{E}\left[\mathbb{E}\left(e^{-\theta\sum_{i=1}^M Y_i}\middle|\, M\right)\right] = \mathbb{E}\left[\mathbb{E}\left(\prod_{i=1}^M e^{-\theta Y_i}\middle|\, M\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^M \mathbb{E}\left(e^{-\theta Y_i}\middle|\, M\right)\right] = \mathbb{E}\left[\prod_{i=1}^M \mathbb{E}\left(e^{-\theta Y_i}\right)\right] = \mathbb{E}\left[\mathbb{E}\left(e^{-\theta Y_i}\right)^M\right] = \phi_M\left(\mathbb{E}\left[e^{-\theta Y_i}\right]\right) \\
&= \frac{(\mu-\lambda)\mathbb{E}\left[e^{-\theta Y_i}\right]}{\mu-\lambda\mathbb{E}\left[e^{-\theta Y_i}\right]} = \frac{\mu(\mu-\lambda)}{\mu(\mu+\theta)-\lambda\mu)} = \frac{\mu-\lambda}{\mu-\lambda+\theta}
\end{aligned}
$$

We have $T \sim \exp(\mu-\lambda)$.

In the $M/M/\infty$ system we find the invariant distribution of $X(t)$ to be Poisson with parameter $\frac{\lambda}{\mu}$.

The $M/M/c/\infty$ system has a distribution that in some sense is a combination of the Poisson and geometric distribution.

Poisson arrivals and general service time $M/G/1$ system, Section 9.3

Sequence of arrivals with inter arrival times $\{S_n; n \geq 0\}$ and service times $\{Y_n; n \geq 0\}$ where S_n are generated by a Poisson process (S_n indpendent exponential), and Y_n are independent with a generic distribution.

$$
\begin{aligned}
\mathbb{P}\{Y_n \leq y\} &= G(y) \\
\mathbb{E}(Y_n) &= \nu \\
\mathbb{V}\text{ar}(Y_n) &= \tau^2
\end{aligned}
$$

The process $\{X(t); t \geq 0\}$ is not Markovian, however, we can make it Markovian by including the elapsed service time in the state description. That process would in general be uncountable infinite in one of the dimensions.

Introduce

$$V_n = \text{Time of } n\text{th departure}$$

the successive times of departures from the system, and define

$$X_n = X(V_n)$$

the queue content immediately after the nth departure, and A_n the number of arrivals during the service of the nth customer. With a little care, this can be formulated as a function of the random variables already introduced.

Define $\delta_n = I_{X_n>0}$, the indicator that there is at least one customer in the system after the nth departure

We have

$$\begin{aligned} X_n &= \begin{cases} X_{n-1} - 1 + A_n & \text{if} \quad X_{n-1} > 0 \\ A_n & \text{if} \quad X_{n-1} = 0 \end{cases} \\ X_n &= \max\left(A_n, X_{n-1} - 1 + A_n\right) = (X_{n-1} - 1)^+ + A_n = X_{n-1} - \delta_{n-1} + A_n \end{aligned}$$

Thus X_n forms a (discrete time) Markov chain with

$$\begin{aligned} P_{0j} &= \alpha_j \\ P_{ij} &= \alpha_{j-i+1}, \quad \text{for } i \geq 1 \\ &\text{with} \\ \alpha_j &= \mathbb{P}\{A_n = j\} = \int_0^\infty \mathbb{P}\{A_n = j | Y_n = y\} \mathrm{d}G(y) = \int_0^\infty \frac{(\lambda y)^j}{j!} e^{-\lambda y} \mathrm{d}G(y) \\ &= \int_0^\infty \frac{(\lambda y)^j}{j!} e^{-\lambda y} g(y) \mathrm{d}y, \quad \text{if } Y \text{ has density } g(y) \end{aligned}$$

The Markov chain $\{X_n; n \geq 0\}$ can be analysed directly or by use of transform methods, we will, however, restrict ourselves to finding the moments, the first moment in particular. That is $\lim_{t\to\infty} \mathbb{E}(X(t)) = L$.

To shorten notation a bit, we rephrase

$$\begin{aligned} X_n &= X_{n-1} - \delta_{n-1} + A_n \\ X' &= X - \delta + A \\ \mathbb{E}(X') &= \mathbb{E}(X - \delta + A) \\ &= \mathbb{E}(X) - \mathbb{E}(\delta) + \mathbb{E}(A) \end{aligned}$$

$$\mathbb{E}(\delta) = \mathbb{E}(A) = \mathbb{E}[\mathbb{E}(A|Y)) = \mathbb{E}(\lambda Y) = \lambda\nu$$

So taking the expectation on both sides gives us the probability that the system is empty.

Now squaring the equation we get

$$\begin{aligned} X'^2 &= (X - \delta + A)^2 = X^2 + \delta^2 + A^2 - 2X\delta + 2(X - \delta)A \\ \mathbb{E}\left(\delta^2\right) &= \mathbb{E}(\delta) = \lambda\nu \\ \mathbb{E}\left(X\delta\right) &= \mathbb{E}(X) = L \\ \mathbb{E}\left(A^2\right) &= \mathbb{E}\left[\mathbb{E}\left(A^2 \middle| Y\right)\right] = \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(A \middle| Y\right) + \mathbb{E}\left(A \middle| Y\right)^2\right] = \mathbb{E}\left[\lambda Y + (\lambda Y)^2\right] = \lambda\nu + \lambda^2(\nu^2 + \tau^2) \end{aligned}$$

So taking expectation on both sides gives us

$$\begin{aligned} 0 &= \lambda\nu + \lambda\nu + \lambda^2(\nu^2 + \tau^2) - 2L + 2(L - \lambda\nu)\lambda\nu \\ L &= \frac{2\lambda\nu + \lambda^2\tau^2 - \lambda^2\nu^2}{2(1 - \lambda\nu)} \end{aligned}$$

with $\rho = \lambda\nu$ we can reformulate as

$$L = \rho + \frac{\lambda^2\tau^2 + \rho^2}{2(1 - \rho)}$$

The mean queue length depends on the service time distribution only through the first two moments of that distribution

Now consider Z_n the number of customers in the system immediately before the nth arrival. The distribution of Z_n must be equal to the distribution of X_n as the X_n can only decrease by one and Z_n can only increase by one.

Finally due to PASTA Z_n has the same distribution as the limiting (invariant) distribution of $X(t)$.