Random Processes III

APP MTH 3016, APP MTH 4116, APP MTH 7056

Lectured by

Prof. Joshua Ross, The University of Adelaide

Course Information

General Information

Lecturer: Prof. Joshua Ross, joshua.ross@adelaide.edu.au

Consulting: Wednesdays and Thursdays from 2pm – 3pm in my office which is in Ingkarni Wardli Building, Office 6.22. However, I cannot guarantee that I will be in my office the entire (or any) of this time unless you schedule a time with me. This can be done by emailing me (joshua.ross@adelaide.edu.au) to confirm a time.

Queries:

- Use the discussion board on MyUni. Other students in the class can also answer any questions here so if you have a look and see a question that you know the answer to, you can answer it. I will keep checking these questions and direct you towards (or simply tell you) the answers.
- You also have the option to email me any *short* questions. If the answer is quite long, I will direct you to book a time to see me.

Tutorials

Tutorials will run in the even weeks up until (and including) Week 10. There will be no tutorial in Week 12. No solutions to these tutorial questions will be uploaded but all questions will be answered in these tutorials (or subsequent lectures if we run out of time).

Assessment

1. Assignments ($\times 5$) in total are worth 15%.
   - Due at 1pm on Fridays in Weeks $3, 5, 7, 9, 12$.
   - Submit (scanned or typed) assignments online.
   - Late assignments will not be accepted.
   - No extensions will be given, but exemptions for medical/compassionate reasons may be granted.
2. Project worth 15%.
   - Due at 1pm Friday in Week 10.
   - More details to follow in lectures.
3. Exam worth 70%.

Lecture 1: An introduction to stochastic modelling in continuous time – Epidemic!

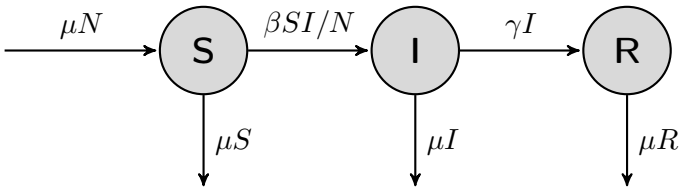
Concepts checklist

At the end of this lecture, you should be able to:

- Describe the difference between *deterministic* and *stochastic* models;
- Understand *why we wish to model* (i.e., understand model *outputs*);
- Appreciate *why stochasticity can be important in modelling*; and,
- Appreciate *why discreteness and stochasticity together can be important in modelling.*

Let's start by considering an outbreak of a disease in a population. For some diseases, each individual in the population can be thought of as being *susceptible* to the disease, *infectious* with the disease, or *recovered* from the disease. For many diseases, for example measles, there is a long period of immunity following infection, meaning that once recovered you can no longer be infected.

This structure can be represented in a *(state) transition diagram* (this is technically not a state transition diagram – you will see why next lecture), with *compartments* representing the number of susceptible (S), infectious (I) and recovered (R) individuals, arrows representing the *events* – additions to, deletions from, or transitions between compartments – and the (instantaneous) *rates* of these events. This (state) transition diagram represents a so-called *compartmental model* of the *process/system.*



One compartmental model that you might specify for this system is a set of ordinary differential equations:

$$\frac{dS}{dt} = \mu(N - S) - \beta SI/N,$$
$$\frac{dI}{dt} = \beta SI/N - (\mu + \gamma)I,$$
$$\frac{dR}{dt} = \gamma I - \mu R,$$

where μ is the per-capita birth/death rate, β is the (effective) transmission rate parameter, γ is the per-capita recovery rate, and N is the population size (i.e., $N = S(0) + I(0) + R(0)$).

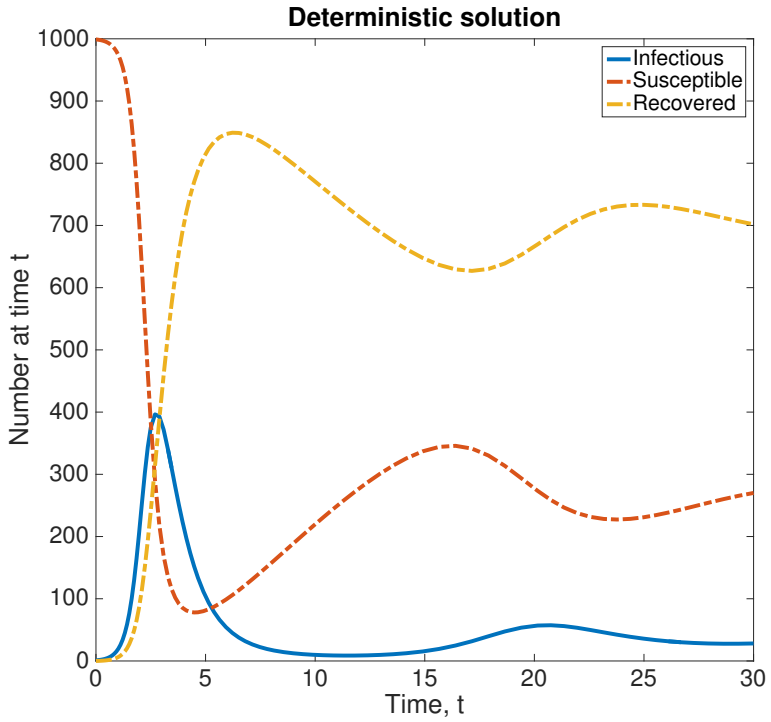


Figure 1: Numerical solution to ODE epidemic model with $\mu = 0.05$, $\beta = 4$, $\gamma = 1$, $N = 1000$, $S(0) = N - 1$, $I(0) = 1$ and $R(0) = 0$ over the period $t = 0$ to $t = 30$.

This is an example of a *deterministic* model of the system. **A deterministic model of a system will predict a single certain outcome, given *inputs* to the system.** Here the inputs are values for the parameters and an initial condition $(S(0), I(0), R(0))$.

The outcome for $\mu = 0.05$, $\beta = 4$, $\gamma = 1$, $N = 1000$, $S(0) = N - 1$, $I(0) = 1$ and $R(0) = 0$ is plotted in Figure 1, from *numerical solution* in MATLAB.

Another compartmental model that one might specify for this system – the type of model that will be the focus of this course – is a *continuous-time Markov chain* (CTMC). This is an example of a *stochastic* model of the system. **A stochastic model of a system will predict a set of possible outcomes, along with their probabilities of occurrence, given the inputs to the system.**

Stochastic, from the Greek *οτόχος* ("stokhos") for *target*, *aim* or *guess*, means *random*; its antonym is *sure*, *certain*, or *deterministic*.

A set of five independent outcomes (*realisations*) for the same parameters as for the deterministic model are plotted in Figure 2, from *numerical simulation* in MATLAB.

Some important features of these realisations, and particularly in comparison to the deterministic (ODE) model outcome, should be noted:

- In some realisations, the infection *fades out*, whereas in the deterministic model we have persistence of the disease. This is important for informing disease management.

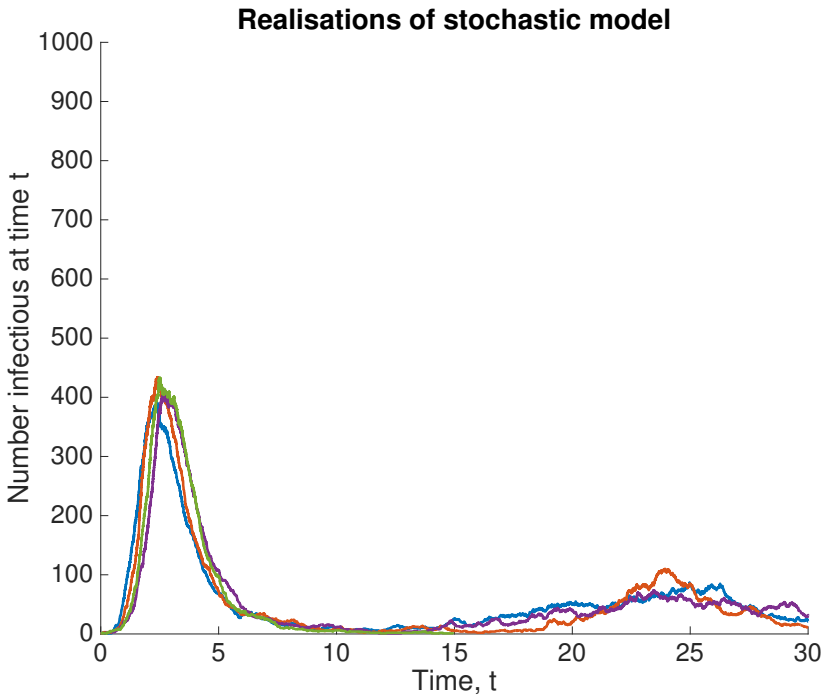


Figure 2: Five realisations of the continuous-time Markov chain epidemic model with $\mu = 0.05$, $\beta = 4$, $\gamma = 1$, $N = 1000$, $S(0) = N - 1$, $I(0) = 1$ and $R(0) = 0$ over the period $t = 0$ to $t = 30$.

- The realisations reflect *discrete* individuals, rather than being modelled as continuous variables. This is more accurate for this (and many other) systems. This also means that stochasticity can be more influential, in particular with respect to phenomena such as fade out.

Some questions that you might use this model to answer are:

- What is the probability of *initial fade out* (i.e., fade out before seeing a large peak)?;
- What is the probability of *epidemic fade out* (i.e., fade out in the trough following the first wave of the infection)?;
- What is the expected time until fade out?; and,
- What is the distribution of the number of infectious individuals at time $t, t \in \mathbb{R}^+$?

The answer to these questions are model *outputs* – the probabilities of the various fade out events, or the expected time to fade out, or the distribution of the number infectious at time $t, t \in \mathbb{R}^+$. Effectively model outputs are what we want to know, i.e., the purpose of modelling is to gain these outputs.

We build mathematical models to translate information that we know, or are willing to assume, into information that we want to know.

So, what is the algorithm used to produce the stochastic realisations? To explain this algorithm, it is helpful to consider the *Events*, their *Rates*, and the *Change in State* that results from each event. The state of this CTMC epidemic model at time t, is $X(t) = (S(t), I(t))$

being the number of susceptible and infectious individuals in the population at time t. The information discussed is given in Table 1.

Event	Rate	Change in State
Birth (of susceptible)	μN	$(S, I) \to (S+1, I)$
Death of susceptible	μS	$(S, I) \to (S-1, I)$
Infection	$\beta SI/N$	$(S, I) \to (S-1, I+1)$
Removal of infectious	$(\gamma+\mu)I$	$(S, I) \to (S, I-1)$

Table 1: CTMC epidemic model Events, Rates and Changes in State.

Note, that we have not modelled the number of recovered individuals here, as we are primarily interested in the dynamics of infectious individuals and the rates are only dependent upon S, I and parameters (and not R). Also, note that the two events, death of an infectious and recovery of an infectious, have been combined into a single event – removal of infectious – as the Change in State is identical between these events, and that we have specified the Rate as the sum of the individual rates.

The algorithm essentially operates as follows. Given the current state (S, I), compute the rate of each event and sum these to also get the *total rate* (TR) of seeing an event. Wait an exponentially-distributed amount of time, with mean 1/TR; this is the time of the next event. Choose the event to occur with probability in proportion to the rate of each event; i.e., e.g., the event Death of Susceptible happens with probability (w.p.) $\mu N/\text{TR}$. Update the state according to which event occurs. Repeat until the infection dies out (i.e., $I = 0$), or the time surpasses the desired time period.

The name *random process* should now be clearer. We have considered a continuous series of events that involve randomness in their timing and type.

Some examples of systems which are typically best modelled with random processes are:

- **Engineering:** Telecommunications, computer networks, industrial processes, dams.
- **Biology:** Evolution, genetics, epidemics, species interaction.
- **Chemistry:** Polymerisation, reactions, bonding.
- **Physics:** Quantum mechanics, statistical mechanics.
- **Economics and Finance:** Portfolio management, financial instruments.
- **Management Science:** Call centres, queues and networks of queues.

The focus in this course is on random processes in continuous time ($t \in \mathbb{R}^+$, henceforth simplified to $t \geq 0$) and discrete in state.

Lecture 2: An introduction to continuous-time Markov chains (CTMCs) – Problematic printers!

Concepts checklist

At the end of this lecture, you should be able to:

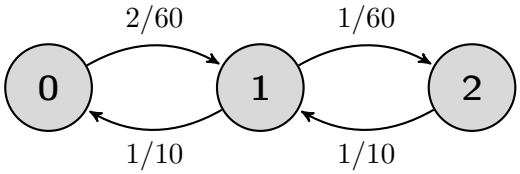
- Define the *state space* of a discrete-state random process;
- Define a *transition rate matrix* from a *state transition diagram*, and vice-versa;
- State the *features/properties of a transition rate matrix*;
- Define a continuous-time Markov chain (CTMC);
- Define a *time-homogeneous* CTMC; and,
- Define a *transition function* of a time-homogeneous CTMC and state its *assumed properties*.

Example 1. Repairman problem, with 2 machines

In the School of Mathematical Sciences we have 2 printers on level 6. From experience, a printer fails, independently of the other printer, on average, every 60 days. It takes, on average, 10 days to repair a single machine when it fails; but there is only one repairman.

We could model this process as a continuous-time Markov chain. If we assume that the *times of failures and repairs are independent and exponentially distributed*, then we can let $X(t)$ be the number of failed printers at time $t \geq 0$. Hence $X(t)$ can be either 0 (both printers working), 1 (1 printer working, 1 not) or 2 (both printers 'on the blink'). The *state space*, $\mathcal{S}$, of the process is then $\mathcal{S} = \{0, 1, 2\}$. **The state space is the set of possible values the process can adopt.**

Given the states, events, and rates, we can write down a *state transition diagram*.



Do you know where all the rates come from? What about the rate 2/60 per day of transitioning from State 0 to State 1? In State 0, we have 2 printers working, each failing independently at rate 1/60. Hence, the rate at which we see one fail can be determined

by considering the distribution of $F = \min\{F_1, F_2\}$, where F_1 and F_2 are independent and identically distributed exponential random variables with rate 1/60:

$$\begin{aligned}
\Pr(F > t) &= \Pr(\min\{F_1, F_2\} > t) \\
&= \Pr(F_1 > t \cap F_2 > t) \\
&= \Pr(F_1 > t)\Pr(F_2 > t) && \text{(independence)} \\
&= \exp(-(1/60)t)\exp(-(1/60)t) && \text{(identically exponential, rate 1/60)} \\
&= \exp(-(2/60)t).
\end{aligned}$$

Hence, the time to the first failure is exponentially distributed with rate 2/60.

Useful for analysing the behaviour of processes such as this, is the *generator* (also called *transition rate matrix* or *Q-matrix*), Q. The generator, Q, corresponding to the process described, is

$$Q = \begin{bmatrix} -2/60 & 2/60 & 0 \\ 1/10 & -7/60 & 1/60 \\ 0 & 1/10 & -1/10 \end{bmatrix}.$$

Some important features to note about the generator are:

- Off-diagonal entries are non-negative, and correspond to the rates of events that change the state of the process;
- Each row sum is zero; and,
- Diagonal entries are non-positive (here negative), which follows from the first two observations.

Now, what defines a random process as being a continuous-time Markov chain?

Definition 1. *A continuous-time Markov chain is a random process which satisfies the* Markov property*:*

$$\Pr(X(t+s) = k \mid X(u) = i, X(s) = j, u < s) = \Pr(X(t+s) = k \mid X(s) = j),$$

for all $s, t \in [0, \infty)$ *and all* $i, j, k \in \mathcal{S}$.

This says that the future path depends on the history $\{X(u), u \leq s\}$ only through the present state $X(s)$. The process is memoryless.

Is the repairman process described a continuous-time Markov chain? Let's think about the behaviour of this process:

- If in State 0, we wait an exponentially-distributed time, T_0, with mean 30 days until we see a printer failure, and the process transitions to State 1;
- Similarly, if in State 2, we wait an exponentially-distributed time, T_2, with mean 10 days until we see a repair of a machine, and the process transitions to State 1; and,
- In State 1,

– (a) if we have entered from State 0, then one printer has not failed. But should the fact that it has not failed while sojourning in State 0 influence the time until it now fails, say, F_2?

– (b) we have two competing events: the repairman is working on fixing the broken printer, while the second printer might fail. Like the two competing events of failure of printers, these are independent and exponentially-distributed, but are not identical as they have different means. Let's say these random variables are $T_{1,0}$ and $T_{1,2}$, respectively, and then the event that occurs is whichever happens first; i.e., we are interested in $\min\{T_{1,0}, T_{1,2}\}$.

Hence, the behaviour of the process is memoryless when in States 0 and 2, but we need to check that it is memoryless when in State 1. Let's consider (a) first.

Since the time to failure is assumed exponentially-distributed, then assuming that the process sojourned in State 0 for t units of time, we have

$$\begin{aligned}
\Pr(F_2 > t+s | F_2 > t) &= \frac{\Pr(F_2 > t+s \cap F_2 > t)}{\Pr(F_2 > t)} \\
&= \frac{\Pr(F_2 > t+s)}{\Pr(F_2 > t)} \\
&= \frac{\exp(-(t+s)/60)}{\exp(-t/60)} \\
&= \exp(-s/60) \\
&= \Pr(F_2 > s).
\end{aligned}$$

Hence, the distribution of the time until failure remains exponential, with the time elapsed in State 1, regardless of its history. It is said to be memoryless – a property for which the exponential distribution is the only continuous distribution to possess.

Definition 2. *A continuous random variable T is said to have a memoryless property if*

$$\Pr(T > t+s | T > t) = \Pr(T > s)$$

for $s, t \geq 0$.

Now, let's return to (b). As before, considering the distribution of $M = \min\{T_{1,0}, T_{1,2}\}$ we have:

$$\begin{aligned}
\Pr(M > t) &= \Pr(\min\{T_{1,0}, T_{1,2}\} > t) \\
&= \exp(-(7/60)t).
\end{aligned}$$

Hence, the time to the next event is exponentially distributed with rate 7/60. This means that the memoryless property is preserved with respect to the time to the next event.

Let us now consider the probabilities of transitioning to each of the states, given that a jump occurs. We have, for some $t > 0$,

$$\begin{aligned}
\Pr(\text{moves to State 0} | \text{moves out of State 1 at time } t) &= \Pr(T_{1,0} < T_{1,2} | M \in [t, t+dt)) \\
&= \frac{\Pr(T_{1,0} \in [t, t+dt) \cap T_{1,2} > t)}{\Pr(M \in [t, t+dt))} \\
&= \frac{(1/10)e^{-t/10}e^{-t/60}}{(7/60)e^{-7t/60}} (\text{independence}) \\
&= 6/7.
\end{aligned}$$

This probability is independent of t. In other words, the time of the next event, and which event occurs first, are independent random variables. Hence, this process is memoryless; i.e., it is a continuous-time Markov chain!

For this particular process, we can write down its *transition function*, $P(s, t+s)$. This can be written as matrix with entries corresponding to the probabilities of transitioning between states from time s to time $t+s$. The transition function is [I got this from Q, and will tell you how soon (like, next week!)]:

$$P(s,t+s) = \frac{1}{25}\begin{bmatrix} 18+3e^{-t/6}+4e^{-t/12} & 6-4e^{-t/6}-2e^{-t/12} & 1+e^{-t/6}-2e^{-t/12} \\ 18-12e^{-t/6}-6e^{-t/12} & 6+16e^{-t/6}+3e^{-t/12} & 1-4e^{-t/6}+3e^{-t/12} \\ 18+18e^{-t/6}-36e^{-t/12} & 6-24e^{-t/6}+18e^{-t/12} & 1+6e^{-t/6}+18e^{-t/12} \end{bmatrix}.$$

Note that this transition function is actually only dependent upon the elapsed time t (and not on the precise times s and $t+s$). It is said to be *time-homogeneous*.

Definition 3. *A continuous-time Markov chain is* time-homogeneous *if*

$$\Pr(X(t+s) = j | X(s) = i) = \Pr(X(t) = j | X(0) = i)$$

for all $i, j \in \mathcal{S}$ *and* $s, t \geq 0$.

The *transition function* of a time-homogeneous CTMC is $P(t) = (P_{ij}(t))_{i,j\in\mathcal{S}, t\geq 0}$.

For such a time-homogeneous CTMC, knowledge of the transition function

$$P_{i,j}(t) = \Pr(X(t) = j | X(0) = i)$$

for all $i, j \in \mathcal{S}$ and $t \geq 0$ would allow to answer any question we'd like about the process! So it would be good to be able to know this for any continuous-time Markov chain; you'll find out how in this course! But for now some *assumed properties*:

$$\sum_{j\in\mathcal{S}} P_{i,j}(t) = 1$$

for each $i \in \mathcal{S}$ and for all $t \geq 0$ (the probabilities across states always sum to one); we then say that the transition function is *honest*. And the probabilities are non-negative,

$$P_{i,j}(t) \geq 0$$

for all $i, j \in \mathcal{S}$ and $t \geq 0$.

Lecture 3: The Poisson process, and the M/M/1 queue – How busy will you keep me, and how big a waiting room?

Concepts checklist

At the end of this lecture, you should be able to:

- Define the Poisson process as a *continuous-time Markov chain* (CTMC); and,
- Define the M/M/1 queue.

Example 2. The Poisson process

Another (important) example of a continuous-time Markov chain (CTMC) is the Poisson process. You have probably seen the Poisson process before, defined as follows.

Definition 4. *Let the sequence $t_1, t_2, \ldots$ be independent exponential random variables with rate λ, and define*

$$\begin{aligned} T_0 &:= 0, \\ T_n &:= t_1 + \cdots + t_n \text{ for } n \geq 1, \\ \text{and} \quad N(t) &:= \max\{n : T_n \leq t\} \text{ for } t \geq 0. \end{aligned}$$

Then, $\{N(t) : t \geq 0\}$ is called the Poisson process.

For example, we can think of the random variables t_n as the times between arrivals of students at a Professor's office (a timely reminder that I have specified office hours!), so

- $T_n = t_1 + \cdots + t_n$ is the arrival time of the n^{th} student, and
- $N(t)$ is the number of arrivals by time t.

In general, the process $\{N(t)\}$ is used to model systems where events of a particular type (often called, points) occur in time, in such a way that the probability of a point occurring in the interval after t is independent of what has happened up to t.

A quantity of interest, to help me plan my day, is the distribution of the number of arrivals I should expect in a day, or a working week. That is, what is the distribution of $N(t)$?

Let us calculate the distribution of $N(t)$. We have $N(t) = n$ if and only if $T_n \leq t < T_{n+1}$; i.e., the n^{th} student arrives before time t but the $(n+1)^{\text{th}}$ after t. Conditioning on the value of $T_n \in [s, s + dt)$ such that $s \leq t$ and noting that $T_{n+1} > t$, we have

$$\begin{aligned} \Pr(N(t) = n) &= \Pr(T_n \leq t < T_{n+1}) \\ &= \int_0^t \frac{\Pr(T_n \leq s)}{ds} \Pr(T_{n+1} > t | T_n \in [s, s + dt)) ds \quad \text{(Law of Total Probability)} \\ &= \int_0^t f_{T_n}(s) \Pr(t_{n+1} > t - s) ds \qquad \text{(independence of } t_{n+1}\text{)}, \end{aligned}$$

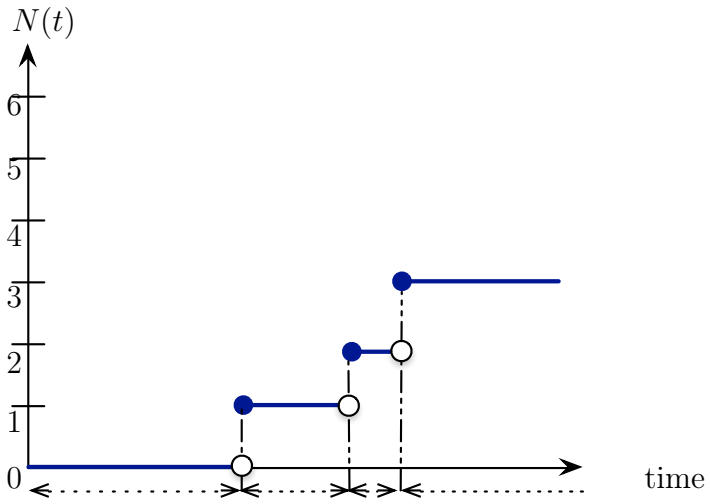


Figure 3: A realisation of the Poisson Process $\{N(t)\}$

where $f_{T_n}(s)$ is the probability density function (pdf) of T_n.

Note that the sum $T_n = t_1 + \cdots + t_n$, where t_i are independent exponential random variables with rate λ, has an Erlang(n, λ) distribution, with pdf

$$f_{T_n}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{n-1}}{(n-1)!} \quad \text{for } t \geq 0.$$

Thus,

$$\begin{aligned}
\Pr(N(t) = n) &= \int_0^t \lambda e^{-\lambda s} \frac{(\lambda s)^{n-1}}{(n-1)!} e^{-\lambda(t-s)} ds \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda t} \int_0^t s^{n-1} ds \\
&= \frac{\lambda^n}{(n-1)!} e^{-\lambda t} \frac{t^n}{n} \\
&= e^{-\lambda t} \frac{(\lambda t)^n}{n!}.
\end{aligned}$$

This is the Poisson distribution with parameter λt. Makes sense to call this the Poisson process, right?

Poisson process as a continuous-time Markov chain

Consider a process $\mathcal{X} = (X(t), t \geq 0)$, with the random variable $X(t)$ denoting the state of the system at time t.

The *state space* is $\mathcal{S} = \{0, 1, 2, \ldots\} = \mathbb{Z}_+$.

If $X(t) = n$, then n *arrivals*/*events* have occurred in the time interval $[0, t]$.

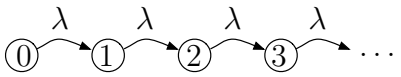
As the events occur at rate λ (i.e., the time between two consecutive events is $\exp(\lambda)$), we say that the *transition rate* $q_{i,i+1}$ from state i to state $i+1$ is λ:

$$\begin{aligned}
q_{i,i+1} &= \lambda \quad \text{for } i \in \mathcal{S}, \\
q_{ij} &= 0 \quad \text{for } j \in \mathcal{S}, j \neq \{i, i+1\}.
\end{aligned}$$

The generator $Q = (q_{ij})_{i,j\in\mathcal{S}}$ is given by

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ 0 & -\lambda & \lambda & 0 & 0 & \cdots \\ 0 & 0 & -\lambda & \lambda & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Given the states, events, and rates, we can write down a state transition diagram:



The Poisson process is also known as the *simple pure birth process*.

The Poisson process is a CTMC, as

$$\begin{aligned} \Pr(X(t+s) = k \mid X(u) = i, X(s) = j, u < s) &= e^{-\lambda t}\frac{(\lambda t)^{k-j}}{(k-j)!} \\ &= \Pr(X(t+s) = k \mid X(s) = j), \end{aligned}$$

for all $j, k \in \mathcal{S}$ and $t \geq 0$.

Further, note that this also equals $\Pr(X(t) = k \mid X(0) = j)$, and hence the process is *time homogeneous*.

This is helpful, as it tells me something about the distribution of the number of students I can expect (once I've estimated λ); but it doesn't tell me if this will annoy Nigel, Matt, Sanjeeva, and possibly others! That will depend on how many are queued for help in the corridor...

Example 3. A continuous-time single-server queue (the M/M/1 queue)

Consider a simple, single-server queue. The state of the process $\mathcal{X} = (X(t), t \geq 0)$ is the number of customers (students) in the queue at time t, including the person being served (helped).

Assume there are arrivals to the queue according to a Poisson process with rate λ; and, services occur as a Poisson process with rate μ whenever there is at least one customer.

• The state space is $\mathcal{S} = \{0, 1, 2, \ldots\} = \mathbb{Z}_+$, where each state $i \in \mathcal{S}$ means there are i people in the system.

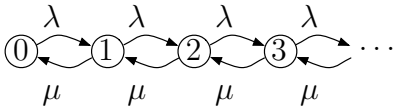
• The transition rates are:

$$\begin{aligned} q_{i,i+1} &= \lambda \quad \text{for } i \in \mathcal{S}, \\ q_{i,i-1} &= \mu \quad \text{for } i \in \mathcal{S}\backslash\{0\}, \\ q_{ij} &= 0 \quad \text{for } j \in \mathcal{S}\backslash\{i, i+1, i-1\}. \end{aligned}$$

Thus, the generator Q is:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix}.$$

- The state transition diagram is



Using the earlier analysis, you can (and should attempt to) show that in states $1, 2, \ldots$, the process waits an exponentially-distributed amount of time with rate $\lambda + \mu$, and at that time sees a decrease in the queue length of one with probability $\mu/(\lambda + \mu)$ and an increase in the queue length by one otherwise.

Is this a good model to inform the possible number of students taking Random Processes III waiting to see me? What physical constraint is missing? How could the model be modified to accomodate that?

Lecture 4: Sample paths and simulation – A simple, useful check / tool using estimation

Concepts checklist

At the end of this lecture, you should be able to:

- Prove the *sample path behaviour* of a CTMC:
  - the state next visited by a CTMC at the time of a state transition is proportional to the rate of visiting that state, and that it is independent of the time of the jump;
  - the time spent in a state before transitioning to another state is exponentially-distributed with rate equal to the sum of the rates of all possible transitions out of that state;
- Produce *realisations* of a CTMC using a computer.

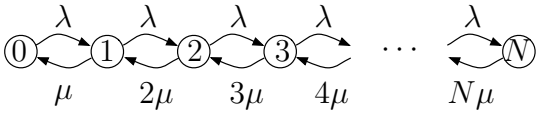
Example 4. N-machine reliability (Repairman problem), or a birth-death process, or an immigration-death process

Let us consider a process $\mathcal{X} = (X(t), t \geq 0)$ which, for example, models the number of N machines currently working, or the number of individuals in a population with a strict population ceiling of size N.

- State space $\mathcal{S} = \{0, 1, 2, \ldots, N\}$, where i is the number of machines currently working.

In terms of machines, as for the repairman problem, we have each machine fails at a constant rate, here μ, and the one repairman repairs each machine one at a time at a constant rate, here λ. In a population context, what does λ represent, and what does μ represent?

- State transition diagram



- The transition rates are

$$\begin{aligned}
\text{(repairing)} \quad & q_{n,n+1} = \lambda \quad \text{for } n = 0, 1, 2, \ldots, N-1, \\
\text{(breaking down)} \quad & q_{n,n-1} = n\mu \quad \text{for } n = 1, \ldots, N, \\
& q_{n,n} = -(n\mu + \lambda) \quad \text{for } n = 0, \ldots, N-1, \\
& q_{N,N} = -N\mu.
\end{aligned}$$

Thus, the generator Q is

$$Q = \begin{bmatrix} -\lambda & \lambda & \cdots & \cdots & \cdots & \cdots & 0 \\ \mu & -(\mu+\lambda) & \lambda & & & & \vdots \\ 0 & 2\mu & -(2\mu+\lambda) & \lambda & & & \vdots \\ \vdots & & 3\mu & -(3\mu+\lambda) & \lambda & & \vdots \\ \vdots & & & & & \ddots & \vdots \\ \vdots & & & \ddots & & \ddots & \lambda \\ 0 & \cdots & \cdots & \cdots & N\mu & & -N\mu \end{bmatrix}.$$

Where does the rate $n\mu$ come from? In the 2-printer example in Lecture 2, we had that the rate of a single failure when both machines are working was twice the rate of a single failure. Extend the calculations there to answer this question! After that, we are back to a process with at most only two events possible in each state; what about a continuous-time Markov chain with an arbitrary number of possible events. What is its behaviour?

Theorem 1. *Consider a continuous-time Markov chain $\mathcal{X}$ on state space $\mathcal{S}$ with generator Q. Then, for all $i \in \mathcal{S}$,*

$$\Pr\textit{(moves to state } j \neq i \mid \textit{leaves state } i \textit{ at time } t) = \frac{q_{ij}}{-q_{ii}}. \tag{1}$$

Proof. Let T_{ij}, for $i, j \in \mathcal{S}$, denote the time to leave i for j, then $T_{ij} \sim \text{Exp}(q_{ij})$. Define $M = \min_{k \neq i, k \in \mathcal{S}} T_{ik}$. Then,

$\Pr(\text{moves to state } j \neq i \mid \text{leaves state } i \text{ at time } t)$

$$\begin{aligned}
&= \frac{\Pr(T_{ij} \in [t, t+dt)) \cap \min_{k \neq i, k \neq j, k \in \mathcal{S}} T_{ik} > t)}{\Pr(M \in [t, t+dt))} \\
&= \frac{\Pr(T_{ij} \in [t, t+dt)) \prod_{k \neq i, k \neq j} \Pr(T_{ik} > t)}{\Pr(M \in [t, t+dt))} \qquad \text{(independence)} \\
&= \frac{\Pr(T_{ij} \in [t, t+dt)) \Pr(Z > t)}{\Pr(M \in [t, t+dt))} \qquad \text{where } Z \sim \exp\left(\sum_{k \neq i, k \neq j, k \in \mathcal{S}} q_{ik} \right) \\
&= \frac{q_{ij}}{q_{ij} + \sum\limits_{k \neq i, k \neq j, k \in \mathcal{S}} q_{ik}} \\
&= \frac{q_{ij}}{\sum\limits_{k \neq i, k \in \mathcal{S}} q_{ik}} \\
&= \frac{q_{ij}}{-q_{ii}}.
\end{aligned}$$

□

Theorem 2. *Consider a continuous-time Markov chain $\mathcal{X}$ on state space $\mathcal{S}$ with generator Q. Once moving to some state $i \in \mathcal{S}$, the time the system stays in i until it moves out of i is exponentially distributed with rate $-q_{ii}$.*

Proof. Note that:

- The time the system stays in i until it moves out of i, is the same as the time until the system moves to some state $j \neq i$.
- The latter time is $\min_{j \neq i, j \in \mathcal{S}} T_{ij} \sim \text{Exp}\left(\sum_{j \neq i, j \in \mathcal{S}} q_{ij}\right)$.
- This implies that $\min_{j \neq i, j \in \mathcal{S}} T_{ij} \sim \text{Exp}(-q_{ii})$, which completes the proof. □

We have discussed earlier that the transition function is a powerful tool that we'd like to evaluate, but our specification of models, and the *sample-path behaviour* has been in terms of rates of events/transitions. Before starting to explore the relationship between these in more detail, we should revisit the algorithm mentioned in Lecture 1 in light of the two theorems we just proved.

Simulating a continuous-time Markov chain

Pseudo-code for simulation of a CTMC:

1. INPUTS: initial state, state $\in \mathcal{S}$; total time to run simulation for, T; and set $t = 0$, $ts = 0$, and $ss =$ state.
2. Calculate the rate of each possible transition from the current state, $q_{\text{state},j}$.
3. Calculate the sum of these rates, q_{state}.
4. While $t < T$ and $q_{\text{state}} > 0$:
5. Generate an exponential random variable with rate q_{state}, te, and update time $t = t + te$ and store $ts = [ts; t]$.
6. Choose the state next visited, j, with probability $q_{\text{state},j}/q_{\text{state}}$, and update the current state, state $= j$ and store $ss = [ss; \text{state}]$.
7. End while

Let's consider Example 1, the problematic printers in maths. I am interested in the probability of both printers not working in 6 months time, given both are currently working. How can I evaluate this probability?

First, I could evaluate from the transition function presented in Lecture 1. The probability both machines are not working in t time units starting with both machines working is

$$(1 + \exp(-t/6) - 2\exp(-t/12))/25.$$

Hence, the desired probability is $(1 + \exp(-180/6) - 2\exp(-180/12))/25 \approx 0.04$.

However, I could also estimate this probability by simulation. Figure 4 below shows the estimate of this probability based upon increasing numbers of simulations.

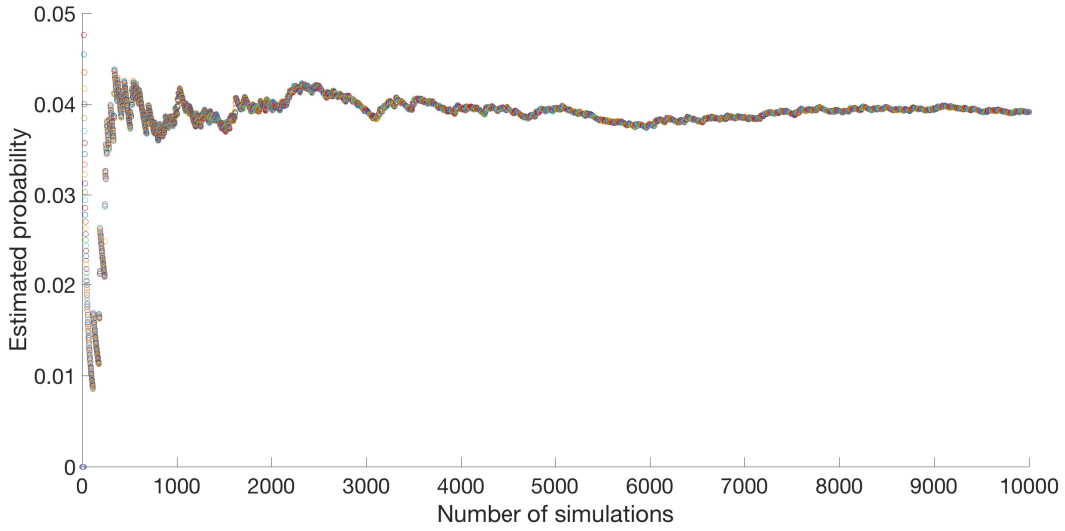


Figure 4: Estimate of the probability of both printers not working after 180 days, given both initially working, versus the number of simulations used.

Do you think $10,000$ simulations is sufficient to estimate the probability? How would you assess this?

In the exact expression for the probability, what happens as $t \to \infty$? What value does this converge to? What is the interpretation of this phenomenon?

Lecture 5: Transition rates to transition functions – The journey begins

---

Concepts checklist

At the end of this lecture, you should be able to:

- Define the *generator* of a CTMC as *limit of the transition function*;
- Define a *conservative* generator;
- Have an intuitive / *physical interpretation of the transition rates* of a CTMC; and,
- *Specify, and derive, the Chapman-Kolmogorov equation* of a CTMC.

As discussed, the transition function is a powerful tool that we'd like to evaluate, but our specification of models, and the *sample-path behaviour* has been in terms of rates of events/transitions. Here we begin our journey of going from transition rates to transition functions.

Definition 5. *The* generator *Q of a continuous-time Markov chain $\mathcal{X}$ has entries (when the limits exist)*

$$q_{ij} := \lim_{h\to 0^+} \frac{P_{ij}(h) - P_{ij}(0)}{h} = \lim_{h\to 0^+} \frac{P_{ij}(h)}{h} \quad (\geq 0) \qquad \textit{for } j \in \mathcal{S}, j \neq i,$$

$$q_{ii} := \lim_{h\to 0^+} \frac{P_{ii}(h) - P_{ii}(0)}{h} = \lim_{h\to 0^+} \frac{P_{ii}(h) - 1}{h} \quad (\leq 0)\,,$$

where $P_{ij}(h) := \Pr(X(t+h) = j \mid X(t) = i)$ is the conditional probability that the system is in state j by the end of the time interval h.

Loosely speaking, $Q = (q_{ij})_{i,j\in\mathcal{S}}$ is the *right-derivative* of the matrix $P(t)$ at the point $t = 0$. Q is sometimes referred to as the infinitesimal generator, or the (instantaneous) transition rate matrix (in particular the latter if $\mathcal{S}$ is finite).

In matrix notation:

$$Q = \lim_{h\to 0^+} \frac{P(h) - I}{h},$$

where $P(h) = (P_{ij}(h))_{i,j\in\mathcal{S}}$ and I is an identity matrix.

Properties

(i) Non-negative off diagonal elements:

$$q_{ij} \geq 0 \text{ for } j \in \mathcal{S}, j \neq i.$$

(ii) Non-positive diagonal elements:
Since we have

$$\sum_{j\in\mathcal{S}} P_{ij}(h) = 1 \text{ for } i \in \mathcal{S} \text{ and } h \in [0,\infty),$$

$$1 - P_{ii}(h) = \sum_{\substack{j\neq i \\ j\in\mathcal{S}}} P_{ij}(h)$$

$$\therefore \lim_{h\to 0^+} \frac{1-P_{ii}(h)}{h} = \lim_{h\to 0^+} \sum_{\substack{j\neq i \\ j\in\mathcal{S}}} \frac{P_{ij}(h)}{h}$$

$$\Rightarrow \quad -q_{ii} \geq \sum_{\substack{j\neq i \\ j\in\mathcal{S}}} \lim_{h\to 0^+} \frac{P_{ij}(h)}{h}$$

$$\Rightarrow \quad -q_{ii} \geq \sum_{\substack{j\neq i \\ j\in\mathcal{S}}} q_{ij}.$$

Note, that if every row sum is zero,

$$\sum_{j\in\mathcal{S}} q_{ij} = 0 \text{ for all } i \in \mathcal{S},$$

then we say that Q is *conservative*. We will deal with conservative generators only.

The **input to our model** are the q_{ij}. So, it will be useful to get some feeling about what these mean physically. Recall that we define for $i, j \in \mathcal{S}$ and $s, t \in [0,\infty)$

$$P_{ij}(t) = \mathbb{P}(X(t+s) = j | X(s) = i)$$

to be the probability of being in j at $t \geq 0$, given the system starts in i.

Physical Meaning

(i) For small h and for $i \neq j$,

$$P_{ij}(h) = q_{ij}h + o(h),$$

where $o(h)$ denotes a function $f(h)$ that satisfies $\lim_{h\to 0} \frac{f(h)}{h} = 0$.

$\equiv$ Pr(the chain moving out of state $i \to j$ in some small time h) $\approx q_{ij}h$.

$\equiv q_{ij}$ is the *instantaneous rate* (in a probabilistic sense) that the chain moves from $i \to j$.

(ii) For small h and $i \in \mathcal{S}$, we have

$$1 - P_{ii}(h) = -hq_{ii} + o(h).$$

$\equiv$ Pr(the chain moving out of state i in some small time h) $\approx (-q_{ii})h$.

$\equiv -q_{ii}$ is the instantaneous rate that the chain moves out of state i.

We now have an intuitive feel for what the entries of the generator represent, and how that relates, loosely, to derivatives of the transition function at time $t = 0$. However, how do we extend this to information about the transition function at any time t. Important to this translation are the Chapman-Kolmogorov equations.

Chapman-Kolmogorov Equation

Theorem 3. *For a continuous-time Markov chain* $(X(t) : t \geq 0)$ *on a state space* $\mathcal{S}$ *and for* $i, j \in \mathcal{S}$*, we have*

$$P_{ij}(t) = \sum_{k \in \mathcal{S}} P_{ik}(u) P_{kj}(t-u) \quad \textit{for} \quad 0 < u < t.$$

This is known as the Chapman-Kolmogorov equation.

In other words, the probability of going from $i \to j$ in time t, is the probability of going from $i \to k$ in time u multiplied by the probability of going from $k \to j$ in time $(t-u)$, summed over all possible states k.

Proof. Consider the CTMC at some time $s+u$ which is such that

$$s < s+u < s+t \quad \text{for} \quad s, u, t \in [0, \infty).$$

The chain must be in some state $k \in \mathcal{S}$ at time $s+u$, thus,

$$\begin{aligned}
P_{ij}(t) &= \Pr(X(s+t) = j | X(s) = i) \\
&= \sum_{k \in \mathcal{S}} \Pr(X(s+t) = j, X(s+u) = k | X(s) = i) \\
&= \sum_{k \in \mathcal{S}} \Pr(X(s+t) = j | X(s+u) = k, X(s) = i) \Pr(X(s+u) = k | X(s) = i) \text{ (L.T.P.)} \\
&= \sum_{k \in \mathcal{S}} \Pr(X(t) = j | X(u) = k) \Pr(X(u) = k | X(0) = i) \text{ (Markov property; time homogeneity)} \\
&= \sum_{k \in \mathcal{S}} P_{kj}(t-u) P_{ik}(u).
\end{aligned}$$

□

In matrix form: $P(t) = P(u)P(t-u)$, where the time-dependent matrix $P(t)$ is given by

$$P(t) = \begin{bmatrix} P_{i_0,i_0}(t) & P_{i_0,i_1}(t) & P_{i_0,i_2}(t) & \cdots \\ P_{i_1,i_0}(t) & P_{i_1,i_1}(t) & P_{i_1,i_2}(t) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}, \quad \text{with } i_0, i_1, i_2, \ldots \in \mathcal{S}.$$

Lecture 6: Kolmogorov differential equations – The key equations

Concepts checklist

At the end of this lecture, you should be able to:

- Derive the *Kolmogorov forward and backward differential equations*;
- Appreciate that the *Kolmogorov forward equations (KFDEs) don't always hold*, and that they *hold for finite-state processes and birth-death processes*; and,
- Understand a variety of approaches to *solve the KFDEs for certain CTMCs*, and effect these possibly with the assistance of a computer.

How do we gain information about the CTMC from the rates q_{ij}? That is, how does $P_{ij}(t)$ relate to Q? The Kolmogorov differential equations are key to answering this.

Kolmogorov Differential Equations

Theorem 4. *The Kolmogorov backward differential equations (KBDEs) of a continuous-time Markov chain are*

$$\frac{\mathrm{d}P_{ij}(t)}{\mathrm{d}t} = \sum_{k\in\mathcal{S}} q_{ik}P_{kj}(t) \qquad \text{for } i,j \in \mathcal{S}.$$

Proof. Starting with the Chapman-Kolmogorov equation,

$$P_{ij}(t+h) = \sum_{k\in\mathcal{S}} P_{ik}(h)P_{kj}(t),$$

we have

$$\begin{aligned}
\lim_{h\to 0^+} \frac{P_{ij}(t+h) - P_{ij}(t)}{h} &= \lim_{h\to 0^+} \frac{\sum_{k\in\mathcal{S}} P_{ik}(h)P_{kj}(t) - P_{ij}(t)}{h} \\
&= \lim_{h\to 0^+} \left[\frac{\sum_{k\in\mathcal{S}} P_{ik}(h)P_{kj}(t)}{h} - \frac{P_{ij}(t)}{h} \right] \\
&= \lim_{h\to 0^+} \left[\left(\sum_{\substack{k\neq i \\ k\in\mathcal{S}}} \frac{P_{ik}(h)}{h} P_{kj}(t) \right) + \frac{P_{ii}(h)P_{ij}(t)}{h} - \frac{P_{ij}(t)}{h} \right] \\
&= \lim_{h\to 0^+} \left[\left(\sum_{\substack{k\neq i \\ k\in\mathcal{S}}} \frac{P_{ik}(h)}{h} P_{kj}(t) \right) - \left(\frac{1 - P_{ii}(h)}{h} \right) P_{ij}(t) \right] \\
&\overset{\clubsuit}{=} \sum_{\substack{k\neq i \\ k\in\mathcal{S}}} q_{ik}P_{kj}(t) + q_{ii}P_{ij}(t) \\
&= \sum_{k\in\mathcal{S}} q_{ik}P_{kj}(t).
\end{aligned}$$

We need to justify the interchange of the *limit* and the *summation* in equality ♣, one such way to do so is by Fatou's Lemma. The details are omitted. □

Theorem 5. *The Kolmogorov forward differential equations (KFDEs) of a continuous-time Markov chain are*

$$\frac{\mathrm{d}P_{ij}(t)}{\mathrm{d}t} = \sum_{k\in\mathcal{S}} P_{ik}(t)q_{kj}.$$

Proof. Starting with the Chapman-Kolmogorov equations

$$P_{ij}(t+h) = \sum_{k\in\mathcal{S}} P_{ik}(t)P_{kj}(h),$$

we have

$$\begin{aligned}
\lim_{h\to 0^+} \frac{P_{ij}(t+h) - P_{ij}(t)}{h} &= \lim_{h\to 0^+} \frac{\left(\sum_{k\in\mathcal{S}} P_{ik}(t)P_{kj}(h)\right) - P_{ij}(t)}{h} \\
&= \lim_{h\to 0^+} \left[\left(\sum_{\substack{k\neq j\\ k\in\mathcal{S}}} P_{ik}(t)\frac{P_{kj}(h)}{h}\right) - P_{ij}(t)\left(\frac{1-P_{jj}(h)}{h}\right)\right] \\
&\overset{\spadesuit}{=} \left(\sum_{\substack{k\neq j\\ k\in\mathcal{S}}} P_{ik}(t)q_{kj}\right) + P_{ij}(t)q_{jj} \\
&= \sum_{k\in\mathcal{S}} P_{ik}(t)q_{kj}.
\end{aligned}$$

In general, the interchange of the lim and $\sum$ at ♠ cannot be justified. The most that can be shown is

$$\frac{\mathrm{d}P_{ij}(t)}{\mathrm{d}t} \geq \sum_{k\in\mathcal{S}} P_{ik}(t)q_{kj}.$$

We can, however, prove equality for *finite-state* processes and *birth-and-death* processes. The details are omitted. □

In matrix form,

$$\begin{aligned}
\text{KFDEs:} \quad & \frac{\mathrm{d}}{\mathrm{d}t}P(t) = P(t)Q \\
\text{KBDEs:} \quad & \frac{\mathrm{d}}{\mathrm{d}t}P(t) = QP(t).
\end{aligned}$$

Remark 1. We can construct continuous-time Markov chains for which the KFDEs do not hold; these Markov chains have weird properties: they can traverse infinitely many states in a finite time. They are used, for example, in modelling nuclear explosions, but are not often encountered in the sort of situation we consider in this course.

We shall now consider how to solve the Kolmogorov differential equations for some of the examples we have considered. It is usually more convenient to work with the KFDEs (when possible), so this is what we will do.

Solving Kolmogorov Differential Equations

When $|\mathcal{S}| < \infty$, the Kolmogorov differential equations are a finite system of linear ordinary first-order differential equations. We could solve these equations using a variety of *differential equation solvers*, for example using `ode45` in `MATLAB`. In this finite case, the solution to the Kolmogorov differential equations is

$$P(t) = \mathrm{e}^{Qt}, \qquad \text{where } \mathrm{e}^{A} := \sum_{n=0}^{\infty} \frac{A^n}{n!}.$$

This gives another potential method of solution, as a *matrix exponential*. This matrix exponential solution holds more broadly, but the details will be omitted here. This can assist the development of theory, however, typically we can still only evaluate the matrix exponential with the assistance of a computer. We might be able to evaluate the general form using an algebra package, such as `Wolfram Alpha`, `Mathematica` or `Maple`, or numerically for specific parameters / values of t, for example in `Matlab`.

Recalling Example 1, the problematic printers, I evaluated the transition function as a matrix exponential using `Wolfram Alpha`; I also checked using the `expm` function in `MATLAB` for a few values of t.

Another approach is to consider the *spectral representation* of Q. If $|\mathcal{S}| = d < \infty$, then assuming the eigenvalues of Q are all distinct[1] and labelled $\lambda_1, \lambda_2, \ldots, \lambda_d$, we have

$$Q = RDL = \lambda_1 M_1 + \lambda_2 M_2 + \ldots + \lambda_d M_d,$$

where, $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, R has columns $r'_1, r'_2, \ldots, r'_d$, L has rows $l_1, l_2, \ldots, l_d$, and so $M_i = r'_i l_i$, with l_i $(1 \times d)$ the left eigenvector of Q corresponding to the eigenvalue λ_i and r'_i $(d \times 1)$ the right eigenvector of Q corresponding to the eigenvalue λ_i, such that $r_i l'_j = \delta_{ij}$. Hence, we have

$$P(t) = \exp(Qt) = e^{\lambda_1 t} M_1 + e^{\lambda_2 t} M_2 + \ldots + e^{\lambda_d t} M_d.$$

Use this approach to construct the transition function for the problematic printers example.

We will now return to the Poisson process, and consider the solution to the KFDEs.

Example 2. Poisson process as a CTMC (continued)

Recall that

$$\begin{aligned} q_{n,n+1} &= \lambda \quad \text{for } n \geq 0, \\ q_{nn} &= \sum_{m \neq n} q_{nm} = -\lambda \quad \text{for } n \geq 0. \end{aligned}$$

Recall also that the KFDEs are

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{ij}(t) = \sum_{k \in \mathcal{S}} P_{ik}(t) q_{kj}.$$

For $i = 0$ and $n > 0$, we have

$$\frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} = \sum_{k \in \mathcal{S}} P_{0k}(t) q_{kn} = P_{0,n-1}(t) q_{n-1,n} + P_{0n}(t) q_{nn}.$$

[1] The theory is more general than this.

Hence,

$$\frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} = \lambda P_{0,n-1}(t) - \lambda P_{0n}(t) \quad \text{for} \quad n > 0,$$

$$\text{and} \quad \frac{\mathrm{d}P_{00}(t)}{\mathrm{d}t} = -\lambda P_{00}(t).$$

We can recursively solve these differential-difference equations, starting with

$$P_{00}(t) = e^{-\lambda t} P_{00}(0) = e^{-\lambda t}.$$

Substituting we have

$$\frac{\mathrm{d}P_{01}(t)}{\mathrm{d}t} = \lambda e^{-\lambda t} - \lambda P_{01}(t),$$

giving

$$P_{01}(t) = (\lambda t) e^{-\lambda t}.$$

Substituting again, we have

$$\frac{\mathrm{d}P_{02}(t)}{\mathrm{d}t} = (\lambda^2 t) e^{-\lambda t} - \lambda P_{02}(t),$$

giving

$$P_{02}(t) = \frac{(\lambda t)^2}{2} e^{-\lambda t}.$$

Continuing, we arrive at

$$P_{0n}(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}.$$

Lecture 7: Kolmogorov differential equations – Keep on solving

Concepts checklist

At the end of this lecture, you should be able to:

- Understand how *generating functions may sometimes assist in the solution of the KFDEs*;
- Solve the KFDEs for the Poisson process using generating functions; and,
- Appreciate that it is too difficult to be able to typically solve the KFDEs.

Example 2. Poisson process as a CTMC (continued)

Recall that

$$q_{n,n+1} = \lambda \quad \text{for } n \geq 0,$$
$$q_{nn} = \sum_{m \neq n} q_{nm} = -\lambda \quad \text{for } n \geq 0,$$

and the KFDEs are

$$\frac{\mathrm{d}}{\mathrm{d}t} P_{ij}(t) = \sum_{k \in \mathcal{S}} P_{ik}(t) q_{kj}.$$

For $i = 0$ and $n > 0$, we have

$$\frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} = \sum_{k \in \mathcal{S}} P_{0k}(t) q_{kn} = P_{0,n-1}(t) q_{n-1,n} + P_{0n}(t) q_{nn}.$$

Hence,

$$\frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} = \lambda P_{0,n-1}(t) - \lambda P_{0n}(t) \quad \text{for} \quad n > 0,$$
$$\text{and} \quad \frac{\mathrm{d}P_{00}(t)}{\mathrm{d}t} = -\lambda P_{00}(t).$$

We shall now consider solving these equations using a generating function approach.

Definition 6. *The generating function $P(z,t)$ for a process with transition probabilities $P_{0n}(t)$ for $n = 0, 1, 2, \ldots$, is given by*

$$P(z,t) = \sum_{n=0}^{\infty} P_{0n}(t) z^n.$$

By the triangle inequality, for $|z| \leq 1$, we have

$$\left| \sum_{n=0}^{\infty} P_{0n}(t) z^n \right| \leq \sum_{n=0}^{\infty} P_{0n}(t) \, |z^n| \leq \sum_{n=0}^{\infty} P_{0n}(t) = 1.$$

Therefore, the generating function $P(z,t)$ is well defined for $|z| \leq 1$.

If we multiply both sides of the equation

$$\frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} = \lambda P_{0,n-1}(t) - \lambda P_{0n}(t)$$

by z^n and sum from $n = 1$ to ∞, we have

$$\sum_{n=1}^{\infty} \frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} z^n = -\lambda \sum_{n=1}^{\infty} P_{0n}(t) z^n + \lambda \sum_{n=1}^{\infty} P_{0,n-1}(t) z^n. \tag{2}$$

We then add $\frac{\mathrm{d}P_{00}(t)}{\mathrm{d}t} z^0 = -\lambda P_{00}(t) z^0$ to Equation (2) to get

$$\sum_{n=0}^{\infty} \frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} z^n = -\lambda \sum_{n=0}^{\infty} P_{0n}(t) z^n + \lambda \sum_{n=1}^{\infty} P_{0,n-1}(t) z^n.$$

This gives us

$$\begin{aligned}
\frac{\mathrm{d}P(z,t)}{\mathrm{d}t} &= -\lambda P(z,t) + \lambda z \sum_{n=1}^{\infty} P_{0,n-1}(t) z^{n-1} \\
&= -\lambda P(z,t) + \lambda z \sum_{n=0}^{\infty} P_{0n}(t) z^n \\
&= -\lambda P(z,t) + \lambda z P(z,t) \\
&= -(\lambda - \lambda z) P(z,t).
\end{aligned}$$

The solution to this linear ordinary differential equation is

$$P(z,t) = c(z) e^{-(\lambda - \lambda z)t}. \tag{3}$$

To find $c(z)$, we need to know the value $P(z,t)$ at $t = 0$. That is,

$$P(z,0) = \sum_{n=0}^{\infty} P_{0n}(0) z^n = 1(z^0) + 0(z^1) + 0(z^2) + \cdots = 1. \tag{4}$$

Using (4) and substituting $t = 0$ into equation (3) yields

$$\begin{aligned}
1 &= c(z) e^{-(\lambda - \lambda z) \times 0} \\
\Rightarrow \quad c(z) &= 1 \\
\Rightarrow \quad P(z,t) &= e^{-(\lambda - \lambda z)t}.
\end{aligned}$$

Note that

$$P(z,t) = e^{-(\lambda - \lambda z)t} = e^{-\lambda t} e^{\lambda z t} = e^{-\lambda t} \sum_{j=0}^{\infty} \frac{(\lambda z t)^j}{j!} = \sum_{j=0}^{\infty} \frac{e^{-\lambda t} (\lambda t)^j}{j!} z^j. \tag{5}$$

Comparing (5) to the generating function,

$$P(z,t) = \sum_{j=0}^{\infty} P_{0j}(t) z^j, \quad \text{which is valid for } |z| \leq 1,$$

we must have that for each j,

$$P_{0j}(t) = \frac{e^{-\lambda t}(\lambda t)^j}{j!}.$$

Thus, the probability that the system is in state j by time t, given that it starts in state 0 at time 0, follows a Poisson distribution with parameter λt. In other words, the number of events that have happened by time t follows a Poisson distribution with parameter λt.

However, taking this just a little bit further, and considering a still relatively simple model, evaluating the transition function explicitly becomes difficult (and quickly becomes infeasible).

Example 3. The M/M/1 queue (continued)

The KFDEs for the M/M/1 queue are

$$\begin{aligned}
\frac{d\,P_{00}(t)}{dt} &= -\lambda P_{00}(t) + \mu P_{01}(t), \text{ and} \\
\frac{d\,P_{0j}(t)}{dt} &= \lambda P_{0,j-1}(t) - (\lambda+\mu)P_{0j}(t) + \mu P_{0,j+1}(t) \quad \text{for } j \geq 1.
\end{aligned}$$

The solution $P_{0j}(t)$ to these equations is given by

$$e^{-(\lambda+\mu)t}\left[\left(\frac{\lambda}{\mu}\right)^{-\frac{j}{2}} I_j\left(2\sqrt{\lambda\mu}t\right) + \left(\frac{\lambda}{\mu}\right)^{-\frac{(j+1)}{2}} I_{j-1}\left(2\sqrt{\lambda\mu}t\right) + \left(1-\frac{\lambda}{\mu}\right)\sum_{\ell=j+2}^{\infty}\left(\frac{\lambda}{\mu}\right)^{-\frac{\ell}{2}} I_\ell\left(2\sqrt{\lambda\mu}t\right)\right],$$

where $I_j(x) = \sum_{m=0}^{\infty} \frac{\left(\frac{x}{2}\right)^{j+2m}}{(j+m)!m!}$ is a Bessel Function of order j.

For most models it is too much to ask for us to calculate the $P_{i,j}(t)$ and so, **we need to look for simpler measures of our CTMC.**

Lecture 8: Evaluating moments, and generating functions

Concepts checklist

At the end of this lecture, you should be able to:

- *Perform calculations involving generating functions and KFDEs*, in particular with respect to simple CTMCs such as the Poisson process; and,
- *Evaluate the moments of simple CTMCs*, such as the Poisson process, via generating functions.

Let us consider evaluating moments of CTMCs, and return to the Poisson process.

Question 1: What is the expected number of events that happened by time t?

Let $N(t)$ be the random variable representing the number of events by time t. Then,

$$\begin{aligned}
\mathbb{E}[N(t)] &= \sum_{j=0}^{\infty} j \Pr(N(t) = j) \\
&= \sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \\
&= e^{-\lambda t} \lambda t \sum_{j=1}^{\infty} \frac{(\lambda t)^{(j-1)}}{(j-1)!} \\
&= e^{-\lambda t} \lambda t \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} \\
&= e^{-\lambda t} \lambda t e^{\lambda t} \\
&= \lambda t.
\end{aligned}$$

Question 2: What is the variance of the number of events that happened by time t?

$$\begin{aligned}
\mathrm{Var}(N(t)) &= \sum_{j=0}^{\infty} j^2 \Pr(N(t) = j) - (\mathbb{E}[N(t)])^2 \\
&= \sum_{j=0}^{\infty} j^2 \frac{e^{-\lambda t}(\lambda t)^j}{j!} - \left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \right)^2 \\
&= \sum_{j=1}^{\infty} j^2 \frac{e^{-\lambda t}(\lambda t)^j}{j!} - \left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \right)^2 \\
&= \sum_{j=1}^{\infty} j(j-1) \frac{e^{-\lambda t}(\lambda t)^j}{j!} + \sum_{j=1}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} - \left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \right)^2 \\
&= \sum_{j=2}^{\infty} j(j-1) \frac{e^{-\lambda t}(\lambda t)^j}{j!} + \sum_{j=1}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} - \left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \right)^2 \\
&= \sum_{j=2}^{\infty} \frac{e^{-\lambda t}(\lambda t)^j}{(j-2)!} + \sum_{j=1}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} - \left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda t}(\lambda t)^j}{j!} \right)^2 \\
&= e^{-\lambda t}(\lambda t)^2 \sum_{j=0}^{\infty} \frac{(\lambda t)^j}{j!} + \lambda t - (\lambda t)^2 \\
&= \lambda t,
\end{aligned}$$

so that the variance of the number of events by time t is identical to the mean λt.

Let us now consider evaluating the mean and variance of the Poisson process directly from the generating function.

First,

$$\begin{aligned}\mathbb{E}[N(t)] &= \sum_{j=0}^{\infty} jP_{0j}(t) \\ &= \sum_{j=0}^{\infty} \left(\left[\frac{\mathrm{d}}{\mathrm{d}z} z^j \right]_{z=1} P_{0j}(t) \right) \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}z} \left(\sum_{j=0}^{\infty} P_{0j}(t) z^j \right) \right]_{z=1} \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}z} P(z,t) \right]_{z=1}.\end{aligned}$$

Hence, we can calculate the expected number of events from the generating function $P(z,t)$ by differentiating the function with respect to z and setting $z = 1$.

Therefore, recalling $P(z,t) = e^{-(\lambda - \lambda z)t}$, the mean number of events is

$$\mathbb{E}[N(t)] = \left[\frac{\mathrm{d}}{\mathrm{d}z} e^{-(\lambda-\lambda z)t} \right]_{z=1} = \left[\lambda t e^{-(\lambda-\lambda z)t} \right]_{z=1} = \lambda t.$$

We can also calculate higher moments of $N(t)$ by taking higher order derivatives.

The variance of $N(t)$ is given by

$$\begin{aligned}\mathrm{Var}(N(t)) &= \sum_{j=0}^{\infty} j^2 P_{0j}(t) - \left(\sum_{j=0}^{\infty} jP_{0j}(t) \right)^2 \\ &= \sum_{j=0}^{\infty} \left(j(j-1)P_{0j}(t) + jP_{0j}(t) \right) - \left(\sum_{j=0}^{\infty} jP_{0j}(t) \right)^2.\end{aligned}$$

The second derivative of $P(z,t)$ with respect to z evaluated at $z = 1$ is

$$\begin{aligned}\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2} \sum_{j=0}^{\infty} P_{0j}(t) z^j \right]_{z=1} &= \left[\sum_{j=0}^{\infty} P_{0j}(t) \frac{\mathrm{d}^2}{\mathrm{d}z^2} z^j \right]_{z=1} \\ &= \left[\sum_{j=0}^{\infty} P_{0j}(t) j(j-1) z^{j-2} \right]_{z=1} \\ &= \sum_{j=0}^{\infty} P_{0j}(t) j(j-1).\end{aligned}$$

Thus,

$$\sum_{j=0}^{\infty} \left(j(j-1)P_{0j}(t) + jP_{0j}(t) \right) = \left[\frac{\mathrm{d}^2}{\mathrm{d}z^2} P(z,t) \right]_{z=1} + \underbrace{\sum_{j=0}^{\infty} jP_{0j}(t)}_{\text{mean} = \lambda t}.$$

Therefore,

$$\mathrm{Var}(N(t)) = \left[\frac{\mathrm{d}^2}{\mathrm{d}z^2}P(z,t)\right]_{z=1} + \lambda t - (\lambda t)^2.$$

Substituting $P(z,t)$ for the Poisson process into the previous expression, we see that

$$\begin{aligned}\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2}P(z,t)\right]_{z=1} + \lambda t - (\lambda t)^2 &= \left[\frac{\mathrm{d}^2}{\mathrm{d}z^2}e^{-(\lambda-\lambda z)t}\right]_{z=1} + \lambda t - (\lambda t)^2 \\ &= \left[\frac{\mathrm{d}}{\mathrm{d}z}\lambda t e^{-(\lambda-\lambda z)t}\right]_{z=1} + \lambda t - (\lambda t)^2 \\ &= \left[(\lambda t)^2 e^{-(\lambda-\lambda z)t}\right]_{z=1} + \lambda t - (\lambda t)^2,\end{aligned}$$

and hence $\mathrm{Var}(N(t)) = \lambda t = \mathbb{E}[N(t)]$.

So far in this lecture, we avoided the need to determine the transition function from the generating function to evaluate the (first two) moments of $N(t)$ and instead worked directly with the explicit generating function. Can we take this back one step further, so that we do not need to evaluate the generating function explicitly, and still evaluate these moments?

Recall that

$$P(z,t) = \sum_{n=0}^{\infty} P_{0n}(t) z^n,$$

and

$$\mathbb{E}[N(t)] = \left[\frac{\mathrm{d}}{\mathrm{d}z}P(z,t)\right]_{z=1}.$$

Hence,

$$\begin{aligned}\mathbb{E}[N(t)] &= \left[\frac{\mathrm{d}}{\mathrm{d}z}\sum_{n=0}^{\infty} P_{0n}(t) z^n\right]_{z=1} \\ &= \sum_{n=0}^{\infty} n P_{0n}(t).\end{aligned}$$

This is a very unsurprising equation! Now, let's differentiate with respect to time, t, to get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N(t)] &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{n=0}^{\infty} n P_{0n}(t) \\ &= \sum_{n=0}^{\infty} n \frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} \\ &= \sum_{n=0}^{\infty} \lambda P_{0n}(t) \qquad \text{(Substituting Poisson KFDEs)} \\ &= \lambda. \qquad \text{(Honest)}\end{aligned}$$

Hence, $\mathbb{E}[N(t)] = \lambda t$.

Let us consider the variance. $\mathrm{Var}(N(t)) = \mathbb{E}(N(t)^2) - \mathbb{E}(N(t))^2$. We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[N(t)^2] &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{n=0}^{\infty} n^2 P_{0n}(t) \\
&= \sum_{n=0}^{\infty} n^2 \frac{\mathrm{d}P_{0n}(t)}{\mathrm{d}t} \\
&= \sum_{n=0}^{\infty} ((n+1)^2 - n^2)\lambda P_{0n}(t) \qquad \text{(Substituting Poisson KFDEs)} \\
&= \lambda \sum_{n=0}^{\infty} (2n+1) P_{0n}(t) \\
&= \lambda(2\mathbb{E}(N(t)) + 1) \qquad \text{(Honest)} \\
&= 2\lambda^2 t + \lambda.
\end{aligned}$$

Hence, $\mathbb{E}(N(t)^2) = \lambda^2 t^2 + \lambda t$. Therefore $\mathrm{Var}(N(t)) = \lambda^2 t^2 + \lambda t - (\lambda t)^2 = \lambda t$.

Lecture 9: Equilibrium distributions and stationarity

Concepts checklist

At the end of this lecture, you should be able to:

- *Define* a *stationary distribution* and *equilibrium distribution* of a CTMC;
- Appreciate that for our purposes *these are equivalent*; and,
- *Solve the global balance equations*, both algebraically for simple CTMCs and with the assistance of a computer otherwise.

Project Description

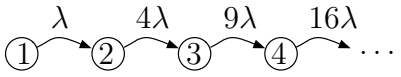
Discussion of the Project.

Moments

Whilst the moments approach seems promising, it also quickly becomes complicated.

Example 5: Quadratic birth process

Consider now a *quadratic (pure) birth process*. This has a state transition diagram



Here, we have (try to show this)

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X(t)] = \lambda\mathbb{E}[X(t)^2].$$

Hence, we have that the mean (first moment) is dependent upon the second moment. This higher-order moment dependence persists, meaning we can not get a closed system of differential equations to solve using this approach.

As stated, explicit solution of the KFDEs is rarely achievable, and moment equations also quickly become infeasible. We now consider another measure of a CTMC which is of interest in its own right, and is typically more tractable.

Stationary and equilibrium distributions

Let's consider a distribution (a probability mass function), $\pi = (\pi_i)_{i\in\mathcal{S}}$ such that if we started the CTMC according to π we have that the distribution of the process is π for all t.

Definition 7. *Assume we have a CTMC $(X(t), t \geq 0)$ with transition function $P(t)$. An $|S|$-dimensional vector $\pi = (\pi_i)_{i\in\mathcal{S}}$ with $\pi_i \geq 0$ for all i and $\sum_{i\in S}\pi_i = 1$ is called a stationary distribution if*

$$\pi = \pi P(t) \quad \textit{for all} \;\; t \geq 0.$$

This appears to be a difficult quantity to evaluate, as it requires knowledge of the transition function. However, let's consider this system of equations, in particular for the cases we are interested in where

$$P(t) = e^{Qt}.$$

We then have

$$\begin{aligned}\pi &= \pi P(t) \\ &= \pi e^{Qt} \\ &= \pi \sum_{n=0}^{\infty} Q^n \frac{t^n}{n!}.\end{aligned}$$

Now, subtract π from each side,

$$0 = \pi \sum_{n=1}^{\infty} Q^n \frac{t^n}{n!},$$

and since this must hold for all $t \geq 0$, it implies that

$$0 = \pi Q^n \quad \text{for all } n \geq 1.$$

Hence, $\pi Q = 0$.

So, we have shown that for CTMCs in which the matrix exponential solution to the KFDEs exists, we can determine the stationary distribution by solving $\pi Q = 0$, with the constraint $\sum_{i \in \mathcal{S}} \pi_i = 1$; much simpler than solving $\pi = \pi P(t)$! In fact,

$$\pi Q = 0, \quad \sum_{i \in \mathcal{S}} \pi_i = 1,$$

is simply a system of linear equations. We can attempt to solve these algebraically, but a variety of algorithms (i.e., numerical methods) also exist; for example, you can use the / (slash operator) in MATLAB.

In general, a solution to $\pi Q = 0, \sum_{i \in \mathcal{S}} \pi_i = 1$ is called an *equilibrium distribution*. As stated, for the CTMCs of interest to us in this course, this is equal to the stationary distribution.

Let us know take a closer look at what is happening to the dynamics under π. Let's differentiate both sides of

$$\pi = \pi P(t)$$

with respect to t, and then substitute the KFDEs. Considering the $(i,j)^{\text{th}}$ entry we have

$$0 = \sum_{i \in \mathcal{S}} \pi_i \sum_{k \in \mathcal{S}} P_{ik}(t) q_{kj} = \sum_{k \in \mathcal{S}} \sum_{i \in \mathcal{S}} \pi_i P_{ik}(t) q_{kj} = \sum_{k \in \mathcal{S}} \pi_k q_{kj} = \pi_j q_{jj} + \sum_{\substack{k \neq j \\ k \in \mathcal{S}}} \pi_k q_{kj}$$

$$\Rightarrow \pi_j \sum_{\substack{k \neq j \\ k \in \mathcal{S}}} q_{jk} = \sum_{\substack{k \neq j \\ k \in \mathcal{S}}} \pi_k q_{kj} \quad \text{(global balance equation)}.$$

We can interpret $\pi_j q_{jk}$ as the (probability) flux from state j to state k. Thus,

$$\sum_{\substack{k \neq j \\ k \in \mathcal{S}}} \pi_j q_{jk} = \sum_{\substack{k \neq j \\ k \in \mathcal{S}}} \pi_k q_{kj}$$

$$\text{Flux out of state } j = \text{Flux into state } j,$$

which we can expect to hold in equilibrium. Hence, the above equations are also referred to as the flux balance equations or the equilibrium equations for a CTMC.

Example 6. Linear pure-death process

The linear pure-death process has state space $\mathcal{S} = \{0, 1, 2, \ldots, N\}$ and for $j = 0, 1, \ldots, N$

$$\begin{aligned} q_{j,j-1} &= j\mu, \\ q_{jj} &= -j\mu. \end{aligned}$$

The equilibrium equations are

$$\pi_N N\mu = 0, \tag{6}$$
$$\pi_j j\mu = \pi_{j+1}(j+1)\mu, \qquad \text{for } j = 1, 2, \ldots, N-1 \tag{7}$$
$$0 = \pi_1 \mu. \tag{8}$$

Equation (7) gives us

$$\begin{aligned} \pi_j &= \frac{j+1}{j}\pi_{j+1} \qquad \text{for } j = 1, 2, \ldots, N-1, \\ &= \left(\frac{j+1}{j}\right)\left(\frac{j+2}{j+1}\right)\pi_{j+2} \\ &\vdots \\ &= \frac{N}{j}\pi_N. \end{aligned}$$

However, $\pi_N N\mu = 0$ implies that $\pi_N = 0$, and thus $\pi_j = 0$ for all $j = 1, 2, \ldots, N$. We have no information about π_0. However, we require $\sum_{i\in\mathcal{S}} \pi_i = 1$ and hence,

$$\sum_{k=0}^{N} \pi_k = 1 \quad \Rightarrow \quad \pi_0 = 1.$$

This corresponds to the intuitive result that if we have no individuals / working machines then we will always remain there. This also corresponds to what we'd intuitively expect to see in the long-term, starting from any number of individuals / working machines: all the components will die / break down as there are no births / repairs. Thus, the CTMC will eventually enter state 0 and remain in state 0 forever. We'll revisit this long term, or limiting behaviour.

Lecture 10: Equilibrium distributions continued

Example 4. Reliability (Birth and Death) (continued)

Here, the components can be repaired; we have the state space $\mathcal{S} = \{0, 1, 2, \ldots, N\}$ and

$$\begin{aligned} \text{for } \ j = 0, 1, \ldots, N-1: \quad & q_{j,j+1} = \lambda, \\ & q_{jj} = -(j\mu + \lambda), \\ \text{for } \ j = 1, \ldots, N: \quad & q_{j,j-1} = j\mu, \\ & q_{NN} = -N\mu. \end{aligned}$$

The equilibrium equations are

$$\pi_N N\mu = \pi_{N-1}\lambda, \tag{9}$$

$$\pi_j j\mu + \pi_j\lambda = \pi_{j-1}\lambda + \pi_{j+1}(j+1)\mu \quad \text{for } j = 1, \ldots, N-1, \tag{10}$$

$$\pi_0\lambda = \pi_1\mu. \tag{11}$$

By (10), we get

$$\pi_j j\mu - \pi_{j-1}\lambda = \pi_{j+1}(j+1)\mu - \pi_j\lambda,$$

which is of the form $A_j = A_{j+1}$, where $A_j = \pi_j j\mu - \pi_{j-1}\lambda$.

$$\Rightarrow A_j = A_1 \quad \text{for all } j = 1, \ldots, N.$$

By (9) and (11), $A_N = A_1 = 0$. Therefore,

$$\pi_j j\mu = \pi_{j-1}\lambda \quad \text{for all } j = 1, 2, \ldots, N,$$

which are known as detailed balance equations (these will be discussed more generally later). Hence,

$$\pi_j = \frac{\pi_{j-1}\lambda}{j\mu} = \frac{\pi_{j-2}\lambda^2}{j(j-1)\mu^2} = \pi_0 \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} \quad \text{for all } j = 1, 2, \ldots, N.$$

We need $\sum_{j=0}^{N} \pi_j = 1$ so that $\sum_{j=0}^{N} \pi_0 \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} = 1$ yields

$$\pi_0 = \left[\sum_{j=0}^{N} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}\right]^{-1},$$

$$\pi_j = \frac{\left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}}{\sum_{i=0}^{N} \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}} \quad \text{for all } j = 1, 2, \ldots, N.$$

Example 3. M/M/1 Queue (Single Server Queue) (cont.)

Recall that $\mathcal{S} = \{0, 1, 2, \dots\}$ and the transition rates are

$$\begin{aligned} q_{j,j+1} &= \lambda, \quad \text{for } j = 0, 1, 2, \dots \\ q_{j,j-1} &= \mu, \quad \text{for } j = 1, 2, \dots \\ q_{jj} &= -(\lambda + \mu), \quad \text{for } j = 1, 2, \dots \\ q_{00} &= -\lambda. \end{aligned}$$

The equilibrium equations are

$$\pi_0 \lambda = \pi_1 \mu \tag{12}$$

$$\pi_j(\lambda + \mu) = \pi_{j-1}\lambda + \pi_{j+1}\mu \quad \text{for } j = 1, 2, \dots. \tag{13}$$

We could solve these equations using the same method as for Example 4, but we can also use generating functions.

Let $P(z) := \sum_{j=0}^{\infty} \pi_j z^j$. If the equilibrium probabilities π_j exist, then $P(z)$ is analytic for $|z| \leq 1$, because

$$\begin{aligned} \left| \sum_{j=0}^{\infty} \pi_j z^j \right| &\leq \sum_{j=0}^{\infty} \pi_j \left| z^j \right| \quad \text{by the triangle inequality} \\ &\leq \sum_{j=0}^{\infty} \pi_j \quad \text{as } |z| \leq 1 \\ &= 1. \end{aligned}$$

We multiply (13) by z^j and sum from $j = 1$ to ∞ and then add (12) to get

$$\begin{aligned} &\sum_{j=0}^{\infty} \pi_j \lambda z^j + \sum_{i=1}^{\infty} \pi_i \mu z^i = \sum_{j=1}^{\infty} \pi_{j-1} \lambda z^j + \sum_{i=0}^{\infty} \pi_{i+1} \mu z^i \\ \Rightarrow\ &\lambda P(z) + \mu(P(z) - \pi_0) = \lambda z P(z) + \frac{\mu}{z}(P(z) - \pi_0) \\ \Rightarrow\ &P(z)\left(\lambda + \mu - \lambda z - \frac{\mu}{z}\right) = \pi_0 \left(\mu - \frac{\mu}{z}\right). \end{aligned}$$

Thus,

$$\Rightarrow \quad P(z) = \frac{\pi_0 \left(\mu - \frac{\mu}{z}\right)}{\lambda + \mu - \lambda z - \frac{\mu}{z}} = \frac{\pi_0(\mu z - \mu)}{(\lambda + \mu)z - \lambda z^2 - \mu} = \frac{\pi_0(\mu z - \mu)}{\left(1 - \frac{\lambda z}{\mu}\right)(\mu z - \mu)} = \frac{\pi_0}{1 - \frac{\lambda z}{\mu}}.$$

As $P(z)$ converges if and only if the π_j exist, we need $\left.\dfrac{\pi_0}{1 - \frac{\lambda z}{\mu}}\right|_{z=1}$ to converge.

Note that $\displaystyle\sum_{j=0}^{\infty} x^j = \frac{1}{1-x}$ if and only if $|x| < 1$

$\Rightarrow$ the equilibrium probabilities π_j exist if and only if $\left|\dfrac{\lambda z}{\mu}\right| < 1$ when $z = 1$

$\Rightarrow \pi_j$ exist if and only if $\lambda/\mu < 1$, which is a natural stability condition, where the service rate is higher than the arrival rate.

In this case,

$$1 = P(z)\Big|_{z=1} = \frac{\pi_0}{1 - \frac{\lambda}{\mu}} \quad \Rightarrow \pi_0 = 1 - \frac{\lambda}{\mu},$$

and consequently

$$P(z) = \frac{1 - \frac{\lambda}{\mu}}{1 - \frac{\lambda z}{\mu}} = \left(1 - \frac{\lambda}{\mu}\right) \sum_{j=0}^{\infty} \left(\frac{\lambda z}{\mu}\right)^j = \sum_{j=0}^{\infty} \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^j z^j \overset{\text{by def}}{=} \sum_{j=0}^{\infty} \pi_j z^j.$$

So, we have for all $j \in \mathcal{S}$

$$\pi_j = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^j \text{ if } \lambda < \mu,$$

and the π_j do not exist, otherwise.

Physically, we have a solution to the equilibrium equations if and only if $\lambda < \mu$.

Lecture 11: Characterisation of States and Limiting Distributions

Concepts checklist

At the end of this lecture, you should be able to:

- Characterise states of a CTMC in terms of *communicating classes*, *irreducibility*, and *recurrence / transience*;
- Understand the relationship between these characteristics and equilibrium probabilities; and,
- State a theorem regarding the existence and uniqueness of a limiting distribution for irreducible, finite-state CTMCs.

Summary of three examples of equilibrium distributions

- In Example 6. Reliability (Pure Death), we have N equilibrium probabilities equal to 0, and one equilibrium probability equal to 1.
- In Example 4. Reliability (Birth and Death), we have $N+1$ equilibrium probabilities, all of which are greater than 0.

  — Both reliability models are finite state space continuous-time Markov chain.
- In Example 3. Single-server queue, we either have
  - all positive equilibrium probabilities if $\lambda < \mu$, or
  - no solution to the equilibrium equations which sums to 1, otherwise.

  — The single server queue is an infinite state space continuous-time Markov chain.

Question: What characteristics of a CTMC lead to these different types of behaviour?

Characterisation of States

Definition 8. *For $i, j \in \mathcal{S}$, state j is said to be* **accessible** *from state i if there is some path of transitions via which the Markov chain can move from state i to state j. In other words, there exists a sequence of states $\{i = i_0, i_1, i_2, \ldots, i_n = j\}$ such that*

$$q_{i_0,i_1} q_{i_1,i_2} \ldots q_{i_{n-1},i_n} > 0.$$

Definition 9. *States i and j are said to* **communicate** *(and we write $i \leftrightsquigarrow j$) if*

1. *$j = i$, or*
2. *j is accessible from i and i is accessible from j.*

Proposition 1. *The relation $\leftrightsquigarrow$, i.e., communication, is an equivalence relation.*

Proof. We need to show that $\leftrightsquigarrow$ is *Reflexive*, *Symmetric* and *Transitive*.

(*i*) *Reflexive* means that $i \leftrightsquigarrow i$. This follows directly from the definition.

(*ii*) *Symmetric* means that if $i \leftrightsquigarrow j$ then $j \leftrightsquigarrow i$. This also follows directly from the definition.

(*iii*) *Transitive* means that if $i \leftrightsquigarrow k$ and $k \leftrightsquigarrow j$ then we have $i \leftrightsquigarrow j$.
This follows because if k is accessible from i and j is accessible from k, then there exists a path from i to j (via k). This implies that j is accessible from i. Similarly, if k is accessible from j and i from k, then i is accessible from j. Hence, $i \leftrightsquigarrow j$.

□

Corollary 1. *The state space $\mathcal{S}$ of a continuous-time Markov chain can be partitioned into* **communicating classes** *$\mathcal{S}_1, \mathcal{S}_2, \ldots$ such that $i, j \in \mathcal{S}_k$ if and only if $i \leftrightsquigarrow j$.*

Example 3. M/M/1 Queue

Here, all states are accessible from every other state. Thus, there is a single communicating class $\mathcal{S} = \{0, 1, 2, \ldots\}$.

Example 6. Linear pure-death process

Recall, in this example we have N individuals, each subject dying after an exponentially distributed amount of time with rate μ. Here, state $n-1$ is accessible from state n, but state n is not accessible from $n-1$. Therefore, states n and $n-1$ do not communicate. Furthermore, each state is in its own communicating class,

$$\Rightarrow \mathcal{S} = \bigcup_{i=0}^{N} \mathcal{S}_i, \text{ where } \mathcal{S}_i = \{i\}.$$

Example 4. N-machine reliability (Birth and Death)

Each state is accessible from every other state – failure, and repair! Thus, there is a single communicating class, which is the whole state space $\mathcal{S} = \{0, 1, \ldots, N\}$.

Let's introduce some further terminology, to label common communicating class structures, and also some properties possessed by states in communicating classes.

Definition 10. *A continuous-time Markov chain is said to be* **irreducible** *if it has a single communicating class, and to be* **reducible** *otherwise.*

Definition 11. *A state is said to be* **recurrent** *if the probability that the continuous-time Markov chain returns to that state after it has left is 1. The state is* **transient** *otherwise.*

Definition 12. *A state that is recurrent is said to be* **positive recurrent** *if the mean return time is finite (or it is an absorbing state). Otherwise, it is called* **null recurrent**.

Within a communicating class, states are either all recurrent or all transient. Recurrence or transience is a property of communicating classes; hence in the irreducible case, recurrence or transience is a property of the CTMC itself.

The classification of states and hence communicating classes depends on the probability that a continuous-time Markov chain returns to a state after it has left it.

Theorem 6. *Consider a communicating class $\mathcal{C}$ and let $i \in \mathcal{C}$. If there exists $j \notin \mathcal{C}$ such that j is accessible from i, then $\mathcal{C}$ is transient.*

Proof. Note that state i cannot be accessible from j, because then j would be in $\mathcal{C}$. Therefore, the probability of returning to i having left it, must be less than 1. □

Theorem 7. *If $\mathcal{C}$ is finite and if for every $i \in \mathcal{C}$ there exists no $j \notin \mathcal{C}$ that is accessible from state i, then $\mathcal{C}$ is recurrent.*

Note, Theorem 7 does not extend to infinite communicating classes. For example, the single server queue with $\lambda > \mu$ is transient, and yet it has a single communicating class.

Now, returning to linking this characterisation of states and equilibrium distributions. We have

Theorem 8. *If j is in a transient communicating class $\mathcal{C}$, then there exists no solution $(\pi_i)_{i\in\mathcal{S}}$ with $\sum_i \pi_i = 1$ and $\pi_j > 0$.*

This theorem shows that equilibrium probabilities for states in the transient communicating classes are equal to zero. This can arise in one of two ways:

1. The solution to the equilibrium equations for π_j is zero, as in Example 6 (Pure Death), for all states $j > 0$.
2. There exists a positive solution $(\pi_i)_{i\in\mathcal{S}}$ to the equilibrium equations, but it is impossible to normalise it such that $\sum_{i\in\mathcal{S}} \pi_i = 1$, as in the single-server queue (Example 3) with $\lambda > \mu$.

Theorem 9. *If j is in a recurrent communicating class $\mathcal{C}$, then there exists two possibilities:*

1. *$\mathcal{C}$ is positive-recurrent: There exists a solution $(\pi_i)_{i\in\mathcal{S}}$ with $\sum_{i\in\mathcal{S}} \pi_i = 1$ to the equilibrium equations, in which $\pi_j > 0$.*
2. *$\mathcal{C}$ is null-recurrent: There exists no solution $(\pi_i)_{i\in\mathcal{S}}$ with $\sum_{i\in\mathcal{S}} \pi_i = 1$ to the equilibrium equations, in which $\pi_j > 0$.*

Examples 6, 4, 3.

1. The communicating class $\mathcal{S} = \{0\}$ in Example 6 is positive-recurrent, since $\pi_0 = 1$ and $\pi_i = 0$ otherwise.
2. The communicating class $\mathcal{S} = \{0, 1, \ldots, N\}$ in Example 4 is positive-recurrent, since $\pi_i > 0$ for all $i \in \{0, 1, \ldots, N\}$.

3. The communicating class $\mathcal{S} = \{0, 1, 2, \dots\}$ in the single-server queue (Example 3)
    - with $\lambda < \mu$ is positive-recurrent,
    - with $\lambda = \mu$ is null-recurrent (will justify this later).

Finally, a theorem regarding the long-term behaviour of a certain class of CTMC.

Theorem 10. *For an irreducible finite-state CTMC $(X(t), t \geq 0)$ with state space S, there exists a unique limiting probability vector, $\pi = (\pi_i)_{i \in S}$, i.e., there exists a unique probability vector π such that*

$$\lim_{t \to \infty} P_{ij}(t) = \pi_j, \ \ \forall i, j \in S.$$

Moreover, that limiting probability vector π is the unique stationary (and equilibrium) probability vector, i.e., if

$$\Pr(X(0) = j) = \pi_j, \ \ \forall j \in S,$$

then

$$\Pr(X(t) = j) = \pi_j, \ \ \forall j \in S \ \text{ and } \ t > 0.$$

Lecture 12: Hitting Probabilities – Our first performance measure

Concepts checklist

At the end of this lecture, you should be able to:

- *derive a system of linear equations for the hitting probability of a particular state* for simple CTMCs; and,
- *solve homogeneous second-order difference equations with constant coefficients*, in order to solve simple hitting probabilities equations.

Hitting Probabilities

So far, we've seen how to calculate

1. time-dependent probabilities for some simple CMTCs, and
2. equilibrium probabilities for some more complex CTMCs.

Now, we would like to calculate the probability that a CTMC ever reaches a given state. This is useful in answering questions we might have about a particular process, and can be used as a *performance measure*, but it is also useful for determining whether a state is recurrent or not.

Example 7. Finite-capacity single-server queue.

A single-server queue with finite capacity N has arrival rate λ and service rate μ. Without loss of generality we assume that N is even and that the system starts half full.

Question: What is the probability that the system empties before it fills up?

We use a slight *modification* of the earlier single-server queue continuous-time Markov chain. We put $\mathcal{S} = \{0, 1, 2, \ldots, N\}$ and the non-zero rates are

$$\begin{aligned} q_{n,n+1} &= \lambda, \quad \text{for } n = 1, 2, \ldots, N-1, \\ q_{n,n-1} &= \mu, \quad \text{for } n = 1, 2, \ldots, N-1, \\ q_{nn} &= -(\lambda + \mu), \quad \text{for } n = 1, 2, \ldots, N-1. \end{aligned}$$

Note: States 0 and N have no transition rate out and hence are now absorbing states.

This new Markov chain is not irreducible, as it has (a) two recurrent communicating classes $\{0\}$ and $\{N\}$ and, (b) one transient communicating class $\{1, 2, \ldots, N-1\}$.

Since the CTMC will be absorbed into one of the states 0 or N,

$$\Pr(\text{the chain visits state 0 before state } N) = \Pr(\text{it ever visits state 0}).$$

We want to know, for states $1, 2, \ldots, N-1$, the probability that the CTMC is absorbed in state 0 rather than state N.

Let f_i be the probability that the chain is absorbed in state 0 given the initial state i:

$$f_i = \Pr(\text{absorbed in state } 0 \mid \text{starts in state } i).$$

Assuming that $i \neq \{0, N\}$ we use a *first step analysis*:

$$\begin{aligned}
f_i &= \Pr(\text{absorbed in state } 0 \mid \text{starts in state } i) \\
&= \Pr(\text{goes to } i+1 \text{ in the next step, gets absorbed in state } 0 \mid \text{ starts in state } i) \\
&\quad + \Pr(\text{goes to } i-1 \text{ in the next step, absorbed in state } 0 \mid \text{ starts in state } i) \\
&= \Pr(X(t_1) = i+1 \cap \text{absorbed in state } 0 \mid X(0) = i) \\
&\quad + \Pr(X(t_1) = i-1 \cap \text{absorbed in state } 0 \mid X(0) = i)
\end{aligned}$$

where t_1 is the first time the Markov chain makes a transition out of i,

$$\begin{aligned}
&= \Pr(\text{absorbed in state } 0 \mid X(0) = i, X(t_1) = i+1) \Pr(X(t_1) = i+1 \mid X(0) = i) \\
&\quad + \Pr(\text{absorbed in state } 0 \mid X(0) = i, X(t_1) = i-1) \Pr(X(t_1) = i-1 \mid X(0) = i) \\
&= \Pr(\text{absorbed in state } 0 \mid X(t_1) = i+1) \frac{\lambda}{\lambda+\mu} + \Pr(\text{absorbed in state } 0 \mid X(t_1) = i-1) \frac{\mu}{\lambda+\mu}
\end{aligned}$$

by the Markov property, and by the facts that $\lambda/(\lambda+\mu)$ is the probability of a transition from i to $i+1$, and $\mu/(\lambda+\mu)$ is the probability of a transition from i to $i-1$,

$$= \Pr(\text{absorbed in state } 0 \mid X(0) = i+1) \frac{\lambda}{\lambda+\mu} + \Pr(\text{absorbed in state } 0 \mid X(0) = i-1) \frac{\mu}{\lambda+\mu}$$

by the time-homogeneity property,

$$= \frac{\lambda}{\lambda+\mu} f_{i+1} + \frac{\mu}{\lambda+\mu} f_{i-1}.$$

In summary, we have

$$f_i = \frac{\lambda}{\lambda+\mu} f_{i+1} + \frac{\mu}{\lambda+\mu} f_{i-1} \quad \text{for } i = 1, \ldots, N-1, \tag{14}$$

with the following boundary conditions $f_0 = 1$ and $f_N = 0$.

Equations (14) are simply a system of linear equations. Try to solve these equations using a standard approach. See if you can write it in a matrix-vector form.

Equations (14) can also be viewed as a homogeneous second-order difference equation with constant coefficients[2]. One way to solve these is to try a solution of the form $f_i = m^i$, which

[2]The *homogeneous* part refers to the property that we can rearrange (14) as

$$\frac{\lambda}{\lambda+\mu} f_{i+1} - f_i + \frac{\mu}{\lambda+\mu} f_{i-1} = 0 \quad \text{for } i = 1, \ldots, N-1,$$

with 0 on the right-hand side, as opposed to an *inhomogeneous* equation

$$\frac{\lambda}{\lambda+\mu} f_{i+1} - f_i + \frac{\mu}{\lambda+\mu} f_{i-1} = c \quad \text{for } i = 1, \ldots, N-1,$$

for some constant $c \neq 0$.

we substitute into (14) to get

$$m^i = \frac{\lambda}{\lambda+\mu}m^{i+1} + \frac{\mu}{\lambda+\mu}m^{i-1}.$$

Thus,

$$\lambda m^2 - (\lambda+\mu)m + \mu = 0 \quad \Leftrightarrow \quad (\lambda m - \mu)(m-1) = 0 \quad \Leftrightarrow \quad m = \frac{\mu}{\lambda}, 1.$$

• If $\mu \neq \lambda$, then the general solution is of the form $f_i = A\left(\frac{\mu}{\lambda}\right)^i + B(1)^i = A\left(\frac{\mu}{\lambda}\right)^i + B$.

Now, we use the boundary conditions see that

$$\begin{aligned} f_0 = 1 \quad &\Rightarrow A + B = 1, \quad \text{so that } B = 1 - A, \\ f_N = 0 \quad &\Rightarrow A\left(\frac{\mu}{\lambda}\right)^N + 1 - A = 0 \quad \Rightarrow A = \frac{1}{1-\left(\frac{\mu}{\lambda}\right)^N} \\ &\Rightarrow f_i = \frac{\left(\frac{\mu}{\lambda}\right)^i - \left(\frac{\mu}{\lambda}\right)^N}{1-\left(\frac{\mu}{\lambda}\right)^N} \\ &\Rightarrow f_{N/2} = \frac{\left(\frac{\mu}{\lambda}\right)^{N/2} - \left(\frac{\mu}{\lambda}\right)^N}{1-\left(\frac{\mu}{\lambda}\right)^N}. \end{aligned}$$

• If $\mu = \lambda$, we have repeated roots of $m = 1$, which implies that the general solution is of the form $f_i = Ai(1) + B(1) = Ai + B$.

Using the boundary conditions:

$$\begin{aligned} f_0 = 1 &\Rightarrow B = 1 \\ f_N = 0 &\Rightarrow AN + 1 = 0 \\ &\Rightarrow \quad A = -\frac{1}{N} \\ &\Rightarrow \quad f_i = 1 - \frac{i}{N}. \end{aligned}$$

In this case, $f_{N/2} = \frac{N/2}{N} = \frac{1}{2}$, which fits with intuition.

Lecture 13: Hitting Probabilities continued

Concepts checklist

At the end of this lecture, you should be able to:

- *Understand that we are seeking the minimal non-negative solution* to the hitting probability equations;
- *Find the minimal non-negative solution of the hitting probability equations* for simple CTMCS; and,
- *State a Theorem regarding the solution of the hitting probability of a particular state for a general CTMC.*

Example 3. M/M/1 Queue

We extend our analysis to the case where the queue has infinite capacity, by removing the upper absorbing state N and setting $\mathcal{S} = \{0, 1, \ldots\}$. Again, let f_i be the probability that the Markov chain ever visits state 0, given that it starts in state i:

$$f_i = \Pr\big(\text{ever visits state } 0 \;\big|\; \text{starts in state } i\big).$$

Then, as for the finite capacity case we have

$$f_i = \frac{\lambda}{\lambda+\mu} f_{i+1} + \frac{\mu}{\lambda+\mu} f_{i-1}, \qquad \text{with } f_0 = 1. \tag{15}$$

Unlike the finite case, we only have one boundary condition, so using just the previous technique we cannot secure a unique solution. Solving equation (15) as in the previous example, we have

$$f_i = \begin{cases} A\left(\frac{\mu}{\lambda}\right)^i + B & \text{for } \mu \neq \lambda, \\ Ai + B & \text{for } \mu = \lambda. \end{cases}$$

As $f_0 = 1$, we have

$$B = \begin{cases} 1 - A & \text{for } \mu \neq \lambda, \\ 1 & \text{for } \mu = \lambda. \end{cases}$$

Thus,

$$f_i = \begin{cases} 1 + A\left[\left(\frac{\mu}{\lambda}\right)^i - 1\right] & \text{for} \quad \mu \neq \lambda, \\ Ai + 1 & \text{for} \quad \mu = \lambda. \end{cases}$$

We shall see later that f_i is the minimal non-negative solution to

$$x_i = \frac{\lambda}{\lambda+\mu} x_{i+1} + \frac{\mu}{\lambda+\mu} x_{i-1}, \qquad \text{subject to } x_0 = 1.$$

- If $\lambda < \mu$: then $(\mu/\lambda)^i - 1 > 0$ for all $i > 0$.
  $\Rightarrow$ the minimal non-negative solution occurs when $A = 0$ and therefore $f_i = 1$ for all i.
- If $\lambda > \mu$: then $(\mu/\lambda)^i - 1 < 0$ for all $i > 0$; as $i \to \infty$, this term approaches -1.
  $\Rightarrow$ the largest value of A for which $1 + A\left[\left(\frac{\mu}{\lambda}\right)^i - 1\right] \geq 0$ for all i is $A = 1$.
  $\Rightarrow f_i = \left(\frac{\mu}{\lambda}\right)^i$ for all i.
- If $\mu = \lambda$: the minimal non-negative solution occurs when $A = 0$ and thus $f_i = 1$ for all i.

$$\text{Hence,} \qquad f_i = \begin{cases} 1, & \text{if} \quad \mu > \lambda, \\ \left(\dfrac{\mu}{\lambda}\right)^i, & \text{if} \quad \mu < \lambda, \\ 1, & \text{if} \quad \mu = \lambda. \end{cases}$$

We can now state that the *(unmodified) single server queue* is recurrent if $\mu \geq \lambda$ and transient otherwise. Let us assume that the Markov chain starts in state 0, then the only state it can go to is state 1.

- If $\lambda \leq \mu$, we return to state 0 with probability 1. Thus by definition, 0 is a recurrent state, and therefore since it is irreducible, the whole Markov chain is recurrent.
- If $\lambda > \mu$, we return to state 0 from state 1 with probability $\frac{\mu}{\lambda} < 1$. Therefore, by definition, state 0 is a transient state and the whole of the Markov chain is transient.

For $\lambda > \mu$ or $\lambda < \mu$, we could argue that this reflects the intuitive fact that the number in the queue will drift "towards ∞" or drift "towards zero". However this doesn't tell us much about the case when $\lambda = \mu$, but the analysis has shown us that the queue is still recurrent in this case.

Minimal non-negative solution:

We now show why the hitting probability f_i is the minimal non-negative solution to the equation

$$x_i = \frac{\lambda}{\lambda + \mu} x_{i+1} + \frac{\mu}{\lambda + \mu} x_{i-1}, \qquad \text{subject to } x_0 = 1.$$

Let $f_{in} = \Pr(\text{reaches state 0 in at most } n \text{ steps} \mid \text{starts in state } i)$. Then

$$f_{i,n+1} = \left(\frac{\lambda}{\lambda + \mu}\right) f_{i+1,n} + \left(\frac{\mu}{\lambda + \mu}\right) f_{i-1,n}, \tag{16}$$

with $f_{0n} = 1$ for all $n \geq 0$ and $f_{i0} = 0$ for all $i \in \mathcal{S} \setminus \{0\}$.

Lemma 1. *Let x_i be* any *non-negative solution to the equation*

$$x_i = \frac{\lambda}{\lambda + \mu} x_{i+1} + \frac{\mu}{\lambda + \mu} x_{i-1}, \qquad \textit{subject to } x_0 = 1. \tag{17}$$

Then, the probability f_{in} of hitting state 0 in n steps or fewer (given the initial state i) satisfies the inequality

$$f_{in} \leq x_i$$

for all $n \geq 0$ and $i \in \mathcal{S}$.

Proof: Clearly, this is true for $n = 0$, so let's assume that it is true for $n = k$. By (16), we have

$$\begin{aligned} f_{i,k+1} &= \left(\frac{\lambda}{\lambda+\mu}\right) f_{i+1,k} + \left(\frac{\mu}{\lambda+\mu}\right) f_{i-1,k} \\ &\leq \left(\frac{\lambda}{\lambda+\mu}\right) x_{i+1} + \left(\frac{\mu}{\lambda+\mu}\right) x_{i-1} = x_i, \end{aligned}$$

so that $f_{i,k+1} \leq x_i$ and the result is proven by induction. □

Lemma 2. *(For Example 3) The probability f_i of hitting state 0, given that the system starts in state i, is the minimal non-negative solution to the equation*

$$x_i = \frac{\lambda}{\lambda+\mu} x_{i+1} + \frac{\mu}{\lambda+\mu} x_{i-1}, \qquad \textit{subject to } x_0 = 1.$$

Proof. Clearly, f_{in} is increasing in n, since we are allowing more and more steps to reach state 0. Therefore, f_{in} is an increasing sequence, which is bounded above and so

$$f_i = \lim_{n\to\infty} f_{in} \qquad \text{exists.}$$

Also since $f_{in} \leq x_i$ for all n we have that

$$\lim_{n\to\infty} f_{i,n} \leq x_i,$$

which implies that $f_i \leq x_i$. We have already seen that f_i is a solution to equation (17) and therefore, it must be the the minimal non-negative solution to equation (17). □

We can state this result for a general CTMC.

Theorem 11. *For a particular state j in a CTMC with generator Q, the probability f_i that the CTMC ever reaches j, given that it starts in state i is given by the minimal non-negative solution to the equations*

$$(-q_{ii})x_i = \sum_{k\in\mathcal{S}, k\neq i} q_{ik}x_k, \;\; i \in S \setminus \{j\},$$

subject to the boundary condition $x_j = 1$.

Lecture 14: Expected Hitting Times – A very useful performance measure

Concepts checklist

At the end of this lecture, you should be able to:

- *derive a system of linear equations* that the expected first hitting times of a state satisfy;
- *state a theorem* regarding the desired solution to this system of equations; and,
- *evaluate* expected first hitting times for simple CTMCs, both analytically and with the assistance of a computer.

Expected Hitting Time

We have shown how to calculate the probability f_i that a continuous-time Markov chain ever reaches state j given that it starts in state i. For the case $f_i = 1$ for all $i \in \mathcal{S}$, it can be of interest to calculate the expected time that this will take.

Let T the time to first reach state j, and $t_i = \mathbb{E}(T|X(0) = i)$ be the expected time (t_i might be infinite) until the process is absorbed in state j given that it starts in state i. Then assuming $i \neq j$, using a first step analysis we can write

$$t_i = -\frac{1}{q_{ii}} + \sum_{\substack{k \neq i \\ k \in \mathcal{S}}} \frac{q_{ik}}{-q_{ii}} t_k, \qquad \text{with } t_j = 0,$$

where

$$\begin{aligned} -\frac{1}{q_{ii}} &= \text{expected time until the next transition,} \\ \frac{q_{ik}}{-q_{ii}} &= \text{probability of jumping to state } k \text{ at the next transition,} \\ t_k = \mathbb{E}(T|X(0) = k) &= \text{expected time to reach } j \text{ given that the process starts in state } k. \end{aligned}$$

Theorem 12. *The expected time to first reach state j starting from state i, t_i, is given by the minimal non-negative solution to the equations*

$$\sum_{k \in \mathcal{S}} q_{ik} t_k = -1 \,, \qquad i \in \mathcal{S} \setminus \{j\},$$

subject to $t_j = 0$. If no non-negative solution exists, then t_i is infinite for all i.

The proof of this result follows by using similar but more complicated methods to those used in the proof of the hitting probability result.

Note, defining Q_{-j} as the generator Q with the jth row and column removed, we have $Q_{-j}t = -\mathbf{1}$ where $t = (t_i)_{i \in \mathcal{S} \setminus \{j\}}$ and $\mathbf{1}$ is a vector of ones.

Example 3. M/M/1 Queue

When $\mu \geq \lambda$, the probability that the single server queue ever visits state 0 given that it starts in state $i > 0$ is equal to 1. Let us consider the expected time t_i until the Markov chain reaches state 0 given that it starts in state i.

$$t_i = \frac{1}{\lambda+\mu} + \left(\frac{\lambda}{\lambda+\mu}\right) t_{i+1} + \left(\frac{\mu}{\lambda+\mu}\right) t_{i-1}, \text{ for } i > 0, \text{ with } t_0 = 0.$$

Step 1. The homogeneous version of this equation has a solution of the form

$$t_i = \begin{cases} A\left(\frac{\mu}{\lambda}\right)^i + B & \text{for } \mu > \lambda, \\ Ai + B & \text{for } \mu = \lambda. \end{cases}$$

In the non-homogeneous case we try a particular solution of the form

$$t_i = \begin{cases} Ci & \text{for } \mu > \lambda, \\ Ci^2 & \text{for } \mu = \lambda. \end{cases}$$

- If $\mu > \lambda$:

$$\begin{aligned} iC &= \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu}(i+1)C + \frac{\mu}{\lambda+\mu}(i-1)C \\ \Rightarrow \quad 0 &= \frac{1}{\lambda+\mu} + \frac{C\lambda}{\lambda+\mu} - \frac{C\mu}{\lambda+\mu} \\ \Rightarrow \quad (\mu-\lambda)C &= 1 \\ \Rightarrow \quad C &= \frac{1}{\mu-\lambda}. \end{aligned}$$

Note: A solution in the non-homogeneous case = a general solution to the homogeneous case + a particular solution to the non-homogeneous case.

Therefore, the general solution of the non-homogeneous equation is

$$t_i = \frac{i}{\mu-\lambda} + A\left(\frac{\mu}{\lambda}\right)^i + B.$$

Note: We do not use the boundary conditions until we have the entire form of solution, and then after that we choose the minimal non-negative solution, if necessary.

Step 2. Using the boundary condition $t_0 = 0$ gives us that $B = -A$; thus,

$$t_i = \frac{i}{\mu-\lambda} + A\left[\left(\frac{\mu}{\lambda}\right)^i - 1\right].$$

Step 3. We now need to find the minimal non-negative solution. Since $\mu > \lambda$, the term in square brackets is always positive and grows much quicker in i than does the first term.

$\Rightarrow$ the minimal non-negative solution occurs when $A = 0$, as we cannot guarantee $t_i > 0$ for any $A < 0$. Hence,

$$t_i = \frac{i}{\mu-\lambda},$$

which tells us that the time until absorption in state 0 is linear in the initial number i of customers present.

- If $\mu = \lambda$: We try a solution of the form $T_i = i^2 C$.

$$i^2 C = \frac{1}{2\mu} + \left(\frac{1}{2}\right)(i+1)^2 C + \left(\frac{1}{2}\right)(i-1)^2 C = \frac{1}{2\mu} + C\left(i^2+1\right)$$
$$\Rightarrow C = -\frac{1}{2\mu}.$$

Therefore the general solution is $t_i = Ai + B - \frac{1}{2\mu} i^2$.

Step 2. Now we use the boundary condition, $t_0 = 0$ to show that $B = 0$ and thus

$$t_i = Ai - \frac{1}{2\mu} i^2.$$

There exists no A such that this is always non-negative (since i^2 grows faster than i) and hence there is no non-negative solution and $t_i = \infty$ for all i.

$\Rightarrow$ When $\mu = \lambda$ the probability of reaching state 0 is 1, but the expected time that this takes is infinity; hence the CTMC is null recurrent.

Lecture 15: Stationary and Reversed-Time Processes

Concepts checklist

At the end of this lecture, you should be able to:

- *define* a stationary CTMC;
- *define* a reversed-time CTMC;
- *state and prove* a theorem regarding the transition rates and equilibrium distribution of the reversed-time process corresponding to a stationary CTMC; and,
- *define* a reversible CTMC.

We will show that some stochastic processes are such that when we consider them in reversed time, the behaviour of the process remains indistinguishable from the forward time behaviour.

Stationary Processes

Definition 13. *A continuous-time Markov chain* $\{X(t)\}$ *is stationary if*

$$\begin{aligned}&\Pr(X(t_1) = i_1, X(t_2) = i_2, \ldots, X(t_n) = i_n)\\ &= \Pr(X(t_1 + \delta) = i_1, X(t_2 + \delta) = i_2, \ldots, X(t_n + \delta) = i_n)\end{aligned}$$

for every positive integer n *and for all* $\delta, t_1, t_2, \ldots, t_n$.

In other words, the joint distribution of $\{X(t_1), X(t_2), \ldots, X(t_n)\}$ is the same as that of $\{X(t_1 + \delta), X(t_2 + \delta), \ldots, X(t_n + \delta)\}$. In particular, when $n = 1$ the distribution of $X(t)$ is the distribution of $X(t + \delta)$.

Roughly speaking, a CTMC (that has an equilibrium distribution) operates in a stationary manner after it has been running for a long time. If we start a Markov chain with its equilibrium distribution, that is, let $\Pr(X(0) = j) = \pi_j$ for all $j \in \mathcal{S}$, then it will have the property

$$\Pr(X(t) = j) = \pi_j \text{ for all } j \in \mathcal{S} \text{ and } t \geq 0.$$

Reversed-time Processes

Definition 14. *For each stationary Markov chain* $\{X(t)\}$*, we define the reversed-time process* $\{X^R(t)\}$ *to be*

$$X^R(t) = X(\tau - t), \quad \text{for arbitrary } \tau.$$

Theorem 13. *If* $X(t)$ *is a stationary Markov chain with transition rates* q_{jk} *and equilibrium distribution* $\boldsymbol{\pi}$*, then the reversed-time process* $X^R(t) = X(\tau - t)$*, is a stationary Markov chain with*

$$\begin{aligned}\textit{transition rates} \quad & q^R_{kj} = \frac{q_{jk}\pi_j}{\pi_k}, \quad \textit{for } k, j \in \mathcal{S}, k \neq j,\\ \textit{and equilibrium distribution} \quad & \pi^R_j = \pi_j.\end{aligned}$$

Proof. (i) Since $\{X(t)\}$ is stationary, the distribution of $X(\tau-t)(=_d X^R(t))$ and the distribution of $X(\tau-t-y)(=_d X^R(t-y))$ are the same for all t. Thus, the reversed process is also stationary.

For $\{X(t)\}$, we have that the future and the past are independent, conditional only on the present state. For $\{X^R(t)\}$,we interchange the concepts of future and past: the past becomes the future, the future becomes the past, and they are still independent from each other, conditional on the present $\Rightarrow X^R(t)$ too has the Markov property.

Consequently, if $X(t)$ is a stationary CTMC, then $X^R(t)$ is also a stationary CTMC.

(ii) For $h > 0$ and $j \neq k$,

$$\begin{aligned}\Pr(X(t)=j, X(t+h)=k) &= \Pr(X(t+h)=k|X(t)=j)\Pr(X(t)=j)\\ &\text{or } \Pr(X(t)=j|X(t+h)=k)\Pr(X(t+h)=k).\end{aligned}$$

Then

$$\begin{aligned}&\lim_{h\to 0^+}\frac{\Pr(X(t+h)=k|X(t)=j)}{h}\Pr(X(t)=j)\\ &\qquad=\lim_{h\to 0^+}\frac{\Pr(X(t)=j|X(t+h)=k)}{h}\Pr(X(t+h)=k).\end{aligned}$$

Thus,

$$q_{jk}\pi_j = q^R_{kj}\pi_k \qquad \text{and} \qquad q^R_{kj}=\frac{q_{jk}\pi_j}{\pi_k}.$$

To verify that the π_j satisfy the equilibrium equations of the reversed-time Markov chain, we substitute the rates we have just established into the global balance equations and see if they are satisfied. That is,

$$\begin{aligned}&\sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_j q^R_{jk} = \sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_k q^R_{kj}\\ \Rightarrow\quad &\sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_j\frac{q_{kj}\pi_k}{\pi_j} = \sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_k\frac{q_{jk}\pi_j}{\pi_k}\\ \Rightarrow\quad &\sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_k q_{kj} = \sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\pi_j q_{jk}.\end{aligned}$$

□

Corollary 2. $q_{jj} = q^R_{jj}$.

Proof.

$$\begin{aligned}-q^R_{jj} &= \sum_{\substack{k\neq j\\ k\in\mathcal{S}}} q^R_{jk} = \sum_{\substack{k\neq j\\ k\in\mathcal{S}}}\frac{q_{kj}\pi_k}{\pi_j}\\ &= \frac{1}{\pi_j}\sum_{\substack{k\neq j\\ k\in\mathcal{S}}} q_{kj}\pi_k = \frac{1}{\pi_j}\sum_{\substack{k\neq j\\ k\in\mathcal{S}}} q_{jk}\pi_j \quad \text{by equilibrium equations}\\ &= \sum_{\substack{k\neq j\\ k\in S}} q_{jk} = -q_{jj}.\end{aligned}$$

□

Example 3. M/M/1 Queue

Recall that $\mathcal{S} = \{0, 1, 2, \dots\}$, and

$$q_{j,j+1} = \lambda \quad \text{and} \quad q_{j+1,j} = \mu \quad \text{for all } j \in \mathcal{S},$$

with equilibrium probabilities

$$\pi_j = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j \quad \text{for } \lambda < \mu.$$

Hence,

$$q^R_{j+1,j} = \frac{\pi_j q_{j,j+1}}{\pi_{j+1}} = \frac{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^j \lambda}{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{j+1}} = \mu$$

and similarly $$q^R_{j,j+1} = \frac{\pi_{j+1} q_{j+1,j}}{\pi_j} = \frac{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{j+1} \mu}{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{j}} = \lambda.$$

Definition 15. *A continuous-time Markov chain is reversible if the reversed-time process has the same transition rates as the forward-time process, that is,*

$$q^R_{jk} = q_{jk} \quad \textit{for all } j \textit{ and } k \in \mathcal{S}.$$

Note: There is a clear distinction here between a reversed-time process and a reversible process: All CTMCs may be looked at in reversed time, but *not all* CTMCs are *reversible*.

We will look at the special properties of reversible processes in the next lecture.

Lecture 16: Reversible Processes

Concepts checklist

At the end of this lecture, you should be able to:

- *Understand (and hence exploit)* relationships regarding reversible processes, reversed-time processes, and detailed-balance equations.

Recall,

Definition 16. *A continuous-time Markov chain is reversible if the reversed-time process has the same transition rates as the forward-time process, that is,*

$$q^R_{jk} = q_{jk} \quad \text{for all } j \text{ and } k \in \mathcal{S}.$$

A *reversible* process has special properties that we will introduce first by way of example and then by formal statement and proof.

Example 8. A general birth-and-death process.

We have the state space $\mathcal{S} = \mathbb{Z}_+ = \{0, 1, 2, \dots\}$ and non-zero rates (note, can be state dependent)

$$\begin{aligned} q_{j,j+1} &= \lambda_j \quad \text{for } j \geq 0, \\ q_{j,j-1} &= \mu_j \quad \text{for } j \geq 1, \end{aligned}$$

and the equilibrium equations are

$$\begin{aligned} &\pi_j (\lambda_j + \mu_j) = \pi_{j+1}\mu_{j+1} + \pi_{j-1}\lambda_{j-1} \quad \text{for } j \geq 1 \\ &\text{with} \quad \pi_0\lambda_0 = \pi_1\mu_1. \end{aligned} \tag{18}$$

Equation (18) can be re-written as

$$\pi_{j+1}\mu_{j+1} - \pi_j\lambda_j = \pi_j\mu_j - \pi_{j-1}\lambda_{j-1}$$

which is of the form $A_{j+1} = A_j$, where $A_j = \pi_j\mu_j - \pi_{j-1}\lambda_{j-1}$ for $j \geq 1$.

From the equilibrium equations we get the boundary equation

$$\pi_1\mu_1 - \pi_0\lambda_0 = 0,$$

which implies that $A_1 = 0$, and hence that $A_j = 0$ for all $j \geq 1$, so that

$$\pi_j\mu_j = \pi_{j-1}\lambda_{j-1} \quad \text{— known as detailed balance equations.}$$

Rearranging (by repeated substitution) these equations reveals

$$\pi_j = \frac{\pi_{j-1}\lambda_{j-1}}{\mu_j} = \pi_0 \frac{\lambda_0}{\mu_1}\frac{\lambda_1}{\mu_2}\cdots\frac{\lambda_{j-1}}{\mu_j} = \pi_0 \prod_{\ell=0}^{j-1} \frac{\lambda_\ell}{\mu_{\ell+1}}.$$

The equilibrium distribution $\boldsymbol{\pi} = \{\pi_0, \pi_1, \pi_2, \dots\}$ then exists if

$$\sum_{j=0}^{\infty} \prod_{\ell=0}^{j-1} \frac{\lambda_\ell}{\mu_{\ell+1}} \quad < \quad \infty.$$

If the equilibrium distribution exists, the reversed-time transition rates are

$$q^R_{j,j+1} = \frac{\pi_{j+1} q_{j+1,j}}{\pi_j} = \frac{\dfrac{\pi_j \lambda_j}{\mu_{j+1}} \mu_{j+1}}{\pi_j} = \lambda_j$$

and

$$q^R_{j+1,j} = \frac{\pi_j q_{j,j+1}}{\pi_{j+1}} = \frac{\dfrac{\pi_{j+1}\mu_{j+1}}{\lambda_j} \lambda_j}{\pi_{j+1}} = \mu_{j+1}.$$

Note that the reversed-time transition rates are identical to the forward-time transition rates,

$$q^R_{j,j+1} = q_{j,j+1} = \lambda_j$$
$$q^R_{j+1,j} = q_{j+1,j} = \mu_{j+1},$$

and therefore this process is reversible.

Theorem 14. *A stationary continuous-time Markov chain is reversible if and only if there exists a collection of numbers $\pi_j > 0$, summing to unity, that satisfies the detailed balance equations given by*

$$\pi_j q_{jk} = \pi_k q_{kj} \quad \text{for all } j, k \in \mathcal{S}.$$

If such a collection of π_j exists, it is the equilibrium distribution of the Markov chain.

Essentially this means that we could *assume* reversibility and then attempt to find a collection of numbers $\pi_j > 0$ summing to unity that satisfy the detailed balance equations. If we can do this we have both the equilibrium probability distribution and the knowledge that the Markov chain is reversible; otherwise we only know that the Markov chain is not reversible.

Proof:

($\Leftarrow$) Assume that the detailed balance equations have a solution. Then by summing over all states $k \neq j$, we get the global balance equations for the CTMC and therefore $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ must be the equilibrium probability distribution.

Then, by using the detailed balance equations, we have

$$q^R_{jk} = \frac{q_{kj}\pi_k}{\pi_j} = q_{jk},$$

showing that the reversed-time transition rates are the same as the forward-time transition rates, and hence that the CTMC is reversible. □

($\Rightarrow$) Assume now that the CTMC is reversible. Then

$$q_{jk} = q^R_{jk} = \frac{q_{kj}\pi_k}{\pi_j} \quad \text{for all } j, k \in \mathcal{S},$$

where $\boldsymbol{\pi} = \{\pi_0, \pi_1, \dots\}$ is the equilibrium distribution of the CTMC and so

$$\pi_j q_{jk} = \pi_k q_{kj}, \quad \text{for all } j, k \in \mathcal{S},$$

which are the detailed balance equations. □

Theorem 15. *Let $X(t)$ be a stationary, not necessarily reversible, continuous-time Markov chain with transition rates q_{jk} and state space $\mathcal{S}$.*

If we can find numbers q^R_{jk} and π_j for $j,k \in \mathcal{S}$ such that

$$\begin{aligned} -q^R_{jj} = \sum_{\substack{k\in S\\ k\neq j}} q^R_{jk} = -q_{jj} = \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} q_{jk} \quad & \text{for all } j \in \mathcal{S}, \quad \text{(equal holding times)} \\ \pi_j q_{jk} = \pi_k q^R_{kj} \quad & \text{for all } j,k \in \mathcal{S}, \quad (\text{"detailed balance" equations satisfied}) \\ \text{with } \sum_{j\in\mathcal{S}} \pi_j = 1 \quad & \text{and} \quad \pi_j > 0 \text{ for all } j \in \mathcal{S}, \quad \text{(p.m.f.)} \end{aligned}$$

then

(*i*) *the q^R_{jk} are the transition rates of the reversed-time process, and*

(*ii*) *$\boldsymbol{\pi} = \{\pi_j\}_{j\in\mathcal{S}}$ is the equilibrium distribution of both the forward and reversed-time processes.*

Proof. For all $j \in \mathcal{S}$, since $-q_{jj} = -q^R_{jj}$ we have

$$\begin{aligned} \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} q_{jk} &= \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} q^R_{jk}, \\ &= \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \frac{\pi_k q_{kj}}{\pi_j}, \\ \Rightarrow \quad \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_j q_{jk} &= \sum_{\substack{k\in\mathcal{S}\\ k\neq j}} \pi_k q_{kj}, \end{aligned}$$

which are the global balance equations. Hence, $\boldsymbol{\pi} = \{\pi_1, \pi_2, \dots\}$ is the equilibrium distribution of both the forward and reversed-time processes. □

This is useful, in particular in the analysis of queues, because we can guess the transition rates of the reversed-time process and then verify the process by using Theorem 15.

Lecture 17: Queueing Systems – Multiple Customer Classes

Concepts checklist

At the end of this lecture, you should be able to:

- *model* (some) multiple customer class queueing systems as k-variate birth-and-death processes;
- *specify* the equilibrium distribution of a stationary k-variate birth-and-death process; and,
- *specify* the equilibrium distribution of a truncated k-variate birth-and-death process (and more broadly reversible process).

Multiple Customer Classes

There are many situations in which different customer streams compete for the service resources available. Sometimes this involves priorities, different resource requirements, reservation, etc. In each of these cases, we have to keep track of the numbers of each type of customer in the system.

The simplest of these kinds of models is when they can be considered as *a set of k independent birth and death processes*, for some $k \geq 2$.

Definition 17. *A k-variate (sometimes called k-dimensional) birth-and-death process has*

- *state space* $\mathcal{S} = \{\boldsymbol{n} = (n_1, n_2, \ldots, n_k) : n_i \in \{0, 1, \ldots\}\}$,
- *arrival rate* $\lambda_i(n_i)$ *for Type* i *customers, when there are currently* n_i *of them, and*
- *service rate* $\mu_i(n_i)$ *of Type* i *customers, when there are* n_i *of them.*

Theorem 16. *A stationary k-variate birth-and-death process has equilibrium distribution*

$$\pi(\boldsymbol{n}) = C \prod_{r=1}^{k} \prod_{\ell=1}^{n_r} \frac{\lambda_r(\ell - 1)}{\mu_r(\ell)}, \qquad \text{where } C \text{ is a normalising constant.} \tag{19}$$

Proof. We will use Theorem 15. First we guess that the process is reversible, that is

$$q^R(\boldsymbol{n}, \boldsymbol{m}) = q(\boldsymbol{n}, \boldsymbol{m}) \quad \text{for all } \boldsymbol{n}, \boldsymbol{m} \in \mathcal{S}.$$

By Corollary 2 on the properties of diagonal elements, we have that $q^R(\boldsymbol{n}, \boldsymbol{n}) = q(\boldsymbol{n}, \boldsymbol{n})$. Then, by Theorem 15 we only need to show that

$$\pi(\boldsymbol{n}) q(\boldsymbol{n}, \boldsymbol{m}) = \pi(\boldsymbol{m}) q^R(\boldsymbol{m}, \boldsymbol{n}) \overset{\text{by assumption}}{=} \pi(\boldsymbol{m}) q(\boldsymbol{m}, \boldsymbol{n}) \quad \text{for all } \boldsymbol{n}, \boldsymbol{m} \in \mathcal{S}.$$

In particular, we need to consider only the cases where

- $\boldsymbol{m} = \boldsymbol{n} + \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is a k-vector of zeros with a 1 in the ith position, and
- $\boldsymbol{m} = \boldsymbol{n} - \boldsymbol{e}_i$.

That is,

$$\begin{aligned}\pi(\boldsymbol{n})q(\boldsymbol{n}, \boldsymbol{n}+\boldsymbol{e}_i) &= \left(C\prod_{r=1}^{k}\prod_{\ell=1}^{n_r}\frac{\lambda_r(\ell-1)}{\mu_r(\ell)}\right)\lambda_i(n_i)\\ &= \left(C\prod_{\substack{r=1\\ r\neq i}}^{k}\prod_{\ell=1}^{n_r}\frac{\lambda_r(\ell-1)}{\mu_r(\ell)}\right)\left(\prod_{\ell=1}^{n_i+1}\frac{\lambda_i(\ell-1)}{\mu_i(\ell)}\right)\mu_i(n_i+1)\\ &= \pi(\boldsymbol{n}+\boldsymbol{e}_i)q(\boldsymbol{n}+\boldsymbol{e}_i, \boldsymbol{n}).\end{aligned}$$

Therefore, the guess was correct, which implies that a k-variate birth-and-death process is reversible, and that its equilibrium distribution is given by (19). □

The following theorem tells what happens if we truncate a reversible process.

Theorem 17. *Consider a reversible continuous-time Markov chain with state space $\mathcal{S}$ and equilibrium distribution $\boldsymbol{\pi}$. For some set $\mathcal{A} \subseteq \mathcal{S}$, we change the transition rates as follows:*

$$\textit{alter } q_{jk} \textit{ to } cq_{jk} \textit{ for all } j \in \mathcal{A} \textit{ and } k \in \mathcal{S}\backslash\mathcal{A} \textit{ for some } c \geq 0.$$

Then, the resulting process is also a reversible continuous-time Markov chain with equilibrium distribution $\overline{\boldsymbol{\pi}}$ given by

$$\overline{\pi}_j = \begin{cases} B\pi_j & \textit{for } j \in \mathcal{A},\\ Bc\pi_j & \textit{for } j \in \mathcal{S}\backslash\mathcal{A},\end{cases} \tag{20}$$

where B is a normalising constant such that

$$B\sum_{j\in\mathcal{A}}\pi_j + Bc\sum_{i\in S\backslash\mathcal{A}}\pi_i = 1.$$

In particular, if $c = 0$ then

$$\overline{\pi}_j = B\pi_j \quad \textit{for } j \in \mathcal{A},$$

and $B = \left(\sum_{j\in\mathcal{A}}\pi_j\right)^{-1}$.

Proof. First we show that with the probabilities $\overline{\boldsymbol{\pi}}$ given by (20), the necessary and sufficient conditions for reversibility are satisfied for the modified process. Then, we show that the probabilities given by (20) must be the equilibrium probabilities of the modified process.

The detailed balance equations,

$$\overline{\pi}_j\bar{q}_{j,k} \;=\; \overline{\pi}_k\bar{q}_{k,j},$$

where $\bar{q}_{j,k}$ are the intensities for the new process.

Since the original process is reversible, we know that

$$\pi_j q_{j,k} \;=\; \pi_k q_{k,j}, \qquad \text{for all } j, k \in S.$$

Consider the following cases

1. $j, k \in A$**:**

$$\begin{aligned} \pi_j q_{j,k} &= \pi_k q_{k,j} \\ \Rightarrow (B\ \pi_j)\, q_{j,k} &= (B\ \pi_k)\, q_{k,j} \\ \Rightarrow\ \bar{\pi}_j \bar{q}_{j,k} &= \bar{\pi}_k \bar{q}_{k,j}. \end{aligned}$$

2. $j \in A,\ k \in (S \backslash A)$**:**

$$\begin{aligned} \pi_j q_{j,k} &= \pi_k q_{k,j} \\ \Rightarrow B\ c\ \pi_j q_{j,k} &= B\ c\ \pi_k q_{k,j} \\ \Rightarrow\ (B\ \pi_j)\ (c\ q_{j,k}) &= (B\ c\ \pi_k)\, q_{k,j} \\ \Rightarrow\ \bar{\pi}_j \bar{q}_{j,k} &= \bar{\pi}_k \bar{q}_{k,j}. \end{aligned}$$

3. $j, k \in (S \backslash A)$**:**

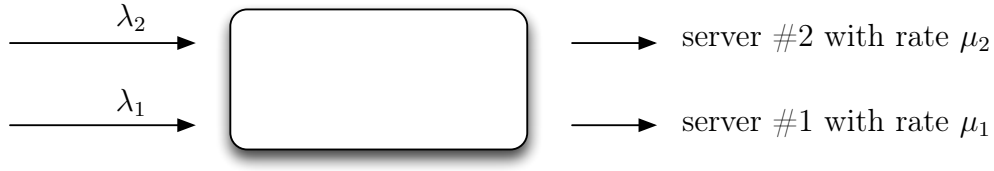
$$\begin{aligned} \pi_j q_{j,k} &= \pi_k q_{k,j} \\ \Rightarrow (B\ c\ \pi_j)\, q_{j,k} &= (B\ c\ \pi_k)\, q_{k,j} \\ \Rightarrow\ \bar{\pi}_j \bar{q}_{j,k} &= \bar{\pi}_k \bar{q}_{k,j}. \end{aligned}$$

Therefore, the new process is reversible with distribution given by the theorem and if $c = 0$, the process has been truncated to the set A and has equilibrium distribution given by $\bar{\pi}_j$.

□

Example 9. Shared Finite Buffer

Consider two independent single-server queues, with arrival rates λ_i and service rates μ_i for $i = 1, 2$. These queues share a common waiting room of size R and customers arriving to a full waiting room are lost. What is the equilibrium distribution for this queueing system?



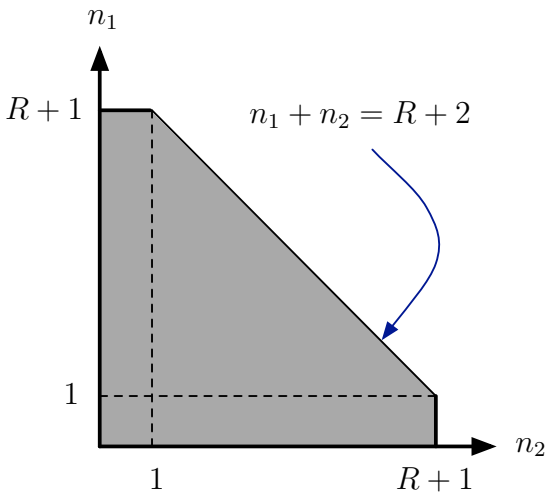
The state space for this system can be represented as follows

Step 1. Letting $[n]^+ = \max(0, n)$, we can write the state space $\mathcal{A}$ as follows

$$\mathcal{A} = \{(n_1, n_2) : [n_1 - 1]^+ + [n_2 - 1]^+ \leq R\}.$$

Step 2. Consider R to be ∞, which implies that the two queues are totally independent, reversible birth-and-death processes and have the joint equilibrium probability distribution given by

$$\pi(n_1, n_2) = \left(1 - \frac{\lambda_1}{\mu_1}\right)\left(1 - \frac{\lambda_2}{\mu_2}\right)\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\ \left(\frac{\lambda_2}{\mu_2}\right)^{n_2}$$



for $\lambda_1 < \mu_1$ and $\lambda_2 < \mu_2$, and with invariant measure $\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{\mu_2}\right)^{n_2}$.

Note: An invariant measure represents any positive multiple of the equilibrium distribution, and thus it is not normalised.

Step 3. We can use Theorem 17, which implies that since the original process was reversible, the truncated process must also be reversible with the same invariant measure.

With the state space $\mathcal{A} = \{(n_1, n_2) : [n_1 - 1]^+ + [n_2 - 1]^+ \leq R\}$, the equilibrium distribution $\overline{\boldsymbol{\pi}}$ is given by

$$\overline{\pi}(n_1, n_2) = D\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{\mu_2}\right)^{n_2} \quad \text{where } D = \left(\sum_{\mathcal{A}}\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{\mu_2}\right)^{n_2}\right)^{-1}.$$

Out of interest, the normalising constant B of the theorem is given by

$$\frac{D}{\left(1-\frac{\lambda_1}{\mu_1}\right)\left(1-\frac{\lambda_2}{\mu_2}\right)} = \frac{1}{\left(1-\frac{\lambda_1}{\mu_1}\right)\left(1-\frac{\lambda_2}{\mu_2}\right)\sum_{\mathcal{A}}\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\left(\frac{\lambda_2}{\mu_2}\right)^{n_2}}.$$

Lecture 18: Queueing Systems - Loss Networks

Concepts checklist

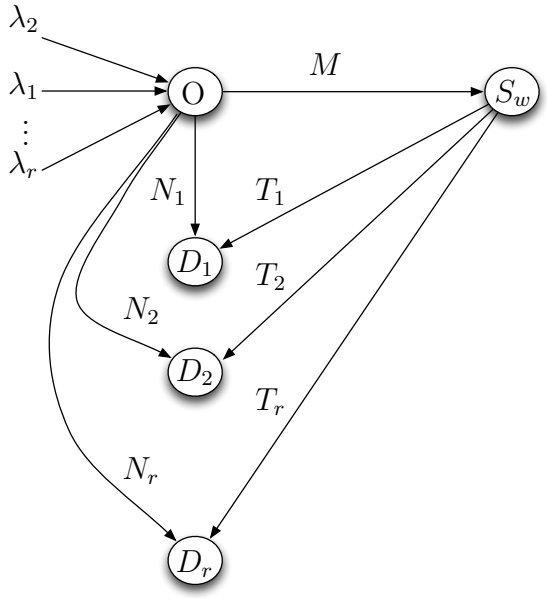
At the end of this lecture, you should be able to:

- *model* and *specify equilibrium distributions* for alternative routing with call packing, and circuit-switched networks; and,
- *state a theorem* and *specify exact* route blocking probabilities for circuit-switched networks.

Example 10. Alternative Routing with Call Packing

There are r Poisson streams of calls sharing a common origin O, which are routed to their destinations D_i for $i \in \{1, 2, \ldots, r\}$ via N_i direct circuits before overflowing onto M common circuits through the switch S_w.

From S_w, there are T_i circuits available to each of the required destinations D_i. We assume arrival rates λ_i and mean call holding times $1/\mu_i$ for each $i \in \{1, 2, \ldots, r\}$.



Let n_i be the number of customers in the system from stream i and let $\boldsymbol{n} = (n_1, n_2, \ldots, n_r)$.

We assume call packing, which means that calls are packed back onto the direct circuits from a common (overflow) link whenever a direct circuit becomes available.

Initially, we consider an infinite number of available circuits for each stream. Hence, every caller gets a circuit and essentially sees an infinite server queue, so the invariant measure is

$$\prod_{i=1}^{r} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} \frac{1}{n_i!}.$$

The number of overflow circuits from the ith direct route is $[n_i - N_i]^+$. Therefore, the state space $\mathcal{A}$ is restricted by

1. $[n_i - N_i]^+ \leq T_i$ and

2. $\sum_{i=1}^{r} [n_i - N_i]^+ \leq M.$

Truncating the state space to $\mathcal{A}$ gives us the equilibrium distribution

$$\pi(n_1, n_2, \ldots, n_r) = C \prod_{i=1}^{r} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} \frac{1}{n_i!}, \quad \text{where } C = \left(\sum_{\mathcal{A}} \prod_{i=1}^{r} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} \frac{1}{n_i!}\right)^{-1}.$$

Example 11. Circuit-Switched Network

Assumptions:

- fixed routing is used (that is, no overflow or alternative routing),
- there are J links in total – link j has c_j circuits, for $1 \leq j \leq J$, and $\boldsymbol{c} = (c_1, c_2, \ldots, c_J)$,
- there are R routes in total – calls requesting route r arrive in a Poisson stream of rate λ_r, for $1 \leq r \leq R$,
- without loss of generality (wlog), call holding times have unit mean,
- route r calls use $A_{j,r}$ circuits on link j.

Let n_r be the number of calls using route r and let $\boldsymbol{n} = (n_1, n_2, \ldots, n_R)$ be the state of the network. Furthermore, if we let $A = \{A_{j,r}\}$ be the $J \times R$ matrix such that $A_{j,r}$ represents the circuit usage on link j for route r, then we can write

$$\begin{aligned} S(\boldsymbol{c}) &= \{\boldsymbol{n} : \sum_{r} A_{j,r} n_r \leq c_j, \quad 1 \leq j \leq J\} \\ &= \{\boldsymbol{n} : A\boldsymbol{n} \leq \boldsymbol{c}\}. \end{aligned}$$

If we let $c_j \longrightarrow \infty$ for all j, then the process is an R-dimensional Birth-and-Death process, which is reversible. Then, by truncating the process by making each c_j finite, we can use Theorem 17 to find the equilibrium probability distribution, which is given by

$$\pi(\boldsymbol{n}) = [G(\boldsymbol{c}, R)]^{-1} \prod_{r=1}^{R} \frac{\lambda_r^{n_r}}{n_r!} \quad \text{for } \boldsymbol{n} \in \mathcal{S}(\boldsymbol{c}),$$

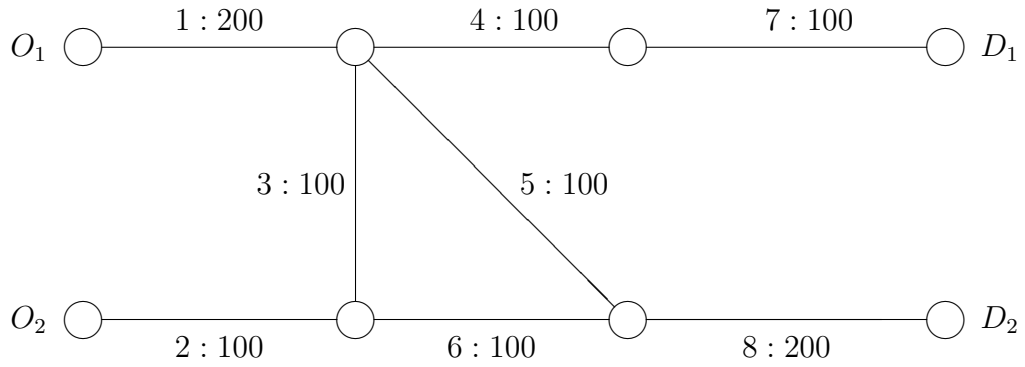
$$\text{where} \qquad \boldsymbol{c} = (c_1, c_2, \ldots, c_J)^\top \text{ and } \mathcal{S}(\boldsymbol{c}) = \{\boldsymbol{n} : \sum_{r} A_{j,r} n_r \leq c_j, \quad 1 \leq j \leq J\}.$$

Here, $[G(\boldsymbol{c}, R)]^{-1}$ is the normalising constant, which is dependent on the state space $\mathcal{S}(\boldsymbol{c})$ defined by $\boldsymbol{c}$. Note that, summing the invariant measure over $\boldsymbol{n} \in \mathcal{S}(\boldsymbol{c})$ gives

$$G(\boldsymbol{c}, R) = \sum_{\boldsymbol{n}: A\boldsymbol{n} \leq \boldsymbol{c}} \prod_{r=1}^{R} \frac{\lambda_r^{n_r}}{n_r!}.$$

An instance

Consider the following (simple, circuit-switched) loss network:



A label on the link is given on the diagram using the legend $a : b$, where a is the link number and b is the number of circuits on link a.

It is required to route traffic of offered load 60, 70, 50, and 80 (Erlangs) respectively between the four origin-destination pairs ($O_1 - D_1$, $O_1 - D_2$, $O_2 - D_1$ and $O_2 - D_2$). This means $\lambda_1/\mu_1 = 60$ Erlangs, $\lambda_2/\mu_2 = 70$ Erlangs, and so on. Note that as stated above we assume $\mu_i = 1$ for $i = 1, 2, 3, 4$.

Assuming that traffic between origin O_i and destination D_j is routed along the shortest possible path and requires a single circuit from each such link, the routes of the loss network are:

Route ID	Route	Links used	Offered loads
1	O_1 – D_1	1, 4, 7	60
2	O_1 – D_2	1, 5, 8	70
3	O_2 – D_1	2, 3, 4, 7	50
4	O_2 – D_2	2, 6, 8	80

The state space $\mathcal{S}$ of the network, in the form $\{\boldsymbol{n} : A\boldsymbol{n} \leq \boldsymbol{c}\}$, where A is the matrix with components A_{jr} being the number of circuits that route r calls use on link j, and $\boldsymbol{c} = (c_1, c_2, \ldots)$ is the vector containing the number c_j of circuits on link j, is

$$\mathcal{S} = \{\boldsymbol{n} : A\boldsymbol{n} \leq \boldsymbol{c}\}$$

where

$$\boldsymbol{c} = (200, 100, 100, 100, 100, 100, 100, 200)^\top$$

and

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Exact Route Blocking Probabilities

The equilibrium distribution can be used to give the exact route blocking probabilities. However, it is not clear which states block a route r call and therefore over which set of states we need to sum to calculate the blocking probability. Fortunately, there is a much simpler formula which requires calculating only the normalising constants for two networks.

Theorem 18. *If* $B_r := \Pr$*(call on route* r *is blocked) and* $\boldsymbol{e}_r := (0, 0, \ldots, 0, 1, 0, \ldots, 0)^\top$*, with* 1 *in the* r*th position, then*

$$B_r = 1 - \frac{G(\mathbf{c} - A\boldsymbol{e}_r, R)}{G(\mathbf{c}, R)}.$$

Proof:

$$\begin{aligned}
&\Pr(\text{call is accepted on route } r)\\
&\quad = \Pr(\text{there are at least } A_{j,r} \text{ circuits available on link } j, \text{ for all } 1 \le j \le J)\\
&\quad = \Pr(\text{the number of circuits in use on link } j \le c_j - A_{j,r}, \text{ for all } 1 \le j \le J)\\
&\quad = \Pr(A\boldsymbol{n} \le \mathbf{c} - A\boldsymbol{e}_r)\\
&\quad = \sum_{\{\mathbf{n}: A\mathbf{n} \le \mathbf{c} - A\mathbf{e}_r\}} [G(\mathbf{c}, R)]^{-1} \prod_{\ell=1}^{R} \frac{\lambda_\ell^{n_\ell}}{n_\ell!}\\
&\quad = \frac{1}{G(\mathbf{c}, R)} \sum_{\{\mathbf{n}: A\mathbf{n} \le \mathbf{c} - A\mathbf{e}_r\}} \prod_{\ell=1}^{R} \frac{\lambda_\ell^{n_\ell}}{n_\ell!}\\
&\quad = \frac{G(\mathbf{c} - A\boldsymbol{e}_r, R)}{G(\mathbf{c}, R)}.
\end{aligned}$$

Therefore,

$$B_r = \Pr\left(\text{call on route } r \text{ is blocked}\right) = 1 - \frac{G(\mathbf{c} - A\boldsymbol{e}_r, R)}{G(\mathbf{c}, R)}. \qquad \square$$

Example 11: Instance of a circuit-switched network

What is the exact expression for the probability that a call attempting to access D_1 from O_1 is accepted?

The set of states $\mathcal{S}_1$ in which a call attempting to access D_1 from O_1 is accepted is

$$\mathcal{S}_1 = \{\boldsymbol{n} : A\boldsymbol{n} \le \boldsymbol{c} - A\boldsymbol{e}_1\}.$$

Hence the probability that a call attempting to access D_1 from O_1 is accepted is given by

$$\frac{G(\boldsymbol{c} - A\boldsymbol{e}_1, R)}{G(\boldsymbol{c}, R)} \quad , \text{ where } G(\boldsymbol{c}, R) \;=\; \sum_{\boldsymbol{n}: A\boldsymbol{n} \le \boldsymbol{c}} \frac{60^{n_1}}{n_1!} \frac{70^{n_2}}{n_2!} \frac{50^{n_3}}{n_3!} \frac{80^{n_4}}{n_4!} \; .$$

Lecture 19: Queueing Systems - Loss Networks and Reduced Load Approximations

Concepts checklist

At the end of this lecture, you should be able to:

- *evaluate* the Erlang Blocking/Loss Formula for Erlang Loss Systems; and,
- *evaluate* approximate blocking probabilities using the Erlang Fixed Point Method (EFPM).

Example 12. The $M/M/N/N$ queue

This queue is known as the Erlang Loss System, which can model a group of N circuits handling a Poisson stream of calls where arrivals which occur when all circuits are occupied are lost. The notation refers to *arrival process/service time distribution/number of servers/capacity*. Hence, calls arrive in a Poisson stream, of some rate λ; if a circuit is available, the call grabs the circuit and holds it for the connection time, which is exponentially distributed with some parameter μ; and the state space is $S = \{0, 1, 2, \ldots, N\}$.

The rates are as follows

$$\lambda_\ell = \begin{cases} \lambda & 0 \leq \ell < N \\ 0 & \ell = N \end{cases}$$
$$\mu_\ell = \ell\mu \qquad \ell \leq N.$$

We have already derived the equilibrium probability distribution $\boldsymbol{\pi}$, where

$$\pi_i = \left[\sum_{\ell=0}^{N} \frac{1}{\ell!}a^\ell\right]^{-1} \frac{1}{i!}a^i \quad \text{for } i \in \{0, 1, 2, \ldots, N\} \text{ and } a = \frac{\lambda}{\mu}.$$

The Erlang-B Formula (also known as *Erlang Blocking Formula* or *Erlang Loss Formula*) $B(N, a) = \pi_N$ is the probability that an arriving call is lost (as given by PASTA – *Poisson Arrivals See Time Averages*), and is also the blocking probability of an arriving call under equilibrium conditions. The Erlang Loss Formula is valid when connection times follow any distribution with mean $1/\mu$. That is, the Erlang Loss formula is insensitive to service time distributions.

Evaluating the (Erlang-B) blocking formula

$$B(N, a) = \pi_N = \frac{1}{N!}a^N \left[\sum_{\ell=0}^{N} \frac{1}{\ell!}a^\ell\right]^{-1}$$

becomes a formidable task when the value of N becomes large, because such things as

$$\frac{100^{200}}{200!}$$

are very difficult to evaluate and use. Therefore, it is best to use an iterative method for calculation, as follows (which is not hard to derive):

$$B(N, a) = \frac{aB(N-1, a)}{N + aB(N-1, a)} \qquad \text{with} \qquad B(0, a) = 1.$$

Reduced Load Approximations

Reduced load approximations are commonly used to find approximate performance measures for stochastic systems which either

1. do not have *closed form* equilibrium distributions, or
2. the equilibrium distribution has a closed form but cannot be numerically calculated because of the size of the system's state space.

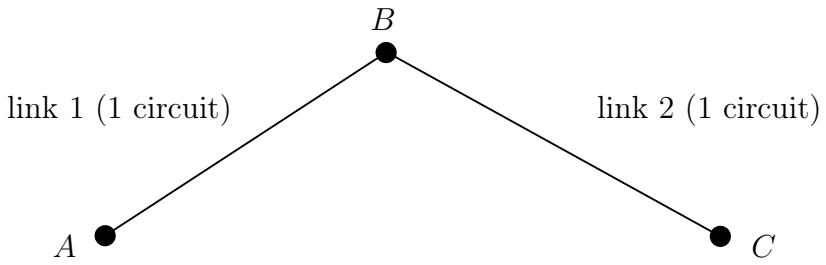
The general idea of a reduced load approximation is

1. consider sections of the network in isolation, and
2. reduce the traffic load offered to that section in accord with calls rejected by the remainder of the network.

The reduced load technique we shall demonstrate is the Erlang fixed point method (EFPM), a network decomposition technique, which can be used to approximate many performance measures of interest for a variety of loss networks.

Example 13. Two-link network

Consider two links with a single circuit on each (as illustrated below), one route only from A to C, which requires both circuits and an offered load of 1 (Erlang).



Obtaining the exact blocking probability:

The system behaves exactly the same as a single link with one circuit and offered load $a = 1$. Therefore, using the Erlang blocking formula we have

$$\Pr(\text{a call is blocked}) = B(1,1) = \left.\frac{a}{1+a}\right|_{a=1} = \frac{1}{1+1} = 0.5 \quad \textit{– the correct probability!}$$

Obtaining an approximation to the blocking probability, using EFPM:

EFPM is a reduced load technique which considers each link in turn and reduces the load offered to that link by the proportion of that load which is blocked at the other links in the network. The method assumes that

- links are **independent**, and
- a call is accepted only if accepted on all links.

Thus, in the above example, if for $i = 1, 2$ we let

$$\alpha_i = \Pr(\text{call is blocked on link } i \text{ under EFPM assumptions}),$$

then $\alpha_1 = \alpha_2$ by symmetry.

The traffic offered to link 1 is $a(1-\alpha_2) = a(1-\alpha_1)$. Thus,

$$\alpha_1 = \Pr(\text{call is blocked on link 1}) = B(1, 1-\alpha_1) = \frac{1-\alpha_1}{1+1-\alpha_1}.$$

That is,

$$\begin{aligned} & 2\alpha_1 - \alpha_1^2 = 1 - \alpha_1, \\ \implies \quad & \alpha_1^2 - 3\alpha_1 + 1 = 0, \end{aligned}$$

which has solution $\alpha_1 \approx 0.38$.

Then we calculate

$$\begin{aligned} \Pr(\text{call is accepted by network}) &= \Pr(\text{accepted by link 1}) \Pr(\text{accepted by link 2}) \\ &\quad \textit{(by our } \textbf{\textit{independence}} \textit{ assumption)}, \\ &\approx (1-0.38)(1-0.38) = 0.62^2. \\ \Rightarrow \Pr(\text{a call is blocked}) &\approx 1 - (0.62)^2 = 0.62. \end{aligned}$$

Even in this case, where **it is blatantly wrong to assume independence** between the links, the approximate blocking probability of 0.62 is *not too far* from the true value of 0.5.

Example 14 - Two link, slightly more complex routing

Consider the same two link network, but with a slightly more complex routing pattern.

Route Number	Route	Offered load (a_i)	Links used
1	AB	0.5E	1
2	BC	0.6E	2
3	AC	0.3E	1, 2

If the vector $\boldsymbol{n} = (n_1, n_2, n_3)$ records the number of calls on routes $1, 2$ and 3 respectively, then we can then write down all the possible states for the network as

- $(0, 0, 0)$ – no calls present
- $(0, 0, 1)$ – one call using route 3, which takes both links
- $(1, 0, 0)$ – one call on route 1, which only uses link 1
- $(0, 1, 0)$ – one call on route 2, which only uses link 2
- $(1, 1, 0)$ – two calls, one using route 1 and the other route 2.

We can then work out the exact equilibrium distribution by:

$\boldsymbol{n}$	Invariant Measure $= \prod_{i=1}^{3} \frac{a_i^{n_i}}{n_i!}$	$\boldsymbol{\pi}(\boldsymbol{n})$
$(0,0,0)$	1	$1/2.7 = 0.3704$
$(0,0,1)$	0.3	$0.3/2.7 = 0.1111$
$(1,0,0)$	0.5	$0.5/2.7 = 0.1852$
$(0,1,0)$	0.6	$0.6/2.7 = 0.2222$
$(1,1,0)$	$0.5 \times 0.6 = 0.3$	$0.3/2.7 = 0.1111$
	Total $= 2.7$	Total $= 1.0$

We can calculate the exact blocking probabilities as follows:

Route	
AB	$0.1111 + 0.1852 + 0.1111 = 0.4074$
BC	$0.1111 + 0.2222 + 0.1111 = 0.4444$
AC	$1.0 - 0.3704 = 0.6296$

Lecture 20: Queueing Systems - Erlang Fixed Point Method

Concepts checklist

At the end of this lecture, you should be able to:

- *state* the Erlang fixed point method, and a numerical procedure to implement it.

Example 14 - Two link, slightly more complex routing, continued

Let us use an approximate technique by assuming now that

- link 1 is blocked with probability α_1, and
- link 2 is blocked with probability α_2.

Then, the demand made on each link using our previous reduced load approach is:

Route	Link 1	Link 2
AB	0.5	0
BC	0	0.6
AC	$0.3(1-\alpha_2)$	$0.3(1-\alpha_1)$
Total	$y_1 = 0.8 - 0.3\alpha_2$	$y_2 = 0.9 - 0.3\alpha_1$

$$\alpha_1 = B(1, y_1) = \frac{y_1}{1+y_1} = \frac{0.8-(0.3)\alpha_2}{1.8-(0.3)\alpha_2}$$

$$\text{and} \quad \alpha_2 = B(1, y_2) = \frac{y_2}{1+y_2} = \frac{0.9-(0.3)\alpha_1}{1.9-(0.3)\alpha_1}.$$

Solving these equations gives (the only sensible solution) as

$$\alpha_1 = 0.4007 \qquad \text{and} \qquad \alpha_2 = 0.4381.$$

Approximate blocking probabilities for the routes can now be calculated and compared with the exact solutions evaluated above.

Blocking on route AC is approximated by $1-(1-\alpha_1)(1-\alpha_2) \approx 0.6633$.

Route	AB	BC	AC
Exact	0.4074	0.4444	0.6296
Approximate	0.4007	0.4381	0.6633

The accuracy of the approximation is significantly better than that which we saw in the previous simpler example. Generally, the more complex the routing and the network, the more accurate the approximation becomes!

The idea behind the EFPM approximation is ...

- For each link j, let α_j be an approximation of the probability that an incoming request for a single circuit on link j is blocked.
- Calls for routes through link j are only accepted if **all** the links on the route are available.
- Consequently, not all requests for circuits on link j can be accepted, even if the link has sufficient circuits available. For example, with a route r through j, a request may be "rejected" because the circuit required on a different link i of r is not available. The probability of this is approximately α_i.

So...

- In order to compensate for the unavailability of the other circuits, the reduced load of requests for circuits on link j is used, instead of the actual load of requests for link j (which is simply $\sum_{\{r:j\in r\}} a_r$).
- If requests for circuits were to arrive as a Poisson process with rate y_j, then
  1. the link blocking probability would be given by $\alpha_j = B(c_j, y_j)$
  2. and the loss probability on route r would satisfy $B_r = 1 - \prod_{j\in\mathcal{J}} (1-\alpha_j)^{A_{j,r}}$ exactly, as a call is accepted if and only if it is accepted on all links.

We now give a mathematical description of the EFPM, which is most commonly used in practice in the telecommunications industry.

Definition 18 (Erlang Fixed Point Method)**.** *For a network comprised of a set of links $j \in \mathcal{J}$, let $\{\alpha_j, j \in \mathcal{J}\}$ be the unique solution to the equations*

$$\alpha_j = B(c_j, y_j), \text{ where } y_j = \sum_{\{r:j\in r\}} a_r \prod_{\{i\in r, i\neq j\}} (1-\alpha_i)$$

- *a_r is the offered load on route r and*
- *$B(\cdot,\cdot)$ is the Erlang Loss Formula.*

The vector $\boldsymbol{\alpha} = (\alpha_j, j \in \mathcal{J})$ is called the Erlang fixed point solution and an approximation for the loss probability on route r is given by,

$$B_r \approx 1 - \prod_{j\in\mathcal{J}} (1-\alpha_j)^{A_{j,r}} ,$$

where $A_{j,r}$ is the number of circuits that route r calls use on link j (here, either 0 or 1)

The Erlang fixed point approximation usually works well if the number of circuits and the associated demands are very high, or if the network is such that the routing is highly diverse. This means that on any link the circuits in use are likely to be from different routes and so likely to be connected via different links.

It is possible to improve the accuracy of the approximation, while still using reduced load techniques, by assuming that larger *modules* of the network, rather than individual links, operate

"independently" of each other. Of course, this will mean that each module is larger and hence more difficult to analyse. However, careful choice of the modules may lead to significantly more accurate results with only a slight increase in complexity.

An example in which the approximation procedure should not be expected to perform well is where a number of small capacity links are arranged one after another in a line like our initial example. Here we would expect considerable dependence between the number of free circuits on adjacent links.

Also note that our model assumes that routes are fixed.

Example 11 (continuing from Lecture 18)

$$\begin{aligned}\Pr(\text{a call on route } O_1 - D_1 \text{ is accepted}) &= 1 - \Pr(\text{a call on route } O_1 - D_1 \text{ is blocked}) \\ &= \frac{G(\boldsymbol{c} - A\boldsymbol{e}_1, R)}{G(\boldsymbol{c}, R)} \\ &= \frac{G(\boldsymbol{c} - A\boldsymbol{e}_1, 4)}{G(\boldsymbol{c}, 4)}\end{aligned}$$

where

$$G(\boldsymbol{c}, 4) = \sum_{\{\boldsymbol{n}: A\boldsymbol{n} \le \boldsymbol{c}\}} \prod_{r=1}^{4} \frac{\lambda_r^{n_r}}{n_r!} = \sum_{\{\boldsymbol{n}: A\boldsymbol{n} \le \boldsymbol{c}\}} \frac{60^{n_1}}{n_1!} \frac{70^{n_2}}{n_2!} \frac{50^{n_3}}{n_3!} \frac{80^{n_4}}{n_4!},$$

$$G(\boldsymbol{c} - A\boldsymbol{e}_1, 4) = \sum_{\{\boldsymbol{n}: A\boldsymbol{n} \le \boldsymbol{c} - A\boldsymbol{e}_1\}} \prod_{r=1}^{4} \frac{\lambda_r^{n_r}}{n_r!} = \sum_{\{\boldsymbol{n}: A\boldsymbol{n} \le \boldsymbol{c} - A\boldsymbol{e}_1\}} \frac{60^{n_1}}{n_1!} \frac{70^{n_2}}{n_2!} \frac{50^{n_3}}{n_3!} \frac{80^{n_4}}{n_4!}.$$

The general formula for the Erlang fixed point equations is:

$$\alpha_j = B(c_j, y_j)$$

where

$$y_j = \sum_{r: j \in r} a_r \prod_{i \in r, i \ne j} (1 - \alpha_i),$$

a_r is the offered load for route r, and

$$B(c_j, y_j) = \frac{y_j B(c_j - 1, y_j)}{c_j + y_j B(c_j - 1, y_j)}.$$

Note that $B(c_j, y_j)$ is the Erlang-B formula, i.e. the probability that a call is lost on a link with c_j circuits and offered load y_j.

So, the eight equations for α_j are

$$\begin{aligned}&\alpha_1 = B(200, y_1), \quad \alpha_2 = B(100, y_2), \quad \alpha_3 = B(100, y_3), \quad \alpha_4 = B(100, y_4), \\ &\alpha_5 = B(100, y_5), \quad \alpha_6 = B(100, y_6), \quad \alpha_7 = B(100, y_7), \quad \alpha_8 = B(200, y_8),\end{aligned}$$

where

$$
\begin{aligned}
y_1 &= 60(1-\alpha_4)(1-\alpha_7) + 70(1-\alpha_5)(1-\alpha_8),\\
y_2 &= 50(1-\alpha_3)(1-\alpha_4)(1-\alpha_7) + 80(1-\alpha_5)(1-\alpha_8),\\
y_3 &= 50(1-\alpha_2)(1-\alpha_4)(1-\alpha_7),\\
y_4 &= 60(1-\alpha_1)(1-\alpha_7) + 50(1-\alpha_2)(1-\alpha_3)(1-\alpha_7),\\
y_5 &= 70(1-\alpha_1)(1-\alpha_8),\\
y_6 &= 80(1-\alpha_2)(1-\alpha_8),\\
y_7 &= 60(1-\alpha_1)(1-\alpha_4) + 50(1-\alpha_2)(1-\alpha_3)(1-\alpha_4),\\
y_8 &= 70(1-\alpha_1)(1-\alpha_5) + 80(1-\alpha_2)(1-\alpha_6).
\end{aligned}
$$

Numerical Procedure

In order to use the EFPM to determine the blocking probabilities in any network, the following numerical procedure can be invoked.

(1) Initially assume $\boldsymbol{\alpha} = (\alpha_j,\ j \in \mathcal{J}) = \mathbf{0}$.

(2) Calculate $\mathbf{y} = (y_j,\ j \in \mathcal{J})$ using the current value of $\boldsymbol{\alpha}$ and

$$y_j = \sum_{\{r: j \in r\}} a_r \prod_{\{i \in r, i \neq j\}} (1-\alpha_i).$$

(3) Calculate $\boldsymbol{\alpha}$ using the Erlang-B formula and $\boldsymbol{y}$ as calculated in Step 2.

(4) Compare the new estimate of $\boldsymbol{\alpha}$ with the previous estimate. If any component differs by more than some specified tolerance then go to Step 2; else

(5) Calculate $\mathbf{B} = (B_r, r \in \mathcal{R})$ according to $B_r = 1 - \prod_{j \in \mathcal{J}} (1-\alpha_j)^{A_{j,r}}$.

END OF TERM

Lecture 21: Burke's Theorem and Jackson Networks

Concepts checklist

At the end of this lecture, you should be able to:

- *state* Burke's Theorem and *explain its importance*, and *use it* to analyse equilibrium behaviour of particular queueing systems;
- *define* an Open Jackson Network; and,
- *state* Jackson's Theorem.

Theorem 19 (Burke's Theorem.)**.** *Consider a queue with a Poisson arrival process of rate λ and exponential service time distribution with parameter $\mu > \lambda$. In equilibrium,*

(*i*) *the departure process from this queue is a Poisson process with parameter λ,*

(*ii*) *the number in the queue at any time t is independent of the departure process prior to t.*

Proof. (*i*) Recall that the queue-length process of a birth-and-death process is reversible. This implies that the reverse process is a continuous-time Markov chain with the same transition rates as the forward time process.

In forward time, arrivals occur in a Poisson process; since the reversed-time Markov chain has the same transition rates, the reversed-time "arrival process" is also a Poisson process. Hence, we have that the forward time departure process is also a Poisson process.

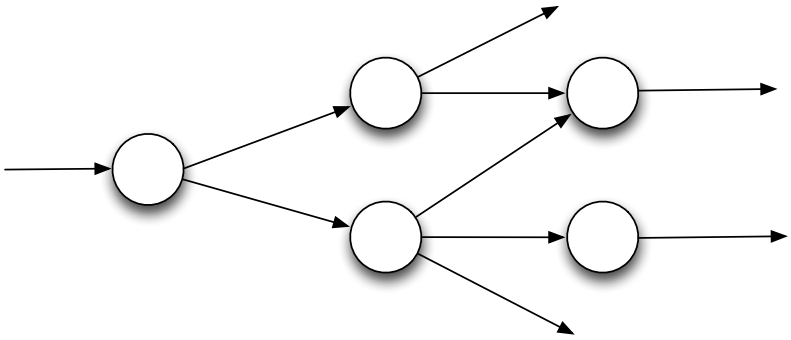
(*ii*) Furthermore, in forward time the state of the queue at time t is independent of future arrivals, which implies that the state is also independent of the past departure process in the reversed-time Markov chain. Because the process is reversible, it is also true for the forward time process that the state is independent of past departure process. □

Note: it is not surprising that the departure rate is λ — because what enters must leave *in equilibrium* for a continuous-time Markov chain to be stable — but what is surprising is the fact that the departure process *in equilibrium* is Poisson. We might have expected a more complicated description of the departure process. (For example, that it is Poisson of rate μ (the service rate) when the queue is busy and of rate 0 when the queue is empty.)

Burke's Theorem is important because it

- allows us to split up the output of one queue and feed it to other queues, where the departure process from one queue is the arrival process to other queues,

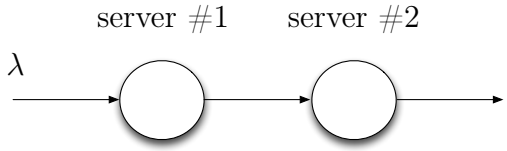
- and tells us that
  1. the arrival process of downstream queues will be Poisson,
  2. the state of downstream queues at time t
     - depends on the departure process of upstream queues before time t, but
     - is independent of the state of the upstream queues at time t.



Essentially, this means that if a network of n-server queues can be ordered from 1 to J in such a way that customers leaving queue $j \in \{1, 2, \ldots, J\}$ are fed into queues $j+1, \ldots, J$ or leave the system, then this network has a product form equilibrium distribution

$$\pi(\boldsymbol{n}) = \pi(n_1, n_2, \ldots, n_J) = \pi_1(n_1)\pi_2(n_2)\ldots\pi_J(n_J).$$

Example 15. Tandem of single-server queues (Feed-Forward)



Let $\boldsymbol{\pi}(n_1, n_2)$ be the equilibrium distribution, where n_i is the level of occupancy of the ith single-server queue in the tandem.

By Burke's Theorem, the states n_1 and n_2 are independent, so $\boldsymbol{\pi}(n_1, n_2) = \pi_1(n_1)\pi_2(n_2)$, where $\pi_i(n_i)$ is the equilibrium distribution of the ith single-server queue.

Since $\pi_i(n_i) = \left(1 - \dfrac{\lambda}{\mu_i}\right)\left(\dfrac{\lambda}{\mu_i}\right)^{n_i}$ for $\lambda < \mu_i$ and $i \in \{1, 2\}$,

$$\pi(n_1, n_2) = \pi_1(n_1)\pi_2(n_2) = \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^{n_1}\left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^{n_2}$$

iff $\lambda < \min\{\mu_1, \mu_2\}$.

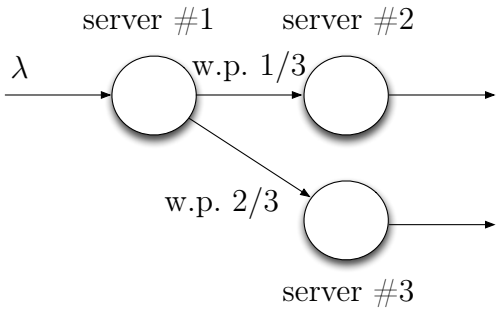
Example 16. Three single-server queues (Feed-forward)

Here, the departure process from queue 1 is probabilistically split: 1/3 to queue 2 and 2/3 to queue 3.

Using Burke's Theorem again, we have

$$\begin{aligned}\pi(\boldsymbol{n}) &= \pi(n_1, n_2, n_3) = \pi_1(n_1)\pi_2(n_2)\pi_3(n_3), \\ &= \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^{n_1}\left(1 - \frac{\frac{1}{3}\lambda}{\mu_2}\right)\left(\frac{\frac{1}{3}\lambda}{\mu_2}\right)^{n_2}\left(1 - \frac{\frac{2}{3}\lambda}{\mu_3}\right)\left(\frac{\frac{2}{3}\lambda}{\mu_3}\right)^{n_3}\end{aligned}$$

iff $\lambda < \min\left\{\mu_1, 3\mu_2, \dfrac{3\mu_3}{2}\right\}$.



We have used Burke's Theorem to show that feed forward queueing networks have a product form equilibrium distribution. We can extend this to networks which have feedback. When output streams are fed back into earlier queues, the input streams to queues are generally non-Poisson and the logic we have used collapses. However the **independence** result survives, as we will see.

The breakthrough dealing with feedback networks has been generally credited to Jackson in 1957, who considered an open network of N s_i-server queues for $1 \leq i \leq N$. He proved that the joint equilibrium distribution for the network is a product over the queues of the equilibrium distributions of the individual queues.

We will call such networks Open Jackson Networks. Jackson generalised this idea further by allowing the arrival rate at the ith queue to be an arbitrary function $\lambda_i(\boldsymbol{n})$ of the total number of customers in the network.

Definition 19 (Open Jackson Network.)**.** *An Open Jackson Network consists of a network of N queues, where at node i of that network,*

- *arrivals come from outside of the network at rate λ_i,*
- *the service rate is $\mu_i(n_i)$ when there are n_i customers in that queue, and*
- *a customer upon completing service will either*
  - *move to queue j with probability γ_{ij}, or*
  - *leave the network with probability $\beta_i = 1 - \sum_j \gamma_{ij}$.*

The state space of the network records the number of customers at a queue but does not distinguish between customers when a service period ends and a customer is removed from the queue.

Theorem 20 (Jackson's Theorem.)**.** *An Open Jackson Network has the following product form equilibrium distribution (provided it can be normalised):*

$$\pi(\boldsymbol{n}) = \pi(n_1, n_2, \ldots, n_N) = \prod_{i=1}^{N} \pi_i(n_i),$$

$$\text{where} \quad \pi_i(n_i) = \pi_i(0) \prod_{\ell=1}^{n_i} \frac{y_i}{\mu_i(\ell)} \text{ is the equilibrium of the } i\text{th queue}$$

and y_i is the average arrival rate to queue i, given by the traffic equations

$$y_i = \lambda_i + \sum_{j=1}^{N} y_j \gamma_{ji}.$$

Note:

- Normalisation depends on whether the constants $\pi_i(0)$ can be found for each $i \in \{1, 2, \ldots, N\}$ such that

$$\sum_{n_i=0}^{\infty} \pi_i(n_i) = 1.$$

- We tacitly assume that the rate into each queue must be the same as the rate out (this implies stability). That is, y_i is both the arrival rate and departure rate from queue i.

The form

$$Q_i(n_i) = \prod_{\ell=1}^{n_i} \frac{y_i}{\mu_i(\ell)}$$

is an invariant measure for the number of customers at queue i if the queue is fed with a Poisson arrival stream of rate y_i.

In fact, y_i is the total *average* arrival rate to queue i but it is, in general, not a Poisson stream and yet the result is as if it is a Poisson stream.

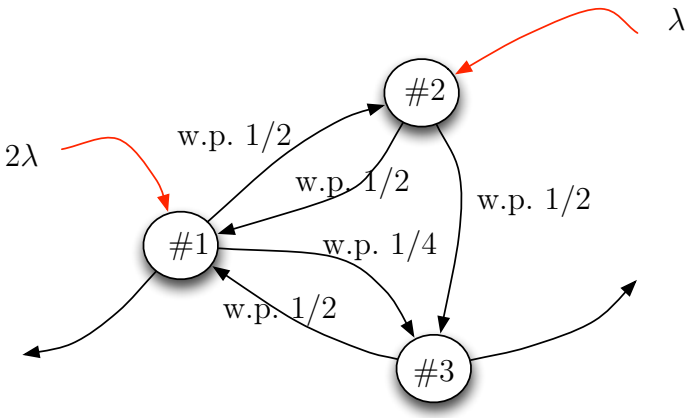
Lecture 22: Jackson Network example, and observed distributions in CTMCs

Concepts checklist

At the end of this lecture, you should be able to:

- *use* Jackson's Theorem to specify the equilibrium distribution of Open Jackson Networks;
- *state* and *prove* relationships between observed distributions by event streams; and,
- *state* PASTA theorem.

Example 17. Three single-server queues with feed-back



Consider a network of three single-server queues, with

- service rates μ_1, μ_2 and μ_3 at each respective queue when there are customers,
- exogenous arrival rates are marked (in red) as rates 2λ and λ,
- transition probabilities between nodes and those which leave the system are also marked.

Here, we have

$$\begin{aligned}
\text{exogenous arrival rates} \quad & \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) = (2\lambda, \lambda, 0), \\
\text{traffic routing probabilities} \quad & \gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{bmatrix} = \begin{bmatrix} 0 & 1/2 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 \end{bmatrix}, \\
\text{network departure probabilities} \quad & \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 0 \\ 1/2 \end{bmatrix}.
\end{aligned}$$

Denote by $\boldsymbol{y} = (y_1, y_2, y_3)$ total average throughput rates; then, the traffic equations are $\boldsymbol{y} = \boldsymbol{\lambda} + \boldsymbol{y\gamma}$, or

$$y_1 = 2\lambda + \frac{1}{2}y_2 + \frac{1}{2}y_3, \tag{21}$$

$$y_2 = \lambda + \frac{1}{2}y_1, \tag{22}$$

$$y_3 = \frac{1}{4}y_1 + \frac{1}{2}y_2. \tag{23}$$

We can solve for each y_i, for example by substituting (22) into (21) and (23) to get

$$y_1 = 2\lambda + \frac{\lambda}{2} + \frac{1}{4}y_1 + \frac{1}{2}y_3 \quad \text{and} \quad y_3 = \frac{1}{4}y_1 + \frac{\lambda}{2} + \frac{1}{4}y_1.$$

These imply

$$y_1 = 2\lambda + \frac{\lambda}{2} + \frac{1}{4}y_1 + \frac{1}{4}y_1 + \frac{\lambda}{4} \Rightarrow \quad y_1 = \frac{11\lambda}{2}, \quad y_2 = \frac{15\lambda}{4}, \quad y_3 = \frac{13\lambda}{4}$$

and so

$$\begin{aligned}
\pi(\boldsymbol{n}) &= \pi(n_1, n_2, n_3) \\
&= \pi_1(n_1)\pi_2(n_2)\pi_3(n_3) \\
&= \left(1 - \frac{11\lambda}{2\mu_1}\right)\left(\frac{11\lambda}{2\mu_1}\right)^{n_1}\left(1 - \frac{15\lambda}{4\mu_2}\right)\left(\frac{15\lambda}{4\mu_2}\right)^{n_2}\left(1 - \frac{13\lambda}{4\mu_3}\right)\left(\frac{13\lambda}{4\mu_3}\right)^{n_3}, \\
&\text{iff } \lambda < \min\left\{\frac{2\mu_1}{11}, \frac{4\mu_2}{15}, \frac{4\mu_3}{13}\right\}.
\end{aligned}$$

Recall David Kendall's notation for queues: A/B/n/m, where A describes the arrival process, B describes the service time distribution, n is the number of servers, m is the number of waiting and service spaces (capacity) = the maximum number of customers that may occupy the system.

A and B can have many forms, some of which are M (Markov or memoryless, i.e., exponential distributions), G (General), and D (Deterministic).

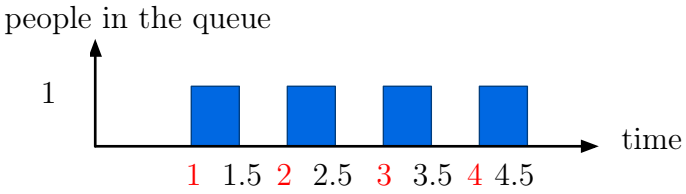
Equilibrium Distributions as seen by Arrivals

Recall that the equilibrium probabilities π_j can be interpreted as either

1. **Ergodic:** The long term proportion of time the system is in state j, or
2. **Limiting:** The probability that, at an arbitrary time point (in equilibrium $\equiv$ far away from any initial conditions), the system is in state j.

Waiting times for customers depend not on 1. or 2., but on the number of customers in the system when the customer arrives. The distribution of this number can differ from that reflected in 1. and 2.

Example 18. D/D/1 queue



Consider a $D/D/1$ queue with inter-arrival time 1 and service time 1/2.

The queue is occupied for half of the time, and so, if the queue is observed at a random moment, the probability that it will have one customer is 1/2:

$$\Rightarrow \Pr\{ \text{ in equilibrium, the system has no customers in it } \} = 1/2.$$

However, arriving customers do not observe the queue at random instants, but come exactly one unit of time after the last customer. In fact, an arriving customer always sees an empty queue:

$$\Rightarrow \Pr\{ \text{ an arriving customer sees no customers } \} = 1.$$

The above simple example shows that arrival time distributions are not always the same as equilibrium distributions.

Questions: Under what circumstances are they the same? If not, when can the arrival time distribution be calculated from the equilibrium distribution?

Let $\mathcal{X}$ be any CTMC with state space $\mathcal{S}$, $P_j(t)$ be the probability that the chain is in state j at time t, and γ_j be the intensity of an event stream which occurs when $\mathcal{X}$ is in state j. These *events* may, or may not, change the state of $\mathcal{X}$. For example they may be arrival or departure points, which do change the state or they may simply be an arbitrary set of points, where no state change is occurring.

We want to observe the state of the Markov Chain at these event points.

Denote by $\{\pi_j^{(E)}, j \in \mathcal{S}\}$ the distribution in equilibrium just before particular event points, and $P_j^{(E)}(t)$ the time-dependent equivalent. We can find general relationships between $\{\pi_j^{(E)}, j \in \mathcal{S}\}$ and $\{\pi_j, j \in \mathcal{S}\}$, and between $P_j^{(E)}(t)$ and $P_j(t)$.

Theorem 21.

$$P_j^{(E)}(t) = \frac{\gamma_j P_j(t)}{\sum_{k\in\mathcal{S}} \gamma_k P_k(t)} \quad \textit{and} \quad \pi_j^{(E)} = \frac{\gamma_j \pi_j}{\sum_{k\in\mathcal{S}} \gamma_k \pi_k}.$$

Proof. Let $X(t)$ be the random variable representing the number in the system at time t.

$$\begin{aligned}
P_j^{(E)}(t) &= \Pr\left(\text{an event, which occurs at time } t, \text{“sees” state } j\right) \\
&= \lim_{h\to 0} \Pr(X(t) = j | \text{ event occurs in } (t, t+h)) \\
&= \lim_{h\to 0} \frac{\Pr(X(t) = j \cap \text{an event in } (t, t+h))}{\Pr(\text{an event in } (t, t+h))} \\
&= \lim_{h\to 0} \frac{\Pr(\text{an event in } (t, t+h) \mid X(t) = j) \Pr(X(t) = j)}{\sum_{k\in\mathcal{S}} \Pr(\text{event in } (t, t+h) \mid X(t) = k) \Pr(X(t) = k)} \\
&= \lim_{h\to 0} \frac{(\gamma_j h + o(h))P_j(t)}{\sum_{k\in\mathcal{S}} (\gamma_k h + o(h))P_k(t)} \\
&= \frac{\gamma_j P_j(t)}{\sum_{k\in\mathcal{S}} \gamma_k P_k(t)}.
\end{aligned}$$

Now, taking the limit as $t \to \infty$ gives

$$\begin{aligned}
\lim_{t\to\infty} \text{LHS} &= \lim_{t\to\infty} P_j^{(E)}(t) = \pi_j^{(E)}, \\
\lim_{t\to\infty} \text{RHS} &= \lim_{t\to\infty} \frac{\gamma_j P_j(t)}{\sum_{k\in\mathcal{S}} \gamma_k P_k(t)} = \frac{\gamma_j \pi_j}{\sum_{k\in\mathcal{S}} \gamma_k \pi_k}.
\end{aligned}$$

□

The next theorem is generally referred to as PASTA, meaning **P**oisson **A**rrivals **S**ee **T**ime **A**verages, but we present it in a more general setting. The theorem tells us that any Poisson stream, whether they are arrivals or not, sees the distribution of the Markov Chain exactly the same as if they watched it over all time.

In particular, the equilibrium distribution at the Poisson time points is exactly the same as the distribution found from the equilibrium equations (that is, the time averaged distribution).

Theorem 22 (PASTA)**.** *If $\gamma_j = \lambda$ for all $j \in \mathcal{S}$ and $P_j(t) > 0$, then*

$$P_j^{(E)}(t) = P_j(t) \quad \text{and} \quad \pi_j^{(E)} = \pi_j \quad \text{for all } j \in \mathcal{S}.$$

Proof. We have already established that

$$P_j^{(E)}(t) = \frac{\lambda P_j(t)}{\lambda \sum_{k\in\mathcal{S}} P_k(t)} \quad \text{for all } j \in \mathcal{S}$$

and since $\sum_{k\in S} P_k(t) = 1$, we have that $P_j^{(E)}(t) = P_j(t)$ for all $j \in \mathcal{S}$. Taking the limits as $t \to \infty$ of the above gives $\pi_j^{(E)} = \pi_j$ for all $j \in \mathcal{S}$. □

Note that we have not assumed that the Poisson stream of arrivals necessarily changes the state of the process when it occurs. Very often this is the case, such as for the $M/M/N$ queue, but for the $M/M/N/N$ queue, for example, the Poisson input keeps happening when the system is full even though they are lost.

By the above result, however, we see that the arrival stream to an $M/M/N/N$ queue also has the PASTA property. Consequently,

$$\begin{aligned}\text{the proportion of calls lost} &= \Pr(\text{an arrival is rejected})\\ &= \Pr\{\text{an arrival sees } N \text{ in the system}\} = \pi_N^{(E)}\\ &= \pi_N \quad \text{by PASTA.}\end{aligned}$$

Earlier in this course, we **assumed** that the probability that a call is lost is π_N but, as noted above, this result relies on the PASTA property and is not usually true when the arrival stream is not Poisson.

Example 20. The Birth and Death Process

Let us now consider the birth and death process and how the new arrivals observe the state of the process. Let $\{\pi_j^{(A)}, j \geq 0\}$ be the equilibrium distribution just before arrival points.

Theorem 23. *For an arbitrary birth and death process the equilibrium distribution as seen by those arrivals that change the state of the process can be written as*

$$\pi_j^{(A)} = \pi_0^{(A)} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell}, \quad \textit{for all } j \geq 0.$$

Note: this is not the same as $\pi_j = \pi_0 \prod_{\ell=1}^{j} \frac{\lambda_{\ell-1}}{\mu_\ell}$.

Proof. Set $\gamma_j = \lambda_j$ for all $j \geq 0$ and then we have by Theorem 21

$$\begin{aligned}\pi_j^{(A)} &= \frac{\lambda_j \pi_j}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} = \frac{\lambda_j \pi_0}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} \prod_{\ell=0}^{j-1} \frac{\lambda_\ell}{\mu_{\ell+1}}\\ &= \frac{\lambda_0 \pi_0}{\sum_{k\in\mathcal{S}} \lambda_k \pi_k} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell} = \pi_0^{(A)} \prod_{\ell=1}^{j} \frac{\lambda_\ell}{\mu_\ell}.\end{aligned}$$

□

Note that if the arrival rate is a constant λ for all j then we have $\pi_j^{(A)} = \pi_j$ (by PASTA).

Example 21. $M/M/N$ queue

We assume that the queue is operating under a first come first served (FCFS) or first in first out (FIFO) discipline. So, the customer will have to wait whenever it arrives to find greater than or equal to N customers already in the system.

(a) **What is the probability that an arriving customer does not have to wait?**

Let W_Q be the random variable for the equilibrium waiting time in the queue (not including the service time). Then, the event $W_Q = 0$ is the event that an arriving customer sees less than N customers in the queue. Hence,

$$\begin{aligned}\Pr(W_Q = 0) &= \sum_{j=0}^{N-1} \pi_j^{(A)} = \sum_{j=0}^{N-1} \pi_j \quad \text{by PASTA} \\ &= \pi_0 \sum_{j=0}^{N-1} \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!}.\end{aligned}$$

(b) **What is the distribution of the customer's waiting time?**

If the customer arrives to find $N + k$ in the system it has to wait for $k + 1$ services before they are served. During this period, all servers are busy and so the service rate is $N\mu$.

The distribution of the waiting time in the $M/M/N$ queue for a customer that arrives to find $N + k$ in the system is the distribution of time until the $(k + 1)st$ event when the event rate is $N\mu$. The density function for this is the Erlang density function. Thus,

$$\begin{aligned}f_{W_Q}(t) = \frac{\mathrm{d}\Pr(W_Q < t)}{\mathrm{d}t} &= \sum_{k=0}^{\infty} \Pr\begin{pmatrix}\text{arrival finds } N+k \\ \text{in the system}\end{pmatrix} \times \begin{pmatrix}\text{density function} \\ \text{for Erlang}(k+1, N\mu)\end{pmatrix} \\ &= \sum_{k=0}^{\infty} \left(\pi_{N+k}^{(A)}\right) \times \left(N\mu \frac{(N\mu t)^k}{k!} e^{-N\mu t}\right) \\ &= \sum_{k=0}^{\infty} \left(\left(\frac{\lambda}{N\mu}\right)^k \pi_N\right) \times \left(N\mu \frac{(N\mu t)^k}{k!} e^{-N\mu t}\right) \\ &= N\mu e^{-N\mu t} \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} \pi_N \\ &= N\mu e^{(\lambda - N\mu)t} \pi_N.\end{aligned}$$

We can replace π_N in the previous expression by

$$\pi_N = \left(1 - \frac{\lambda}{N\mu}\right) C\left(N, \frac{\lambda}{\mu}\right),$$

where

$$C\left(N, \frac{\lambda}{\mu}\right) = \Pr(\text{all servers are busy}) = \sum_{j=N}^{\infty} \pi_j.$$

Using this definition of $C\left(N, \frac{\lambda}{\mu}\right)$ (known as the Erlang C formula), we can write

$$f_{W_Q}(t) = \frac{\mathrm{d}\Pr(W_Q < t)}{dt} = C\left(N, \frac{\lambda}{\mu}\right)(N\mu - \lambda)e^{-(N\mu-\lambda)t}.$$

Note that $(N\mu - \lambda)e^{-(N\mu-\lambda)t}$ is an exponential density function with parameter $(N\mu - \lambda)$, which is the difference between the maximum service rate and the arrival rate.

The last equation gives the density function for the equilibrium waiting time of a customer that arrives to an $M/M/N$ queue and by integrating this from t to ∞ we obtain the probability that the waiting time is greater than t. That is,

$$\begin{aligned} 1 - F_{W_Q}(t) = \Pr(W_Q > t) &= C\left(N, \frac{\lambda}{\mu}\right)\int_t^\infty (N\mu - \lambda)e^{-(N\mu-\lambda)u}\mathrm{d}u \\ &= C\left(N, \frac{\lambda}{\mu}\right)e^{-(N\mu-\lambda)t}. \end{aligned}$$

(c) **What is the mean waiting time of an arriving customer?**

$$\begin{aligned} \mathbb{E}[W_Q] &= C\left(N, \frac{\lambda}{\mu}\right) \times \begin{pmatrix} \text{Mean of an exponential random} \\ \text{variable with parameter } (N\mu - \lambda) \end{pmatrix} \\ &= \frac{C\left(N, \frac{\lambda}{\mu}\right)}{N\mu - \lambda}. \end{aligned}$$

(d) **What is the conditional waiting time of an arriving customer, given that they have to wait?**

$$\begin{aligned} \Pr(W_Q > t \mid W_Q > 0) &= \frac{\Pr(W_Q > t, W_Q > 0)}{\Pr(W_Q > 0)} \\ &= \frac{\Pr(W_Q > t)}{\Pr(W_Q > 0)} \\ &= \frac{C\left(N, \frac{\lambda}{\mu}\right)e^{-(N\mu-\lambda)t}}{C\left(N, \frac{\lambda}{\mu}\right)} \\ &= e^{-(N\mu-\lambda)t}. \end{aligned}$$

(e) **What is conditional expectation of waiting time, given that they have to wait?**

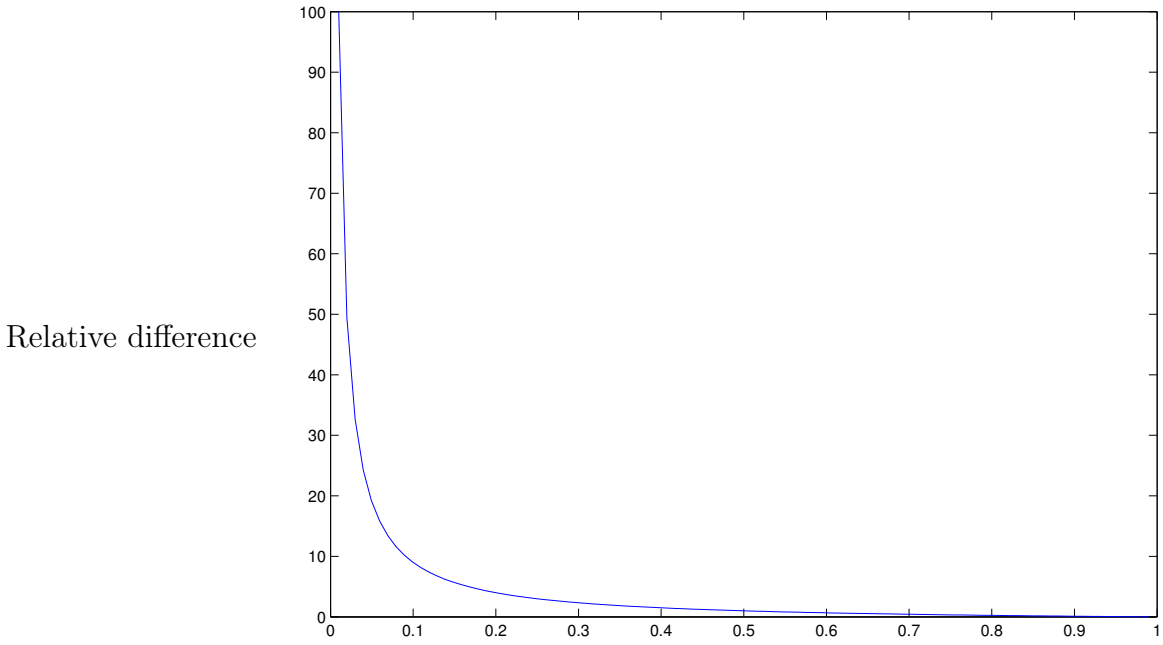
$$\mathbb{E}[W_Q \mid W_Q > 0] = \begin{pmatrix} \text{Mean of an exponential random} \\ \text{variable with parameter } (N\mu - \lambda) \end{pmatrix} = \frac{1}{N\mu - \lambda}.$$

For a single server ($N = 1$), we have

$$\mathbb{E}[W_Q] = \frac{\lambda}{\mu(\mu - \lambda)} \quad \text{and} \quad \mathbb{E}[W_Q \mid W_Q > 0] = \frac{1}{\mu - \lambda}.$$

For a single server ($N = 1$), where $\mu = 1$, plotting the relative difference

$$\frac{\mathbb{E}[W_Q \mid W_Q > 0] - \mathbb{E}[W_Q]}{\mathbb{E}[W_Q]} \quad \text{against } 0 < \lambda < 1 \quad \text{yields}$$



Notice for low offered load that the expected waiting time can be such that the conditional waiting time, given that you have to wait, may be hundreds of times the expected waiting time.

Lecture 24: Observed distributions in CTMCs – Little's Law and Pollaczek-Khinchin mean value formulae

Concepts checklist

At the end of this lecture, you should be able to:

- *state* and *apply* Little's Law; and,
- *state* and *apply* the Pollaczek-Khinchin mean value formulae.

Little's Law

For the $M/M/N$ queue, there is a relation between the average number of waiting customers in the queue and the mean waiting time. The average number of waiting customers (not in service) $\mathbb{E}[L_Q]$ is given by

$$\mathbb{E}[L_Q] = \lambda \mathbb{E}[W_Q]. \tag{24}$$

Equality (24), which is an example of Little's Law, holds in systems much more complex than the $M/M/N$ queue. To see this, let Γ be any section of a queueing system. Assume that the queueing system is stationary (in equilibrium) and let

- $\overline{L}(\Gamma)$ be the average number of customers in Γ,
- $\overline{W}(\Gamma)$ be the average time a customer spends in Γ and
- $\overline{\lambda}(\Gamma)$ be the average number of customers entering Γ per unit time.

Then Little's Law gives a relationship between $\overline{L}(\Gamma)$, $\overline{W}(\Gamma)$ and $\lambda(\Gamma)$.

Theorem 24 (Little's Law.)**.** *In equilibrium,*

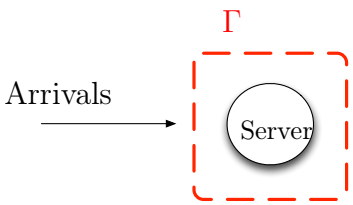
$$\overline{L}(\Gamma) = \overline{\lambda}(\Gamma)\overline{W}(\Gamma). \tag{25}$$

Example 22. $M/G/1/1$ queue

Consider a single-server queue with general service times with mean $1/\mu$, and assume that blocked calls are lost.

Goal: determine the probability that the server is busy.

Consider Γ to be the server:



The input rate $\overline{\lambda}(\Gamma)$ into Γ is $\lambda(1 - \Pr(\text{server busy}))$.

The average time $\overline{W}(\Gamma)$ in Γ is given by $\overline{W} = 1/\mu$.

The average number $\overline{L}(\Gamma)$ of customers in service is given by

$$\overline{L} = 1 \times \Pr(\text{server busy}) + 0 \times (1 - \Pr(\text{ server busy})).$$

By Little's Law, we have

$$\Pr(\text{server busy}) = \frac{\lambda(1 - \Pr(\text{server busy}))}{\mu},$$

which implies

$$\Pr\left(\text{server busy}\right) = \frac{a}{1+a}, \quad \text{where} \quad a = \frac{\lambda}{\mu}.$$

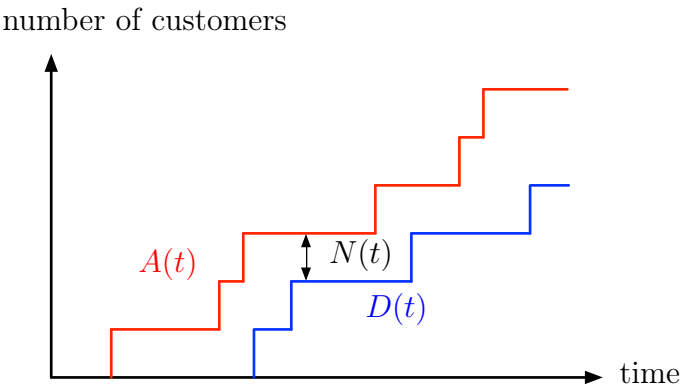
Comparing this with the exponential service time case we see that

$$\Pr\left(\text{server busy}\right) = B(1, a) = \frac{a}{1+a}.$$

A more general result can be shown to apply:

The equilibrium distribution for the $M/G/N/N$ system, including the probability that a call is blocked can be shown to depend on the service time distribution only through its mean. That is, the exponential distribution with mean $1/\mu$ gives the correct equilibrium distribution for an $M/G/N/N$ queue with a general service time distribution with mean $1/\mu$.

Proof. Let $A(t)$ be the number of arrivals to Γ in $(0, t)$, $D(t)$ be the number of departures from Γ in $(0, t)$, and $N(t) = A(t) - D(t)$ be the number in Γ at time t.



On the time interval $[0, t)$, let $\overline{\lambda_t}(\Gamma)$ be the mean arrival rate in $[0, t)$, then

$$\overline{\lambda_t}(\Gamma) = A(t)/t.$$

Let $S(t)$ be the total area between curves $A(t)$ and $D(t)$ up to time t, then

$$S(t) = \int_0^t N(u)du = \text{ cumulative time spent in } \Gamma \text{ up to time } t.$$

Then,

$$\overline{W}_t(\Gamma) = \text{ mean time each customer spends in } \Gamma = \frac{S(t)}{A(t)}$$

$$\text{and} \quad \overline{L}_t(\Gamma) = \text{ average number of customers in } \Gamma \text{ during } [0, t) = \frac{S(t)}{t}.$$

Hence,

$$\overline{L}_t(\Gamma) = \overline{\lambda}_t(\Gamma)\overline{W}_t(\Gamma).$$

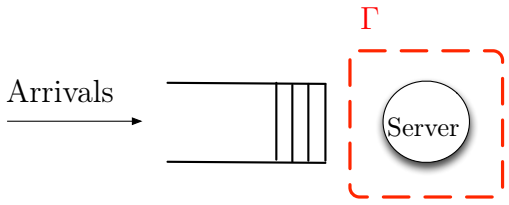
Letting $t \to \infty$, we have $\overline{L}(\Gamma) = \overline{\lambda}(\Gamma)\overline{W}(\Gamma)$ as required. □

Note: We have made no assumption about Poisson arrivals, correlations between customers behaviour, the number of servers, exponential service times etc. The queueing system is completely arbitrary and yet the expression is valid.

Example 23. $M/G/1$ queue

Consider a single-server queue with general service times and an infinite capacity.

Goal: to determine the probability that the server is busy.



Here, we have $\overline{W} = 1/\mu$ and the input rate into $\Gamma = \lambda$. Using Little's law,

$$\overline{L} = \Pr\,(\text{server busy}) = \lambda\overline{W} = a.$$

Therefore, the probability π_0 that the $M/G/1$ system is empty, is $\pi_0 = 1 - a$.

Pollaczek-Khinchin mean value formulae for an arbitrary $M/G/1$ queue

Let q_n be the number of customers left behind by the departure of the nth customer of an $M/G/1$ queue, and let v_n be the number of customers that arrive during the service of the nth customer.

Then we have

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & \text{if } q_n > 0 \\ v_{n+1} & \text{if } q_n = 0 \end{cases},$$

which can be combined to get

$$q_{n+1} = q_n - \mathbb{1}_{\{q_n>0\}} + v_{n+1}, \quad \text{where } \mathbb{1}_{\{q_n>0\}} = \begin{cases} 1 & \text{if } q_n > 0 \\ 0 & \text{if } q_n = 0 \end{cases}.$$

Now, take the expectations of v_n and q_n. Assume that we can take limits as $n \to \infty$ and that $\mathbb{E}[v_n] \to \mathbb{E}[v]$ and $\mathbb{E}[q_n] \to \mathbb{E}[q]$. Then, from before we get

$$\begin{aligned} \mathbb{E}[q] &= \mathbb{E}[q] - \mathbb{E}[\mathbb{1}_{\{q>0\}}] + \mathbb{E}[v], \\ \Rightarrow \mathbb{E}[\mathbb{1}_{\{q>0\}}] &= \mathbb{E}[v]. \end{aligned}$$

On the other hand, as

$$\mathbb{E}[\mathbb{1}_{\{q>0\}}] = 0.\Pr\{q = 0\} + 1.\Pr\{q > 0\} = \Pr\{q > 0\},$$

so we have $\Pr(q > 0) = \mathbb{E}[v]$. In words, the probability that the queue is non-empty at a departure instant is equal to the expected number of arrivals in a service.

Note: if $\mathbb{E}[v] > 1$ there will, on average, be more than one arrival in each service, the queue will be unstable, and the above assumptions about the legality of taking limits will not hold.

We get more information from $q_{n+1} = q_n - \mathbb{1}_{\{q_n>0\}} + v_{n+1}$ by squaring it:

$$q_{n+1}^2 = q_n^2 + \mathbb{1}_{\{q_n>0\}}^2 + v_{n+1}^2 - 2q_n\mathbb{1}_{\{q_n>0\}} + 2q_n v_{n+1} - 2\mathbb{1}_{\{q_n>0\}}v_{n+1}.$$

Note that $\mathbb{1}_{\{q_n>0\}}^2 = \mathbb{1}_{\{q_n>0\}}$ and $q_n\mathbb{1}_{\{q_n>0\}} = q_n$. Then, by the independence of the number of arrivals and the queue length we have

$$\mathbb{E}[q_n v_{n+1}] = \mathbb{E}[q_n]\mathbb{E}[v_{n+1}] \quad \text{and} \quad \mathbb{E}[\mathbb{1}_{\{q_n>0\}}v_{n+1}] = \mathbb{E}[\mathbb{1}_{\{q_n>0\}}]\mathbb{E}[v_{n+1}].$$

We then take expectations, and use the above result that $\mathbb{E}[\mathbb{1}_{\{q>0\}}] = \mathbb{E}[v]$ to show that

$$\begin{aligned} \mathbb{E}[q^2] &= \mathbb{E}[q^2] + \mathbb{E}[v] + \mathbb{E}[v^2] - 2\mathbb{E}[q] + 2\mathbb{E}[q]\mathbb{E}[v] - 2\mathbb{E}[v]^2 \\ \Rightarrow \quad \mathbb{E}[q] &= \mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])}. \end{aligned}$$

It can be argued, using a graph of arrivals and departures, that the departure points leaving n customers in the queue can be matched one-to-one with the arrival points which find n customers in the queue and therefore that the distribution as left behind by departures is the same as the distribution seen by arrivals.

We can therefore appeal to PASTA to conclude that

$$\mathbb{E}[q] = \mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])} \tag{26}$$

is also the mean queue length at any arbitrary time.

Using Little's Law, we can further get an expression for the mean time spent by a customer in the queue as

$$\mathbb{E}[w] = \frac{1}{\lambda}\left[\mathbb{E}[v] + \frac{\mathbb{E}[v^2] - \mathbb{E}[v]}{2(1-\mathbb{E}[v])}\right]. \tag{27}$$

Now, using $\Pr(v = k) = \displaystyle\int_0^\infty \frac{(\lambda t)^k}{k!}e^{-\lambda t}\mathrm{d}B(t)$ — k arrivals during a service time — we can show that

$$\mathbb{E}(v^2) - \mathbb{E}(v) = \lambda^2\mathbb{E}[Y^2],$$

where $\mathbb{E}[Y^2]$ is the second (non-central) moment of the service time distribution $B(t)$.

Equations (26) and (27) can now be written in the equivalent forms

$$\mathbb{E}[q] = a + \frac{\lambda^2\mathbb{E}[Y^2]}{2(1-a)}, \tag{28}$$

$$\text{and} \quad \mathbb{E}[w] = \frac{1}{\lambda}\left[a + \frac{\lambda^2\mathbb{E}[Y^2]}{2(1-a)}\right]. \tag{29}$$

Equations (28) and (29) are known as the *Pollaczek-Khinchin mean value formulae* for the queue length and waiting time respectively. They give expressions for the mean queue length and the mean waiting time for an arbitrary $M/G/1$ queue.

Lecture 25: Point processes and renewal processes

Concepts checklist

At the end of this lecture, you should be able to:

- *define* a stationary Poisson process; and,
- *define* a Renewal Process.

1870 – Seidel considered the occurrence of thunderstorms.

1889 – Von Bortkiewicz gave a systematic account of phenomena that fit the Poisson distribution; the most famous example being that of the number of deaths from horse kicks in the Prussian cavalry.

1943 – Palm systematically described (as a generalisation of the Poisson process) the input to a service system. *A powerful concept:* the notion of a regeneration point or a time instant at which the system reverts to a specified state with the property that the future evolution is independent of how that state was reached.

$\Rightarrow$ Poisson process is characterised by the property that every instant is a regeneration point, which is different to other processes where the commencement of a new inter-event time is the only regeneration point.

Two distributions are necessary for describing a stationary point process:

1. the distribution of time to the next event from an arbitrary time, and
2. the distribution of time to the next event from an arbitrary event of the process.

Ideal examples of point processes: the Poisson process and the renewal processes.

By a point process, we refer to some method of randomly allocating points to intervals of the real line (this can easily be extended to hyper-rectangles in n-dimensional Euclidean space).

A basic approach to point processes: to consider counting the number of events in intervals or regions of various types.

Definition 20. *The stationary Poisson process on the line is completely defined by the following equation,*

$$\Pr\left(N(a_i, b_i] = n_i,\ i = 1, \ldots, k\right) = \prod_{i=1}^{k} \frac{[\lambda(b_i - a_i)]^{n_i}}{n_i!} e^{-\lambda(b_i - a_i)}. \tag{30}$$

where $N(a_i, b_i]$ is the number of events in the open interval $(a_i, b_i]$ for $a_i < b_i \leq a_{i+1}$.

Properties:

1. the number of events in each finite interval $(a_i, b_i]$ is Poisson distributed.
2. the number of points in disjoint intervals are independent random variables.
3. the process is stationary as the distributions only depend on the lengths $b_i - a_i$ of the intervals, not the value of the actual end-points.

By (30), the mean $M(a,b]$ and variance $V(a,b]$ of the number of events in $(a,b]$ are given by

$$M(a,b] = \lambda(b-a) = V(a,b]. \tag{31}$$

The parameter λ can therefore be interpreted as the mean rate or mean density of points of the process.

Furthermore, $\Pr(N(0,\tau] = 0) = e^{-\lambda\tau}$, which is the probability of finding no points in the interval of length τ. This probability may also be interpreted as

1. the probability that the random interval extending from the origin to the first point to the right has a length exceeding τ, or
2. the survivor function for the length of this interval,

and shows that the interval under consideration has an exponential distribution.

From stationarity, the same result applies to

- the length of the interval to the next point immediately to the right from any arbitrarily chosen origin, and
- the length of the interval to the first point immediately to the left of any arbitrarily chosen origin.

In queueing terms, these are called the *forward and backward recurrence times*.

Hence, for a Poisson process the forward and backward recurrence times have an exponential distribution with parameter λ. Using the independence property, this distribution extends to the distribution of time between two consecutive points in the process, and it may be shown that successive intervals are independently distributed with exponential distributions.

Let t_k be the time from the origin $t_0 = 0$ until the kth point of the point process to the right of the origin, then the following events are equivalent: $\{t_k \geq x\}$ and $\{N(0,x] \leq k\}$, so that their respective probabilities are the same, with the probability of the latter being given by equation (30). That is,

$$\Pr(t_k > x) = \Pr(N(0,x] < k) = \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!} e^{-\lambda x},$$

$$\text{or} \quad \Pr(t_k \leq x) = 1 - \Pr(N(0,x] < k) = 1 - \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!} e^{-\lambda x}.$$

Differentiating this expression we get the corresponding density

$$\begin{aligned} f_k(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\Pr(t_k \le x) \\ &= \lambda\sum_{j=0}^{k-1}\frac{(\lambda x)^j}{j!}e^{-\lambda x} - \lambda\sum_{j=0}^{k-1}\frac{j(\lambda x)^{j-1}}{j!}e^{-\lambda x} \\ &= \lambda\sum_{j=0}^{k-1}\frac{(\lambda x)^j}{j!}e^{-\lambda x} - \lambda\sum_{j=0}^{k-1}\frac{(\lambda x)^{j-1}}{(j-1)!}e^{-\lambda x} \\ &= \lambda\sum_{j=0}^{k-1}\frac{(\lambda x)^j}{j!}e^{-\lambda x} - \lambda\sum_{j=0}^{k-2}\frac{(\lambda x)^j}{j!}e^{-\lambda x} \\ &= \lambda\frac{(\lambda x)^{k-1}}{(k-1)!}e^{-\lambda x}. \end{aligned}$$

This is the Erlang density, which is therefore the sum of the lengths of the k random intervals $(t_0, t_1]$, $(t_1, t_2]$, $\ldots$, $(t_{k-1}, t_k]$, which are independently and identically distributed according to the exponential distribution with parameter λ. This gives an indirect proof that the result of the sum of k independent exponential random variables has the Erlang distribution.

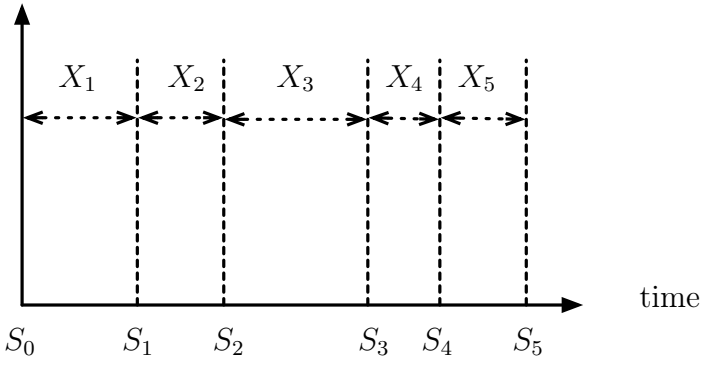
We have defined and analysed the Poisson process, which assumes that points arrive in a memoryless stream. This assumption is reasonable in cases where a large population all have a small chance of generating a point. For example, in communications, this might be requesting a particular route in a communications network. No matter how many requests have accrued in the recent past there are still so many potential callers in the population that future calls arrive independently of the past history.

However, the Poisson process is not ideal for many cases, where the inter-event time distribution is not exponential.

Renewal Processes

The class of renewal processes can be viewed as a natural generalisation of the Poisson process:

- In a Poisson process, the state moves from i to $i+1$ at a jump and the times X_i between jumps are independent and exponentially distributed.
- A renewal process is similar, as the times X_i between jumps are independent and identically distributed, but they are not necessarily exponentially distributed.



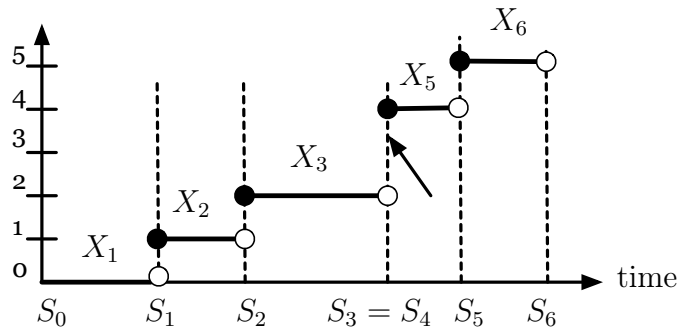


Figure 5: Counting and waiting time processes

Definition 21 (Renewal Processes.). *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) non-negative random variables with*

$$\Pr\{X_i \le x\} = F(x) \quad \text{if } x \ge 0.$$

We assume that $F(0) < 1$ and that $\mathbb{E}[X_i] = \mu < \infty$. Let

$$\begin{aligned} S_n &= X_1 + X_2 + \cdots + X_n \quad \text{(with } S_0 = 0\text{)} \\ &= \text{waiting time to the } n\text{th event} \\ \text{and} \quad N(t) &= \sup\{n : S_n \le t\} \\ &= \text{number of events before time } t. \end{aligned}$$

Then, $\{N(t), t > 0\}$ is called the counting process and $\{S_n, n \ge 1\}$ is called the waiting time process. The process $\{N(t)\}$ is referred to as a renewal process.

Note: If $F(0) > 0$, then two events can occur simultaneously. This is reflected in Figure (5).

Examples of renewal processes.

1. Poisson process (X_i's are exponential).
2. Replacement models: where a component with lifetime distribution $F(x)$ is replaced every time it breaks down (e.g., lightbulbs, or Professors!). Then the number of replacements before time t gives the counting process and the time until the nth replacement gives the waiting time process.
3. Breakdown and repair models: where the event times could be either
   (a) times of breakdown
   (b) times of repair.

   In both cases, the inter-event time random variable is the sum of a breakdown random variable and a repair random variable.
4. Sometimes renewal processes can be "embedded" in more complex stochastic processes. For example, consider a queueing system with exponential inter-arrival times, but general (non-memoryless) service times. Then, consider the points at which successive idle periods begin. The system is memoryless at these points and the times between them are iid, even though this distribution may be very complicated. These are the regeneration points mentioned at the start of the lecture.

Lecture 26: Preliminary materials – Riemann-Stieltjes Integration

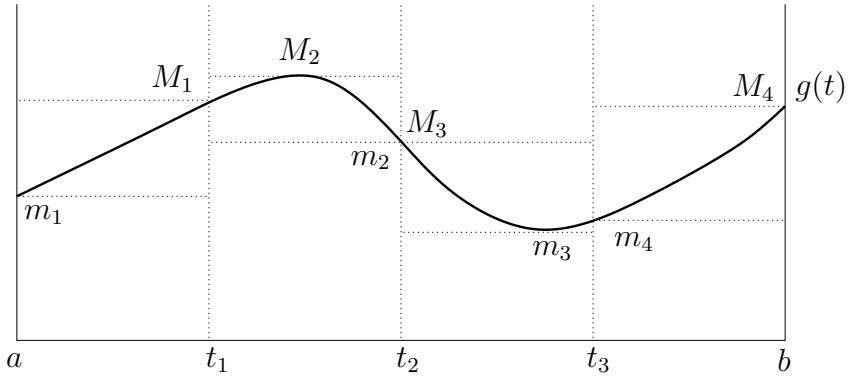
Concepts checklist

At the end of this lecture, you should be able to:

- *Understand* how the Riemann-Stieltjes integral generalises the Riemann integral; and,
- *evaluate* Riemann-Stieltjes integrals.

Riemann-Stieltjes Integration

In Maths I Calculus, you used the Riemann integral to calculate the area under some function $g(t)$ on some finite interval $[a, b]$. The formal definition of the Riemann integral is in terms of upper and lower sums of rectangular areas. That is, we partition $[a, b]$ into n sub-intervals with end points $a = t_0 < t_1 < \cdots < t_n = b$, and let M_i and m_i be the supremum and infimum of $g(t)$ on sub-interval i.



We know that

$$\sum_{i=1}^{n} m_i \left(t_i - t_{i-1}\right) \leq \text{ area under } g(t) \leq \sum_{i=1}^{n} M_i \left(t_i - t_{i-1}\right),$$

where

- the LHS is known as the lower (Riemann) sum with respect to the partition $\mathcal{P} = (t_0, t_1, \ldots, t_n)$, and
- the RHS is the upper (Riemann) sum with respect to $\mathcal{P}$.

Definition 22. *The Riemann integral exists if*

$$\sup_{\text{all partitions } \mathcal{P}} \sum_{i=1}^{n} m_i \left(t_i - t_{i-1}\right) = \inf_{\text{all partitions } \mathcal{P}} \sum_{i=1}^{n} M_i \left(t_i - t_{i-1}\right) = \ell,$$

in which case we write $\int_a^b g(t)\mathrm{d}t = \ell$.

Definition 23. *A function $f(t)$ is said to be right-continuous at a point τ, if for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$|f(t) - f(\tau)| < \varepsilon \quad \textit{for all} \quad \tau < t < \tau + \delta.$$

That is,

$$\lim_{t \to \tau^+} f(t) = f(\tau).$$

Definition 24. *A right-continuous function is a function that is right-continuous at all points.*

Definition 25. *A function $f(t)$ is monotonically non-decreasing if*

$$f(x) \leq f(y) \quad \textit{for all} \quad x < y.$$

Distribution functions are examples of monotonically non-decreasing functions.

Note: Similar definitions exist for left-continuous and monotonically non-increasing functions.

If $\alpha(t)$ is a right-continuous, monotonically non-decreasing function on the interval $[a, b]$, define the lower and upper sums with respect to a partition $\mathcal{P} = \{t_0, t_1, \ldots, t_n\}$ (where $t_0 = a$ and $t_n = b$) and the function $\alpha(t)$ as

$$\sum_{i=1}^{n} m_i(\alpha(t_i) - \alpha(t_{i-1})) \quad \text{and} \quad \sum_{i=1}^{n} M_i(\alpha(t_i) - \alpha(t_{i-1})).$$

Definition 26. *The Riemann-Stieltjes integral of $g(t)$ with respect to $\alpha(t)$ over $[a, b]$ exists if*

$$\sup_{\textit{all partitions } \mathcal{P}} \sum_{i=1}^{n} m_i\big(\alpha(t_i) - \alpha(t_{i-1})\big) = \inf_{\textit{all partitions } \mathcal{P}} \sum_{i=1}^{n} M_i\big(\alpha(t_i) - \alpha(t_{i-1})\big) = \ell,$$

in which case we write

$$\int_a^b g(t) \mathrm{d}\alpha(t) = \ell.$$

If $\alpha(t)$ is *differentiable* on $[a, b]$ then

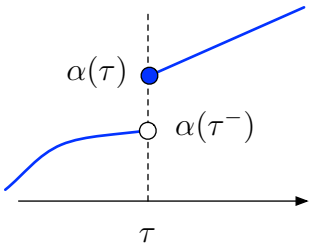
$$\int_a^b g(t) \mathrm{d}\alpha(t) = \int_a^b g(t) \frac{\mathrm{d}\alpha(t)}{\mathrm{d}t} \mathrm{d}t.$$

That is, the Riemann-Stieltjes integral is just the ordinary Riemann integral of the function

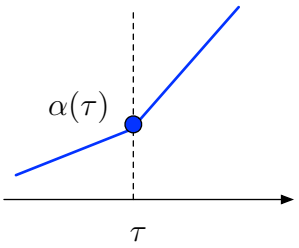
$$g(t) \frac{\mathrm{d}\alpha(t)}{\mathrm{d}t}.$$

However, the Riemann-Stieltjes integral is more general than the Riemann integral because it allows us to integrate with respect to non-differentiable $\alpha(t)$:

- The right-continuous function $\alpha(t)$ may have discontinuities. As it is monotonically non-decreasing, its left-hand limits must exist. That is, $\lim_{t\to\tau^-} \alpha(t) = \alpha(\tau^-)$.

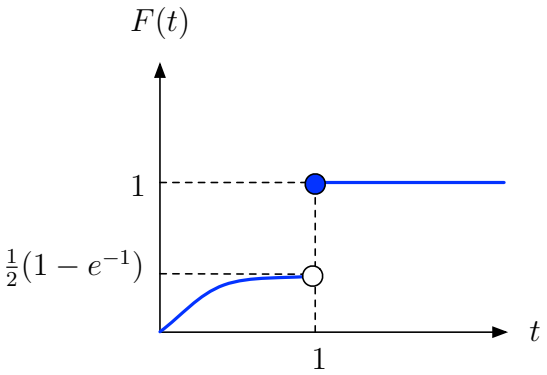


- The function $\alpha(t)$ may also be continuous but not differentiable at τ.



Example 24.

$$F(t) = \begin{cases} 0 & \text{if} \quad t < 0, \\ \frac{1}{2}(1 - e^{-t}) & \text{if} \quad 0 \le t < 1, \\ 1 & \text{if} \quad t \ge 1. \end{cases}$$



Now consider (a little roughly!)

$$\begin{aligned} \int_0^\infty g(t)\mathrm{d}F(t) &= \lim_{N\to\infty} \sum_{n=1}^{N} M_n \left[F(t_n) - F(t_{n-1})\right] \quad \text{for an appropriate partition } t_0, t_1, \ldots, \\ &= \lim_{\varepsilon\to 0^+} \int_0^{1-\varepsilon} g(t)\frac{\mathrm{d}F(t)}{\mathrm{d}t}\mathrm{d}t + g(1)\left[F(1) - F(1-\varepsilon)\right] + \int_{1+\varepsilon}^{\infty} g(t)\frac{dF(t)}{\mathrm{d}t}\mathrm{d}t, \\ &= \int_0^\infty g(t)\frac{\mathrm{d}F(t)}{\mathrm{d}t}\mathrm{d}t + g(1)\left[F(1) - F(1-)\right]. \end{aligned}$$

Extensions to improper integrals:

$$\lim_{b\to\infty} \int_0^b \quad \text{or} \quad \int_0^\infty$$

follow the same line as for standard Riemann integrals.

Two minor problems with the definition:

1. If $\alpha(t)$ has a discontinuity at a, we want to include it in any integral that starts (or finishes) at a. That is, we want to include the mass $(\alpha(a) - \alpha(a^-))$ at a. Thus, we interpret the lower limit of the integration to be to the left of a (and the upper limit to be to the right of b), and so we could write $\int_{a^-}^{b^+} g(t)\mathrm{d}\alpha(t)$.

2. There is a problem if $g(t)$ and $\alpha(t)$ have a discontinuity at the same point τ, because the upper and lower sums always differ by at least

$$\big[g(\tau) - g(\tau^-)\big]\big[\alpha(\tau) - \alpha(\tau^-)\big].$$

This can really only be resolved by resorting to more advanced measure theory. However, we can get over this by regarding the correct contribution of this point to the integral as

$$g(\tau)\big(\alpha(\tau) - \alpha(\tau^-)\big).$$

Examples (Exercises in class)

Calculate the Laplace-Stieltjes transform

$$\int_0^\infty \mathrm{e}^{-st}\mathrm{d}F(t)$$

with respect to the following distribution functions $F(t)$, where $\lambda > 0$ and $\rho \in [0, 1)$.

(a)

$$\begin{aligned} F(t) &= \begin{cases} 0 & \text{for} \quad t < 0 \\ 1 - e^{-\lambda t} & \text{for} \quad t \geq 0 \end{cases} \\ \Rightarrow \widehat{F}(s) &= \int_0^\infty e^{-st} dF(t) \\ &= \int_0^\infty e^{-st}\lambda e^{-\lambda t}\mathrm{d}t \\ &= \lambda \int_0^\infty e^{-(s+\lambda)t}\mathrm{d}t \\ &= \frac{\lambda}{(s+\lambda)}. \end{aligned}$$

(b)

$$\begin{aligned} F(t) &= \begin{cases} 0 & \text{for} \quad t < 0 \\ 1 - \rho\, e^{-\lambda t} & \text{for} \quad t \in [0, \infty) \end{cases} \\ \Rightarrow \quad \widehat{F}(s) &= \int_0^\infty e^{-st}\mathrm{d}F(t) \\ &= (1-\rho)e^{-s.0} + \lambda\rho \int_0^\infty e^{-(s+\lambda)t}\mathrm{d}t \\ &= (1-\rho) + \frac{\lambda\rho}{(s+\lambda)}. \end{aligned}$$

(c)

$$F(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ \rho & \text{for} \quad t \in [0, \lambda) \\ 1 & \text{for} \quad t \geq \lambda \end{cases}$$

$$\begin{aligned} \Rightarrow \quad \widehat{F}(s) &= \int_0^\infty e^{-st} \mathrm{d}F(t) \\ &= \rho e^{-s.0} + (1-\rho)e^{-s\lambda} \\ &= \rho + (1-\rho)e^{-s\lambda}. \end{aligned}$$

Lecture 27: Laplace-Stieltjes Transforms, Convolution Theorem and the Renewal Function

Concepts checklist

At the end of this lecture, you should be able to:

- *Define and evaluate* Laplace-Stieltjes transforms;
- *State* the Convolution Theorem and *apply* it appropriately;
- *Define* the renewal function (or mean-value function).

Laplace-Stieltjes Transform

In renewal theory, we make extensive use of Laplace-Stieltjes transforms.

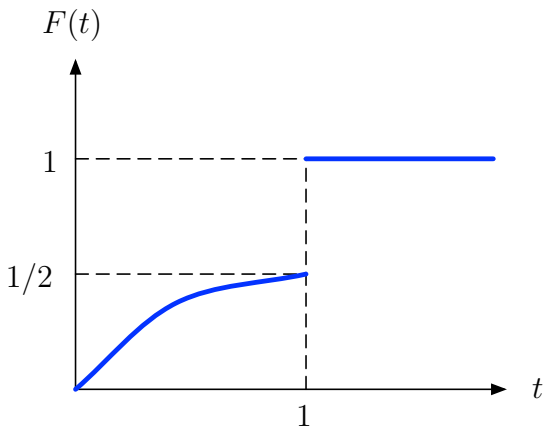
Definition 27. *Let X be a non-negative random variable with distribution function $F(x)$. Then, the Laplace-Stieltjes transform of X is given by*

$$\widehat{F}(s) = \int_0^\infty e^{-sx} \mathrm{d}F(x).$$

Note: $\widehat{F}(s) = \mathbb{E}(e^{-sX})$.

Example 25.

$$F(t) = \begin{cases} 0 & \text{if} \quad t < 0 \\ \frac{1}{2}(1 - e^{-t}) & \text{if} \quad 0 \leq t < 1 \\ 1 & \text{if} \quad t \geq 1 \end{cases}$$



Then,

$$\begin{aligned}\widehat{F}(s) &= \int_0^\infty e^{-st}\mathrm{d}F(t)\\ &= \int_0^1 e^{-st}\left(\frac{1}{2}e^{-t}\mathrm{d}t\right) + \left[1-\frac{1}{2}(1-e^{-1})\right]e^{-s} + \int_1^\infty e^{-st}\big(0\mathrm{d}t\big)\\ &= -\frac{1}{2}\frac{1}{s+1}e^{-(s+1)t}\Big|_0^1 + \left[\frac{1}{2}+\frac{1}{2}e^{-1}\right]e^{-s}\\ &= \frac{1}{s+1}\left(\frac{1}{2}-\frac{1}{2}e^{-(s+1)}\right) + \left[\frac{1}{2}e^{-s}+\frac{1}{2}e^{-(s+1)}\right].\end{aligned}$$

Convolution Theorem

The most important result on Laplace-Stieltjes transforms for renewal theory is the Convolution Theorem.

Definition 28. *If X and Y are independent random variables then the random variable $Z = X+Y$ is known as the convolution of X and Y. By conditioning on the value of Y, we can see that Z has the distribution function*

$$F_Z(z) = \int_0^z F_X(z-y)\mathrm{d}F_Y(y),$$

where $F_W(\cdot)$ is the distribution function for the random variable W, for $W = X, Y, Z$.

Theorem 25 (Convolution Theorem)**.** *For independent random variables X and Y, the random variable $Z = X+Y$ has the Laplace-Stieltjes transform*

$$\widehat{F}_Z(s) = \widehat{F}_X(s)\widehat{F}_Y(s).$$

More generally, for independent random variables X_i where $i \in \{1, 2, \ldots, n\}$, the random variable $Z := \sum_{i=1}^{n} X_i$, has the Laplace-Stieltjes transform

$$\widehat{F}_Z(s) = \prod_{i=1}^{n}\widehat{F}_{X_i}(s).$$

Example 26. Distribution of waiting time

If we define S_n to be the time until the nth event in a stochastic process, $S_n = X_1+X_2+\cdots+X_n$, where all the X_i are i.i.d., then

$$F_n(t) = \Pr(S_n \le t) \quad \text{is the waiting time distribution function.}$$

Then, using the Convolution Theorem we have

$$\widehat{F}_n(s) = \prod_{i=1}^{n}\widehat{F}_{X_i}(s) = \Big(\widehat{F}(s)\Big)^n,$$

where $\widehat{F}(s)$ is the Laplace-Stieltjes transform of the distribution function of each of the random variables X_i.

Remark: The main purpose of renewal theory is to derive information about the counting process and the waiting time process from the inter-event time distribution $F(t)$. The above result is an example of this for the waiting time process $F_n(t)$.

Letting $P_n(t) = \Pr(N(t) = n)$, and since the events $\{N(t) < n\}$ and $\{S_n > t\}$ are equivalent, we have

$$\begin{aligned} P_n(t) &= \Pr(N(t) \geq n) - \Pr(N(t) \geq n+1) \\ &= \Pr\left(S_n \leq t\right) - \Pr\left(S_{n+1} \leq t\right) \\ &= F_n(t) - F_{n+1}(t). \end{aligned}$$

Definition 29. *The renewal function (or mean-value function) $M(t)$ is defined as*

$$M(t) = \mathbb{E}[N(t)] = \sum_{n=0}^{\infty} nP_n(t).$$

$\equiv M(t)$ *is the expected number of events that have occurred by time t.*

Note: It can also be shown (not trivial, and omitted) that if $F(0) < 1$, then $M(t) < \infty$ for $t > 0$.

Letting $\widehat{M}(s) = \int_0^{\infty} e^{-st} \mathrm{d}M(t)$, we have

$$\widehat{M}(s) = \sum_{n=1}^{\infty} \left(\widehat{F}(s)\right)^n,$$

and since $\hat{F}(s) < 1$ because

$$\widehat{F}(s) = \int_0^{\infty} e^{-st} \mathrm{d}F(t) < \int_0^{\infty} 1 \mathrm{d}F(t) = 1,$$

and $F(0) < 1$ means that there must be some contribution to both of the above integrals for a positive value of t, for which $e^{-st} < 1$ for $s > 0$, we have

$$\widehat{M}(s) = \frac{\widehat{F}(s)}{1 - \widehat{F}(s)}.$$

Example 27. Poisson process

$$\begin{aligned} F(t) = 1 - \mathrm{e}^{-\lambda t} \quad &\Rightarrow \widehat{F}(s) = \frac{\lambda}{\lambda + s} \\ &\Rightarrow \widehat{M}(s) = \frac{\dfrac{\lambda}{\lambda + s}}{1 - \dfrac{\lambda}{\lambda + s}} = \frac{\lambda}{s} \\ &\Rightarrow M(t) = \lambda t, \end{aligned}$$

which is exactly what we expect, as $N(t)$ is Poisson distributed with parameter λt.

Lecture 28: Renewal Theory I

Concepts checklist

At the end of this lecture, you should be able to:

- *State, prove and use* the theorem for the renewal equation satisfied by the renewal function;
- *State and apply* the one-to-one correspondence between the renewal function and the inter-arrival distribution; and,
- *State, prove and use* the theorem for the generalised renewal equation.

Theorem 26. *The renewal function $M(t)$ satisfies the renewal equation*

$$M(t) = F(t) + \int_0^t M(t-x)\mathrm{d}F(x). \tag{32}$$

Proof. Let X_1 be the random variable representing the time of the first renewal. We condition on X_1, and count the expected number of renewals thereafter. We have

$$\mathbb{E}[N(t)\big|X_1 = x] = \begin{cases} 0 & \text{if} \quad x > t, \\ 1 + M(t-x) & \text{if} \quad x \leq t. \end{cases}$$

In the second case of the equation (above), we have the first renewal + the expected number since the first renewal, since the probabilistic structure begins anew at the instant of the first renewal.

Then,

$$\begin{aligned} M(t) &= \mathbb{E}[N(t)] \\ &= \int_0^\infty \mathbb{E}\left[N(t)\big|X_1 = x\right] \mathrm{d}F(x) \\ &= \int_0^t [1 + M(t-x)]\mathrm{d}F(x) \\ &= F(t) + \int_0^t M(t-x)\mathrm{d}F(x). \end{aligned}$$

□

Note: Much of the power of renewal theory derives from the method of reasoning used in the previous proof: conditioning on the time of the first renewal.

Theorem 27. *The only solution to the renewal equation (32), which is bounded on finite intervals, is given by*

$$M(t) = \sum_{n=1}^{\infty} F_n(t). \tag{33}$$

Proof. Taking the Laplace-Stieltjes transform of equation (32), we have

$$\begin{aligned} \widehat{M}(s) &= \widehat{F}(s) + \widehat{M}(s)\widehat{F}(s) \\ \Rightarrow \quad \widehat{M}(s) &= \frac{\widehat{F}(s)}{1-\widehat{F}(s)}. \end{aligned}$$

We have $\widehat{F}(s) < 1$, and so using the identity for the sum of a geometric series we have that

$$\begin{aligned} \widehat{M}(s) &= \sum_{n=1}^{\infty} \left[\widehat{F}(s)\right]^n \\ \Rightarrow M(t) &= \sum_{n=1}^{\infty} F_n(t), \end{aligned}$$

where $F_n(t)$ is the distribution function for the convolution of n random variables with distribution function $F(t)$.

□

Corollary 3.

$$\widehat{M}(s) = \frac{\widehat{F}(s)}{1-\widehat{F}(s)},$$

and consequently

$$\widehat{F}(s) = \frac{\widehat{M}(s)}{1+\widehat{M}(s)}.$$

Proof. This follows directly from the proof of the previous Theorem. □

Note: This corollary shows that there is a one-to-one correspondence between $M(t)$ and $F(t)$, so if we know one we can determine the other. Consequently, the Poisson process is the only renewal process having a linear mean-value function, $M(t) = \lambda t$.

Example 28.

Evaluate the renewal function corresponding to the lifetime distribution

$$F(t) = 1 - e^{-2t}(1+2t).$$

We have

$$\begin{aligned} \widehat{F}(s) &= \int_0^{\infty} e^{-st} 4te^{-2t} dt \\ &= \frac{4}{(s+2)^2}. \end{aligned}$$

Hence, we have

$$\widehat{M}(s) = \frac{4}{s(s+4)},$$

and (using the table of Laplace Transforms*)

$$M'(t) = 1 - e^{-4t},$$

and thus the renewal function is

$$M(t) = \frac{1}{4}(e^{-4t} - 1) + t.$$

**Note*: But you can see this via partial fractions:

$$\frac{4}{s(s+4)} = \frac{1}{s} - \frac{1}{(s+4)},$$

and so $M'(t) = 1 - e^{-4t}$.

Generalised Renewal Equation

When we consider renewal processes where the first lifetime is different from the rest, the generalised renewal equation for $H(t)$ (the expected number of events by time t) arises:

$$H(t) = G(t) + \int_0^t H(t-y)\mathrm{d}F(y), \tag{34}$$

where $G(t)$ is the distribution of the first lifetime, and $F(t)$ is the distribution of each of the subsequent lifetimes. In the renewal equation (32), $F(t)$ and $G(t)$ are the same function.

Theorem 28. *The solution to the generalised renewal equation is*

$$H(t) = G(t) + \int_0^t G(t-y)\mathrm{d}M(y),$$

where $M(t)$ is the solution to equation (32).

Proof.

$$\begin{aligned}
H(t) &= G(t) + \int_0^t G(t-y)\mathrm{d}M(y) \\
\Rightarrow \widehat{H}(s) &= \widehat{G}(s) + \widehat{G}(s)\widehat{M}(s) \\
&= \widehat{G}(s) + \widehat{G}(s)\,\frac{\widehat{F}(s)}{1-\widehat{F}(s)} \qquad \text{by Corollary 3} \\
&= \widehat{G}(s)\left[1 + \frac{\widehat{F}(s)}{1-\widehat{F}(s)}\right] \\
&= \frac{\widehat{G}(s)}{1-\widehat{F}(s)}.
\end{aligned}$$

Thus,

$$\widehat{H}(s) = \widehat{G}(s) + \widehat{H}(s)\widehat{F}(s) \quad \Rightarrow H(t) = G(t) + \int_0^t H(t-y)\mathrm{d}F(y).$$

□

Lecture 29: Renewal Theory II – Renewal Theorems

Concepts checklist

At the end of this lecture, you should be able to:

- *State, prove, and use* a Corollary for the expected time from the start of the process until the first event after time t; and,
- *state and use* a Theorem regarding the almost sure convergence of the long-term average number of events per unit time.

Note, in the Proof of Theorem (28), that no assumptions were made about the functions involved except the existence of their Laplace-Stieltjes Transforms.

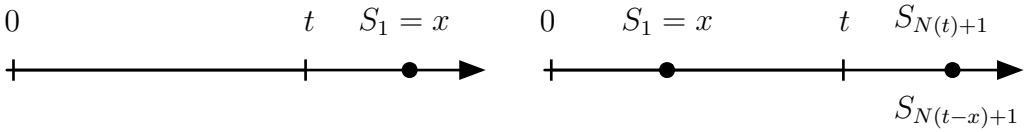
Corollary 4. *For any renewal process, the expected time from the start of the process until the first event after time t,*

$$\mathbb{E}[S_{N(t)+1}] = \mu(M(t)+1), \quad \textit{where } \mu \textit{ is the mean inter-event time.}$$

Proof. Let $H(t) = \mathbb{E}[S_{N(t)+1}]$, then

$$\mathbb{E}(S_{N(t)+1} \mid X_1 = x) = \begin{cases} x, & \text{if} \quad t < x, \\ x + \mathbb{E}[S_{N(t-x)+1}], & \text{if} \quad t \geq x. \end{cases}$$

The two cases correspond to the left and the right figures below



For $t \geq x$, we have by definition that $H(t-x) = \mathbb{E}[S_{N(t-x)+1}]$ and so

$$\begin{aligned} H(t) &= \int_0^\infty H(t \mid X_1 = x) \mathrm{d}F(x) \\ &= \int_t^\infty x \mathrm{d}F(x) + \int_0^t (x + H(t-x))\, \mathrm{d}F(x) \\ &= \int_0^\infty x \mathrm{d}F(x) + \int_0^t H(t-x) \mathrm{d}F(x) \\ &= \mu + \int_0^t H(t-x) \mathrm{d}F(x) \end{aligned}$$

so that $H(t)$ satisfies the generalised renewal equation (34) with $G(t) = \mu$. By Theorem 28, the solution is given by

$$H(t) = \mu + \int_0^t \mu \mathrm{d}M(y) = \mu(1 + M(t)).$$

□

Example 29. Engaged Signals and Reattempts

The behaviour of telephone users upon receiving an engaged signal can be modelled roughly as follows:

- with probability $1-p$ the caller does not attempt to retry,
- with probability p, the caller waits a time X with distribution $F(x)$ and then tries again.
- Upon retrying, the caller gets the engaged signal with probability r; in which case the process is repeated,
- or gets through with probability $1-r$, and the process ends.

A telephone company might want to know the average number of re-attempts $R(t)$ that occur during the time interval $[0,t)$ after the first engaged signal at time 0. We have,

$$R(t \,|\text{reattempts after first engaged at time } x) = \begin{cases} 0 & \text{if} \quad x > t \\ 1 + rR(t-x) + (1-r) \times 0 & \text{if} \quad x \leq t \end{cases}$$

whereas $R(t \,|\text{no reattempt}) = 0$. Then removing the conditioning of a reattempt and the time of the first wait yields

$$\begin{aligned} R(t) &= \int_0^t [p\big(1 + rR(t-x)\big) + (1-p) \times 0] \mathrm{d}F(x) \\ &= pF(t) + pr \int_0^t R(t-x) \mathrm{d}F(x). \end{aligned}$$

Taking Laplace-Stieltjes transforms, we get

$$\widehat{R}(s) = p\widehat{F}(s) + pr\widehat{R}(s)\widehat{F}(s) \quad \Rightarrow \quad \widehat{R}(s) = \frac{p\widehat{F}(s)}{1 - pr\widehat{F}(s)}.$$

Inverting this transform (not always easy!) gives $R(t)$. Inversion is easy analytically if $F(t) = 1 - e^{-\alpha t}$ since $\widehat{F}(s) = \frac{\alpha}{\alpha+s}$ then

$$\begin{aligned} \widehat{R}(s) &= \frac{p\alpha}{\alpha + s - pr\alpha} \\ &= \frac{p\alpha}{s + \alpha(1-pr)} \\ &= \frac{p}{(1-pr)} \frac{\alpha(1-pr)}{s + \alpha(1-pr)} \\ \text{and therefore} \quad R(t) &= \frac{p}{(1-pr)} \left(1 - e^{-\alpha(1-pr)t}\right). \end{aligned}$$

Theorem 29. *Let $\{N(t) : t \geq 0\}$ be a counting process. Then,*

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mu} \quad \textit{with probability } 1,$$

where μ is the mean time between events.

There are many sorts of convergence of random variables (see Additional Materials on MyUni). In Theorem 29, the phrase "with probability 1" (or, equivalentally "almost surely") means that:

of all the possible sample paths of the process $N(t)$*, the quantity* $\dfrac{N(t)}{t}$ *converges to* $\dfrac{1}{\mu}$ *as* $t \to \infty$ *for a* set *of realisations* **and** *this set has probability* 1 *of happening.*

This doesn't necessarily mean that $N(t)/t \to 1/\mu$ for all realisations, but the set of realisations on which this does not happen, has probability 0.

Example 30. A counting process

Let the time between two consecutive events, $i-1$ and i, be $X_i = \begin{cases} 0 & \text{with probability } 1/2, \\ 2 & \text{with probability } 1/2. \end{cases}$

Then $\mathbb{E}[X_i] = 1$. If we consider the counting process $N(t)$, then according to Theorem 29,

$$\lim_{t\to\infty} \frac{N(t)}{t} = 1 \quad \text{with probability 1.}$$

A possible realisation of the process, however, is that the lifetime $X_i = 2$ *every* time, in which case

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{2}.$$

The probability of this realisation is

$$\lim_{n\to\infty} \left(\frac{1}{2}\right)^n = 0,$$

which is, of course, the probability of any particular realisation.

Additional Materials: Convergence of Random Variables

Sample Paths and Random Variables — Revisited

Sample paths

When we observe a stochastic process, we are actually running a random experiment, where the sample space Ω of the experiment is the set of all possible outcomes. For example:

- In a discrete time Markov chain with state space $\mathcal{S}$, the sample space Ω is the set of all sequences $\{x_i\}_{i=1}^{\infty}$ with $x_i \in \mathcal{S}$.
- In a continuous time Markov chain with state space $\mathcal{S}$, the sample space Ω is the set of all right-continuous functions $x(t)$ which map $t \in [0, \infty) \to \mathcal{S}$.

A single element ω of Ω is known as a sample path of the process. There are uncountably many sample paths (even if the state space is finite). Therefore, in general the probability of any given sample path occurring is 0, just like the probability of picking a real number randomly between 0 and 1. There are also sets of sample paths that have probability zero.

Random variables

A random variable defined on a stochastic process is a mapping from Ω to $\mathbb{R}$ (or sometimes a subset of $\mathbb{R}$). For example:

- In a discrete-time Markov chain with state space $\mathcal{S} = \{0, 1, \ldots, n\}$, the function

$$\begin{aligned} X_j : \Omega &\to \mathbb{R} \\ (x_i)_{i=1}^{\infty} &\to x_j \end{aligned}$$

  is the random variable that gives the state at time point j.

- In a continuous-time Markov chain with state space $\mathcal{S} = \{0, 1, \ldots, n\}$, the function

$$\begin{aligned} X_s : \Omega &\to \mathbb{R} \\ x(t) : t \in [0, \infty) &\to x_s \end{aligned}$$

  is the random variable that gives the state at time point s.

Types of Convergences

Let $Y_1, Y_2, \ldots$ be a sequence of random variables defined on the sample space of a stochastic process and Y be another random variable defined on the same sample space. When we consider the convergence of the sequence of random variables Y_n to Y (written $\lim_{n\to\infty} Y_n = Y$), we could mean one of at least three things.

1. **Convergence with probability** 1 **(almost sure convergence):** If there is a set $\Omega^* \subseteq \Omega$ that has probability 1 (which may not contain all the sample paths as we saw before) such that

$$\lim_{n\to\infty} Y_n(\omega) \to Y(\omega) \quad \text{for all } \omega \in \Omega^*$$

then we say

$$\lim_{n\to\infty} Y_n = Y \text{ with probability 1.}$$

2. **Convergence in probability:** If for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[\omega : \left|Y_n(\omega) - Y(\omega)\right| > \epsilon\right] = 0,$$

then we say $\lim_{n\to\infty} Y_n = Y$ in probability.

3. $\mathcal{L}_p$ **– convergence, or convergence in the** p**th mean:** The set of random variables Y_n converges to the random variable Y in the pth mean or $\mathcal{L}_p$ if

$$\lim_{n\to\infty} \mathbb{E}\left[\left|Y_n - Y\right|^p\right] = 0,$$

where the expectation is taken with respect to the probability measure associated with the sample paths.

The most important case is when $p = 2$, when we say $Y_n \to Y$ in $\mathcal{L}_2$ (or mean square).

The three types of convergence are related as follows:

- convergence with probability 1 $\longrightarrow$ convergence in probability
- convergence in $\mathcal{L}_p \longrightarrow$ convergence in probability.

Convergence in probability is weaker than the other two, which do not imply each other.

Laws of Large Numbers

Theorem (Weak law of large numbers.)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables and suppose that $\mathbb{E}[Y_i] = \mu < \infty$. Then, the sequence of random variables*

$$\overline{Y}(n) = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} \quad \textit{converges in probability to } \mu.$$

That is, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[|\overline{Y}(n) - \mu| > \varepsilon\right] = 0.$$

Theorem (Strong law of large numbers.)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of independent, identically distributed random variables and suppose that $\mathbb{E}[Y_i] = \mu < \infty$. Then, the sequence of random variables*

$$\overline{Y}(n) = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} \quad \textit{converges to } \mu \textit{ with probability } 1.$$

That is,

$$\Pr\left[\lim_{n\to\infty} \overline{Y}(n) = \mu\right] = 1.$$

Lecture 30: Renewal Theory – Renewal Theorems, and the Bus Paradox

Concepts checklist

At the end of this lecture, you should be able to:

- *State and use* the Basic Renewal Theorem and Blackwell's Renewal Theorem, and the Elementary Renewal Theorem.

Theorem 30 (Basic Renewal Theorem)**.** *Let $F(t)$ be the distribution function of a positive random variable with mean $\mu < \infty$, and assume that $F(t)$ is not lattice. (That is, it does not have all of its points of increase at multiples of some δ.)*

Suppose that $H(t)$ is a solution of the generalised renewal equation

$$H(t) = G(t) + \int_0^t H(t-y)\mathrm{d}F(y), \quad \textit{where } G(t) \textit{ is integrable,}$$

then

$$\lim_{t\to\infty} H(t) = \frac{1}{\mu}\int_0^\infty G(t)\mathrm{d}t. \tag{35}$$

Note: *If F is lattice, then equation (35) is valid for $t = n\delta$ with $n \in \mathbb{N}$.*

We now state another form of this theorem in terms of the renewal function $M(t) = \mathbb{E}\big[N(t)\big]$.

Theorem 31 (Blackwell's Renewal Theorem)**.** *Let F be the distribution function of a positive random variable with mean $\mu < \infty$, which is not lattice, then*

$$\lim_{t\to\infty}\left[M(t) - M(t-h)\right] = \frac{h}{\mu} \quad \textit{for } h > 0. \tag{36}$$

Proof:

$$\text{For } h > 0, \text{ if we let } \; G(y) \; = \begin{cases} 1 & \text{if} \quad 0 \le y < h \\ 0 & \text{if} \quad y \ge h \end{cases}$$

(which is integrable) and insert into the generalised renewal equation (34), then we can then use Theorem 28 to get for $t > h$ that

$$\begin{aligned} H(t) &= G(t) + \int_0^t G(t-y)dM(y) \\ &= 0 + \int_{t-h}^t 1dM(y) \\ &= M(t) - M(t-h)\ . \end{aligned}$$

Then by Theorem 30,

$$\begin{aligned}\lim_{t\to\infty}\big[M(t)-M(t-h)\big] &= \lim_{t\to\infty} H(t)\\ &= \frac{1}{\mu}\int_0^\infty G(t)dt\\ &= \frac{1}{\mu}\int_0^h 1dt\\ &= \frac{h}{\mu}.\end{aligned}$$

Corollary 5 (Elementary Renewal Theorem)**.** *If $F(t)$ is not lattice, then*

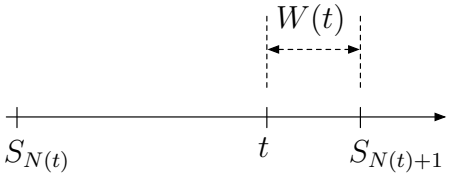
$$\lim_{t\to\infty}\frac{M(t)}{t} = \frac{1}{\mu}.$$

In words: *the rate of change of the renewal function $M(t)$ approaches $1/\mu$ as $t\to\infty$.*

Forward recurrence time & the Bus Paradox

Goal: Find the equilibrium distribution of the time until the next event.

We define $W(t) = S_{N(t)+1} - t$ and $H(z,t) = \Pr\{W(t) > z\}$.



To get an equation for $H(z,t)$, we again use the renewal argument and condition on the time of the first renewal. That is, we consider

$$\Pr(W(t) > z | X_1 = x).$$

There are three cases:

Case 1: The first arrival occurs before time t, i.e., $0 \le x \le t$:

$$\Pr\{W(t) > z | X_1 = x\} = H(z, t-x),$$

because the process might as well have started at time x.

Case 2: The first arrival occurs after time t, but before time $t+z$, i.e., $t \le x \le t+z$:

$$\Pr\{W(t) > z | X_1 = x\} = 0, \quad \text{because we know that } W(t) = x - t < z.$$

Case 3: The first arrival occurs after time $t+z$, i.e., $t+z < x$:

$$\Pr\{W(t) > z | X_1 = x\} = 1, \quad \text{because we know that } W(t) = x - t > z.$$

Therefore, $H(z,t)$ is given by

$$\begin{aligned} H(z,t) &= \int_0^\infty \Pr(W(t) > z | X_1 = x) \mathrm{d}F(x) \\ &= \int_0^t H(z, t-x)\mathrm{d}F(x) + \int_{z+t}^\infty 1 dF(x) \\ &= 1 - F(z+t) + \int_0^t H(z,t-x)\mathrm{d}F(x) \end{aligned}$$

which is a generalised renewal equation with $G(t) = 1 - F(z+t)$ and $H(t) = H(z,t)$. By Theorem 30, the solution is therefore given by

$$H(z,t) = (1 - F(z+t)) + \int_0^t (1 - F(z+t-y))\mathrm{d}M(y).$$

• Now, as we are looking for the equilibrium distribution, let $t \to \infty$. It is clear that $\big(1 - F(z+t)\big) \to 0$, but what about the second term? This is much more difficult to directly evaluate.

However, the Basic Renewal Theorem tells us that

$$\lim_{t\to\infty} H(z,t) = \frac{1}{\mu}\int_0^\infty G(t)\mathrm{d}t = \frac{1}{\mu}\int_0^\infty (1 - F(z+t))\mathrm{d}t,$$

as long as

$$\int_0^\infty G(t)\mathrm{d}t = \int_0^\infty (1 - F(z+t))\mathrm{d}t < \infty.$$

Let us start by considering

$$\begin{aligned} \int_0^\infty \big(1 - F(u)\big)\mathrm{d}u &= \int_0^\infty \left[\int_u^\infty dF(y)\right]\mathrm{d}u \\ &= \int_0^\infty \left[\int_0^y 1\mathrm{d}u\right]\mathrm{d}F(y) \\ &= \int_0^\infty y\ dF(y) = \mu \qquad \text{(the mean)}. \end{aligned} \tag{37}$$

Therefore,

$$\begin{aligned} \int_0^\infty G(t)\mathrm{d}t &= \int_0^\infty \big(1 - F(z+t)\big)\mathrm{d}t \\ &\le \int_0^\infty \big(1 - F(t)\big)\mathrm{d}t \quad \text{since } F(t+z) \ge F(t) \text{ for } z \ge 0, \\ &= \mu < \infty. \end{aligned}$$

Hence,

$$\lim_{t\to\infty} H(z,t) = \frac{1}{\mu}\int_0^\infty (1 - F(z+t))\mathrm{d}t = \frac{1}{\mu}\int_z^\infty (1 - F(u))\mathrm{d}u.$$

Therefore, by equation (37),

$$\Pr\{W \leq z\} = 1 - \frac{1}{\mu}\int_z^\infty \big(1 - F(u)\big)\mathrm{d}u = \frac{1}{\mu}\int_0^z \big(1 - F(u)\big)\mathrm{d}u.$$

- So,

$$\begin{aligned}
\mathbb{E}[W] &= \int_0^\infty z \,\mathrm{d}\Pr\{W \leq z\} \\
&= \frac{1}{\mu}\int_0^\infty z[1 - F(z)]\mathrm{d}z \\
&= \frac{1}{\mu}\int_0^\infty z\left[\int_z^\infty 1\mathrm{d}F(u)\right]\mathrm{d}z \\
&= \frac{1}{\mu}\int_0^\infty \left[\int_0^u z\mathrm{d}z\right]\mathrm{d}F(u) \\
&= \frac{1}{\mu}\int_0^\infty \frac{u^2}{2}\mathrm{d}F(u) \\
&= \frac{1}{2\mu}\int_0^\infty u^2\mathrm{d}F(u) \\
&= \frac{1}{2\mu}\left[\mu^2 + \sigma^2\right],
\end{aligned}$$

where σ^2 is the variance of the waiting time distribution.

Hence, we have shown that the expected time to the next renewal is given by $\dfrac{\mu^2 + \sigma^2}{2\mu}$.

If the variance, σ^2, is large, it is possible for this to be bigger than μ, so that the average wait until the next arrival can be greater than the average time between arrivals.

This is known as the bus paradox, which is caused by the fact that an arbitrary time point is more likely to fall into a large interval.